Vincent G. Duffy (Ed.)

Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management

LNCS 15791

16th International Conference, DHM 2025
Held as Part of the 27th HCI International Conference, HCII 2025
Gothenburg, Sweden, June 22–27, 2025
Proceedings, Part I

Part I



Springer

Lecture Notes in Computer Science **15791**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Vincent G. Duffy
Editor

Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management

16th International Conference, DHM 2025
Held as Part of the 27th HCI International Conference, HCII 2025
Gothenburg, Sweden, June 22–27, 2025
Proceedings, Part I

Springer

Editor
Vincent G. Duffy
Purdue University
West Lafayette, IN, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-93501-5 ISBN 978-3-031-93502-2 (eBook)
https://doi.org/10.1007/978-3-031-93502-2

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Foreword

The HCI International (HCII) conference was founded in 1984 by Gavriel Salvendy (Purdue University, USA, Tsinghua University, P.R. China, and University of Central Florida, USA) and the first event of the series, "1st USA-Japan Conference on Human-Computer Interaction", was held in Honolulu, Hawaii, USA, 18–20 August. Since then, HCI International is held jointly with several Thematic Areas and Affiliated Conferences, with each one under the auspices of a distinguished international Program Board and under one management and one registration. Twenty-seven HCI International Conferences have been organized so far (every two years until 2013, and annually thereafter).

Last year, we celebrated 40 years since the establishment of the HCII conference, which has been a hub for presenting groundbreaking research and novel ideas and collaboration for people from all over the world. Over the years, this conference has served as a platform for scholars, researchers, industry experts, and students to exchange ideas, connect, and address challenges in the ever-evolving HCI field. The conference has evolved itself, adapting to new technologies and emerging trends, while staying committed to its core mission of advancing knowledge and driving change.

The 27th International Conference on Human-Computer Interaction, HCI International 2025 (HCII 2025), was held as an 'on-site' conference at the Gothia Towers Hotel and Swedish Exhibition & Congress Centre, in Gothenburg, Sweden, on June 22–27, 2025, with the additional option for 'on-line' participation. It incorporated the 21 thematic areas and affiliated conferences listed below.

A total of 7972 individuals from academia, research institutes, industry, and government agencies from 92 countries submitted contributions. 1430 papers and 355 posters (as short research papers) are included in the volumes of the proceedings published just before the start of the conference, and which are listed below. The contributions thoroughly cover the entire field of human-computer interaction, highlight the evolving role of computers in diverse contexts, and demonstrate how HCI research is shaping and improving user experiences across a wide range of domains, influencing technological progress and its effective integration into various sectors.

The HCII conference also offers the option of presenting 'Late Breaking Work', both for papers and posters, with the corresponding proceedings volumes published after the conference. Full papers are included in the 'HCII 2025 - Late Breaking Papers' volumes of the proceedings published in the Springer LNCS series, while 'Poster Extended Abstracts' are included as short research papers in the 'HCII 2025 - Late Breaking Posters' volumes published in the Springer CCIS series.

I would like to thank the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences for their contribution towards the high scientific quality and overall success of the HCI International 2025 conference. Their manifold support including paper reviews (via a single-blind review process, with a minimum of two reviews per submission), session organization, and their willingness to act as goodwill ambassadors for the conference is most highly appreciated.

This conference would not have been possible without the continuous and unwavering support and advice of Gavriel Salvendy, founder, General Chair Emeritus, and Scientific Advisor. For his outstanding efforts, I would like to express my sincere appreciation to Abbas Moallem, Communications Chair and Editor of HCI International News.

June 2025 Constantine Stephanidis

HCI International 2025 Thematic Areas and Affiliated Conferences

- HCI: Human-Computer Interaction Thematic Area
- HIMI: Human Interface and the Management of Information Thematic Area
- EPCE: 22nd International Conference on Engineering Psychology and Cognitive Ergonomics
- AC: 19th International Conference on Augmented Cognition
- UAHCI: 19th International Conference on Universal Access in Human-Computer Interaction
- CCD: 17th International Conference on Cross-Cultural Design
- SCSM: 17th International Conference on Social Computing and Social Media
- VAMR: 17th International Conference on Virtual, Augmented and Mixed Reality
- DHM: 16th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- DUXU: 14th International Conference on Design, User Experience, and Usability
- C&C: 13th International Conference on Culture and Computing
- DAPI: 13th International Conference on Distributed, Ambient and Pervasive Interactions
- HCIBGO: 12th International Conference on HCI in Business, Government and Organizations
- LCT: 12th International Conference on Learning and Collaboration Technologies
- ITAP: 11th International Conference on Human Aspects of IT for the Aged Population
- AIS: 7th International Conference on Adaptive Instructional Systems
- HCI-CPT: 7th International Conference on HCI for Cybersecurity, Privacy and Trust
- HCI-Games: 7th International Conference on HCI in Games
- MobiTAS: 7th International Conference on HCI in Mobility, Transport and Automotive Systems
- AI-HCI: 6th International Conference on Artificial Intelligence in HCI
- MOBILE: 6th International Conference on Human-Centered Design, Operation and Evaluation of Mobile Communications

List of Conference Proceedings Volumes Appearing Before the Conference

1. LNCS 15766, Human-Computer Interaction - Part I, edited by Masaaki Kurosu and Ayako Hashizume
2. LNCS 15767, Human-Computer Interaction - Part II, edited by Masaaki Kurosu and Ayako Hashizume
3. LNCS 15768, Human-Computer Interaction - Part III, edited by Masaaki Kurosu and Ayako Hashizume
4. LNCS 15769, Human-Computer Interaction - Part IV, edited by Masaaki Kurosu and Ayako Hashizume
5. LNCS 15770, Human-Computer Interaction - Part V, edited by Masaaki Kurosu and Ayako Hashizume
6. LNCS 15771, Human-Computer Interaction - Part VI, edited by Masaaki Kurosu and Ayako Hashizume
7. LNCS 15772, Human-Computer Interaction - Part VII, edited by Masaaki Kurosu and Ayako Hashizume
8. LNCS 15773, Human Interface and the Management of Information: Part I, edited by Hirohiko Mori and Yumi Asahi
9. LNCS 15774, Human Interface and the Management of Information: Part II, edited by Hirohiko Mori and Yumi Asahi
10. LNCS 15775, Human Interface and the Management of Information: Part III, edited by Hirohiko Mori and Yumi Asahi
11. LNAI 15776, Engineering Psychology and Cognitive Ergonomics: Part I, edited by Don Harris and Wen-Chin Li
12. LNAI 15777, Engineering Psychology and Cognitive Ergonomics: Part II, edited by Don Harris and Wen-Chin Li
13. LNAI 15778, Augmented Cognition: Part I, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
14. LNAI 15779, Augmented Cognition: Part II, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
15. LNCS 15780, Universal Access in Human-Computer Interaction: Part I, edited by Margherita Antona and Constantine Stephanidis
16. LNCS 15781, Universal Access in Human-Computer Interaction: Part II, edited by Margherita Antona and Constantine Stephanidis
17. LNCS 15782, Cross-Cultural Design: Part I, edited by Pei-Luen Patrick Rau
18. LNCS 15783, Cross-Cultural Design: Part II, edited by Pei-Luen Patrick Rau
19. LNCS 15784, Cross-Cultural Design: Part III, edited by Pei-Luen Patrick Rau
20. LNCS 15785, Cross-Cultural Design: Part IV, edited by Pei-Luen Patrick Rau

21. LNCS 15786, Social Computing and Social Media: Part I, edited by Adela Coman and Simona Vasilache
22. LNCS 15787, Social Computing and Social Media: Part II, edited by Adela Coman and Simona Vasilache
23. LNCS 15788, Virtual, Augmented and Mixed Reality: Part I, edited by Jessie Y. C. Chen and Gino Fragomeni
24. LNCS 15789, Virtual, Augmented and Mixed Reality: Part II, edited by Jessie Y. C. Chen and Gino Fragomeni
25. LNCS 15790, Virtual, Augmented and Mixed Reality: Part III, edited by Jessie Y. C. Chen and Gino Fragomeni
26. LNCS 15791, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Part I, edited by Vincent G. Duffy
27. LNCS 15792, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Part II, edited by Vincent G. Duffy
28. LNCS 15793, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Part III, edited by Vincent G. Duffy
29. LNCS 15794, Design, User Experience, and Usability: Part I, edited by Martin Schrepp
30. LNCS 15795, Design, User Experience, and Usability: Part II, edited by Martin Schrepp
31. LNCS 15796, Design, User Experience, and Usability: Part III, edited by Martin Schrepp
32. LNCS 15797, Design, User Experience, and Usability: Part IV, edited by Martin Schrepp
33. LNCS 15798, Design, User Experience, and Usability: Part V, edited by Martin Schrepp
34. LNCS 15799, Design, User Experience, and Usability: Part VI, edited by Martin Schrepp
35. LNCS 15800, Culture and Computing: Part I, edited by Matthias Rauterberg
36. LNCS 15801, Culture and Computing: Part II, edited by Matthias Rauterberg
37. LNCS 15802, Distributed, Ambient and Pervasive Interactions: Part I, edited by Norbert A. Streitz and Shin'ichi Konomi
38. LNCS 15803, Distributed, Ambient and Pervasive Interactions: Part II, edited by Norbert A. Streitz and Shin'ichi Konomi
39. LNCS 15804, HCI in Business, Government and Organizations: Part I, edited by Fiona Fui-Hoon Nah and Keng Leng Siau
40. LNCS 15805, HCI in Business, Government and Organizations: Part II, edited by Fiona Fui-Hoon Nah and Keng Leng Siau
41. LNCS 15806, Learning and Collaboration Technologies: Part I, edited by Brian K. Smith and Marcela Borge
42. LNCS 15807, Learning and Collaboration Technologies: Part II, edited by Brian K. Smith and Marcela Borge
43. LNCS 15808, Learning and Collaboration Technologies: Part III, edited by Brian K. Smith and Marcela Borge

44. LNCS 15809, Human Aspects of IT for the Aged Population: Part I, edited by Qin Gao and Jia Zhou
45. LNCS 15810, Human Aspects of IT for the Aged Population: Part II, edited by Qin Gao and Jia Zhou
46. LNCS 15811, Human Aspects of IT for the Aged Population: Part III, edited by Qin Gao and Jia Zhou
47. LNCS 15812, Adaptive Instructional System: Part I, edited by Robert A. Sottilare and Jessica Schwarz
48. LNCS 15813, Adaptive Instructional System: Part II, edited by Robert A. Sottilare and Jessica Schwarz
49. LNCS 15814, HCI for Cybersecurity, Privacy and Trust: Part I, edited by Abbas Moallem
50. LNCS 15815, HCI for Cybersecurity, Privacy and Trust: Part II, edited by Abbas Moallem
51. LNCS 15816, HCI in Games, edited by Xiaowen Fang
52. LNCS 15817, HCI in Mobility, Transport and Automotive Systems: Part I, edited by Heidi Krömker
53. LNCS 15818, HCI in Mobility, Transport and Automotive Systems: Part II, edited by Heidi Krömker
54. LNAI 15819, Artificial Intelligence in HCI: Part I, edited by Helmut Degen and Stavroula Ntoa
55. LNAI 15820, Artificial Intelligence in HCI: Part II, edited by Helmut Degen and Stavroula Ntoa
56. LNAI 15821, Artificial Intelligence in HCI: Part III, edited by Helmut Degen and Stavroula Ntoa
57. LNAI 15822, Artificial Intelligence in HCI: Part IV, edited by Helmut Degen and Stavroula Ntoa
58. LNCS 15823, Human-Centered Design, Operation and Evaluation of Mobile Communications: Part I, edited by June Wei and George Margetis
59. LNCS 15824, Human-Centered Design, Operation and Evaluation of Mobile Communications: Part II, edited by June Wei and George Margetis
60. CCIS 2522, HCI International 2025 Posters - Part I, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
61. CCIS 2523, HCI International 2025 Posters - Part II, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
62. CCIS 2524, HCI International 2025 Posters - Part III, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
63. CCIS 2525, HCI International 2025 Posters - Part IV, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
64. CCIS 2526, HCI International 2025 Posters - Part V, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
65. CCIS 2527, HCI International 2025 Posters - Part VI, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy

66. CCIS 2528, HCI International 2025 Posters - Part VII, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy
67. CCIS 2529, HCI International 2025 Posters - Part VIII, edited by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy

https://2025.hci.international/proceedings



Preface

Software representations of humans, including aspects of anthropometry, biometrics, motion capture, and prediction, as well as cognition modeling, are known as Digital Human Models (DHM) and are widely used in a variety of complex application domains where it is important to foresee and simulate human behavior, performance, safety, health, and comfort. Automation depicting human emotion, social interaction, and functional capabilities can also be modeled to support and assist in predicting human response in real-world settings. Such domains include medical and nursing applications, work, education and learning, ergonomics and design, as well as safety and risk management. The 16th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management (DHM 2025), an affiliated conference of the HCI International Conference 2025, encouraged papers from academics, researchers, industry, and professionals, on a broad range of theoretical and applied issues related to Digital Human Modeling and its applications.

This year's conference contributions highlighted interdisciplinary advancements in digital human technologies aimed at enhancing health, safety, and human performance across various domains. Key topics included innovative systems for digital mental health, intelligent triage, medication support, and building trust in AI-driven care, all of which demonstrated the effectiveness of human-centered tools in improving usability, acceptance, and patient outcomes. Since nowadays AI plays a critical role in digital human modeling for safety and risk management, discussions revolved around applications such as simulation for risk analysis, emergency planning, human detection, and generative systems to bolster safety and performance in high-stakes environments. Research in biomechanics and ergonomics focused on physical support systems, posture analysis, and ergonomic tools, including exoskeletons, biomechanics education, and sensor-based risk detection to enhance safety in occupational and training settings. Additionally, studies on user experience design for public-space products emphasized sustainability through inclusive design, usability, and cultural revitalization. The conference also addressed digital health monitoring, exploring sensing technologies for emotional and physical health tracking and evaluating wearables, mobile health platforms, and body scanners for proactive health management. Furthermore, papers introduced healthcare innovations that support rehabilitation, clinical decision-making, and user training by integrating AI methods with UX principles to improve outcomes in care and therapy settings. Contributions on immersive technologies for modeling, simulation, and behavior analysis featured avatars, AR/VR environments, and AI-driven interfaces, while the interplay between human behavior and technology design was examined through persuasive systems, emotional responses to AI, and adaptive interface design.

Three volumes of the HCII 2025 proceedings are dedicated to this year's edition of the DHM conference focusing on topics related to:

- Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management - Part I: Digital Human Modeling for Healthcare and Wellbeing; AI and Digital Human Modeling in Safety and Risk Management; and Biomechanics, Ergonomics, and Risk Mitigation
- Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management - Part II: User Experience Design for Sustainable Products and Public Spaces; and Wearable and Digital Health Monitoring
- Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management - Part III: Healthcare and Rehabilitation Innovation; Augmented and Virtual Reality for Health, Well-being, and Digital Human Modeling; and Behavioral Modeling and Human-Technology Interaction

The papers of these volumes were accepted for publication after a minimum of two single-blind reviews from the members of the DHM Program Board or, in some cases, from members of the Program Boards of other affiliated conferences. I would like to thank all of them for their invaluable contribution, support and efforts.

June 2025 Vincent G. Duffy

16th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management (DHM2025)

Program Board Chair: **Vincent G. Duffy**, *Purdue University, USA*

Karthik Adapa, *UNC-Chapel Hill, USA*
Sandra Alemany, *Biomechanics Institute of Valencia, Spain*
Alhamzah Alnoor, *Southern Technical University, Iraq*
Giuseppe Andreoni, *Politecnico di Milano, Italy*
Alfredo Ballester, *Instituto de Biomecánica de Valencia, Spain*
Joao Barata, *University of Coimbra, Portugal*
Matthew Bennett, *Humanetics, USA*
Josephine Bolaji, *Florida State University, USA*
Andre Calero Valdez, *University of Lübeck, Germany*
Yaqin Cao, *Anhui Polytechnic University, P.R. China*
Damien Chablat, *CNRS, France*
H. Onan Demirel, *Oregon State University, USA*
Yi Ding, *Anhui Polytechnic University, P.R. China*
Martin Fleischer, *EASE, Germany*
Ravindra Goonetilleke, *Khalifa University, United Arab Emirates*
Wenbin Guo, *University of Florida, USA*
Sogand Hasanzadeh, *Purdue University, USA*
Xueting Huang, *Swinburne University of Technology, Australia*
Aditya Jayadas, *Oklahoma State University, USA*
Bochen Jia, *University of Michigan, USA*
Genett Isabel Jimenez Delgado, *Institución Universitaria de Barranquilla, Colombia*
Jari Kaivo-oja, *University of Turku, Finland*
Chutisant Kerdvibulvech, *National Institute of Development Administration, Thailand*
Nikhil Krishnaswamy, *Colorado State University, USA*
Nicola Francesco Lopomo, *Politecnico di Milano, Italy*
Kapil Madathil, *Clemson University, USA*
Mufti Mahmud, *King Fahd University of Petroleum and Minerals, Saudi Arabia*
Siu Shing Man, *South China University of Technology, P.R. China*
Sumit Mandal, *Oklahoma State University, USA*
Alexander Mehler, *Goethe University Frankfurt, Germany*
Jörg Miehling, *Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*
Kristen Morris, *Colorado State University, USA*
Peter Nickel, *Institute for Occupational Safety and Health of the German Social Accident Insurance, Germany*
Miguel Ortiz-Barrios, *Universidad de la Costa CUC, Colombia*
Thaneswer Patel, *North Eastern Regional Institute of Science and Technology, India*
Martin Pohlmann, *Humanetics, USA*

James Pustejovsky, *Brandeis University, USA*
Ziwen Qiu, *North Carolina State University, USA*
Qing-Xing Qu, *Northeastern University, P.R. China*
Xingda Qu, *Shenzhen University, P.R. China*
Xiaodong Qu, *George Washington University, USA*
Arto Reiman, *University of Oulu, Finland*
Ann-Malin Schmidt, *Dresden University of Technology, Germany*
Deep Seth, *Mahindra University, India*
Yichao Shi, *Georgia Tech, USA*
Thitirat Siriborvornratanakul, *National Institute of Development Administration, Thailand*
Beatriz Sousa Santos, *University of Aveiro, Portugal*
Lesley Strawderman, *Mississippi State University, USA*
Leonor Teixeira, *University of Aveiro, Portugal*
Renran Tian, *North Carolina State University, USA*
Joseph Timoney, *Maynooth University, Ireland*
Vinay V. Panicker, *National Institute of Technology Calicut, India*
Dustin Van der Haar, *University of Johannesburg, South Africa*
Sonia Verdugo-Castro, *Universidad de Salamanca, Spain*
Baicun Wang, *Zhejiang University, P.R. China*
Kuan Yew Wong, *Universiti Teknologi Malaysia, Malaysia*
Shuping Xiong, *KAIST, South Korea*
Rachel Zuanon, *State University of Campinas, Brazil*

The full list with the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences of HCII 2025 is available online at:

http://www.hci.international/board-members-2025.php

HCI International 2026 Conference

The 28th International Conference on Human-Computer Interaction, HCI International 2026, will be held jointly with the affiliated conferences at the Montréal Convention Centre (Palais des congrès de Montréal), in Montreal, Canada, 26–31 July 2026. It will cover a broad spectrum of themes related to Human-Computer Interaction, including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer (part of Springer Nature) in a multi-volume set. More information will become available on the conference website: https://2026.hci.international/.

General Chair
Prof. Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
Email: general_chair@2026.hci.international

https://2026.hci.international/

Contents – Part I

Digital Human Modeling for Healthcare and Wellbeing

Integrating a Student Wellbeing Tool for Mentors 3
Nina Bergdahl, Jeanette Sjöberg, and Jalal Nouri

Verifying Scientific Validity and Enhancing Response Motivation in Digital Human-Based Mental Health Assessment: A Mixed-Methods Study 22
Chi-Cheng Lao, Litun Tan, Jing Liang, Ling Zeng, Zongze Li, Yitian Huang, Zi-Yi Guo, Yan-Dian Xu, Tianyang Tan, Xuetao Tian, Jun Liu, and Keliang Chen

Exploring Satisfaction and Comprehension of Pill Identification Applications: Insights on Design Improvements from Users with Varying Medication Usage Frequencies 41
Yu Syuan Lin and Meng-Cong Zheng

Evaluating Usability, Acceptance, and Trust in the BEAMER Adherence Intelligence Visualization Platform: A Preliminary Study 53
Beatriz Merino-Barbancho, Miguel Rujas, Rodrigo Martín Gómez del Moral Herranz, Kristina Livitckaia, Konstantina Pantelidou, Elena Castellano, Konstantina Kostopoulou, Maria Fernanda Cabrera, Maria Teresa Arredondo, and Giuseppe Fico

Constructing A Conversation-First Dialogue Model for Comprehensive, Evidence-Based Counseling for HPV Vaccination for Young Adults 72
Lu Tang, Chenying Weng, Serena Hou, Lara S. Savas, Ana C. Neumann, Nicole Moore, Jane Hamilton, Daniel Rhee, Cui Tao, Licong Cui, and Muhammad "Tuan" Amith

Research on User Experience Design of Internet Healthcare and Intelligent Triage Systems 85
You Yang Wu, Zi Xin Zhou, Lin Guo Jin, Zi Hong Ye, and Yong Yan Guo

How Social Cues in Depression Intervention AI Chatbots Affect Human-Computer Trust 97
Shuo Xu and Tiancong Ma

AI and Digital Human Modeling in Safety and Risk Management

Statistical 3D Body Shape Predictions for Standardisation of Digital Human Modelling Tools 121
Erik Brolin, Estela Pérez Luque, and Aitor Iriondo Pascual

Enhancing Virtual Restorative Environment with Generative AI: Personalized Immersive Stress-Relief Experiences 132
Kuei-Yuan Chang and Chang-Franw Lee

The Visibility and Non-intrusiveness of Peripheral Vision Information During the Use of AR Devices: The Effects of Transparency Levels, Color, and Font Size 145
Ding Chen, Kaige Wang, and Jinyun Cao

Design of a Hierarchical Federated Generative Learning Based Smart Home System 160
Paulo Drefahl, Kevin Kostage, Sean Peppers, Wenbin Guo, Luca Mazzola, and Chengyi Qu

ExpertGen: A Comparative Analysis of User Performance, Cognitive Workload, and Trust in Domain-Tailored Generative AI 178
Nina Jiang, Wei Zhou, Sogand Hasanzadeh, and Vincent G. Duffy

Assessing Vision Obstruction and Safety in E-Scooter Accidents: An Integrated Approach Using Digital Human Modeling and Computer Vision 191
Gabrielle B. Joffe, Yitong Bu, and H. Onan Demirel

The Application of Interactive Narrative in the Health Knowledge Visualization 210
Tian Lei, Yan Deng, Rui Yang, and Zhaoxu Li

Designing Automation for Enhancing Multitasking Performance 227
Jiaxin Li, Wenxin Xiao, and Ji-Eun Kim

Establishing a Fire Risk Map Based on Planned Spatial Layouts and Environment in Rural Planning: A Case Study of a Stereotype Village in China 238
Changda Ma, Yichao Shi, Chunlan Wang, and Patrick Kastner

Enhancing Human Detection in Post-disaster Scenarios Through Generative Adversarial Networks 257
Chengyi Qu, Xiaodong Hao, Feiyu Ma, Jizhuang Hui, Wenbin Guo, and Hong Wang

Biomechanics, Ergonomics, and Risk Mitigation

Research on Ergonomic Evaluation and Design Improvement of Grain Analyzer Based on Jack 275
Yu Chen and Meiling Zhou

Evaluating Multi-User VR for Ergonomic Assessments Using a DHM Software 289
Soureesh De, Henrik Söderlund, Jose Maria van der Ploeg Feriche, and Björn Johansson

Exploring Physical Demand Metric: Slouching Score in Augmented Reality-Based Biomechanics Education 301
Will Mastrantuono, Jung Hyup Kim, Siddarth Mohanty, Sara Mostowfi, Yuanyuan Gu, Fang Wang, Danielle Oprean, Yi Wang, and Kangwon Seo

Assessment of Ergonomic Posture in Augmented Reality Environments Using Slouching Score 312
Siddarth Mohanty, Jung Hyup Kim, Sara Mostowfi, Yuanyuan Gu, Fang Wang, Danielle Oprean, and Kangwon Seo

Biomechanical Evaluation of a Hybrid Upper-Body Exoskeleton During Lifting and Overhead Tasks in Occupational Settings: A Preliminary Study 323
Alessandro Piol, Emilia Scalona, Martina Mosso, Fatehia Bushara Garma Bushara, Maria Lucia Cavallo, Giacomo Valli, Noemi Pintori, Tiwana Varrecchia, Giorgia Chini, Alberto Ranavolo, Luca Falciati, Debora Brignani, Francesco Negro, and Nicola Francesco Lopomo

Impact of Exoskeleton Support on Discharge Statistics of Motor Units During Quasi-Isometric Tasks 342
Giacomo Valli, Martina Mosso, Maria Lucia Cavallo, Alessandro Piol, Emilia Scalona, Tiwana Varrecchia, Giorgia Chini, Alberto Ranavolo, Luca Falciati, Debora Brignani, Nicola Francesco Lopomo, and Francesco Negro

Decision Trees and IMU Sensors for Risk Prediction in Manual Material Handling 353
Hardik Vora, Fatemeh Davoudi Kakhki, and Armin Moghadam

Simulation and Analysis of Cervical Biomechanics and Injury Risk in Fighter Pilots Under Different Seatback Angles 365
Jiatao Wang, Chuancai Sun, Weisheng Jiang, Qianxiang Zhou, and Zhongqi Liu

Effects of Shared Control Modalities on Task Performance and User Experience in Robotic Teleoperation: A Path-Tracking Case Study 378
Dewei Zhu, Uhjin Hur, Seonghyeok Park, and Shuping Xiong

Author Index 391

Digital Human Modeling for Healthcare and Wellbeing

Integrating a Student Wellbeing Tool for Mentors

Nina Bergdahl[1,2(✉)], Jeanette Sjöberg[1], and Jalal Nouri[2]

[1] Halmstad University, Kristian IV:s väg 3, 302 50 Halmstad, Sweden
nina.bergdahl@hh.se
[2] Stockholm University, Borgarfjordsgatan 12, 164 55 Kista, Sweden

Abstract. Student wellbeing is recognised as a critical factor in supporting learning outcomes. However, when teachers are assigned expanded roles as mentors to monitor and support wellbeing, they often lack the necessary support to fulfil these responsibilities effectively. This study explores the tensions and behaviours that arise when teachers implement a novel wellbeing tool designed to provide insights into students' wellbeing. Adopting a mixed-methods approach, data were gathered through a survey and interviews with teachers (n = 39) and analysed through the lenses of Activity Theory (AT) and the Theory of Planned Behaviour (TPB). The findings reveal that integrating the wellbeing tool introduced new practices, resulting in primary and secondary tensions concerning roles, implementation, and use. The integration of TPB uncovered psychological factors influencing teachers' intentions and behaviours, complementing the systemic insights from AT. The study contributes to the broader conversation on enhancing educational practices to support student wellbeing, fundamental to achieving positive learning outcomes.

Keywords: Student Wellbeing · Teacher Mentorship · Wellbeing Tools · Technology Integration

1 Introduction

1.1 Student Wellbeing

Student wellbeing has emerged as a critical factor in education [1], not only because low levels of wellbeing can undermine academic achievements [2, 3] but because even though schools have an obligation to support student wellbeing, schoolchildren are experiencing recurring mental and somatic health issues [4]. Digital wellbeing tools may offer opportunities for teachers [5] and users (students) [6] to monitor wellbeing through real-time feedback and data-driven insights and identify students needing human interaction and support to promote wellbeing [7]. Although these tools offer potential, their successful implementation relies heavily on teachers [8, 9] and their skills to effectively use and act on insights provided. Furthermore, psychological factors can influence teachers' intentions and behaviours in relation to adopting wellbeing tools [10]. Student wellbeing has emerged as a critical factor in education, significantly influencing learning outcomes,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 3–21, 2025.
https://doi.org/10.1007/978-3-031-93502-2_1

engagement, and overall development [1]. Wellbeing is included as a significant indicator of the quality of school life [2] and has also been linked to educational performance [11]. Low levels of wellbeing, on the other hand, can undermine both academic achievements [3] and future success [2]. Teachers play a pivotal role in this process, being uniquely positioned to support and enhance student wellbeing through their daily interactions and pedagogical practices [8] and many teachers also believe that positive emotions are important for helping students to achieve academic success, and support for wellbeing as an educational issue span varied cultural, social and economic contexts across Asia, Latin America, North America, Australia and Europe. Turning to Sweden, a survey on student wellbeing revealed that many schoolchildren experience recurring mental and somatic health issues, such as irritability, moodiness, sleep difficulties, headaches, or stomach aches—symptoms that are more prevalent among girls than boys [4]. Digital wellbeing tools, such as mobile apps and online platforms, can facilitate real-time feedback, personalised interventions, and data-driven insights into students' emotional states building. For example [5] demonstrated that digital mental health interventions could reduce symptoms of anxiety and depression among adolescents. Other tools aimed at supporting student wellbeing, offer opportunities for teachers to identify and respond to students' emotional and mental health needs [5, 12]. To this end, data gathering is important to identify students' mental and emotional health and track whether supportive measures are working [13]. However, the successful implementation of these tools depends heavily on teachers, who may lack the necessary support and training to effectively utilise these technologies [14] as skill development has lagged behind [15]. Or they may encounter other kinds of challenges, stemming from tensions between their traditional roles and new responsibilities [9]. Furthermore, psychological factors, such as attitudes towards technology, perceived behavioural control, and subjective norms, can influence teachers' intentions and behaviours regarding the adoption of wellbeing tools [10]. [9] called for research to better understand influences on teachers' choice of practices in relation to (general) technology integration and adoption. Turning to wellbeing tools more specifically, researchers have noted that there is a scarcity of studies on wellbeing tools within the context of (higher) education [16] and mobile phone-based interventions for youth and adolescents [6]. Our society is currently engaged in a debate about whether technological innovation can positively impact student well-being while respecting the fundamental need for human interaction, with calls to develop monitor and response applications to meet the needs of youth and adolescents, as technologies has been found to "lessens the stress of engagement for some students" [8, 17]. Researchers have call for studies on the integration of wellbeing tools in educational settings [16]. This study aims to fill this gap, and we raise the following research question: What tensions arise when teachers implement a data-driven wellbeing tool in K-12 schools, and how do these tensions affect teacher intentions and behaviours?

2 Background

2.1 Wellbeing and Mentors

While UNESCO emphasises the importance of creating environments that foster not only academic success but also the physical, social, and emotional development of students [18]: there is no single definition of wellbeing. There is a broad concept of wellbeing, however, and this covers two dimensions: the emotional: referring to the presence of positive, and absence of negative affective states) and cognitive [19, 20]. In recognition of wellbeing's impact on learning, and the education system's outsized role in shaping the emotional and mental health of young people, the OECD has measured student wellbeing as part of its Program for International Student Assessment (PISA) ranking since 2015 and defines wellbeing as "the psychological, cognitive, social and physical functioning and capabilities that students need to live a happy and fulfilling life" [2, 60–64]. Emotional wellbeing is fundamental to students' overall development and significantly impacts their academic performance and social interactions. To foster wellbeing, schools are to promote caring pedagogies, that emphasise supportive relationships, empathy, and emotional connections in the educational settings [21]. Teachers who implement caring pedagogies, support students' emotional needs through daily interactions and instructional practices [8]. Positive emotional interactions between teachers and students influence wellbeing [22] and may directly impact on the student working memory [23]. Such interactions can be included in directed mentoring-interventions, and these can be structured differently: for example, as one-to-one mentoring, group mentoring, or a peer mentoring model [24, 25]. Mentoring can also be aimed towards different cohorts of students. For example, the Big Brothers/Big Sisters initiative in the USA aimed to support and connect young individual with adults who would provide caring attention [26] while STEM mentorship programs were launched to support gifted students [27] or underrepresented students with an aim to stimulate female interest in STEM subjects [28]. In the United States the development of youth mentorship programs began during the twentieth century [26]. In Sweden, student mentors became common following legislative amendments in the 1990s [29]. In addition to the generic teacher responsibilities like teaching, assessing, communicating with students, parents and colleagues. Because wellbeing, is influenced by the climate, teachers were expected to direct some effort to caring pedagogies and empower their students, supporting social interaction and relationship building. Why schools introduced a scheduled a weekly mentor time which would typically include activities related to providing guidance, support, and supervision from a mentor-teacher to a select group of students. In addition to the weekly mentoring time, the mentors have an extra responsibility to engage in relationship building talks with their select group of students, to ensure their healthy levels of wellbeing and progression, or to initiate interventions in case of need [30]. According to a report, Swedish primary school teachers spend approximately 9% of their total work-related time on mentoring and class leadership, support and guidance, supervision, responsibility, and student care [30]. The part of caring pedagogies made up of "small talk or guidance with one or a couple of students about teaching" accounted for about 1% of the teachers' total work time [14, 42]. The report notes that, it is unclear what caring pedagogies exactly include, and therefore it can be hard to calculate the exact

time devoted. For example [30] suggest that caring pedagogies also occur outside the classroom, including comforting, supervising breaks, educational lunches, and excursions. In short, the demands on teachers to know how their mentor-group is progressing and how they are feeling when they are in school, is increasing.

2.2 Perspectives on Technology Integration in Education

Throughout the digitalisation of dedicated technologies have been developed to support teachers, but these technologies might influence organisational hierarchy and power distribution [31] as digital systems change the ways in which actors within organisations get and use information [32]. While some technologies can streamline operations and improve efficiency, they may also influence roles and practices [31]. School leaders, teachers, and students may have different intentions and desired outcomes, where teachers can experience a mismatch between policy and classroom realities [9]. [33] found that one reason that teachers did not integrate (general educational) technologies into their teaching was due to excessive workload, whilst factors like infrastructure, professional development and technical support were important, they particularly emphasise the need for teachers to be involved in the decision-making processes. [34] discuss how teachers' beliefs, confidence, and cultural contexts influence their willingness to integrate technology into their teaching practices. They emphasise that teachers must adjust to new demands and responsibilities when adopting innovative tools, which can affect their attitudes and intentions. However, teachers' conceptions will still influence their technology acceptance, with positive beliefs leading to higher adoption rates [35]. Turning to wellbeing tools [6] conducted a review and found that most wellbeing tools offer some kind of support to the user. These tools use trace data or self-reported data, as (learning) analytics can providing data about students with an aim to support them [36]. Results from students' perspective is disperse, on the one hand, studies in the UK and Colombia found that wellbeing tools can contribute to increased understanding and emotional awareness for the students [37] and be promising when combining in-app and human support [38]. While other studies show that the students feel that the tools do not meet their needs, are sceptical about the advice proposed and are unmotivated to use them e.g., [6, 16]. Other kinds of wellbeing tools aim to identify students needing human support, and do not aim to replace human interaction [7]. As wellbeing tools collate data, teachers need not only to be willing to adopt them, but they also need data literacy skills, which has been found lacking [15] meaning that adequate training on how to interpret and act on the insights should be improved [14]. While these results present important aspects they do not explore integration tensions and influences, which may also influence effectiveness and use of wellbeing tools.

2.3 Theoretical Frameworks

To contribute with insights, this study employs a dual theoretical framework comprising Activity Theory (AT) and the Theory of Planned Behaviour (TPB). AT is a socio-cultural framework that examines human practices as complex, socially situated activities [39, 40]. AT focuses on the interactions between individuals (subjects), their goals (objects), and the tools or artefacts they use within a community, governed by rules and division

of labour [41]. In the context of education, AT provides a lens to understand how the introduction of new technologies, such as wellbeing tools, can disrupt existing activity systems, leading to contradictions and tension [31]. This study examines the systemic tensions that arise when teachers integrate a wellbeing tool into their practice. It explores how the tool (artefact) influences teachers' roles (subjects) and their objective of supporting student wellbeing (object), considering the community (students, colleagues, administrators), rules (policies, norms), and division of labour (responsibilities). TPB, on the other hand, posits that an individual's behaviour is determined by their intention to perform the behaviour, which in turn is influenced by attitudes, subjective norms, and perceived behavioural control [10]. Intentions are shaped by three factors: attitudes toward the behaviour, subjective norms, and perceived behavioural control. TPB has been widely used to predict and understand behaviours in various contexts, including technology adoption [35]. In educational technology integration, TPB helps to explain how teachers' beliefs and perceptions influence their willingness to adopt new tools [42]. Teachers' attitudes toward the wellbeing tool, perceptions of support from colleagues and administration (subjective norms), and confidence in their ability to use the tool effectively (perceived behavioural control) can all impact their intention to implement it. While AT provides a lens to examine systemic contradictions within the educational setting [39], TPB offers insights into the psychological factors influencing teachers' intentions and behaviours [10]. By combining these theoretical perspectives, we explore tensions, teacher roles and how teachers' beliefs, attitudes towards the wellbeing tool, perceptions of support, and confidence in their abilities affect their adoption and utilisation of the technology.

3 Method

3.1 Participants and Context

In Sweden, most schools use a learning management system and have internet access with BYOD settings; teachers and students typically have at least one laptop or tablet. Classrooms are equipped with projectors and may also have smart boards and other digital devices. While teachers use various digital resources during class, there are few technologies available to support their roles as mentors. This study is set within the context of a spearhead integration of a digital resource: a wellbeing tool designed to assist teachers in supporting and overseeing student wellbeing. While the teachers access the tool online, the students access the tool via their mobile phone application. We interviewed teachers on their experiences using this digital tool. Data was collected from 13 Swedish upper secondary schools, between January and April 2023, using a web-based survey and in-depth interviews. Employing purposive sampling [43] we only invited teachers who used the tool. Teachers ($n = 39$) responded to the survey, which explored background demographics and expectations and perceived value using Likert scale questions (where "not at all" and "all the time" reflected the values 1–5). We then conducted in-depth interviews ($n = 6$) with teachers. A total of 204 min were recorded and transcribed verbatim. The open-text answers and the transcriptions of the interviews were analysed using AT. All participants provided informed consent before data was collected. This study follows ethical guidelines [44].

3.2 Data Analysis

Statistical analysis on the survey data was conducted using Python 3.10.10 and the packages Pandas, Matplotlib, and Seaborn. The interviews were transcribed and collated with survey data to be coded using the AS in Fig. 1. The data was then mapped using the lens of AT [39]: based on the model proposed by Engeström [45], the "subject" in our AT analysis represents the teachers and their activities, the "artefact" is the wellbeing tool, and the overt "object" is student wellbeing. The "community" is multifaceted, encompassing micro-level participants such as teacher mentors and students, meso-level entities like researchers, developers, and student health counsellors, and macro-level stakeholders such as school leaders and decision-makers. The "rules" encompass local norms, routines, and legislation, while the "division of labour" defines the responsibilities of different parties in undertaking the activities aimed at achieving the object. It is significant to note that the object's understanding may not be uniformly explicit across all participants. It can be explicit or subconscious, shared or individual, consistent, or fragmented, stable, or emerging [45].

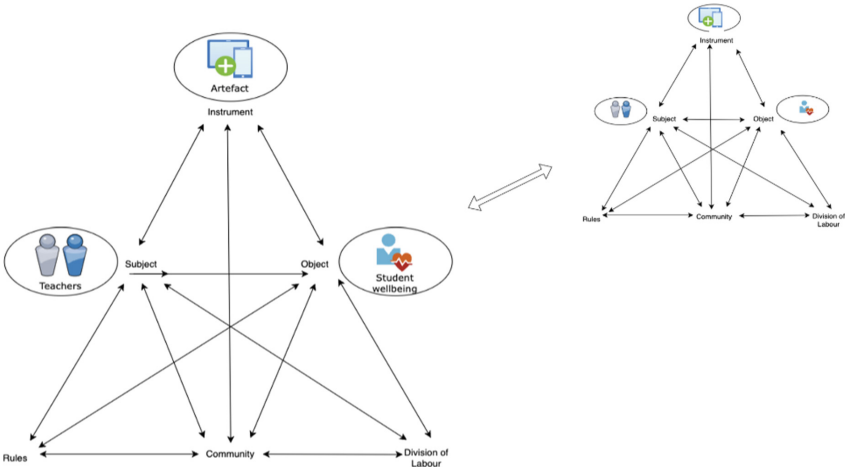


Fig. 1. Interaction between Activity Systems.

Figure 1 shows interaction between Activity Systems (AS) in the spearhead implementation of the wellbeing tool. While Subjects and objects only have primary contradictions, instruments, rules, and division of labour have primary and secondary contradictions (marked as one and two arrows distance respectively). Activities and their associated objectives may challenge and transform the existing norms (rules) and allocations of responsibilities (division of labour). The depiction of two AT triangles represents the dynamic and sometimes contradictory interactions between different ASs within this educational context [45]. The double triangle model reflects the realities of educational environments where multiple ASs coexists, interact, and negotiate their positions. It emphasises the importance of considering the complexity and dynamism of real-world settings when integrating novel tools like the wellbeing tool in this study. The process

of mapping within the AS involved several systematic steps to analyse the data: 1. We employed latent content analysis to code the qualitative data. 2. We categorised the activities and objectives identified, grouping similar themes and actions to understand the patterns within the data. 3. We explored primary contradictions (tensions within a single element) and secondary contradictions (tensions between elements) by examining relationships such as between the subject (teachers) and the tool (wellbeing app), or between the division of labour and rules. 4. We analysed the data through the lens of TPB by identifying expressions related to attitudes, subjective norms, and perceived behavioural control in parallel. We then integrated the findings from both theoretical frameworks to explore systemic tensions and psychological factors.

4 Results

This section presents the findings of the study, integrating both AT and the TPB to explore the systemic tensions and psychological factors influencing teachers' adoption and use of the wellbeing tool.

4.1 Expected and Perceived Values

The correlation heatmap (Fig. 2) provides a visual representation of the correlation between different Likert scale questions in the survey data. Initial expectations of the tool (A) moderately align with current perceptions (B, $r = 0.31$), and these perceptions are strongly associated with the desire to continue using the tool (C, $r = 0.62$). Users who express a desire to continue using the tool (C) generally hold positive views on its questions (D, $r = 0.46$), and those who appreciate the questions also value the categories in the tool's overview (E, $r = 0.69$). This strong correlation between the tool's questions and categories indicates that these elements are integral to user satisfaction and suggest that positive attitudes toward the tool influence teachers' behavioural intentions. Interestingly, while there is a slight tendency for users who appreciate the categories (E, $r = -0.021$) and those with higher initial expectations (A, $r = -0.043$) to prefer less frequent questionnaires (F), users who have positive opinions about the tool's questions (D, $r = 0.12$) are slightly more comfortable with answering wellbeing questions once a week. Furthermore, users with higher initial expectations (A) are more likely to want to continue using the tool (C, $r = 0.41$). Current perception of using the tool (B) have a moderate positive correlation with both opinions on the tool's questions (D, $r = 0.49$), and opinions on the categories (E, $r = 0.43$), suggesting that users who have a positive perception of the tool also tend to have positive opinions on the underlying categories and questions. The findings indicate that, within our data, the frequency of using the tool (once a week) was generally acceptable to teachers and did not adversely affect their attitudes toward the tool.

4.2 The Subjects' Primary Contradiction

Implementing the tool required teachers to engage in additional activities beyond their usual practice, revealing tensions related to their perceived behavioural control—a construct in TPB referring to individuals' perception of their ability to perform a given

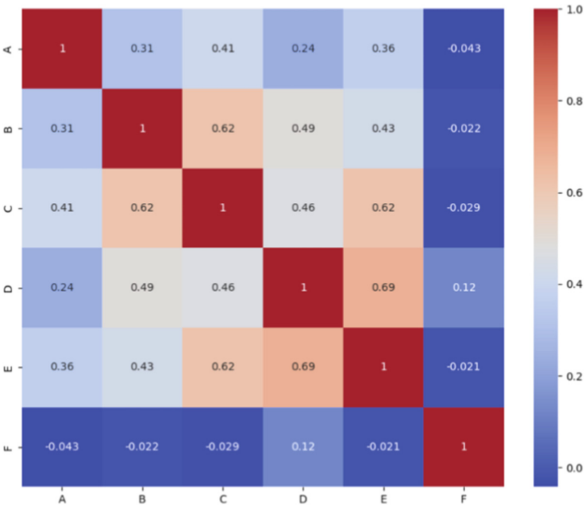


Fig. 2. Correlation heatmap. Figure 2 shows a correlation heatmap with the following questions A: "What were your expectations before using the wellbeing tool?" B: "What is your current perception of using the wellbeing tool?" C: "I would like to continue using the wellbeing tool." D: "What do you think about the questions in the wellbeing tool?" E: "What do you think about the categories in the wellbeing tool's dashboard?" F:15. What do you think about using the wellbeing questions once a week? The values in the correlation matrix range from −1 to 1 (−1 is perfect negative, 1 is a perfect positive, and 0 is no correlation)

behaviour [10]. These tensions stem from systemic contradictions in the elements of the activity system, such as the instrument, objective, division of labour, and rules and norms.

Instrument. The introduction of the wellbeing tool as an instrument in the educational setting elicited varied attitudes among teachers. Some teachers recognised its potential to enhance their mentoring role, which positively influenced their attitudes and intentions to use the tool:

> ... then I would have liked us to create a better routine around how it should be used so that it could be used fully and that the students would have received a different introduction to the tool that made them also understand the purpose of it. (Interview, teacher 3)
>
> I see that it facilitates my role as a mentor to gather information, get an overview and be able to measure trends. (Interview, teacher 1)

There is a notable contradiction between the tool's intended and actual use, which varies among teachers. This discrepancy points to a potential contradiction. Several of the teachers talked about the merits of the tool to gain insights to information that they might otherwise not have access to. While several teachers expressed positive attitudes are crucial determinants of behavioural intentions according to TPB [10, 35]. However, the

lack of consistent routines and student understanding indicates challenges in perceived behavioural control.

Objective. A significant tension emerged regarding the objective of using the tool, affecting teachers' attitudes and perceived behavioural control. Teachers grappled with the expanded role of supporting student wellbeing, which sometimes conflicted with their professional identity and available resources:

> It is important to convey that we are not all at our best all the time and that it is ok to feel bad or have a bad week. The students have become more aware of how they feel, which is usually positive, but in our class it has also become that some feel sorrier for themselves even though everything is fine. (Survey)
>
> What do you do with the answers you get? Therein lies the biggest challenge. (Survey)
>
> Where is the limit for my job here? And my mentorship is 40 minutes a week and in addition to all other work. I also have to work, like this, it's so much, it's so much as a mentor role, not only as a teacher […] there are so many different types of social and psychological problems in young people today that makes you sometimes feel a little…where is the limit for what is my profession? (Interview, teacher 3)
>
> I didn't become a teacher to talk about wellbeing with teenagers. I don't feel super comfortable with it. I don't feel competent doing it. I am very worried about whether I can handle that well (Interview, teacher 1).

A clear contradiction arises here between teachers' professional identities as educators, primarily responsible for imparting academic knowledge, and their newfound roles as support systems for student wellbeing. On the other hand, there is a risk in that by focusing on this the students will reflect a little too much and thus also focus negatively and think they feel worse than they actually do. These reflections reveal that negative attitudes toward the tool arise when teachers feel that it blurs professional boundaries or adds to their workload, which can decrease their intention to use it [10].

Division of Labour. The unclear division of labour contributed to contradictions that affect subjective norms and perceived behavioural control. Teachers expressed uncertainty about their responsibilities concerning student wellbeing, influencing their intention to engage with the tool:

> How much time should this take compared to other tasks? Also, this creates pressure on me as a teacher to take responsibility for how the students feel, which can be a good and a bad thing. Certainly, good for the students, but the risk of burnout also increases when I start caring for their wellbeing after my working day. (Survey)
>
> I'm not entirely convinced that it should be a task for me, but then we are in for a deeper discussion about what mentorship actually is. (Interview, teacher 1)

These concerns highlight the impact of subjective norms—teachers may feel that there is no collective agreement or support for this expanded role, which can reduce the social

pressure to adopt the tool [10, 34]. Additionally, perceived behavioural control is affected by concerns about workload and capacity.

Rules and Norms. The absence of clear guidelines and institutional support influenced teachers' subjective norms and perceived behavioural control. Without explicit expectations, teachers were uncertain about the importance of using the tool:

> It requires clear anchoring in school and recurring work regarding follow-up. Clearer tools and tips on how this work might look would be needed here. (Survey)
>
> [School leader] offered anyone who was interested in starting with it, so it is not a focused practice, nor any time set aside […] I know that there are several other teachers who also use this to varying extents. But … [there were] no guidelines around that. (Interview, teacher 1)

This lack of institutional support can weaken subjective norms, reducing the likelihood of teachers adopting the tool [42]. It also impacts perceived behavioural control, as teachers may feel they lack the necessary support and resources.

4.3 Alternative objectives

When subjects have different formal or informal objectives, this may influence their actions and cause contradictions. In this section, we explore objects that the respondents referred to, other than the formal (to support student wellbeing). These objectives do not necessarily replace the formal objective and may be regarded as complementary. We identified the following five objectives: Accessing and recognising insights, Metacognitive support, Challenging mentoring traditions, Data literacy for teachers, and Data-driven decision-making, elaborated on below.

Accessing and Recognising Insights. The ability to gain new insights into student wellbeing positively influenced teachers' attitudes toward the tool:

> There are potential benefits in detecting ill-health, but for our part it hasn't made a big difference compared to how we have worked before. I think the big advantage is in how it shapes students' thinking by focusing on wellbeing somewhat more clearly than before (Survey)
>
> I don't have to use my gut feeling here, because I can look at how the students have answered, and then we talk about it. So, I feel that it is of course some help for me as a mentor to have this overview. (Interview, teacher 2)
>
> The data itself allows us teachers to discover things that we had not had access to before (Interview, teacher 5)

Positive attitudes enhance the intention to use the tool [10]. However, some teachers also expressed uncertainty about handling these insights, affecting their perceived behavioural control.

Support for Wellbeing. The tool facilitated discussions on wellbeing, impacting subjective norms and attitudes:

> We [the teachers] talked a lot about our approach to support wellbeing, of course, but then we could also go out with emails to their teachers and say, right now it's tough for this class: is it possible to be flexible, is it possible to even out the workload a bit? But then, we also invite students to be involved, influence and we talk about what it means to influence and to take the chance when the teacher says, "what do you want to work on now"? Do you take the chance to influence? And yes, then there were some who did it, but others didn't, so then I also talked about what one's responsibilities are. I thought it was a good discussion. (Interview, teacher 2)

> If a student is starting to feel better, it's great fun to see it, and also for the student to get statistics on this. "But now I see here that now, now I'm looking at trends where you're now on the way up" and that can also give or be a pushing effect which kind of leads to this with what I see that things are starting to happen here now. So maybe you will be encouraged to continue [making on effort] just for today. (Interview, teacher 4)

The teachers emphasised the importance of communication and collaboration when addressing challenges. Their strategies included emailing other teachers to discuss workloads and scheduling flexibility. Such collaborative efforts and positive outcomes can strengthen subjective norms by creating a supportive environment, enhancing the intention to continue using the tool [34].

Challenging Mentoring Traditions. Implementing the tool challenged traditional practices, affecting teachers' attitudes and perceived behavioural control:

> It is becoming difficult to get the students to fill in each week. They need to notice more of what results it can give in order for them to be motivated to continue. All my planning and my way of working need to be based on [the wellbeing tool], which it has not done now, in order for it to be completely relevant. (Survey)

> The mentoring task is in some way without a distant limit, so it can be filled with anything possible, for example, that you should hold a theme day with your students. So, I think it's probably mostly like that how you find the time to implement those changes, which can be difficult. But we have a superb student health team, so I think there is also a lot of collaboration there. We have regular student health meetings where we talk in the class. Now this also becomes a wellbeing app as a natural part of it, so I think that if you want to come to some kind of change, you also need a little help from others at the school. (Interview, teacher 2)

The teacher emphasises the importance of aligning planning and teaching methods with the insights gained from the tool to make it truly relevant. Implementing changes and finding time to incorporate them can be challenging, but collaboration with the student health team and other teachers can facilitate this process.

Data Literacy for Teachers. Challenges in data analysis impacted teachers perceived behavioural control:

> Difficult to know how to analyse the statistics and know what the next step should be (Survey)

> What should we do with the information and what should we do with all the data? And where is the time for this so that we will have time to actually do it. (Interview, teacher 3)
>
> Maybe some people don't want to answer, you could say some maybe mark very, very, that is, the same value every week. And then you get the impression that maybe they don't really think, or they really don't have a problem because they really have the lowest level of anxiety and depression and the highest value of wellbeing and motivation and everything. And then it's either that they're dark for me or that they're actually very carefree people. It's hard to tell sometimes. (Interview, teacher 5)

There is a concern among teachers about finding the time as well as to understand the data and determining the appropriate next steps. Limited confidence in data literacy skills can reduce perceived behavioural control [15].

Data-Driven Decision Making. Some teachers saw the tool as valuable for advocating institutional changes, positively influencing their attitudes:

> But we have never displayed any [wellbeing] charts to the whole class. However, we have said that we have seen that "many of you who answer says that there is a lot of workload and stress" and things like this now…when we make changes, we try to connect it to that. […] This should help us get a basis and make those decisions. (Interview, teacher 1)
>
> I think it is useful for school management […] for those deciding on the school structure, senior managers and politicians. When teachers raise things that we see as problems, it is usually pushed aside with "There is no money for it", but I undeniably think it will have an effect as I can say, "Look. Here we have 300 young people who feel very badly. They answer it week after week. Should we do something different, or should we just continue as we have been doing before?" (Interview, teacher 1)

Tools like this are valued because they provide tangible evidence of the need for change, particularly when resources are involved. When teachers have numerical data to support their discussions and decision-making processes, their concerns have increased credibility. The teachers highlight that even if individual teachers would not use the tool, it would still hold significant value for school administration, structural planning, senior management, and policymakers.

4.4 Secondary Contradictions

Secondary contradictions arose in elements like the instrument, rules, and division of labour, affecting teachers' attitudes, subjective norms, and perceived behavioural control.

Features. Teachers observed that the tool's features provided benefits, influencing their attitudes:

> [We get] frequency, trends, across school equality. [This] contributes to documentation. Long-term statistics. (Survey)
>
> Clear [data] during parent-teacher meeting. Good to get an overview of the student group. (Survey)
>
> The difference is that we ask all the students this way. Instead of just asking a few [students], we ask everyone every week. In normal cases you might not pay attention to those who, uh, feel bad. […] Perhaps, that we get a broader overview (Interview, teacher 5)

Teachers highlighted the benefits of gaining quick overviews of student wellbeing, which can help to ensure equal support and a more instrumental understanding of any potential issues. Perceived time constraints and the tool's integration into practice affect perceived behavioural control and attitudes, influencing the intention to use it.

Efficiency and Work Optimisation. While the tool aimed to optimise work, some teachers were concerned about its impact on perceived behavioural control:

> This tool gives me a good overview, I think, so that I can still see if there is someone who crosses over a lot of red, so of course I can try to find that student. But there is definitely a possibility that I will see things that I might not have time to discover when we have an 80-minute, quite packed lesson. (Interview, teacher 2)
>
> Maybe the students […] don't pay so much attention to it because they don't see that we talk about it [wellbeing] […] then they think: "Well, we're only spending a minute on this, so it can't be particularly important", and that, in turn, can cause them not to prioritise using the tool. (Interview, teacher 5)

Many teachers expressed a desire to utilise the tool as a time-saving method while acknowledging the responsibility involved. However, some teachers also remarked, that when time is optimised, students may not perceive the aim as substantial or important, and while teachers in an instant gain insight, they might not have the time to act.

Replacing Teachers. There were concerns about the tool potentially replacing personal interactions, impacting attitudes:

> Another disadvantage may well be that it is seen as a substitute for direct contact with the students and monitoring of their wellbeing. (Survey)
>
> I think that sometimes it can be easier to enter an answer than to say straight out "'m not feeling well" or "I'm stressed". […] but it's telling someone face to face that they [the students] are unused to. Then it's easier to click in. (Interview, teacher 3)
>
> We're aiming for this tool not to replace personal interaction […] in the best of worlds, maybe this can initiate a conversation earlier than it would have otherwise. (Interview, teacher 5)

The use of the tool as a time-saving measure presents potential drawbacks, such as the risk of substituting direct student contact and close monitoring of their wellbeing. While

some teachers note that students might find it more comfortable to express their emotions through the tool rather than in face-to-face interactions, as also noted by [17].

Transforming Educational Practices. Teachers reflected on the broader changes in educational practices, affecting their attitudes and perceived behavioural control:

> [I] never imagined that I would have this kind of conversation or control. […] Spending this amount of time on it and thinking about it [wellbeing] so much. At the same time, it is a vanishingly small part of my work. It barely exists on paper. It is just something that is included. (Interview, teacher 1)
>
> It's not at all the same job I started with 25 years ago! So much administrative work has been added […]. But above all, I think that the mentorship is growing. More young people suffer from mental illness and then we also see more and more diagnoses. There are many more who are unwell, and more who have diagnoses. There weren't nearly as many before [we maybe had] 25%, I would say. (Interview, teacher 2)

Teacher roles have changed significantly over time, especially when it comes to administrative work and mentorship, as mental health issues have become more prevalent among students. Adjusting to these changes can impact teachers' attitudes toward new tools and their confidence in managing additional responsibilities, influencing their intention to adopt innovations [34]. Moreover, it represents a broader challenge in education, as teachers are left to balance the traditional role with a changing role for the mentorship with growing mental health concerns.

5 Discussion

Our research sought to understand the specific tensions arising from the implementation of a wellbeing tool in schools and the consequent effects on teacher intentions and behaviours. The data revealed that the introduction of the tool disrupted existing activity systems, leading to both primary and secondary contradictions within the educational context. The integration of the wellbeing tool introduced new artefacts into the teachers' activity system, challenging established practices and norms [39]. Teachers grappled with the dual role of educator and mentor, a shift that blurred traditional professional boundaries. This finding resonates with previous assertions [9] that technology integration can exacerbate tensions between policy expectations and classroom realities. The contradictions identified align with the concept of expansive learning, where tensions serve as catalysts for systemic change [41]. For instance, the unclear division of labour and lack of institutional guidelines led to confusion and frustration among teachers. This echoes earlier findings [31] showing that new technologies can disrupt organisational hierarchies and power distributions, necessitating a re-evaluation of roles and responsibilities. Furthermore, the wellbeing tool functioned as more than just an artefact; it influenced other elements of the activity system, such as community norms and the division of labour. This multifaceted impact supports Spinuzzi and Guile's (2019) observation that advanced technologies can act simultaneously as platforms and environments, challenging traditional AT frameworks.

From the TPB perspective, teachers' attitudes, subjective norms, and perceived behavioural control significantly influenced their intentions to adopt and utilise the wellbeing tool [10]. Positive attitudes were observed when teachers recognised the tool's potential to enhance their mentoring capabilities and gain valuable insights into student wellbeing. This aligns with [35] who found that positive beliefs about technology facilitate higher adoption rates. However, perceived behavioural control emerged as a substantial barrier. Teachers expressed concerns about their ability to interpret data effectively, manage increased workloads, and navigate their expanded roles without adequate support. This lack of confidence hindered their intention to use the tool, corroborating [15] emphasis on the importance of data literacy training for educational professionals. Subjective norms also played a critical role. Extending research that has pointed to a need for more support; [34] we found that the absence of clear institutional expectations was demotivating for adoption, and teachers who despite this integrated the tool, reported feeling isolated in their efforts. While previous studies have addressed the challenges of adopting educational technologies [14, 33] our research provides a nuanced understanding of how these challenges specifically manifest in the context of wellbeing tools [7]. found that learning analytics tools can support teachers in identifying at-risk students but also noted the necessity for these tools to complement rather than replace human interaction. Our study supports this view, with teachers expressing concerns about the wellbeing tool potentially substituting personal engagement with students. This underscores the delicate balance between leveraging technology for efficiency and maintaining the relational aspects essential to education. Moreover, while [17] observed that technology can reduce engagement stress for some students, our findings suggest that this benefit may be offset by teachers' discomfort and lack of preparedness to handle the emotional complexities revealed by such tools. This dichotomy highlights the need for comprehensive support systems that address both student and teacher needs.

Emerging technologies are introducing new dynamics within educational activity systems, provoking shifts in roles and responsibilities. [46] explored these shifts in the context of augmented and virtual reality technologies, suggesting that such technologies can foster new forms of agency and create "augmented" subjects with capabilities surpassing those of traditional roles. Our findings extend this notion by demonstrating that wellbeing technologies in education, like the tool studied, can assume multiple and parallel roles due to their multifaceted functionalities. This reflects the tensions and contradictions highlighted in AT when new artefacts disrupt existing systems [31, 39]. In particular, our study reveals that the wellbeing tool not only functions as an artefact (instrument) but can also influence other elements of the activity system, such as the division of labour and community norms. Furthermore [39] discussions on the multifaceted roles of artefacts are important here; the wellbeing tool can assume multiple functions within the activity system, influencing other elements such as rules and division of labour. By integrating the TPB [10] with AT, we contribute a novel perspective that accounts for both systemic and psychological dimensions. The psychological factors influencing teachers' adoption and utilisation of the tool—such as attitudes, subjective norms, and perceived behavioural control—interact with systemic tensions to shape outcomes. This integrated approach addresses calls in the literature for frameworks that capture the complexities of technology integration in education [9, 34]. Our theoretical contribution

lies in demonstrating how the multiplicity of roles assumed by advanced educational technologies necessitates an expansion of traditional AT frameworks. We suggest that incorporating psychological perspectives provides more insights into the influences of technologies in educational practices.

5.1 Limitations, Conclusion and Future Research

Despite the valuable insights gained, this study has several limitations. The sample size was relatively small (n = 39), and participants were from a specific context within Swedish upper secondary schools, which may limit the generalisability of the findings. Additionally, the spearhead implementation approach means that the experiences may not reflect those in a fully integrated system. In the context of AT, as proposed by this wellbeing tool can be seen as a new artefact introduced into the system of activity that is the educational setting. Teachers, the subjects of this system, are to use this tool (the artefact) to achieve a common goal: enhancing students' self-awareness and skills (the object), thereby contributing to their overall wellbeing. This decentralised spearhead implementation approach presents opportunities for expansive learning, a key concept in AT theory [39, 41]. By being given the freedom to experiment with and explore the tool's usage, educators engage in a learning process beyond the traditional acquisition of knowledge. They are exploring new ways of thinking and working (expanding) as they encounter tensions and contradictions that arise from introducing this new tool into their established routines and practices. While tensions, here addressed as contradictions, are crucial elements AT theory, essential for driving change and development, we found that the tool may take on the roles of different elements in the AS. Expanding the affordances of previous technologies, the tool provokes change and is a potential source of learning and development that could be guided systematically to better harness insights, and support wellbeing. We therefore suggest that future research explores shared agency and roles between humans and advanced technologies.

Disclosure of Interests. The authors declare having no competing interests.

References

1. World Health Organization [WHO].: Adolescent Mental Health (2021). www.who.int/news-room/fact-sheets/detail/adolescent-mental-health
2. OECD: PISA 2015 Results (Vol. III): Students' Wellbeing. OECD Publishing, Paris (2017). https://doi.org/10.1787/9789264273856-en
3. OECD: PISA 2022 Results (Vol.II): Learning During and From Disruption. OECD Publishing, Paris (2022). https://doi.org/10.1787/a97db61c-en
4. Public Health Agency of Sweden: School children's health habits in Sweden 2021/22. In: Skolbarns hälsovanor i Sverige (2023) (Art no: 22228). https://bit.ly/3BVLg0D
5. Wasil, A.R., Venturo-Conerly, K.E., Shingleton, R.M., Weisz, J.R.: A review of popular smartphone apps for depression and anxiety: assessing the inclusion of evidence-based content. Beh. Res. Ther. **123**, 103498 (2019). https://doi.org/10.1016/j.brat.2019.103498

6. Goldberg, S.B., Lam, S.U., Simonsson, O., Torous, J., Sun, S.: Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLoS Digit. Health. **1**(1), e0000002 (2022). https://doi.org/10.1371/journal.pdig.0000002
7. Henriquez, V.A., Guerra, J., Scheihing, E.: The impact of an academic counselling learning analytics tool: evidence from 3 years of use. Br. J. Edu. Technol. **55**(5), 1884–1899 (2024). https://doi.org/10.1111/bjet.13474
8. Brackett, M.A., Katulak, N.A.: Emotional intelligence in the classroom: skill-based training for teachers and students. In: Ciarrochi, J., Mayer, J.D. (eds.) Applying Emotional Intelligence: A Practitioner's Guide, pp. 1–27. Psychology Press (2013). https://doi.org/10.4324/9781315782935
9. Abedi, E.A.: Tensions between technology integration practices of teachers and ICT in education policy expectations: implications for change in teacher knowledge, beliefs and teaching practices. J. Comput. Educ. **11**, 1215–1234 (2023). https://doi.org/10.1007/s40692-023-00296-6
10. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991)
11. Papamitsiou, Z., Economides, A.A., Giannakos, M.N.: Fostering learners' performance with on-demand metacognitive feedback. In: In 14th European Conference on Technology Enhanced Learning, EC-TEL, pp. 423–435. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-29736-7_32
12. Dolev, N., Leshem, S.: Developing emotional intelligence competence among teachers. Teach. Dev. **21**(1), 21–39 (2017). https://doi.org/10.1080/13664530.2016.1207093
13. The Economist Intelligence Unit Limited: Emotion and cognition in the age of AI (White paper). In: Microsoft (2019) https://bit.ly/4eKQ3k4
14. Eickelmann, B., Vennemann, M.: Teachers' attitudes and beliefs regarding ICT in teaching and learning in European countries. Eur. Edu. Res. J. **16**(6), 733–761 (2017). https://doi.org/10.1177/1474904117725899
15. Henderson, J., Corry, M.: Data literacy training and use for educational professionals. J. Res. Innov. Teach. Learn. **14**(2), 232–244 (2021). https://doi.org/10.1108/JRIT-11-2019-0074
16. Bannigan, G., McGrath, D., Matthews, J., Duignan, C.: Facilitated collaborative group co-design of hypothetical digital tools to understand university students' wellbeing support needs and requirements. SN Comput. Sci. **5**, 668 (2024). https://doi.org/10.1007/s42979-024-03014-w
17. Gross, B., Hamilton, L.: Student Mental Health and Well-Being: A Review of Evidence and Emerging Solutions. Center on reinventing public education, Arizona State University (2023) https://crpe.org/wp-content/uploads/SEL-Report-2023-Final.pdf
18. UNESCO.: Going global with the Happy Schools framework: supporting holistic school happiness to improve teaching, learning, and wellbeing. Working paper, No. 12 (2022). https://doi.org/10.54675/TFXO4953
19. Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L.: Subjective wellbeing: three decades of progress. Psychol. Bull. **125**(2), 276–302 (1999)
20. Ryff, C.: Psychological wellbeing revisited: advances in science and practice of Eudaimonia. Psychother. Psychosom. **83**(1), 10–28 (2013). https://doi.org/10.1159/000353263
21. Darling-Hammond, L., Cook-Harvey, C.M.: Educating the Whole Child: Improving School Climate to Support Student Success. Learn Policy Inst (2018)
22. Ruzek, E.A., Hafen, C.A., Allen, J.P., Gregory, A., Mikami, A.Y., Pianta, R.C.: How teacher emotional support motivates students: the mediating roles of perceived peer relatedness, autonomy support, and competence. Learn. Instr. **42**, 95–103 (2016). https://doi.org/10.1016/j.learninstruc.2016.01.004

23. Vandenbroucke, L., Spilt, J., Verschueren, K., Baeyens, D.: Keeping the spirits up: the effect of teachers' and parents' emotional support on children's working memory performance. Front. Psych. **8**, 512 (2017). https://doi.org/10.3389/fpsyg.2017.00512
24. Nowell, L., Norris, J.M., Mrklas, K., White, D.E.: A literature review of mentorship programs in academic nursing. J. Prof. Nurs. **33**, 334–344 (2017). https://doi.org/10.1016/j.profnurs.2017.02.007
25. Wright-Harp, W., Cole, P.A.: A mentoring model for enhancing success in graduate education. Contemp. Issues. Commun. Sci. Disord. **35**, 4–16 (2008) https://bit.ly/3Y5T8En
26. Freedman, M.: The Kindness of Strangers: Reflections on the Mentoring Movement. Public/Private Ventures (1992)
27. Kim, M.: Intensive learning experience: development of STEM mentorship program for high school gifted students. Gift. Child Today. **44**(4), 228–235 (2021). https://doi.org/10.1177/10762175211030522
28. Okwan, P., Omojola, J., Kambhampati, M.: Mentoring K-12 students in STEM education. In: Craig, C.D. (ed.) Mentoring in STEM Through a Female Identity Lens: Heroes Make a Difference for Women, pp. 75–92. Vernon Press (2024)
29. Government Official Reports. (SOU): An upper secondary education for all: Measures to ensure that all young people start and complete an upper secondary education. In: En gymnasieutbildning för alla. Åtgärder för att alla unga ska påbörja och slutföra en utbildning (2016) ISBN: 9789138245217
30. The Swedish National Agency of Education: Teachers' professional day: A national survey of primary school teachers' time use. In: Lärarnas yrkesvardag - En nationell kartläggning av grundskollärares tidsanvändning (2013) https://bit.ly/3Ymhx9Y
31. Allen, D.K., Brown, A., Karanasios, S., Norman, A.: How should technology-mediated organizational change be explained? A comparison of the contributions of critical realism and activity theory. MIS Q. **37**(3), 835–854 (2013)
32. Mingers, J., Mutch, A., Willcocks, L.: Critical realism in information systems research. MIS Q. **37**(3), 795–810 (2013)
33. Marwan, A., Sweeney, T.: Using activity theory to analyse contradictions in English teachers' technology integration. Asia Pac. Educ. Res. **28**(2), 115–125 (2019). https://doi.org/10.1007/s40299-018-0418-x
34. Ertmer, P.A., Ottenbreit-Leftwich, A.T.: Teacher technology change: how knowledge, confidence, beliefs, and culture intersect. J. Res. Technol. Educ. **42**(3), 255–284 (2010). https://doi.org/10.1080/15391523.2010.10782551
35. Teo, T., Zhou, M.: The influence of teachers' conceptions of teaching and learning on their technology acceptance. Int. Learn. Environ. **25**(4), 513–527 (2016). https://doi.org/10.1080/10494820.2016.1143844
36. Long, P., Siemens, G.: Penetrating the fog: analytics in learning and education. Educ. Rev. **46**(5), 31–40 (2011)
37. Kadirvelu, B., Bellido Bel, T., Wu, X., Burmester, V., Ananth, S., Cabral, C.C., Branco, B., Girela-Serrano, B., Gledhill, J., Di Simplicio, M., Nicholls, D., Faisal, A.A.: Mindcraft, a mobile mental health monitoring platform for children and young people: development and acceptability pilot study. JMIR Form. Res. **7**, e44877. https://doi.org/10.2196/44877
38. Gómez-Restrepo, C., Sarmiento-Suárez, M.J., Alba-Saavedra, M., Calvo-Valderrama, M.G., Rincón-Rodríguez, C.J., Bird, V.J., Priebe, S., van Loggerenberg, F.: Development and implementation of DIALOG+S in the school setting as a tool for promoting adolescent mental well-being and resilience in a post-armed conflict area in colombia: exploratory cluster randomized controlled trial. JMIR Form. Res. **7**, e46757 (2023). https://doi.org/10.2196/46757
39. Engeström, Y.: Learning by Expanding, 2nd edn. Cambridge University Press, Cambridge (2015)

40. Kaptelinin, V., Nardi, B.A.: Acting with Technology: Activity Theory and Interaction Design, 1st edn. MIT Press, Cambridge, MA (2006)
41. Engeström, Y.: Expansive learning at work: toward an activity theoretical reconceptualization. J. Educ. Work. **14**(1), 133–156 (2001). https://doi.org/10.1080/13639080020028747
42. Holden, H., Rada, R.: Understanding the influence of perceived usability and technology self-efficacy on teachers' technology acceptance. J. Res. Technol. Educ. **43**(4), 343–367 (2011). https://doi.org/10.1080/15391523.2011.10782576
43. Bryman, A.: Social Research Methods, 5th edn. Oxford University Press, Oxford (2016). https://doi.org/10.1017/CBO9781107415324.004
44. The Swedish Research Council.: Good Research Practice 2024 (2024)
45. Engeström, Y.: Studies in Expansive Learning: Learning What Is Not Yet There. Cambridge University Press, Cambridge (2016)
46. Karanasios, S., Nardi, B., Spinuzzi, C., Malaurent, J.: Moving forward with activity theory in a digital world. Mind Cult. Act. **28**(3), 234–253 (2021). https://doi.org/10.1080/10749039.2021.1914662

Verifying Scientific Validity and Enhancing Response Motivation in Digital Human-Based Mental Health Assessment: A Mixed-Methods Study

Chi-Cheng Lao[1], Litun Tan[2], Jing Liang[1], Ling Zeng[1], Zongze Li[2], Yitian Huang[2], Zi-Yi Guo[2], Yan-Dian Xu[2], Tianyang Tan[3], Xuetao Tian[1(✉)], Jun Liu[2(✉)], and Keliang Chen[2(✉)]

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Beijing Normal University, Beijing 100875, China
chichenglao@mail.bnu.edu.cn

[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
{liujun,chenkl}@bupt.edu.cn

[3] School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

Abstract. Online mental health surveys have long been challenged by respondent carelessness, a problem that has persisted for over a decade without a comprehensive solution. To tackle these issues, we designed and implemented the Digital Human-Based Interactive Mental Health Assessment System (DHMHAS) with college students as the pilot group, and employed mixed methods to validate its scientific credibility and interaction experience to comprehensively explore whether the system may retain scientific validity while increasing motivation to answer questions. We successfully confirmed that DHMHAS possesses robust psychometric properties. In particular, the system performed better than traditional methods in terms of criterion validity. Furthermore, DHMHAS enhanced the college students' interactive experience by providing a high level of social presence and system usability, reducing cognitive load, and eliciting high levels of response differentiation in answers. They also offered overwhelmingly positive feedback on the system, recognizing its benefits for personal growth and development. Compared to traditional methods, DHMHAS offered significant improvements in delivering positive emotional experiences, deepening cognitive understanding, promoting immersive engagement, and increasing the accuracy of responses. This research lays a scientific foundation for developing innovative assessment tools and offers directions for optimizing future AI-based mental health evaluation methods.

Keywords: careless response · digital human · human-AI interaction · user experience design

Keliang Chen, Jun Liu, and Xuetao Tian are the first, second, and third corresponding au-thors, respectively.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 22–40, 2025.
https://doi.org/10.1007/978-3-031-93502-2_2

1 Introduction

The main source of mental health problems among college students is academic pressure [1, 2]. It has been an important issue in the current Chinese educational system [3, 4]. To solve these problems caused by academic stress, the first step is to evaluate college students with academic stress. Although traditional psychological assessment has been applied in wide-speed fields to assess the levels of mental health in individuals, the lack of motivation to answer questions (also called careless responses) is still a tricky challenge in recent decades [5–7]. This challenge could cause a series of issues in data quality, subsequently harming the actual levels of academic stress and mental health, and affecting the judgment and intervention of psychological counselors. To address these challenges, we designed and implemented a digital human-based interactive mental health assessment system (DHMHAS) targeted at college students. Due to the extensive anthropomorphism of digital humans in DHMHAS, college students can have a highly immersed interactive experience to improve the quality of their responses. Moreover, DHMHAS uses natural language to interact with individuals which can offer richer information [8, 9], enabling assessment outcomes to be more authentic and accurately reflect real-world contexts and behaviors (Fig. 1).

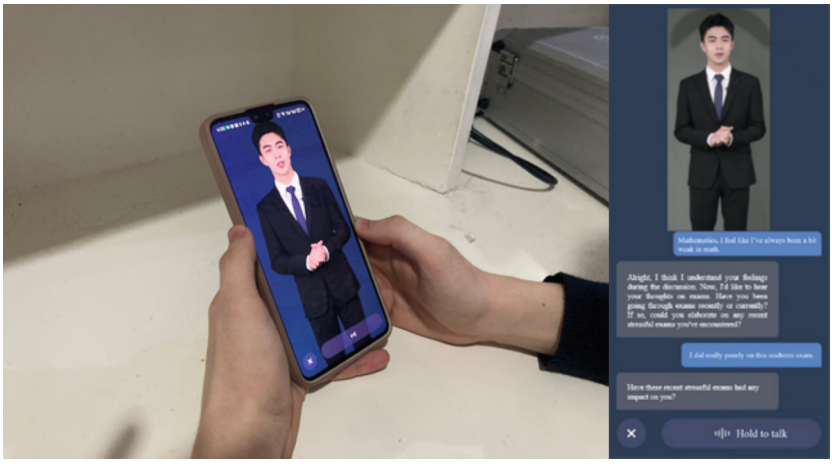


Fig. 1. On the left side of the image is a photo of people using the Digital Human-Based Interactive Mental Health Assessment System (DHMHAS), while on the right side is the user interface of DHMHAS.

Specifically, the digital human in DHMHAS would engage in conversations with students about academic stress questions using natural language and multi-turn dialogue. In the process of the interaction with DHMHAS and students, their answers will be processed through a semantic analysis and classification parts to evaluate their levels of academic stress. Ultimately, an academic stress assessment report would be generated for psychological consultants to understand individual academic stress and potential mental health problems. Additionally, we validated and explored research questions

during the development of the system to strengthen its scientific validity in psychological measurement, providing a scientific and effective academic stress management platform. The system implementation and validation are as follows in Fig. 2.

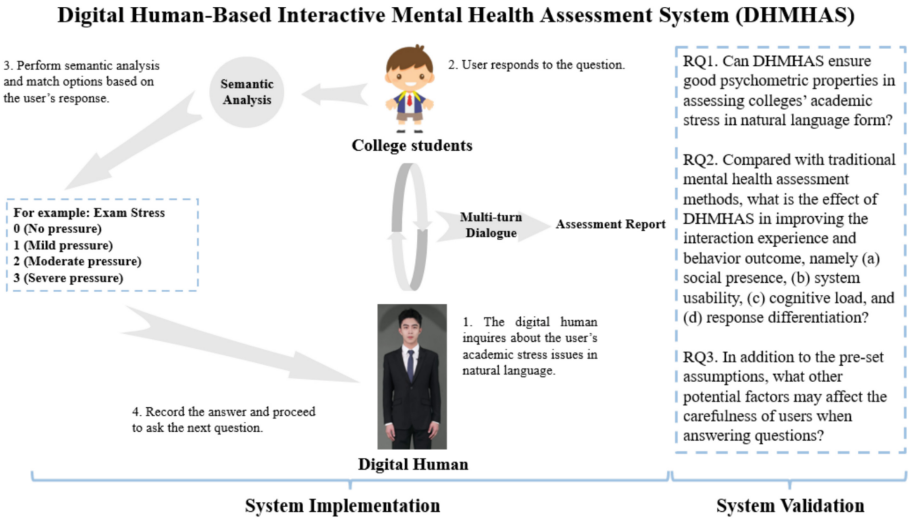


Fig. 2. The implementation and validation of DHMHAS.

2 Related Work

2.1 Limitations of Traditional Mental Health Assessment

Due to the prevalence of the Internet, the most widely used method for mental health assessments at present is through online questionnaire responses [10]. Although this method allows for a rapid understanding of an individual's mental health status, it has a major limitation: Traditional mental health assessment methods are tedious and can easily reduce individuals' motivation to continue answering questionnaires, leading to careless responses. Careless responses from individuals can introduce errors in assessment results, consequently affecting the data quality [6, 11]. In other words, this can result in questionnaire data inaccurately reflecting an individual's mental health status. In such cases, parents, teachers, and mental health professionals may not be able to monitor and address group concerns, potentially leading to serious societal consequences.

Despite the issues being long identified, a universal solution on how to genuinely enhance individuals' motivation to answer questions seriously has yet to be found, spanning over a decade [5–7, 12]. Specifically, previous researchers often used warnings to alert individuals to answer carefully or directly provided external incentives [13, 14] to motivate individuals to respond. However, the latest research indicated that presenting warning instructions to participants is no longer effective in reducing individuals' careless responses [12]. Another study [7] attempted to use social exchange theory, offering

participants discounts on food and beverages in exchange for careful responses, but the study results showed that providing rewards did not effectively reduce careless responses. In other words, the quest for finding methods to eradicate individuals' careless responses remains a long and challenging journey. Before delving into solutions, it is crucial to grasp the underlying reasons for low motivation to respond. This understanding enables us to explore effective strategies to enhance motivation for answering.

2.2 The Reasons for Low Response Motivation in Traditional Mental Health Assessment

Prior research has extensively examined the reasons for individuals' low motivation to respond when completing questionnaires, suggesting a variety of possible factors [13]. The most comprehensive study in this area was conducted by [11], who noted that questionnaire design (e.g., questionnaires that are too long), respondent characteristics (e.g., personality traits), and contextual factors (e.g., environmental distractions) may influence motivation to respond.

In terms of questionnaire design, researchers frequently encounter the challenge of striking a balance between the length of the questionnaire and the richness of the data collected. This trade-off can result in an excessive number of questions, leading to the phenomenon of "bubble hell" (i.e., presenting the questionnaire in the form of a large number of questions) [7], which often evokes negative emotions in the user when completing the questionnaire [15]. Consequently, in the absence of a positive interaction experience, it may be an unrealistic expectation to expect respondents to answer carefully.

In examining the characteristics of respondents, while some studies have indicated that individuals with diminished responsibility and self-control may exhibit less effort in questionnaire completion [16, 17], the correlation between personality traits (e.g., conscientiousness) and carefulness was not particularly strong ($r = .03–.17$). Therefore, personality traits may not be the primary determinant of motivation to respond.

In terms of contextual factors, the absence of direct interaction [18] between the respondents and the researcher may be a potential cause of low individual motivation to respond. Humans have an innate need for social interaction [19], and the lack of interactivity may result in respondents not receiving immediate feedback, which may reduce their engagement and motivation to answer the questionnaire.

In conclusion, following the exclusion of personality factors, questionnaire design and contextual factors should be considered as potential areas for improvement, to increase users' motivation to respond. It is regrettable that, despite the undertaking of studies to suggest improvements in two factors to increase motivation to answer, the effectiveness of these attempts remains limited. For example, [20] attempted to enhance respondents' motivation to answer by establishing an avatar to simulate the researcher's supervision. However, this approach did not yield the anticipated outcomes. Potential explanations for this phenomenon are a dearth of interactivity, an absence of realism in the avatar's image, and persisting deficiencies in the questionnaire design, which collectively contribute to an unsatisfactory interaction experience [21]. This, in turn, maybe a pivotal factor in reducing users' motivation to respond to the questionnaire.

2.3 Potential Innovative Methods for Enhancing Respondent Motivation

Regarding questionnaire design, by adopting a one-question-one-answer presentation, the cognitive load of the user can be effectively reduced, which in turn enhances the usability of the system. According to Cognitive Load Theory (CLT, [22]), intrinsic cognitive load (ICL) is related to the inherent complexity of the task and extraneous cognitive load (ECL) is related to the presentation of the task. In questionnaire answering, a design similar to "bubble hell", may affect the individual's perception of the difficulty of the task itself and the efficiency of information processing. Specifically, in terms of ICL, the inherent complexity of the task may lead individuals to perceive increased difficulty in carefully answering the questionnaire. On the ECL side, lengthy text requires individuals to sift through and process large amounts of information to find key elements to recall and fill in the answers. Both loads consume the user's limited cognitive resources [23]. Once these resources are depleted, individuals are less likely to focus on the task and exhibit low motivation to answer the questionnaire. Therefore, a new mental health assessment tool with a one-question-one-answer format can reduce the amount of information that users need to process at any given time, thereby reducing the ICL and ECL, resulting in a perceived high usability and a good response experience.

In terms of social interaction, it is crucial to enhance the anthropomorphism of avatars, which includes aspects such as their appearance, body movements and speech. Through a multimodal approach, i.e., combining visual, auditory, and other sensory experiences, it is possible for users to feel a higher level of positive experience when interacting with others, also known as social presence. Social Presence (SP) refers to the "degree of interaction with others and the significance of interpersonal relationships that follow," where these "others" can be real people or intelligent forms generated by computer-mediated communication (CMC), such as virtual human agents [24, 25]. In general, increasing the anthropomorphic features of virtual humans enables individuals to perceive a high level of SP [26, 27] because it provides humans with a rich sensory experience, allowing users to immerse themselves in the interaction and have a good interactive experience. Therefore, virtual humans in novel mental health assessment tools with highly anthropomorphic features enable users to perceive high levels of SP, and users are likely to have a more positive user experience when using the tool, and this experience also positively affects motivation for answering the questions and improves the accuracy of responses Specially, the improvement of motivation in responses could be used [28] to calculate.

2.4 Dhmhas

To address the limitations of the traditional mental health assessment methods, we designed a novel system based on natural interaction for assessing mental health using a digital human (DHMHAS) to measure individual academic stress (the specific assessment process has been shown in "introduction"). The system mainly integrates (a) 3D digital human generation, (b) voice cloning, (c) semantic analysis, and (d) speech recognition technology, as shown in Fig. 3.

Based on the ER-NeRF [29] 3D digital human generation technology, we proposed a digital human face-changing solution to address the limitations of ER-NeRF that only

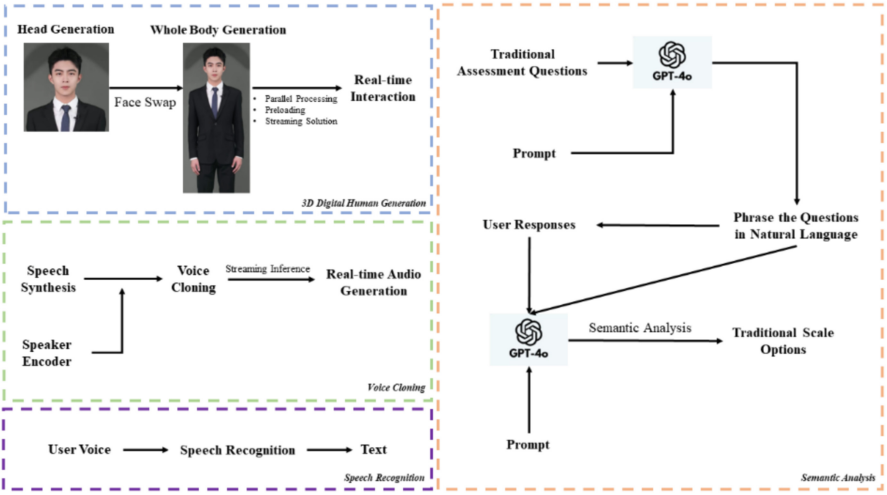


Fig. 3. Technical implementation diagram.

generated the head, and the head was not dynamically scaled under control. Through the alignment algorithm, that is, fixing the non-face area and calculating the corresponding face position, the overall generation of the digital human head, torso, and limbs was achieved, thereby optimizing the appearance of the digital human. To solve the problem of the slow rendering speed of ER-NeRF and the slow generation speed of digital humans after the face-changing operation, which leads to the lag of digital humans, we maximized the use of GPU performance, preload, and selected advanced streaming solutions suitable RTMP [30] for this system to improve the generation efficiency, to meet the requirements of real-time interaction while ensuring high-quality image generation.

Voice cloning technology was used to generate audio with speaker timbre characteristics. Based on the VITS [31] speech synthesis model, we introduced a speaker encoder to characterize the speaker timbre characteristics and achieve the transformation from speech synthesis to voice cloning. To solve the problem of slow reasoning of long texts in the VITS model, we proposed a streaming reasoning solution, which receives text input segment by segment and generates speech in real-time without waiting for all text input to be completed and sent to the digital human module, driving the real-time rendering and generation of the digital human.

Semantic analysis technology was mainly used to generate natural language questions about academic pressure and imitate the ability of humans to understand user intentions to match the options of traditional psychological assessment scales with the user's answers. Specifically, using GPT-4o, combined with a single-sample learning strategy and a psychological expert assessment strategy to construct a prompt project, traditional psychological assessment questions were converted into natural language expressions and output to users. Subsequently, using the semantic analysis capability of GPT-4o, the user's answers were processed to achieve a matching of options in the traditional psychological assessment scale.

Speech recognition technology converted the user's input speech into text for GPT-4o to analyze semantics. The speech recognition technology used in this system is FunASR [32]. FunASR is a basic speech recognition toolkit that provides convenient scripts and tutorials, supports inference and fine-tuning of pre-trained models, and has significant recognition effects on Mandarin Chinese.

By optimizing and integrating digital human appearance, rendering speed, and audio generation speed comprehensively, we aim to enhance the anthropomorphic qualities of digital humans to maximize the sense of social presence for users. In this process, users engage in a question-and-answer format, as well as employing natural language for cognitive processing, which serves to reduce the perceived cognitive load and enhance the overall interaction experience.

2.5 Scientific Validity in DHMHAS

Before exploring the prospective advantages of DHMHAS in bolstering response motivation, it is imperative to ascertain that the transition from a conventional questionnaire to a natural language questioning format does not introduce inaccuracies. Conversely, this is a point that scholars in non-psychology tend to overlook when developing systems. It is therefore essential to rigorously validate the scientific validity of the system. This process comprises the following key steps [33, 34]: (a) content validity analysis; (b) item analysis; (c) reliability analysis; (d) convergent validity; (e) discriminant validity; and (f) criterion-related validity. Collectively, these steps provide a comprehensive assessment of the scientific validity of DHMHAS.

Content validity evaluates the relevance and representativeness of a test's content, specifically its coverage of the intended subject matter, themes, or topics [33]. The assessment process is typically conducted by experts in the field of psychology, who validate content validity by calculating two key metrics: the item-level content validity index (I-CVI) and the scale-level content validity index (S-CVI) [35]. To assess the consistency of scores, the Kappa coefficient (K^*) [36] is calculated as a measure of consistency across raters. This ensures the reliability of the scores, thereby enhancing the accuracy and validity of the content validity assessment. In particular, I-CVI greater than or equal to .78 indicates good content validity. The evaluation criteria for K^* are: .40 to .59 is considered fair, .60 to .74 is deemed good, and greater than .74 is considered excellent.

Item analysis is a process of employing statistical techniques to examine the responses of subjects to test items to evaluate and screen the items themselves. It is typically conducted before evaluating the reliability and validity of a given test. Item-total correlation is a frequently employed statistical instrument in the context of item analysis. The item-total correlation coefficient measures the correlation between the score of a single item and the total score of the entire test. It is used to evaluate the contribution of the item to the overall test performance [34]. In general, the item-total correlation coefficient should fall within the range of .2 to .4, with values above .4 being preferable.

Reliability analyses are used to assess the consistency and stability of measurement instruments. Commonly used reliability coefficients are Cronbach's alpha, developed by

[36], which measures the consistency of each question in a test to ensure the reliability of the test.

The concept of validity is pivotal in determining whether a test accurately measures the intended construct. Validity can be categorized into three sub-types: convergent validity, discriminant validity, and criterion-related validity. Convergent validity ensures that a test measures the same constructs as another test specifically designed for those constructs. In contrast, discriminant validity demonstrates that scores on a test are uncorrelated with scores from a test measuring an unrelated construct. Criterion-related validity evaluates how well test scores correlate with an external criterion, which could be actual behavior, achievements, or other recognized benchmarks. If users are asked questions in a natural language format by a digital human and receive scores assessing academic stress, these scores should exhibit a high correlation with traditional mental health assessments that measure the same academic stress for convergent validity. For discriminant validity, these scores should show weak correlations with constructs that measure completely different attributes, such as social presence. Regarding criterion-related validity, these scores should correlate to some extent with external standards, such as depression, which is often linked to academic stress [37, 38].

3 The Present Study

While DHMHAS presents the potential to address traditional mental health assessments, it is still necessary to address the following 3 research questions to solidify the system's scientific rigor and practicality:

RQ1. Can DHMHAS ensure good psychometric properties in assessing colleges' academic stress in natural language form?
RQ2. Compared with traditional mental health assessment methods, what is the effect of DHMHAS in improving the interaction experience and behavior outcome, namely (a) social presence, (b) system usability, (c) cognitive load, and (d) response differentiation?
RQ3. In addition to the pre-set assumptions, what other potential factors may affect the carefulness of users when answering questions?

We employed 2 studies to verify the scientific validity and explore the interactive experiences in DHMHAS. Study 1 consisted of two experiments, that addressed RQ1. In particular, experiment 1 aimed to confirm whether the natural language version of academic stress can accurately measure the construct using an expert perspective (i.e., context validity). Experiment 2 used participants' data to verify the psychometric properties of DHMHAS to fully ensure the quality of the items (i.e., item analysis), reliability, and validity (including convergent validity, discriminant validity, and criterion-related validity).

Study 2 used mixed methods to explore the benefits of improving response motivation in DHMHAS compared to traditional mental health assessment, which addressed RQ2 and RQ3. First, we compared two methods (i.e., traditional mental health assessment and DHMHAS) on the four key interactive experience variables and behavior outcomes: social presence, system usability, cognitive load, and response differentiation. Second, we used qualitative interviews with DHMHAS users to supplement the

evidence for improving response motivation in DHMHAS. In particular, we focused on the features that attracted users to the system, the usefulness of the system, and how the system affected the psychology of users, including emotions, cognitions, wills, and behaviors. Through these analyses, we aimed to gain a deeper understanding of the potential strengths of DHMHAS and demonstrated their value in improving the motivation of responses. It is hoped that a possible theoretical model can be constructed for this purpose.

4 Method

4.1 Participants

We recruited 5 psychological experts in experiment 1 of study 1. Specifically, the participants included 1 male doctoral student in clinical psychology, 2 male doctoral students in educational psychology, 1 female doctoral student in educational psychology, and 1 female university teacher in psychology. In experiment 2 of study 1, we recruited 39 college students (sophomore: 41.00%, junior: 33.33%, senior: 25.6%), comprising 15 male students (38.50%) and 24 female students (61.50%).

In Study 2, participants were recruited through a student pool from Beijing Normal University. A total of 79 college students aged between 18 and 25 were recruited. All participants volunteered to participate and signed informed consent before the experiment. Participants were randomly assigned to two experimental groups: In the DHMHAS group, 41 participants comprised college students (sophomore: 39.02%, junior: 34.15%, senior: 26.83%). 24 (58.54%) were female and 17 (41.46%) were male. In the Survey group, we recruited 38 participants who comprised college students (freshmen: 7.9%, sophomores: 34.2%, juniors: 44.7%, seniors: 13.2%). 11 (28.95%) were female and 27 (71.05%) were male.

4.2 Apparatus and Procedure

This study has been approved by the Beijing Normal University Ethics Committee (BNU202406210122). In experiment 1 of study 1, we employed WènjuànXīng (a type of survey platform; https://www.wjx.cn/) to collect the content validity evaluation by psychology experts. In particular, each page would display 4 elements: (a) a definition of the "stress" concept, (b) a definition of the "academic stress" concept, (c) the original author's question logic, and (d) the natural language version of academic stress. An example of evaluating content validity is "How relevant is this question (in natural language) to the original question?". In experiment 2 of study 1, we employed the academic stress questions recognized by experts in DHMHAS and conducted a more comprehensive psychological measurement verification of the system. Before the experiment began, participants would be asked to go to a quiet, distraction-free space to experiment. Once it was confirmed that the participants were in suitable conditions for the experiment, they would be introduced to the usage of DHMHAS. When the participants received the response from the digital human "All interview has ended, we will provide you with feedback as soon as possible," it signified the end of the assessment session. Following

this, participants would need to respond to (a) the original scale of academic stress [39], (b) the SP scale [40], (c) PHQ-9 [41]. Upon completion of the answers, the experiment concluded. An example of a natural language version of an academic stress question could be shown in Fig. 4.

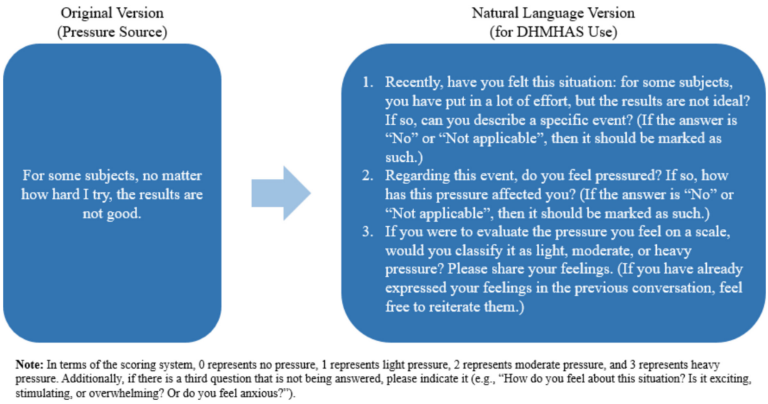


Fig. 4. An example of the natural language version of academic stress.

In Study 2, we recruited college students and randomly assigned them to two groups: the DHMHAS group and the Survey group. For the DHMHAS group, participants first signed an informed consent form and completed a pretest to collect demographic information and their experiences with artificial intelligence. They were instructed to participate in a quiet, distraction-free environment. Participants received an installation package for DHMHAS along with a user guide. They were briefed on the experiment's purpose, app usage, and instructions before using the DHMHAS app. After the experiment, participants completed several questionnaires: the SP scale [40], the ICL and ECL scale (for measuring cognitive loads) [42], and the System Usability scale (for measuring the usability of DHMHAS) [43]. In the Survey group, participants used traditional psychological assessment methods to answer questions. Similar to the DHMHAS group, demographic information and familiarity with artificial intelligence were collected. Participants were asked to respond to questions about academic stress in a quiet environment. After completing these tasks, they filled out the same questionnaires as the DHMHAS group. Finally, in terms of the qualitative interviews, participants in the DHMHAS group would undergo a brief interview focusing on their interactive experiences with DHMHAS, the questions are as follows:

- Are there any features or designs that impressed you?
- What help or impact does this product have had on you?
- In comparison to filling out traditional questionnaires, what advantages do you think this digital agent assessment method has?

4.3 Analysis

In experiment 1 of study 1, we conducted data analysis using content validity. In experiment 2 of Study 1, we employed the item-total correlation coefficient in item analysis.

In terms of reliability, we utilized Cronbach's alpha (α) to evaluate the consistency of academic stress scores obtained by digital humans through questions asked in natural language form. In terms of convergent validity, discriminant validity, and criterion-related validity, we assessed the validity of our measures by calculating Pearson correlations between the natural language version and the original version of academic stress, SP (for discriminant validity), and depression (for criterion-related validity), respectively. In Study 2, we utilized an independent t-test to explore the differences in interactive experiences (i.e., SP, ICL, ECL, system usability, and response differentiation [28]) between the DHMHAS group and the Survey Group. Finally, a Grounded Theory [44] would be used with first, second, and third levels for qualitative interview coding.

In terms of analysis software, we conducted data analysis using SPSS 27.0 in Study 1 and Study 2. Excel was employed for interview coding in Study 2.

5 Results

5.1 Study 1

Content Validity. Five experts conducted a content validity assessment on natural language items regarding academic stress. All items met the content validity criteria, with an Item-Content Validity Index (I-CVI) greater than .78 for all questions, and a Kappa coefficient exceeding .74.

Psychological Measurement Item Analysis. The correlations between each item and the total score range from .40 to .67, all greater than or equal to .4.

Reliability. Regarding internal consistency reliability measured by Cronbach's alpha, the natural language version of academic stress items demonstrated good reliability ($\alpha = .77$).

Convergent Validity, Discriminant Validity, and Criterion-Related Validity. Regarding convergent and discriminant validity, the natural language version of academic stress exhibited a high linear correlation with the original version of the academic stress ($r = .82$, $p < .001$) and a low linear correlation with SP ($r = .28$, $p = .10$), respectively. In terms of criterion-related validity, the natural language version of academic stress was significantly correlated with depression ($r = .62$, $p < .001$). Specifically, the correlation between the natural language version of academic stress and depression was higher than that between the original version of academic stress and depression ($r = .56$, $p < .001$), as shown in Table 1.

5.2 Study 2

Differences in Interactive Experience and Careless Behavior. *Social presence.* A significant difference was observed in participants' SP ($p = .02$, $t = -2.30$, Cohen's $d = 1.05$) between the Survey group ($M = 6.75$, $SD = 1.20$) and the DHMHAS group ($M = 7.23$, $SD = .89$).

Table 1. Convergent validity, discriminant validity, and criterion-related validity.

	M	*SD*	1	2	3
1 Academic Stress (Natural Language Version)	13.06	5.38	–		
2 Academic Stress (Original Version)	14.44	5.28	.82***	–	
3 Social Presence (SP)	48.72	9.62	.28	.29	–
4 Depression	5.81	4.58	.62***	.56***	−.09

Note. $N = 36$, *$p < .05$, **$p < .01$, ***$p < .001$

Cognitive load. Compared to the Survey group ($M = 4.74$, $SD = 1.30$), the DHMHAS group ($M = 2.89$, $SD = 1.30$) had significantly lower ICL ($p < .001$, $t = 6.30$, Cohen's $d = 1.30$) and ECL ($p < .001$, $t = 8.21$, Cohen's $d = 1.36$).

System usability. The DHMHAS group ($M = 78.29$, $SD = 10.22$) demonstrated higher system usability compared to the Survey group ($M = 67.04$, $SD = 19.16$), and this difference was significant ($p = .002$, $t = 3.29$, Cohen's $d = 15.19$).

Carefulness behavior. The DHMHAS group ($M = .69$, $SD = .19$) demonstrated higher response differentiation compared to the Survey group ($M = .56$, $SD = .08$), and this difference was significant ($p < .001$, $t = 3.71$, Cohen's $d = .15$).

Interview Results 1: Impressive Features and Designs in Digital Humans. Digital Human emerged as the most mentioned, capturing the majority of participants' attention towards impressive features or designs, reaching 73.08%. Among these, the Appearance (31.58%), Speech (42.11%), Body Language (13.16%), and Facial Expressions (13.16%) of the digital human were identified as the most noteworthy aspects.

- Appearance: "…this digital human was quite real." (Participant #27)
- Speech: *"The character above would exhibit certain actions, along with speech, to make you feel like you were experiencing interaction with him."* (Participant #1)
- Body Language: *"I thought his action was quite impressive. When you were speaking, its hands moved, giving the feeling that it resembled a real person, making you feel like the other party was seriously listening to your narration as you responded. Additionally, it nods and such, making you feel that what you were saying was being acknowledged and heard."* (Participant #34)
- Facial Expressions: *"…when he answered me, his mouth would also move, which surprised me."* (Participant #19)

Interview Results 2: System Usefulness. The majority of responses revolved around the topic of *Personal Growth and Development*, accounting for 90.91% of the total mentions. There were some examples of second-level and third-level coding:

- Emotional Management (20.00%): "After the conversation just now, I felt that some analysis was done on my recent state, which also led me to reflect on some aspects of myself." (Assisting in reflecting on one's emotions or state; Participant #1)

- Emotional Support (40.00%): "I felt that I could express my thoughts more effectively. Some things that I couldn't express to friends or family around me, I could share with that artificial intelligence to seek help." (Accompanying listening; Participant #2)
- Self-Exploration (32.50%): "By having it go over things with you, you could understand your shortcomings and what troubled you. You could speak them out to it. This helped you see your deficiencies and also recognize your strengths." (Understanding oneself; Participant #33)
- Problem Solving (7.50%): "I felt that this product's features were quite good; it could help me solve some issues in my daily life that I couldn't handle on my own. For example, when it came to my emotions or when I felt mentally uncomfortable, I didn't feel like I was alone. I didn't feel as lonely, and I could consult it for advice, that was the approach I took." (Resolving issues; Participant #19)

Interview Result 3: The Advantages of DHMHAS Based on Psychology Variables. The vast majority of responses (94.55%) were focused on key psychological aspects: Emotions (32.73%), Cognitions (21.82%), Wills (29.09%), and Behaviors (10.91%).

Regarding Emotions, participants perceived that DHMHAS could help in avoiding boredom, tedium, and fatigue (50.00%), experiencing novelty (11.11%), having fun (22.22%), feeling friendly and affable due to the digital human's responses (11.11%), and feeling be empathized (5.56%):

- Avoiding boredom, tedium, and fatigue: "…It won't be as dull. It won't be as boring as the traditional way of filling out questionnaires." (Participant #18)
- Experiencing novelty: "One of its advantages is that it is innovative and more easily accepted by young people." (Participant #3)
- Having fun: "I think it might be more interesting than traditional questionnaires. And if you simply fill out a questionnaire, you might find it a bit dull and boring." (Participant #24)
- Feeling friendly and affable: "…Not as rigid as a questionnaire, it has a tone that feels like a psychological counselor. This may make you feel quite close to it, allowing you to share any issues you might have. You might feel like confiding in it and seeking advice or solutions from it." (Participant #10)
- Feeling empathized: "The questionnaire format fails to grasp my feelings adequately, seeming to lack the empathetic emotion with me, making the questionnaire format too monotonous." (Participant #14)

Concerning *Cognitions*, participants believed that DHMHAS could facilitate *easy and deep problem understanding* (66.67%), *deep self-understanding* (25.00%), and *easier retrieval of key information* (8.33%):

- Easy and deep questions understanding: "The advantage was the ability to delve deeper into understanding (questions). It added pressure and categorizes it into mild, moderate, and severe levels. It also followed up with questions about the reasons behind these aspects. In contrast, questionnaires were superficial." (Participant #43)

- Deep self-understanding: "Sometimes he asked me what aspects or areas of study I felt stressed about, and I might have found subjects like physics more challenging, especially science subjects. Sometimes I also reflected on myself and felt that I might not have been good enough in some aspects. I procrastinated a bit when doing homework, and occasionally I felt a little regretful, thinking that I could do better and needed to make some improvements." (Participant #10)
- Easier retrieving of key information: "…I felt that conducting surveys in this manner was beneficial. He could read it out to me, and I thought this was quite helpful in quickly grasping the important information within the instructions" (Participant #35)

Regarding *Wills*, participants found that this interactive approach could enhance focus, engagement, and concentration (62.50%), elicit relaxed responses (18.75%), and boost motivation to answer questions (18.75%):

- Enhancing focus, engagement, and concentration: "Sometimes, when filling out surveys, details could be overlooked. Digital human narration could capture your attention and help you better understand the questions!" (Participant #33)
- Eliciting relaxed responses: "...His expressions and his movements made you feel like you were interacting with a... anyhow, not like dealing with a machine or a questionnaire. In short, his expressions, his movements, that kind of thing. They would make you feel quite relaxed as if you were communicating with a real human professor."
- Boosting motivation to answer questions: "The advantage was... Honestly, the person on the other end made me more willing to respond a bit because they gave off a dynamic vibe. That format involved a person with body language and some other elements. I wasn't sure if there were facial expressions, but it felt like there might have been a hint of expression. It truly was because... Compared to questionnaires, it certainly was more likely to stimulate the idea of participants responding in this way, I believed." (Participant #22)

On the topic of Behaviors, participants noted that this could lead to more accurate and comprehensive responses (100.00%). For example: *"I felt that in this assessment method (DHMHAS), my responses would be more authentic."* (Participant #4).

6 Discussion

6.1 Scientific Validity of DHMHAS

The experiment in study 1 provides preliminary evidence of the content validity of the DHMHAS in psychological measurement, demonstrating that DHMHAS aligns with psychometric properties. Results indicated that the natural language version of academic stress could measure the original construct [35] in college students' groups.

In study 2, building upon the findings of study 1, a further analysis of the scientific rigor of the natural language version of academic stress was conducted from theoretical perspectives. Results suggested that most evidence points towards the psychological scientific validity of the system. Specifically, (a) item analysis: item-total

correlations ranged from .40 to .67; (b) reliability: the natural language version of academic stress demonstrated good reliability ($\alpha = 0.77$); (c) convergent validity: high correlation between the original version and the natural language version of academic stress; (d) discriminant validity: low correlation between the natural language version of academic stress and social presence; (e) criterion validity: significant and moderate correlation between the natural language version of academic stress and depression, consistent with prior research [38].

A new exploration is that the study found that the criterion-related validity of the natural language version of academic stress with depression was higher compared to the original version. It may indicate that this measurement tool is more effective in assessing individual academic stress. This is because natural language allows individuals to think about and respond to questions from an ecological perspective [8, 9], leading to more accurate measurements and a closer association with the negative emotions often associated with stress, such as depression. In summary, for RQ1, DHMHAS may be superior in terms of accuracy and reliability. The system is not only scientifically sound but also demonstrates better criterion-related validity.

6.2 Improvement in Motivation to Respond

Study 2, building on the established scientific validity of the DHMHAS system, delves deeper into its potential to enhance the sincerity of student responses. The effectiveness of DHMHAS is verified through comparative analysis of interaction experiences and response variability indices, supplemented by qualitative interviews to address research question 2. The findings indicate that DHMHAS significantly improves the level of earnestness in user responses.

Compared to traditional methods, DHMHAS offered a higher degree of SP due to its natural language interaction capabilities. Qualitative interview results revealed that students were notably impressed with digital humans and experienced a sense of engaging communication, which was stimulated by the highly anthropomorphized digital humans. Additionally, the perceived social presence elicited by DHMHAS led to lower intrinsic and extrinsic cognitive loads during interactions. This reduced cognitive load may lead them to perceive DHMHAS as more user-friendly, thereby enhancing their willingness to respond actively, such as the focused and relaxed responses mentioned in interviews. Behaviorally, this willingness to respond actively was reflected in higher response variability, which is not the result of random responses but rather a thoughtful deliberation by users to respond more accurately and comprehensively, fitting the real-life scenarios.

Furthermore, in the process of measuring academic stress, we found that the reduction in cognitive load was not merely achievable through interface design but required the integration of social interaction elements. According to Maslow's hierarchy of needs, humans have fundamental social needs [45]. When these needs are met, positive experiences emerge. Thus, when students answered surveys through the DHMHAS, the system engaged in dialogue using natural language with highly anthropomorphized digital humans, making them feel as though they were interacting with real people [26, 27]. Since humans are inherently skilled at social interactions, this mode of communication reduces the perceived task difficulty, thereby decreasing the internal cognitive load compared to traditional methods of answering psychological health surveys. Regarding

external cognitive load, users can directly contemplate and recall in response to questions posed by digital humans, without the need to search for key information within lengthy surveys, significantly lightening their cognitive burden. Therefore, this digital-human interaction not only satisfies users' social needs but also optimizes their experience in answering psychological health surveys, making the task feel more relaxed and natural.

Given these potential advantages, we recommend considering the design of a truly human-centered survey response product in future practical activities. Such an interaction form can enhance users' willingness to respond earnestly, thus improving the user experience. Additionally, we hope to continue exploring the boundaries of this survey response method. For instance, research is needed to determine for which populations this interactive method can enhance the answering experience and for which populations traditional survey methods might be more appropriate. We also need to explore the optimal level of anthropomorphism to enhance the answering experience, providing more possibilities for future development.

6.3 Other Factors Influencing Survey Quality and Proposed Theoretical Models

In addition to the variables, we anticipated affecting individual response experiences, interview results have also revealed the advantages of DHMHAS in terms of perceived technological usefulness and other psychological variables, such as positive emotional experiences (e.g., avoiding boredom) and a deeper cognitive understanding of questions. These factors could also be key in influencing whether users are willing to continue responding earnestly.

Based on the Technology Acceptance Model (TAM) [46], when a technology is perceived as useful by an individual, it enhances their positive attitude toward the system, thereby increasing the likelihood of them engaging in positive behaviors. From the current interview results, users could better understand their academic issues and reflect on areas needing improvement after interacting with the system, indicating that they found the system helpful and were thus more willing to engage in the dialogue process. In terms of emotional variables, DHMHAS addresses the dullness associated with traditional psychological assessments [15]. In cognitive terms, the questioning style of DHMHAS has ecological validity, prompting users to provide more detailed and accurate responses [8, 9].

Integrating these quantitative and interview findings, we have constructed a potential theoretical model based on the Theory of Planned Behavior [47] and the TAM [46]. This model explores how to enhance users' positive experiences and strengthen their motivation to respond through highly anthropomorphized means such as digital humans, thereby improving data quality. This model primarily involves three core variables: antecedent variables are individual perceptions of the system (e.g., anthropomorphism and social presence), potential mediating variables are psychological factors, including emotions, cognition, and intentions, which are significant research variables that may interact and influence each other, ultimately affecting the user's actual behavior—whether they are willing to respond earnestly to surveys, as shown in the response variability in Fig. 5.

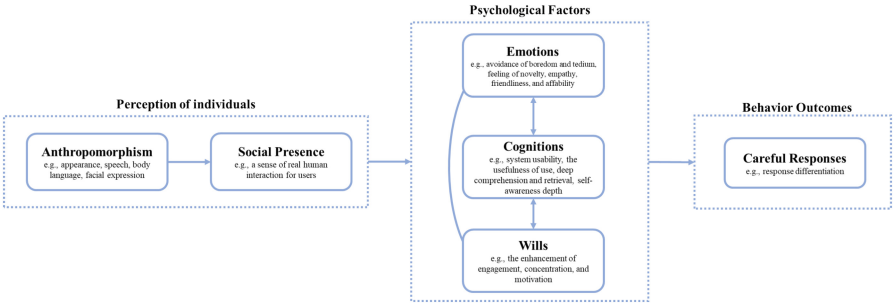


Fig. 5. A theoretical framework for preventing careless responses.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant 62306044, in part by National Social Science and Technology Fund (BGA230254). To enhance the readability of the text, we utilized GPT-4 for grammar checking and refinement.

References

1. Karatekin, C.: Adverse childhood experiences (ACEs), stress and mental health in college students. Stress. Health. **34**, 36–45 (2018). https://doi.org/10.1002/smi.2761
2. Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., Wilens, T.: College students: mental health problems and treatment considerations. Acad. Psychiatry. **39**, 503–511 (2015). https://doi.org/10.1007/s40596-014-0205-9
3. Wu, D., Yu, L., Yang, T., Cottrell, R., Peng, S., Guo, W., Jiang, S.: The impacts of uncertainty stress on mental disorders of Chinese college students: evidence from a nationwide study. Front. Psychol. **11** (2020). https://doi.org/10.3389/fpsyg.2020.00243
4. Zhang, J., Zheng, Y.: How do academic stress and leisure activities influence college students' emotional well-being? A daily diary investigation. J. Adolesc. **60**, 114–118 (2017). https://doi.org/10.1016/j.adolescence.2017.08.003
5. Francavilla, N.M., Meade, A.W., Young, A.L.: Social interaction and internet-based surveys: examining the effects of virtual and in-person proctors on careless response. Appl. Psychol. **68**, 223–249 (2019). https://doi.org/10.1111/apps.12159
6. Meade, A.W., Craig, S.B.: Identifying careless responses in survey data. Psychol. Methods. **17**, 437–455 (2012). https://doi.org/10.1037/a0028085
7. Ward, M.K., Meade, A.W.: Applying social psychology to prevent careless responding during online surveys. Appl. Psychol. Int. Rev. **67**, 231–263 (2018). https://doi.org/10.1111/apps.12118
8. Kjell, O.N.E., Kjell, K., Garcia, D., Sikström, S.: Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. Psychol. Methods. **24**, 92–115 (2019). https://doi.org/10.1037/met0000191
9. Kjell, O.N.E., Sikström, S., Kjell, K., Schwartz, H.A.: Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. Sci. Rep. **12**, 3918 (2022). https://doi.org/10.1038/s41598-022-07520-w

10. Tao, Y., Zou, X., Tang, Q., Hou, W., Wang, S., Ma, Z., Liu, G., Liu, X.: Mapping network connection and direction between anxiety and depression symptoms across the early, middle, and late adolescents: insights from a large Chinese sample. J. Psychiatr. Res. **169**, 174–183 (2024). https://doi.org/10.1016/j.jpsychires.2023.11.035
11. Ward, M.K., Meade, A.W.: Dealing with careless responding in survey data: prevention, identification, and recommended best practices. Annu. Rev. Psychol. **74**, 577–596 (2023). https://doi.org/10.1146/annurev-psych-040422-045007
12. Brühlmann, F., Memeti, Z., Aeschbach, L.F., Perrig, S.A.C., Opwis, K.: The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. Behav. Res. Methods. **56**, 5862–5875 (2024). https://doi.org/10.3758/s13428-023-02321-z
13. Gibson, A.M., Bowling, N.A.: The effects of questionnaire length and behavioral consequences on careless responding. Eur. J. Psychol. Assess. **36**, 410–420 (2020). https://doi.org/10.1027/1015-5759/a000526
14. Toich, M.J., Schutt, E., Fisher, D.M.: Do you get what you pay for? Preventing insufficient effort responding in MTurk and student samples. Appl. Psychol. **71**, 640–661 (2022). https://doi.org/10.1111/apps.12344
15. Ashley, M., Shaughnessy, K.: Predicting insufficient effort responding: the relation between negative thoughts, emotions, and online survey responses. Can. J. Behav. Sci. Rev. Can. Sci. Comport. **55**, 198 (2021)
16. Barber, L.K., Barnes, C.M., Carlson, K.: Random and systematic error effects of insomnia on survey behavior. Acad. Manag. Proc. **2013**, 11338 (2013). https://doi.org/10.5465/ambpp.2013.11338abstract
17. Ward, M.K., Meade, A.W., Allred, C.M., Pappalardo, G., Stoughton, J.W.: Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. Comput. Hum. Behav. **76**, 417–430 (2017). https://doi.org/10.1016/j.chb.2017.06.032
18. Johnson, J.A.: Ascertaining the validity of individual protocols from web-based personality inventories. J. Res. Personal. **39**, 103–129 (2005). https://doi.org/10.1016/j.jrp.2004.09.009
19. Lieberman, M.D.: Social: Why Our Brains Are Wired to Connect. Broadway Books, Portland, OR (2013)
20. Ward, M.K., Pond III, S.B.: Using virtual presence and survey instructions to minimize careless responding on internet-based surveys. Comput. Hum. Behav. **48**, 554–568 (2015). https://doi.org/10.1016/j.chb.2015.01.070
21. Tsai, W.-H.S., Liu, Y., Chuan, C.-H.: How chatbots' social presence communication enhances consumer engagement: the mediating role of parasocial interaction and dialogue. J. Res. Interact. Mark. **15**, 460–482 (2021). https://doi.org/10.1108/JRIM-12-2019-0200
22. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**, 257–285 (1988). https://doi.org/10.1016/0364-0213(88)90023-7
23. Kahneman, D.: Attention and Effort. Prentice-Hall, Hoboken, NJ (1973)
24. Biocca, F., Harms, C., Burgoon, J.K.: Toward a more robust theory and measure of social presence: review and suggested criteria. Presence Teleoperators Virtual Environ. **12**, 456–480 (2003). https://doi.org/10.1162/105474603322761270
25. Gefen, D., Straub, D.: Managing user trust in B2C e-services. E-Serv. J. **2**, 7–24 (2003). https://doi.org/10.2979/ESJ.2003.2.2.7
26. Han, M.C.: The impact of anthropomorphism on consumers' purchase decision in Chatbot commerce. J. Internet Commer. **20**, 46–65 (2021). https://doi.org/10.1080/15332861.2020.1863022
27. Konya-Baumbach, E., Biller, M., von Janda, S.: Someone out there? A study on the social presence of anthropomorphized chatbots. Comput. Hum. Behav. **139**, 107513 (2023). https://doi.org/10.1016/j.chb.2022.107513

28. Heerwegh, D., Loosveldt, G.: Face-to-face versus web surveying in a high-internet-coverage population: differences in response quality. Public Opin. Q. **72**, 836–846 (2008). https://doi.org/10.1093/poq/nfn045
29. Wang, X., Ruan, T., Xu, J., Guo, X., Li, J., Yan, F., Zhao, G., Wang, C.: Expression-aware neural radiance fields for high-fidelity talking portrait synthesis. Image Vis. Comput. **147**, 105075 (2024). https://doi.org/10.1016/j.imavis.2024.105075
30. Nurrohman, A., Abdurohman, M.: High Performance Streaming Based on H264 and Real Time Messaging Protocol (RTMP). In: 2018 6th International Conference on Information and Communication Technology (ICoICT), pp. 174–177 (2018). https://doi.org/10.1109/ICoICT.2018.8528770
31. Kim, J., Kong, J., Son, J.: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (2021). https://arxiv.org/abs/2106.06103
32. Sun, W.-R., Li, J.-H., Liu, L., Kevrekidis, P.G.: The Instabilities Beyond Modulational Type in a Repulsive Bose-Einstein Condensate with a Periodic Potential (2023). https://arxiv.org/abs/2305.00133
33. Cohen, R.J., Swerdlik, M.E., Phillips, S.M.: Psychological Testing and Assessment: an Introduction to Tests and Measurement. Mayfield Publishing, Mountain View, CA (1996)
34. Miller, L.A., Lovler, R.L.: Foundations of Psychological Testing: a Practical Approach. Sage Publications, Thousand Oaks, CA (2018)
35. Lynn, M.R.: Determination and quantification of content validity. Nurs. Res. **35**, 382–385 (1986). https://doi.org/10.1097/00006199-198611000-00017
36. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960). https://doi.org/10.1177/001316446002000104
37. Jayanthi, P., Thirunavukarasu, M., Rajkumar, R.: Academic stress and depression among adolescents: a cross-sectional study. Indian Pediatr. **52**, 217–219 (2015). https://doi.org/10.1007/s13312-015-0609-y
38. Jiang, S., Ren, Q., Jiang, C., Wang, L.: Academic stress and depression of Chinese adolescents in junior high schools: moderated mediation model of school burnout and self-esteem. J. Affect. Disord. **295**, 384–389 (2021). https://doi.org/10.1016/j.jad.2021.08.085
39. Li, H., Boey, K.W.: Development of stress scale for college student. Chin. J. Appl. Psychol. **8**, 27–32 (2002)
40. Lee, K.M., Peng, W., Jin, S.-A., Yan, C.: Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human–robot interaction. J. Commun. **56**, 754–772 (2006). https://doi.org/10.1111/j.1460-2466.2006.00318.x
41. Kurt, K., Spitzer Robert, L.: The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. **32**, 509–515 (2002). https://doi.org/10.3928/0048-5713-20020901-06
42. Klepsch, M., Schmitz, F., Seufert, T.: Cognitive Load Measurement—Naïve Rating Questionnaire (2019). https://doi.org/10.1037/t70903-000
43. Brooke, J.: SUS: A "Quick and Dirty" Usability Scale (1996)
44. Glaser, B., Strauss, A.: Discovery of Grounded Theory: Strategies for Qualitative Research. Routledge, London (2017)
45. Maslow, A.H.: A theory of human motivation. Psychol. Rev. **50**, 370–396 (1943). https://doi.org/10.1037/h0054346
46. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**, 319–340 (1989). https://doi.org/10.2307/249008
47. Ajzen, I.: The Theory of Planned Behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991)

Exploring Satisfaction and Comprehension of Pill Identification Applications: Insights on Design Improvements from Users with Varying Medication Usage Frequencies

Yu Syuan Lin(✉) and Meng-Cong Zheng

Department of Industrial Design, National Taipei University of Technology, Tapiei 10608, Taiwan
linlin1625@gmail.com

Abstract. With the advancement of modern medical technology, the public now has quicker access to medications. Among these advancements, drug identification applications provide more accurate drug information to the public and healthcare providers. The COVID-19 pandemic has further accelerated their usage. Nevertheless, medication errors remain a frequent cause of medical incidents. Previous studies on drug identification systems have primarily focused on professional medical contexts, with limited attention given to their use in the daily lives of the general public. Therefore, this study aims to: (1) evaluate how users with varying medication frequencies utilize drug identification applications; (2) examine the impact of pill information visualization on these application interfaces; and (3) investigate how users identify pill information based on the appearance or textual attributes of pill.

The study employed simulated experiments with participants categorized as high-experience and low-experience medication users, who completed surveys for analysis. Following task completion, semi-structured interviews were conducted. The results indicate that high-experience medication users significantly improved their understanding of drugs through the current pill identification systems. However, low-experience users faced excessive information overload, making it difficult to quickly absorb specialized pill knowledge. Regarding pill information, low-experience users relied more heavily on pill images, whereas high-experience users placed greater emphasis on technical terms and professional information. Both technical terms and pill appearance are critical components of pill identification systems. Thus, designing systems that deliver visually improved interfaces to facilitate quick comprehension for both high- and low-experience users is an essential consideration for future development.

Keywords: Pill identification mobile applications · Mobile healthcare · User experience

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 41–52, 2025.
https://doi.org/10.1007/978-3-031-93502-2_3

1 Introduction

Self-medication is a treatment of health problems to oneself by use of medicines that are bought freely in drug stores for their own initiative without medical advice [1]. According to CDC data published in 2021, 37.0% of adults used telemedicine in 2020, including provider visits, online prescriptions, and mental health services [2]. The COVID-19 pandemic has significantly impacted global healthcare systems, highlighting the importance of advancements in artificial intelligence and big data technologies. By integrating medicine, pharmacy, and information technology, highly efficient and accurate drug identification systems have been developed. These systems not only enhance the efficiency of clinical healthcare but also provide robust support for individual home health management. The limited transparency of information throughout the chain of supply often results in an insufficient transparency of data and challenges in establishing the authenticity of Information pertaining to medications [3]. Pill identification systems also possess the capability to detect counterfeit and substandard medications, preventing them from entering the market and reaching patients, thereby safeguarding patient health and safety. These systems provide detailed medication instructions and information on side effects, assisting patients in correctly understanding and using medications. This, in turn, enhances medication adherence and therapeutic outcomes.

Taiwan's comprehensive National Health Insurance system has led to a growing demand for medication usage. According to the 2022 National Health Insurance medical statistics, the number of prescriptions dispensed in Taiwan reached 107,654,057 cases [4]. However, this increasing demand is accompanied by risks of drug confusion and misuse, stemming from a lack of understanding of medication information. It is estimated that in Taiwan, at least 300 deaths annually are attributed to traffic accidents caused by "drugged driving" [5].

Medication-related harm in New Zealand general practice after weighting was 73.9 harms per 1000 patient years; the incidence rate of potentially preventable medication-related harm was 15.6 harms per 1000 patient–years [6]. The development of pill identification technology can effectively mitigate these risks, ensuring that patients use medications correctly and access accurate drug information. This contributes to improved therapeutic outcomes and enhanced safety.

Patients and their spouses or carers will in many cases be the first to notice any observable problem that arises from a medication error. In some cases, adverse events are immediate or well known. However, other events may be delayed or unfamiliar and patients may not realize that the problem has anything to do with the medication they are taking [7].

This study examines how individuals with varying levels of medication experience interpret and comprehend the information provided by pill identification applications, including pill appearance, indications, contraindications, side effects, and drug-drug interactions (DDIs). Although it is currently challenging to determine whether usability and user experience impact the accuracy of the symptom checkers, enhancements and optimizations in symptom screening and the interface of medical advice could contribute to increased trust and user experience [8]. We aim to propose recommendations for improving the information design of future drug identification applications. These enhancements are expected to help the public use medications more effectively and

effortlessly, promote accurate pill identification, and highlight the potential impact of visual design on pill identification systems.

Research on pill identification not only holds significant academic value but also carries substantial social and economic implications. It has the potential to drive advancements in medical technology and improve societal health outcomes.

2 Research Method

2.1 Pre-study Questionnaire Survey

This study begins with a questionnaire survey to investigate the prevalence and current usage of existing drug identification applications. The survey includes the following aspects: (1) the general public's medication frequency, (2) prior use of pill identification systems, (3) typical scenarios in which drug information is searched, (4) types of devices used for queries, (5) common methods for searching drug information, and (6) ranking of the importance of drug information. Based on the criteria set by National Health Insurance Administration, Ministry of Health and Welfare, where non-daily routine medications are defined as being used at most once per month, the threshold for high-experience medication use is defined as using medication at least once per month. The questionnaire examines the impact of drug information on high-experience and low-experience medication users. Additionally, the survey explores differences in drug information search behaviors between high-experience and low-experience medication users, as well as their levels of emphasis on drug information.

2.2 Procedure

The experiment will utilize random sampling to assign 32 participants into two groups: 16 high-experience participants and 16 low-experience participants. Each participant will operate either the sample WebMD APP or the sample Medscape APP. Before and after the simulated tasks, participants are required to complete the Drug Health Education Information Scale to assess the impact of operating the pill identification system on drug health education. The simulated task consists of seven operational tasks, with a total duration of approximately 40 min. Upon completing the simulated tasks, participants will fill out a satisfaction questionnaire and rank the importance of pill information. Following the completion of these tasks, a semi-structured interview will be conducted to evaluate participants' perceptions of the pill identification system.

2.3 Simulated Task

We selected two drug identification applications, the sample WebMD APP and the sample Medscape APP, based on download frequency and user ratings. A total of 32 participants with varying levels of medication experience were recruited and divided into two groups: 16 high-experience participants and 16 low-experience participants. These participants were invited to perform simulated task experiments using the two applications.

The simulated tasks required participants to utilize pill appearance features within the application interfaces to search for pill information and answer questions related to the

use of pill identification applications. The tasks were divided into seven subtasks, which included querying pill names, pill ingredients, pill indications, pill contraindications, drug-drug interactions (DDI), and pill side effects. In the seventh task, participants switched to the other pill identification system to perform similar queries, enabling a comparative analysis of interface usability and operational differences between the two systems (Fig. 1).

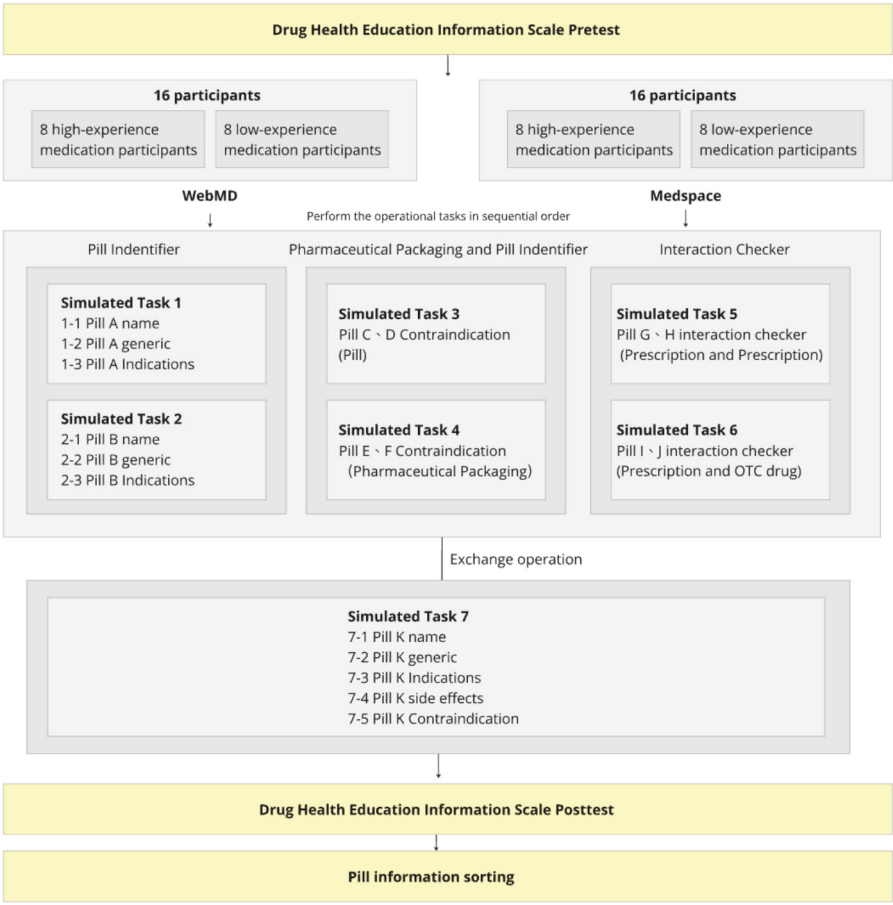


Fig. 1. Simulated Task Procedure.

2.4 Experimental Sample

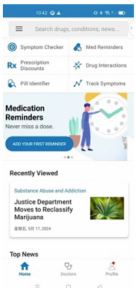
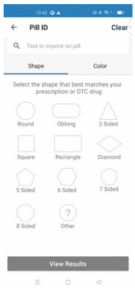
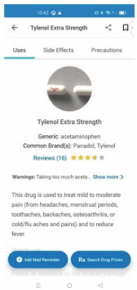
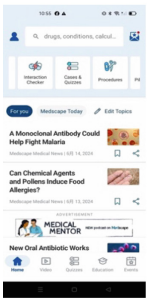
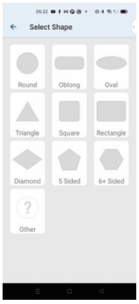
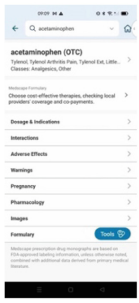
Based on the analysis and integration conducted in 2024 through the App Store and Play Store, WebMD APP and Medscape APP were selected as experimental samples for this study. The selection was guided by five criteria: inclusion of a pill indention

system, provision of drug-drug interaction information, availability of comprehensive pill information, exceeding five million downloads, and high user ratings.

The experimental samples simulate the process of querying pill information and provide photos that match the size of actual pill and pill packaging. The simulated task begins upon opening the interfaces of the sample WebMD APP and sample Medscape APP. Participants are not timed during the operation process and may proceed until completing the task questions or opting to forgo answering.

The sample interfaces include features for pill indentifier, pill information, and interaction checker, ensuring a comprehensive simulation experience for the participants (Table 1).

Table 1. Sample APP Interface.

Sample	Home page	Pill Indentifier	Pill information	Interaction Checker
WebMD				
Medscape				

2.5 Scale

Drug Health Education Information Scale Before and after the tasks, participants were required to complete the Drug Health Education Information Scale [9] as a basis for efficacy analysis. The scale evaluates medication use, drug identification app operation, usability, and satisfaction. The scale includes the following items: (1) Perceived

difficulty and accuracy in following proper medication procedures. (2) Agreement that supplemental health education Information tools assist in understanding medications being used or consumed. (3) Understanding the indications and precautions for the medications being used or consumed. (4) Awareness that drug identification apps provide timely and accurate usage instructions and precautions. (5) Knowledge that drug information is readily accessible via mobile phones or tablets. (6) Proper use of pill identification apps. (7) Understanding that pill identification apps retrieve data via online connections without additional costs.

Satisfaction Survey A satisfaction survey based on the Drug Information Satisfaction Scale [9] was used to assess difficulty, behavior, usability, and overall satisfaction. The survey items included: (1) Clarity of the interface display, including image quality and font size. (2) Convenience and timeliness of obtaining pill information. (3) Helpfulness of the information in understanding medication usage and precautions. (4) Effectiveness in alleviating difficulties in correct medication use or consumption. (5) Overall satisfaction with the integration of drug identification apps into health education.

Importance Ranking of Drug Information Participants were also asked to rank the importance of various types of pill information, including pill indications, dosages, drug-drug interactions, pill appearance filtering, related clinical cases, pill side effects, medication use during pregnancy, and pill images. This ranking aimed to assess the prioritization of information by participants with differing levels of experience.

3 Results

3.1 Pre-study Questionnaire Survey Results

This study surveyed 180 participants, comprising 37 individuals with high medication experience and 143 individuals with low medication experience. The results indicate that smartphones are the most commonly used devices for searching drug-related information, with 88% of respondents reporting their primary reliance on smartphones for this purpose. Accordingly, the sample operations in this study were designed based on smartphone interfaces.

The questionnaire allowed respondents to select multiple scenarios for searching drug information. Among the 180 participants, 132 reported searching for information when they were unclear about a drug's side effects or usage timing. This suggests that limited healthcare resources during medical consultations may prevent individuals from thoroughly understanding their prescribed medications.

Analysis of information prioritization revealed that only 5% of respondents expressed an interest in understanding drug-drug interactions (DDIs). However, all participants are potentially at risk of DDIs, particularly elderly patients, for whom the incidence rate of DDIs reaches 54%. This finding underscores the urgent need to enhance public awareness of DDIs.

Further analysis demonstrated that participants with low medication experience ($M = 3.5946$, $p = 0.007$) rated image-based searches using non-specialized tools (e.g., Google or ChatGPT) significantly higher than those with high medication experience

(M = 3.8112). This indicates that low-experience users are more inclined to leverage user-friendly search engines for image-assisted queries. Conversely, participants with high medication experience (M = 4.0000, p = 0.017) scored significantly higher in using professional drug-related websites for information retrieval compared to low-experience participants (M = 3.7692), reflecting their greater reliance on professional drug information sources.

3.2 Results of Participant Sample Operations

At this stage, data analysis was conducted for the experimental samples WebMD APP and Medscape APP. The experimental results were analyzed using the chi-square test to evaluate the performance of 32 participants during simulated tasks querying information within the pill identification systems. The queried information included manufacturer pill names, ingredients, indications, contraindications, drug-drug interactions (DDIs), and side effects.

According to the chi-square test results, participants using WebMD APP achieved statistically significant outcomes in querying the indications for Pill A (p = 0.001) and the manufacturer name for Pill B (p = 0.007). Additionally, in Task 7, participants switched systems. Those who initially used the Medscape APP sample and then operated WebMD APP for the first time showed significant performance in querying the manufacturer name of Pill K (p = 0.006) and the active ingredient name (p = 0.014).

However, participants using the Medscape APP sample demonstrated significantly better performance than WebMD APP users in Task 3 and Task 4, specifically when querying contraindications via pill appearance (p = 0.002) and querying contraindications using pill packing images (p = 0.012).

3.3 Scale Results

Understanding of Drug Health Education Information by Participants According to the paired sample t-test results of pre- and post-assessments, participants showed significant changes in their proficiency and accuracy awareness of drug usage before and after the simulation. After the simulation, participants more strongly agreed with the statement "The proficiency and accuracy of drug usage steps is a major concern" (t = −6.226, p = 0.006), indicating an increased perception of challenges related to drug usage. Additionally, after the simulation, participants significantly increased their support for "agreeing that other auxiliary educational methods can help understand the drugs being used or taken" (t = −3.304, p = 0.002). These results suggest that the simulation not only increased participants' awareness of the challenges in the drug usage process but also emphasized the importance of supplementary educational information in enhancing drug knowledge and correct usage.

An analysis of the post-test data from the drug health education scale was conducted, comparing the differences in responses to drug identification system information between high-experience and low-experience groups using an independent sample t-test. The results showed that high-experience participants performed significantly better on several indicators than low-experience participants. Specifically, in the rating for " Awareness that drug identification apps provide timely and accurate usage instructions

and precautions," high-experience participants exhibited significantly higher responses ($t = 1.665$, $p = 0.005$); in " Proper use of pill identification apps," the high-experience group also had significantly higher ratings than the low-experience group ($t = 2.390$, $p = 0.019$); furthermore, in the evaluation of " Knowledge that drug information is readily accessible via mobile phones or tablets.," high-experience participants showed a significant statistical difference ($t = 2.119$, $p = 0.035$). Overall, high-experience participants scored significantly higher in drug health education, indicating that the drug identification system responds more effectively to high-experience users.

Participant Satisfaction with Sample Operation Based on the statistical analysis, high-experience participants showed significantly higher satisfaction than low-experience participants regarding the pill identification system's ability to improve pill usage challenges. In the item " Effectiveness in alleviating difficulties in correct medication use or consumption.," high-experience participants scored significantly higher ($t = 3.823$, $p = 0.001$), indicating stronger affirmation of the system's assistance in correct pill usage. Additionally, in the item " Helpfulness of the information in understanding medication usage and precautions.," high-experience participants also showed a significant increase in satisfaction ($t = 2.517$, $p = 0.002$). In summary, high-experience participants demonstrated significantly higher overall satisfaction with the pill identification system compared to low-experience participants, highlighting the system's potential advantage in meeting the needs of the low-experience group.

Importance Ranking of Pill Information by High- and Low-Experience Participants According to the results of the Mann-Whitney U analysis on the importance ranking of information in the simulated task questionnaire, participants with high-experience were significantly more aware that pill are not 100% safe and should be used with caution, compared to those with low-experience. They were more proactive in seeking information about "medication-related case reports" ($P = .028$), from which they could learn how to use medication properly, manage side effects, and avoid drug interactions in order to protect their health. This result highlights differences in information needs between the two groups. Low-experience participants tended to perceive lower risk associated with case reports, which may lead them to overlook the importance of medication safety and increase the risk of drug-related problems.

3.4 Summary

This study examines the impact of simulated tasks and pill identification systems on medication health education. The findings reveal that participants developed a deeper awareness of the challenges associated with medication use following the simulation and expressed stronger support for supplementary educational methods. Furthermore, significant differences were observed in the effectiveness and satisfaction with the drug recognition system based on participants' levels of experience. High-experience participants outperformed their low-experience counterparts across various metrics, including proper use of the drug recognition system and timely access to medication information, while also exhibiting greater satisfaction with the system's ability to alleviate challenges in medication use. Conversely, low-experience participants placed higher importance on intuitive visualized information and drug indications, reflecting their reliance on visual

and application-oriented information. Overall, the simulation and pill recognition system significantly enhanced participants' understanding of medication use. The observed differences in information needs and system evaluations between experience levels provide valuable insights for the development of more inclusive medication education tools (Table 2).

Table 2. Data table highlighting significant findings between high-experience and low-experience participants in this study.

Scale		High experience (M)	Low experience (M)	t-value	p-value
Pre-study Questionnaire Survey Results	Image searches using search engines such as Google and ChatGPT	3.5946	3.8112	−1.117	0.007**
	Drug-related websites (e.g., Ministry of Health and Welfare websites, major hospital websites)	4.0000	3.7692	1.268	0.017*
Drug Health Education Information Scale Pretest	Awareness that drug identification apps provide timely and accurate usage instructions and precau-tions	4.2500	3.5625	1.665	0.005**
	Proper use of pill identification apps	3.9375	3.0000	2.390	0.019*
	Knowledge that drug information is readily accessible via mobile phones or tablets	4.5625	3.9375	2.119	0.035*

(*continued*)

Table 2. (*continued*)

Scale		High experience (M)	Low experience (M)	t-value	p-value
Satisfaction Survey	Effectiveness in alleviating difficulties in correct medication use or consumption	4.0000	2.5625	3.823	0.001***
	Helpfulness of the information in understanding medication usage and precautions	3.9375	2.9375	2.517	0.002**

Note: *p < 0.05. **p < 0.01. ***p < 0.001

4 Discussion

During the semi-structured interviews, participants evaluated the experimental samples, WebMD APP and Medscape APP samples. Among them, 75% expressed a preference for WebMD APP sample, emphasizing its simplicity and clarity. WebMD APP sample was praised for its clear presentation of pill images, use of simpler vocabulary, and straightforward display of pill names. Conversely, Medscape APP sample was commended for providing section headers, more detailed professional information, and a well-structured presentation of contraindication data.

A total of 51% of participants, regardless of their level of experience, considered pill images essential for verifying their medication information. Some participants suggested integrating pill images into systems for identifying drug interactions and appearances. Regarding information design, 40% of participants noted that the current content contained excessive information and professional jargon, while 21% recommended organizing the information into tables and using simpler, more conversational language to cater to the general public.

In terms of interface design, 40% of participants preferred WebMD APP's single-page layout for better comprehension, as it allowed continuous reading without navigation interruptions. Although Medscape APP presented content with section headers, participants suggested minimizing the need for repeated page switching during information retrieval to prevent disruptions in reading flow.

Finally, in the satisfaction survey, 66% of participants agreed that searching for pill based on their appearance was an effective method for identifying medications.

5 Conclusion

This study aims to investigate the impact of drug identification system information on the experience of the general public, particularly examining the relationship between interface design and the evaluation of individuals with varying levels of experience with medications. Based on the experimental results, several areas requiring improvement in existing drug identification systems were identified: (1) **Inadequate Adaptability for Low-Experience Users**: While high-experience participants demonstrated higher efficiency and satisfaction with the system, low-experience users provided relatively lower evaluations of the system's overall performance, indicating room for improvement in meeting the needs of this group. (2) **Lack of Intuitive Visual Information**: Low-experience users emphasized the importance of pill images and intuitive presentations in their information needs. Current system designs have yet to fully address these requirements, potentially causing difficulties in drug identification. (3) **Limited Presentation of Application-Oriented Information**: The study revealed that low-experience users place greater emphasis on drug indications. However, existing systems may not sufficiently highlight the timeliness and accessibility of such information, limiting their effectiveness for these users. (4) **Insufficient Educational Support Features**: Although the systems have shown success in enhancing users' awareness of proper medication use, their educational support features require further enhancement, particularly in alleviating the concerns of low-experience users and providing more guided content.

Future system improvements could incorporate the strengths of WebMD APP and Medscape APP samples: (1) Combining Simplicity and Professionalism: Integrate WebMD APP's simple and clear interface with drug image displays alongside Medscape APP's structured section headers and detailed professional information to design a system that balances usability and expertise. (2) Tiered Information Presentation: Offer graded presentation options catering to different user groups, such as a professional terminology version for experienced users and a simplified version for the general public. (3) Enhanced Drug Image Features: Optimize the application of pill images, expanding their functionality beyond identification to include queries for drug interactions and contraindications, improving the accuracy of appearance-based identification. (4) Data Presentation Optimization: Organize pill information into visualized tables, emphasizing key data such as indications, contraindications, and interactions, while providing detailed options to minimize excessive technical jargon. (5) Educational Guidance: Include instructional aids for low-experience users where appropriate, enhancing educational outcomes.

Future developments should also consider application scenarios and functional expansion to support public education and awareness campaigns. These efforts could enhance the overall level of drug literacy, reduce medication-related harm, and alleviate healthcare burdens.

Limitations. This study faced several limitations. The sample size of 32 participants may not be sufficient to represent a broader user population, especially across different age groups, educational backgrounds, or levels of professional knowledge. Additionally, the experiment utilized an English interface, which could introduce linguistic limitations for participants more accustomed to Chinese. Furthermore, given the frequent updates

to drug information, the systems' ability to reflect the latest data may have influenced user experience and the study's conclusions.

References

1. Swastika, W., et al.: Preliminary study of multi convolution neural network-based model to identify pills image using classification rules. In: 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), vol. 8, pp. 376–380. IEEE, Surabaya, Indonesia (2019). https://doi.org/10.1109/ISITIA.2019.8937272
2. US Digital Pharmacy Market Size and Forecast (2020-2030): Regional Share, Trends, and Growth Opportunity Analysis Coverage: by Drug Type (Prescription Medicine and Over-the-Counter Medicine), Product, Platform, and Gender (Male and Female). The Insight Partners (2023)
3. Kalpana, R., Priya, P.I., Manju, M., Anitha, R., Venusamy, K.: Implementing a unique identification system for improved traceability and prevention of counterfeit drugs in the supply chain. In: 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, pp. 1175–1181. IEEE (2023). https://doi.org/10.1109/ICSSAS57918.2023.10331781
4. https://dep.mohw.gov.tw/dos/lp-5103-113-xCat-y112.html
5. https://today.line.me/tw/v2/article/vX6k6RQ
6. Leitch, S., et al.: Medication-related harm in New Zealand general practice: a retrospective records review. Br. J. Gen. Pract. **71**(709), e626–e633 (2021). https://doi.org/10.3399/BJGP.2020.1126
7. Britten, N.: Medication errors: the role of the patient. Br. J. Clin. Pharmacol. **67**(6), 646–650 (2009). https://doi.org/10.1111/j.1365-2125.2009.03421.x
8. Lin, W.A., Zheng, M.-C., Wang, L.-J.: Exploring the usability and user experience of the symptom checker interface. In: 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) (2024). https://doi.org/10.54941/ahfe1004836
9. C.-H. Huang, Effectiveness Analysis of Drug Health Education Information by QR Code Importation, 2016

Evaluating Usability, Acceptance, and Trust in the BEAMER Adherence Intelligence Visualization Platform: A Preliminary Study

Beatriz Merino-Barbancho[1(✉)], Miguel Rujas[1], Rodrigo Martín Gómez del Moral Herranz[1], Kristina Livitckaia[2], Konstantina Pantelidou[2], Elena Castellano[3], Konstantina Kostopoulou[4], Maria Fernanda Cabrera[5], Maria Teresa Arredondo[5], and Giuseppe Fico[1]

[1] Life Supporting Technologies Research Group, Universidad Politécnica de Madrid, Avda Complutense 30, 28040 Madrid, Spain
bmerino@lst.tfo.upm.es
[2] Centre for Research and Technology Hellas (CERTH), Thessaloniki, Central Macedonia, Greece
[3] MySphera, Valencia, Spain
[4] i-sprint, Brussels Area, Belgium
[5] Universidad Politécnica de Madrid, Madrid, Spain
{chiqui,mta}@lst.tfo.upm.es

Abstract. The BEAMER Adherence Intelligence Visualization Platform (AIVP) is a digital health tool designed to support healthcare professionals in managing patient adherence to treatment. This preliminary study evaluates the platform's usability, acceptance, and trust among healthcare professionals and researchers using a mixed-methods approach. Data were collected through face-to-face (n = 11) and online (n = 9) evaluations, and analyzed using the System Usability Scale (SUS), Technology Acceptance Model (TAM), and Human-Computer Trust Model (HCTM). Results showed an overall SUS score of 70%, indicating acceptable usability, with higher scores in online evaluations (74.72%) compared to face-to-face (66.14%). TAM scores reflected positive acceptance (74.25%), and HCTM scores demonstrated moderate trust (67.64%), with consistent results across evaluation methods. Differences between evaluation formats suggest the platform may be particularly suited for remote healthcare settings. These findings highlight the BEAMER AIVP's potential as a promising tool for improving medication adherence while identifying areas for enhancement, such as reducing complexity and building trust. Future research should focus on refining the platform and exploring its integration into clinical workflows.

Keywords: Usability · Acceptance · Trust

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 53–71, 2025.
https://doi.org/10.1007/978-3-031-93502-2_4

1 Introduction

Lack of adherence to treatment is a key obstacle to improve clinical outcomes, especially in patients with chronic diseases such as cardiovascular disease or diabetes [1]. According to the World Health Organization (WHO), this issue not only impacts patient health but also places a considerable burden on healthcare systems around the world [2].

In recent years, digital health technologies have emerged as promising solutions to improve treatment adherence and patient engagement [3]. These tools, which include mobile health apps, telehealth platforms, and smart devices, offer various features such as reminders, real-time monitoring, and personalized feedback to support patients in managing their health more effectively [4].

Recent studies have shown that digital health interventions can improve adherence rates [3]. However, the effectiveness of these tools varies and not all of them address patient behavioral aspects, therefore stronger evidence is needed to support their widespread implementation in clinical care. The BEAMER Adherence Intelligence Visualization Platform (AIVP) aims to address this gap by offering a comprehensive solution that leverages advanced visualization techniques to help healthcare professionals monitor and improve patient adherence.

This preliminary study aims to assess the usability, acceptance, and trust of the BEAMER Adherence Intelligence Visualization Platform among healthcare professionals and researchers. Using a mixed-methods approach that combines both face-to-face and online evaluations, we aim to collect evidence-based information on a variety of human factors including user experience.

2 Materials and Methods

2.1 BEAMER

BEAMER (BEhavioral and Adherence Model for improving quality, Effectiveness, and efficiency in treatment ecosystems pRoject) [5] is a European funded initiative (grant agreement No 101034369), involving 28 public and private European partners from academia, civil society, and industry. The project aims to develop a disease-agnostic model that segments the population based on actionable factors and predicts treatment adherence behavior to help build digital health solutions and promote better outcomes for all.

2.2 Adherence Intelligence Visualization Platform (AIVP)

The AIVP is a versatile, modular platform designed for use by healthcare professionals (HCPs) and recruited patients. This platform utilizes advanced visualization techniques to present adherence data and insights in a comprehensive format for HCPs to clearly view patients' adherence profiles and their corresponding information cards, facilitating personalized support and adherence management.

2.3 Study Design

This study employed a mixed-methods approach, combining face-to-face and online assessments to evaluate the ease of use, acceptability, and trust of the BEAMER AIVP among healthcare professionals and researchers. All evaluations occurred in 3 phases: 1) meeting organization and full explanation of the project and the tool; 2) user exploration of the tool; 3) participant response to questionnaires.

The dynamics of the online evaluations consisted of approximately 60′ meetings on a 1:1 basis with each participant. The face-to-face evaluation was carried out jointly with all the participants in groups of 3–4 people in the same room. The face-to-face evaluation took place at the Ciudad Jardín health center in Madrid belonging to the Madrid Health Service (SERMAS) [6].

2.4 Data Collection

Three standardized questionnaires were used to gather data:

1. **System Usability Scale (SUS)** [7]**:** A 10-item questionnaire used to assess the platform's usability.
2. **Technology Acceptance Model (TAM)** [8]**:** Used to measure user acceptance of the technology, focusing on perceived usefulness and ease of use**.**
3. **Human-Computer Trust Model (HCTM)** [9]**:** Applied to evaluate user trust in the system, considering factors such as reliability and competence.

2.5 Data Analysis

Upon collecting the various questionnaires, their answers were analyzed by categorizing them into three distinct groups: face-to-face evaluations, online evaluations, and overall evaluations. This approach allowed for independent analysis of the results based on the two different evaluation methods, while also enabling the extraction of generalized conclusions.

Demographic variables were processed using a standard descriptive statistical analysis, which involved calculating the mean and standard deviation for numerical variables, as well as the frequency and relative percentage for categorical variables. The total SUS scores were computed using the standard SUS methodology: subtracting 1 from the scores of odd-numbered questions, subtracting the scores of even-numbered questions from 5, summing the results, and multiplying by 2.5 to generate the final score. For both the TAM and HCTM, the score was derived by calculating the weighted average of the responses to obtain a score out of 100. Additionally, a descriptive analysis, along with box plots and radar charts, was performed for each questionnaire to facilitate interpretation of the overall distribution and the per-question distribution of the scores obtained.

3 Results

This section presents the results derived from the different analyses conducted. To enhance clarity and ease of interpretation, the results are organized by questionnaire.

3.1 Participant Demographics

3.1.1 Demographics

Tables 1, 2, and 3 present the results of the descriptive statistical analyses for the face-to-face, online, and overall assessments.

Table 1. Face-to-face demographics results.

Numerical characteristic	Mean	Standard deviation
Age	44.50	13.91
Number of years working	20.55	18.58
Categorical Characteristic	N	%
Gender		
Female	11	100.00%
Male	0	00.00%
Country		
Germany	0	00.00%
Spain	11	100.00%
Italy	0	00.00%
Portugal	0	00.00%
Greece	0	00.00%
The Netherlands	0	00.00%
Norway	0	00.00%
Role in the organisation		
HCP/Medical staff	11	100.00%
Researcher/Academia	0	00.00%
ML Experience		
Yes	3	27.27%
No	8	72.73%

The face-to-face evaluation group consisted of 11 participants, with an average age of 44.50 years (SD = 13.91) and an average of 20.55 years of professional experience (SD = 18.58). All participants were Spanish women working as HCPs or medical staff, and 72.73% had no prior experience with ML or AI technologies. In contrast, the online evaluation group comprised 9 participants with an average age of 46 years (SD = 9.06) and an average of 20.78 years of professional experience (SD = 8.81). This group had twice as many men as women, with participants from 7 different countries, predominantly Spain and Portugal. Of these participants, 55.56% worked in research or academia, while 44.44% were HCPs or medical staff. Additionally, 88.89% of this group had previous experience with ML systems. When combining both groups for the overall evaluation, the

Table 2. Online demographics results.

Numerical characteristic	Mean	Standard deviation
Age	46.00	9.06
Number of years working	20.78	8.81
Categorical characteristic	N	%
Gender		
Female	3	33.33%
Male	6	66.67%
Country		
Germany	1	11.11%
Spain	2	22.22%
Italy	1	11.11%
Portugal	2	22.22%
Greece	1	11.11%
The Netherlands	1	11.11%
Norway	1	11.11%
Role in the organisation		
HCP/Medical staff	4	44.44%
Researcher/Academia	5	55.56%
ML Experience		
Yes	8	88.89%
No	1	11.11%

total sample included 20 participants, with an average age of 45.17 years and 20.65 years of professional experience. Of this combined group, 70% were women, and Spain was the most represented country (65%). Additionally, 75% were HCPs or medical staff and only 55% had prior experience with ML systems.

3.2 System Usability Scale results

3.2.1 SUS

The descriptive analysis of SUS scores across the three different analyses are presented in Tables 4, 5, and 6.

For the face-to-face group, the mean total SUS score is 66.14% (SD = 12.06), with the lowest mean score recorded for question 2. The online group, on the other hand, achieved a higher mean total score of 74.72% (SD = 9.39), with question 6 receiving the lowest score. When combining both groups, the overall mean score reaches 70% (SD = 11.53), with both questions 2 and 6 identified as the lowest scoring items.

Table 3. Overall demographics results.

Numerical characteristic	Mean	Standard deviation
Age	45.17	11.69
Number of years working	20.65	14.64
Categorical characteristic	N	%
Gender		
Female	14	70.00%
Male	6	30.00%
Country		
Germany	1	5.00%
Spain	13	65.00%
Italy	1	5.00%
Portugal	2	10.00%
Greece	1	5.00%
The Netherlands	1	5.00%
Norway	1	5.00%
Role in the organisation		
HCP/Medical staff	15	75.00%
Researcher/Academia	5	25.00%
ML Experience		
Yes	11	55.00%
No	9	45.00%

Table 4. Face-to-face SUS Results.

	SUS-1	SUS-2	SUS-3	SUS-4	SUS-5	SUS-6	SUS-7	SUS-8	SUS-9	SUS-10	Score
Count	11	11	11	11	11	11	11	11	11	11	11
Mean	3.45	2	3.64	2.73	4	2.27	3.55	2.45	3.45	2.18	66.14
Std	0.82	0.77	1.21	0.90	0.45	0.47	0.82	1.04	1.04	0.75	12.06
Min	2	1	1	2	3	2	2	1	2	1	45
25%	3	2	3.5	2	4	2	3	2	3	2	57.5
50%	3	2	4	2	4	2	4	2	3	2	70
75%	4	2	4	3.5	4	2.5	4	3	4	3	75
max	5	4	5	4	5	3	5	4	5	3	85

Additionally, Figs. 1 and 2 provide further insight by illustrating radar plots for individual question scores and box plots of the total scores.

Table 5. Online SUS results.

	SUS-1	SUS-2	SUS-3	SUS-4	SUS-5	SUS-6	SUS-7	SUS-8	SUS-9	SUS-10	Score
Count	9	9	9	9	9	9	9	9	9	9	9
Mean	3.89	2.11	4.22	2.33	4.33	1.78	3.78	1.89	3.78	2	74.72
Std	0.60	0.78	0.44	1.41	0.5	0.67	0.83	0.60	0.44	0.5	9.39
Min	3	1	4	1	4	1	2	1	3	1	65
25%	4	2	4	1	4	1	4	2	4	2	67.5
50%	4	2	4	2	4	2	4	2	4	2	70
75%	4	2	4	3	5	2	4	2	4	2	80
max	5	4	5	5	5	3	5	3	4	3	92.5

Table 6. Overall SUS results.

	SUS-1	SUS-2	SUS-3	SUS-4	SUS-5	SUS-6	SUS-7	SUS-8	SUS-9	SUS-10	Score
Count	20	20	20	20	20	20	20	20	20	20	20
Mean	3.65	2.05	3.9	2.55	4.15	2.05	3.65	2.2	3.6	2.1	70
Std	0.75	0.76	0.97	1.15	0.49	0.60	0.81	0.89	0.82	0.64	11.53
Min	2	1	1	1	3	1	2	1	2	1	45
25%	3	2	4	2	4	2	3	2	3	2	64.38
50%	4	2	4	2	4	2	4	2	4	2	70
75%	4	2	4	3.25	4	2	4	3	4	2.25	76.25
max	5	4	5	5	5	3	5	4	5	3	92.5

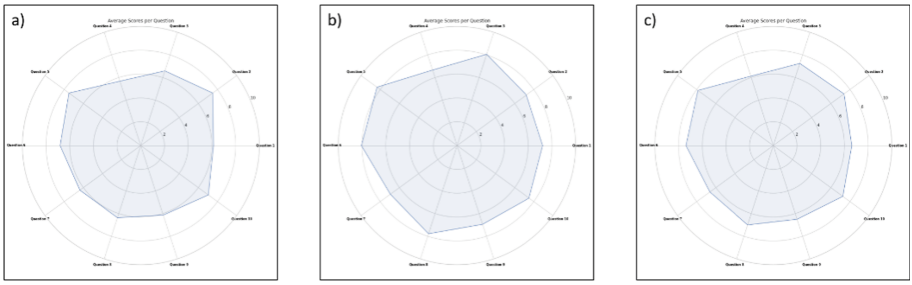


Fig. 1. Radar Plot Comparison of SUS Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

As illustrated, the score for each question is consistently lower in the face-to-face evaluation compared to the online assessment, which is reflected in the overall scores. This trend also extends to the total SUS scores, where the face-to-face group exhibits lower mean scores and greater variability compared to the online group.

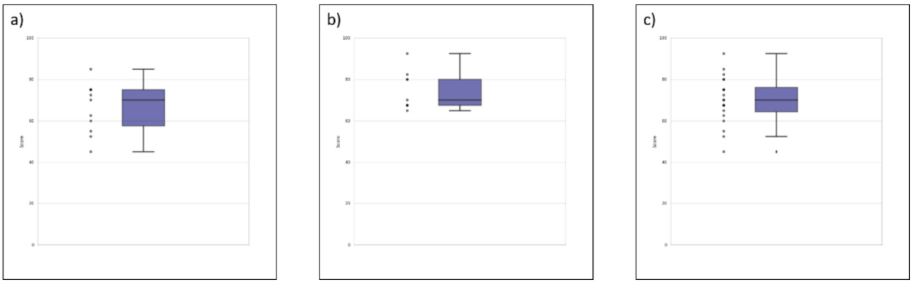


Fig. 2. Box Plot Comparison of SUS Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

3.3 Technology Acceptance Model results

3.3.1 TAM

The descriptive analysis of TAM scores across the three different analyses are presented in Tables 7, 8, and 9.

In the face-to-face assessment, the average TAM score was 72.27% (SD = 13.23), with the highest mean scores observed for questions concerning perceived usefulness and ease of use (mean > 4.00). The online assessment exhibited slightly higher acceptance, with an average score of 76.67% (SD = 5.77). This group showed particularly strong agreement on the perceived effectiveness of the technology, scoring above 4.00 on several items. Combining the results from both groups, the overall average TAM score was 74.25% (SD = 10.55). The combined data show a consistent appreciation for the technology's benefits across all questions, with scores generally above 3.00 and several items averaging above 4.00.

Additionally, Figs. 3 and 4 provide further insight by illustrating radar plots for individual question scores and box plots of the total scores.

The radar plots highlight the distribution of scores across different questions, showing higher scores in the online assessment across most dimensions. The box plots illustrate the central tendency and dispersion of the total scores, showing a higher overall acceptance in the online format compared to the face-to-face format.

Table 7. Face-to-face TAM results.

	TAM-1	TAM-2	TAM-3	TAM-4	TAM-5	TAM-6	TAM-7	TAM-8	TAM-9	TAM-10	TAM-11	TAM-12	Score	%
Count	11	11	11	11	11	11	11	11	11	11	11	11	11	11
Mean	2.81	3.64	3.09	3.64	3.64	3.64	4	3.83	4.09	3.82	4	3.18	43.36	72.27
Std	1.17	1.03	0.94	1.03	1.03	0.92	1	0.87	0.70	1.17	0.77	1.08	7.94	13.23
Min	1	2	2	2	2	2	2	2	3	2	2	1	26	43.33
25%	2	3	2	3	3	3.5	3.5	3.5	4	3	4	2.5	40	66.67
50%	3	4	3	4	4	4	4	4	4	4	4	4	46	76.67
75%	3.5	4	4	4	4	4	5	4	4.5	5	4	4	48.5	80.83
max	5	5	4	5	5	5	5	5	5	5	5	4	52	86.67

Table 8. Online TAM results.

	TAM-1	TAM-2	TAM-3	TAM-4	TAM-5	TAM-6	TAM-7	TAM-8	TAM-9	TAM-10	TAM-11	TAM-12	Score	%
Count	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	3.89	4	3.44	3.78	3.56	3.89	4.11	3.78	3.78	3.67	4.22	3.89	46	76.67
Std	0,60	0,5	0,53	0,44	0,53	0,33	0,60	0,67	0,83	1,12	0,44	0,78	3,46	5,77
Min	3	3	3	3	3	3	3	3	2	1	4	3	42	70
25%	4	4	3	4	3	4	4	3	4	4	4	3	43	71,67
50%	4	4	3	4	4	4	4	4	4	4	4	4	45	75
75%	4	4	4	4	4	4	4	4	4	4	4	4	48	80
max	5	5	4	4	4	4	5	5	5	5	5	5	52	86,67

Table 9. Overall TAM results.

	TAM-1	TAM-2	TAM-3	TAM-4	TAM-5	TAM-6	TAM-7	TAM-8	TAM-9	TAM-10	TAM-11	TAM-12	Score	%
Count	20	20	20	20	20	20	20	20	20	20	20	20	20	20
Mean	3,3	3,8	3,25	3,7	3,6	3,75	4,05	3,8	3,95	3,75	4,1	3,5	44,55	74,25
Std	1,08	0,83	0,79	0,80	0,82	0,72	0,83	0,77	0,76	1,12	0,64	1	6,33	10,55
Min	1	2	2	2	2	2	2	2	2	1	2	1	26	43,33
25%	2,75	3,75	3	3	3	4	4	3	4	3	4	3	43	71,67
50%	3,5	4	3	4	4	4	4	4	4	4	4	4	46	76,67
75%	4	4	4	4	4	4	5	4	4	4,25	4	4	48,25	80,42
max	5	5	4	5	5	5	5	5	5	5	5	5	52	86,67

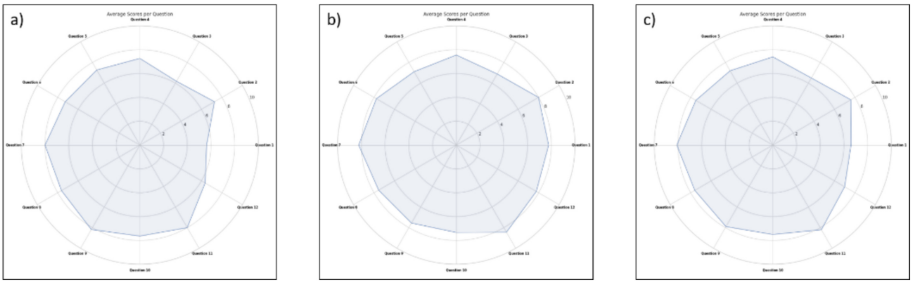


Fig. 3. Radar Plot Comparison of TAM Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

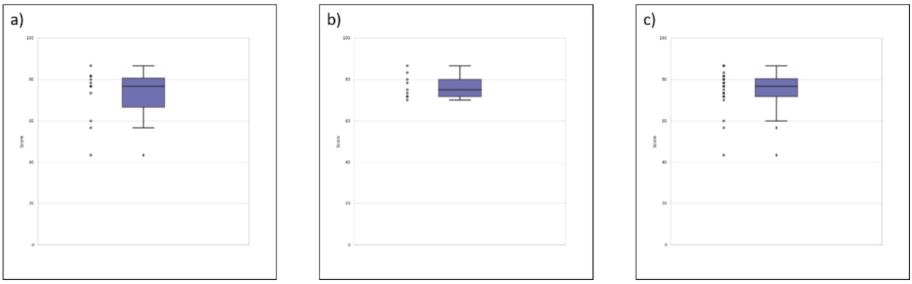


Fig. 4. Box Plot Comparison of TAM Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

3.4 Human-Computer Trust Model

3.4.1 HCTM

The descriptive analysis of HCTM scores across the three different analyses are presented in Tables 10, 11, and 12.

For the face-to-face group, the average trust score was 67.27% (SD = 7.41), with the highest trust ratings observed in the last item, indicating a strong trust in the reliability of the system used. The online group showed a slightly higher trust score average of 68.08% (SD = 3.97), with consistent trust levels across most questions. Combining the results from both groups, the overall trust score average settled at 67.64% (SD = 5.98), reflecting a stable trust level in the technology's reliability and usability across all participants.

Additionally, Figs. 5 and 6 provide further insight by illustrating radar plots for individual question scores and box plots of the total scores.

These plots reveal that while online participants tended to show slightly higher trust scores, the overall distribution indicates a solid foundation of trust in the technology used among all participants.

Table 10. Face-to-face HCTM results.

	HCTM-1	HCTM-2	HCTM-3	HCTM-4	HCTM-5	HCTM-6	HCTM-7	HCTM-8	HCTM-9	HCTM-10	HCTM-11	Score	%
Count	11	11	11	11	11	11	11	11	11	11	11	11	11
Mean	2,27	3,18	2	3,55	3,64	3,45	3,73	3,73	3,55	3,91	4	37	67,27
Std	0,47	0,98	0,45	0,69	0,81	0,82	0,47	0,79	0,69	0,94	1	4,07	7,41
Min	2	2	1	2	2	2	3	2	2	2	2	27	49,09
25%	2	2,5	2	3	3	3	3,5	3,5	3	3,5	3,5	36,5	66,36
50%	2	3	2	4	4	4	4	4	4	4	4	37	67,27
75%	2,5	4	2	4	4	4	4	4	4	4,5	5	38,5	70
max	3	5	3	4	5	4	4	5	4	5	5	42	76,36

Table 11. Online HCTM results.

	HCTM-1	HCTM-2	HCTM-3	HCTM-4	HCTM-5	HCTM-6	HCTM-7	HCTM-8	HCTM-9	HCTM-10	HCTM-11	Score	%
Count	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	2,78	3,11	2,11	3,67	3,89	3,56	3,89	3,67	3,56	3,33	3,89	37,44	68,08
Std	0,67	0,78	0,60	0,50	0,33	0,53	0,60	0,50	0,73	0,87	0,33	2,19	3,97
Min	2	2	1	3	3	3	3	3	3	2	3	33	60
25%	2	3	2	3	4	3	4	3	3	3	4	37	67,27
50%	3	3	2	4	4	4	4	4	3	4	4	38	69,09
75%	3	4	2	4	4	4	4	4	4	4	4	38	69,09
max	4	4	3	4	4	4	5	4	5	4	4	41	74,55

Table 12. Overall HCTM results.

	HCTM-1	HCTM-2	HCTM-3	HCTM-4	HCTM-5	HCTM-6	HCTM-7	HCTM-8	HCTM-9	HCTM-10	HCTM-11	Score	%
Count	20	20	20	20	20	20	20	20	20	20	20	20	20
Mean	2,50	3,15	2,05	3,60	3,75	3,50	3,80	3,70	3,55	3,65	3,95	37,20	67,64
Std	0,61	0,88	0,51	0,60	0,64	0,69	0,52	0,66	0,69	0,93	0,76	3,29	5,98
Min	2	2	1	2	2	2	3	2	2	2	2	27,00	49,09
25%	2	2,75	2	3	3,75	3	3,75	3	3	3	4	36,75	66,82
50%	2	3	2	4	4	4	4	4	4	4	4	37,50	68,18
75%	3	4	2	4	4	4	4	4	4	4	4	38,25	69,55
max	4	5	3	4	5	4	5	5	5	5	5	42,00	76,36

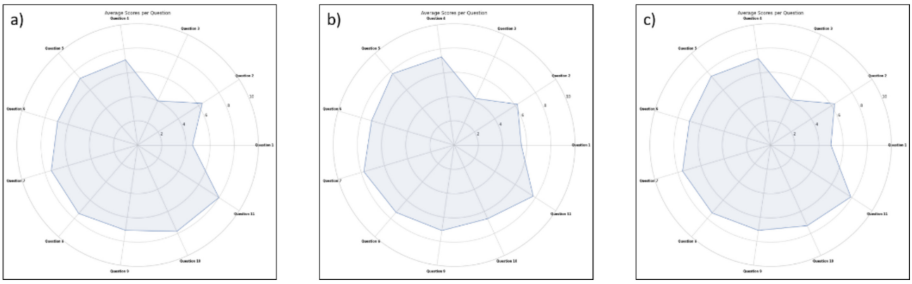


Fig. 5. Radar Plot Comparison of HCTM Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

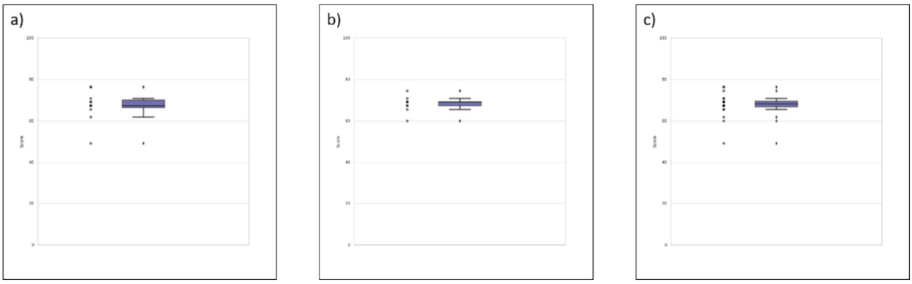


Fig. 6. Box Plot Comparison of HCTM Scores: (**a**) Face-to-Face, (**b**) Online, (**c**) Overall.

4 Discussion

The highlights of some of the aspects that were revealed during the study are as follows:

4.1 Usability

The overall SUS score of 70% indicates that the BEAMER AIVP has acceptable usability, falling within the industry standard benchmark of 68 (SD 12.5) for digital health applications. This score suggests that while the platform is functional, there is room for improvement in its user interface and overall user experience. Interestingly, the online group reported higher usability (74.72%) compared to the face-to-face group (66.14%). This discrepancy could be attributed to several factors:

- **Familiarity with digital tools:** The online group, which had a higher proportion of participants with prior experience in machine learning systems (88.89% vs 27.27%), may have found the platform more intuitive due to their technological background.
- **Context of use**: The online setting might have provided a more relaxed environment for participants to explore the platform, potentially leading to a more positive usability assessment.

The lower scores on specific SUS questions, particularly questions 2 and 6, could indicate potential areas for enhancement in the platform's complexity and consistency. These findings align with previous research highlighting the importance of reducing complexity

and enhancing consistency in digital health applications to improve user experience and adoption rates.

4.2 Acceptance

The TAM results show a generally positive acceptance of the BEAMER AIVP, with an overall score of 74.25%. This score suggests that healthcare professionals perceive the platform as both useful and easy to use, which are critical factors in the adoption of new technologies in healthcare settings. The online group demonstrated slightly higher acceptance (76.67%) compared to the face-to-face group (72.27%). This trend is consistent with the usability findings and may be attributed to similar factors. The higher acceptance among online participants could indicate that the BEAMER AIVP is particularly well-suited for monitoring adherence in healthcare settings. The strong agreement on items related to perceived usefulness and ease of use (mean scores >4.00) is particularly encouraging, as these factors have been identified as key determinants of technology adoption in healthcare. This suggests that the BEAMER AIVP has successfully addressed some of the primary concerns that often hinder the adoption of new technologies in clinical practice.

4.3 Trust

The HCTM scores indicate a moderate level of trust in the BEAMER AIVP, with an overall score of 67.64%. This score suggests that while healthcare professionals generally trust the platform, there is still room for improvement in building confidence in its reliability and effectiveness. The similarity in trust levels between face-to-face (67.27%) and online (68.08%) groups suggests that the platform's perceived reliability is consistently moderate across different user contexts, aligning with initial expectations. This consistency is important, as trust is a crucial factor in the adoption and continued use of digital health technologies. The highest trust ratings observed in items related to the system's reliability are particularly encouraging, as reliability is a critical factor in healthcare technology adoption. However, the moderate overall trust score indicates that further efforts may be needed to enhance healthcare professionals' confidence in the BEAMER AIVP, particularly in areas such as data security, accuracy of recommendations, and integration with existing clinical workflows.

4.4 Implications for Digital Health Solutions

The findings of this study have several implications for the development and implementation of digital health tools aimed at improving treatment adherence:

- **User-centered design:** The varying usability scores between online and face-to-face groups highlight the importance of considering different user contexts and preferences in the design of digital health platforms. Future iterations of the BEAMER AIVP should focus on improving usability for a diverse range of users, including those with varying levels of technological expertise as well as different user roles and permissions.

- **Integration with clinical workflows:** Despite the overall positive acceptance, the moderate trust scores suggest that efforts should be made to better integrate the BEAMER AIVP with existing clinical workflows. This could involve collaborating closely with healthcare professionals to ensure that the platform complements, rather than disrupts, their current practices.
- **Personalization:** The higher acceptance scores for perceived usefulness suggest that the BEAMER AIVP successfully addresses a real need in clinical practice. Future developments should focus on enhancing the platform's ability to provide personalized recommendations and interventions, which has been shown to improve the effectiveness of digital health interventions for medication adherence.
- **Evidence-based approach:** The trust score may indicate a need for more robust and explicit evidence regarding the results presented, particularly in demonstrating measurable improvements in patient outcomes. Additionally, there is a need to offer clearer insights into how the underlying algorithms function, including their methodologies, decision-making processes, and accuracy. This includes explaining how patients are grouped by these algorithms to ensure that the process is both clinically meaningful and aligned with evidence-based standards. Clearer communication and validation of these aspects are essential to enhance trust and adoption among clinicians, patients, and stakeholders.

Generally, the BEAMER AIVP shows promise as a tool for improving patient adherence management, with overall positive usability, acceptance, and trust scores.

5 Conclusions

Preliminary evaluation of BEAMER's Adherence Intelligence Visualization Platform demonstrates promising results in terms of usability, acceptance and trust among healthcare professionals and researchers. The platform shows particular strength in its perceived usefulness and ease of use, as demonstrated by high TAM scores. However, the study also identifies opportunities for improvement, particularly in reducing complexity and increasing the consistency of the user interface.

Acknowledgments. The project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101034369. This joint undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme, the European Federation of Pharmaceutical Industries and Associations (EFPIA) and Link2Trials. This communication reflects the views of the authors and neither the IMI nor the European Union, EFPIA, or Link2Trials are liable for any use that may be made of the information contained herein.

We would like to thank all the healthcare professionals involved, especially those GP from Ciudad Jardín Madrid.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Danielson, E., Melin-Johansso, C., Modanloo, M.: Adherence to treatment in patients with chronic diseases: from alertness to persistence. Int. J. Commun. Based Nurs. Midwif. **7**(4), 248 (2019). https://doi.org/10.30476/IJCBNM.2019.81303.0
2. Kisa, A., Sabaté, E., Nuño-Solinís, R., Karkashian, C.: Adherence to long-term therapies : evidence for action. J. Nurs. Scholarsh. **35**(3), 207 (2003)
3. Ridho, A., et al.: Digital health technologies to improve medication adherence and treatment outcomes in patients with tuberculosis: systematic review of randomized controlled trials. J. Med. Internet Res. **24**(2), e33062 (2022). https://doi.org/10.2196/33062
4. Pistorius, C.: Developments in Emerging Digital Health Technologies, vol. 1, p. 17. DeltaHedron Innovation Insight (2017)
5. BEAMER Project—Improving Health Adherence Solutions. https://beamerproject.eu/. Accessed 30 Jan 2025
6. Salud | Comunidad de Madrid. https://www.comunidad.madrid/servicios/salud. Accessed 30 Jan 2025
7. Brooke, J.: SUS—A Quick and Dirty Usability Scale, pp. 189–194 (1996)
8. Davis, F.: A Technology Acceptance Model for Empirically Testing New End-User Information Systems (1985)
9. Gulati, S., Sousa, S., Lamas, D.: Human-Computer Trust Scale (2020). https://doi.org/10.1037/t75524-000

Constructing A Conversation-First Dialogue Model for Comprehensive, Evidence-Based Counseling for HPV Vaccination for Young Adults

Lu Tang[1], Chenying Weng[1], Serena Hou[2], Lara S. Savas[3], Ana C. Neumann[4], Nicole Moore[4], Jane Hamilton[5], Daniel Rhee[3], Cui Tao[6], Licong Cui[7], and Muhammad "Tuan" Amith[5,8,9(✉)]

[1] Department of Communication and Journalism, Texas A&M University, College Station, TX 77843, USA
ltang@tamu.edu
[2] Emory University, Atlanta, GA, USA
[3] School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA
[4] School of Dentistry, University of Texas Health Science Center of Houston, Houston, TX, USA
[5] Department of Psychiatry and Behavioral Science, University of Texas Health Science Center at Houston, Houston, TX, USA
muamith@utmb.edu
[6] Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL, USA
[7] McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA
[8] Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, Houston, TX, USA
[9] Department of Internal Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract. The human papillomavirus (HPV) vaccine has been shown to prevent several types of cancers; however, vaccination rates remain suboptimal. Effective patient-provider communication is key to promoting vaccination but is often hindered by various clinical barriers. Speech bots could be a promising solution by addressing common barriers and facilitating HPV vaccination uptake. We used a "conversation first" design approach to develop the interaction for speech bots to deliver HPV vaccination counseling to young adults. Two counseling strategies were developed and piloted: motivational interviewing (MI) and the theory-based approaches. Wizard of Oz experiments were conducted with young female adults (n = 24). Preliminary evidence shows that MI yielded higher usability than the theory-based approach. The theory-based approach significantly improved HPV vaccine beliefs and attitudes. These findings led to the development of a hybrid conversational model for future pilot testing and the eventual creation of a computable format for automated speech bots.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 72–84, 2025.
https://doi.org/10.1007/978-3-031-93502-2_5

Keywords: conversational agents · chat bots · cancer · patient-provider communication · health belief model · motivational interviewing

1 Introduction

Human papillomavirus (HPV) is a group of viruses responsible for various cancers and genital warts. Most sexually active individuals contract HPV at least once in their lifetime. HPV vaccination can prevent over 90% of HPV-related cancers and certain precancers. HPV vaccines are recommended for 11 and 12-year-old children but may be started as early as age 9 and as late as age 26. For adults aged 27 to 45, patients and clinicians are encouraged to make decisions together [1].

Despite the proven effectiveness and safety of the HPV vaccine, HPV vaccination rates in the United States have plateaued in recent years [2]. Healthy People 2030 set the goal of an 80% vaccination rate among adolescents aged 13–17 in the U.S., yet only 62.6% of this group were up to date with their HPV vaccinations in 2022 and this number decreased to 61.4% in 2023 [3]. Notably, 2022 marked the first year since 2013 that the HPV vaccination rate among this group failed to increase. Among young adults aged 18 to 26, HPV vaccination rates stagnated at around 50% between 2019 and 2022 [4]. Major barriers to HPV vaccination include limited access, lack of knowledge, fear, vaccine misinformation, and the lack of physician recommendation [5].

Physician counseling has proven to be the most effective method for promoting HPV vaccination [6]. However, clinical appointments are typically short, and physicians do not always have enough time or persuasion skills to promote vaccines. Speech bots offer a solution by simulating clinician-patient communication about HPV and HPV vaccination. These embodied conversational agents actively engage patients in a manner comparable to speaking with a healthcare provider. Overall, the primary goal of a speech bot is not to replace clinicians but to complement their efforts in a clinical environment, nudging patient to engage in meaningful discussions with their healthcare providers.

To design the conversation between the bot and the user, we employed a "conversation first" approach [7]. This method focuses on the design of natural language interaction using words in short interactive utterances rather than relying on visual interfaces org other modalities. It emphasizes using words as the primary means of dialogue management irrespective of devices or machine types [8]. The ultimate goal is to create human-like conversational experience, allowing users to interact with the bot in a way that mirrors natural human communication.

Conversations can potentially be modeled as finite state models or network representations [9, 10], where an utterance is denoted as a node and the directional links denote multiple routes that a conversation can be carried towards. Counseling strategies, such as motivational interviewing [11], provide structured and methodical frameworks that can be formalized and reused within these models. This approach allows for greater flexibility and scalability in designing conversational systems.

Conversation design can take different approaches. One frequently used approach is motivational interview (MI). It is a collaborative counseling style that encourages individuals to find intrinsic motivation to change their behavior [11] and has been shown to

effectively reduce vaccine hesitancy [12, 13]. Developed by William Miller and Stephen Rollick, MI helps individuals explore and resolve ambivalence about changing health behaviors (e.g., substance abuse, exercise, eating habits, etc.). This strategy enables patients to reflect on their own actions without feeling pressured or directed by the provider. The key strength of this approach is its empathic and patient-centric nature, allowing patients to self-assess their behavior, which has been proven effective across a variety of health applications.

Another well-established approach to health education is the theory-based approach. Past research has consistently shown that theory-based health interventions are typically more effective than non-theory based ones in promoting vacation and other health behaviors [14]. This approach focuses on presenting medical knowledge consistent with psychological or behavioral theories (such as the Health Belief Model [15, 16]) to change recipients' health attitudes and behaviors toward better health outcomes [17].

In this study, we developed and tested two speech bots for HPV vaccine promotion based on the MI approach and the Theory-Based approach. We wanted to answer the following three RQs: (1) *What is the overall usability of a speech bot promoting HPV vaccination?* (2) *Do speech bots taking the MI approach and the Theory-based approach differ in usability?* (3) *Is the speech bot effective in changing users' HPV vaccine-related beliefs and attitudes?*

2 Methods

2.1 Wizard of OZ Experiments

We employed the Wizard of OZ [18] experiment to test the two speech bots. The data collection involved two software components: a natural language interface and a remote desktop controller. Participants/users interact with the natural language interface under the impression that the machine is an operational automated speech bot, while a "wizard" controlled the natural language interface through the remote desktop controller. The natural language interface was developed in iOS SDK and designed to interface seamlessly with the remote desktop application controller. It was deployed on an Apple iPad and performed text-to-speech (i.e., converting text from the remote desktop application into speech) and speech-to-text transcription (i.e., transcribing user speech user into text). The desktop application was developed in Java and served as the remote controller of the natural language interface. It enabled the "wizard" to transmit text (from the scripts of the dialogue model) and collect the utterance exchanges in a chat log format to allow for subsequent analysis. See Fig. 1 for the layout of the Wizard of OZ experiment.

2.2 Script Development

2.2.1 Motivational Interviewing Script

We developed a conversation model incorporating strategies rooted in motivational interviewing and other evidence-based theories, such as the Transtheoretical model (TTM) [19] and the Health Belief Model [15, 16]. The network-based script ("model") was



Fig. 1. Layout for Wizard of OZ experiments.

developed with input from an MI expert (JH) and outlined using a graphical tool with draw.io.

The MI script included two phases: an exploratory phase and an informational phase. The script's skeletal structure is modeled after the TTM, directing patients from pre-contemplation and ambivalence about a health topic toward behavior change and steps to action. Throughout the stages, interaction was developed and embedded based on techniques utilized in MI counseling, such as the OARS technique. The OARS technique includes four components: (1) asking an *open-ended question* to empower the patient, (2) providing an *affirmation* statement to build trust, (3) *reflective listening* to validate the patient's statement, and (4) *summarizing* to track the patient's engagement. Another technique we embedded are *readiness scales* to measure and inquire the likelihood of the patient's personal feelings about behavioral change. Lastly, we also incorporated the Health Belief Model within the script to help tailor the input of the patient's perception of the HPV vaccine. Figure 2 shows a simplified example of these techniques and their applications [30].

2.2.2 Theory-Based Script

A second script was developed based on scientific information about HPV and HPV vaccine, grounded in two health behavior theories: the Health Belief Model (HBM) [15, 16] and the Theory of Planned Behavior (TPB) [20]. Both theories aim to promote individuals' health behaviors by creating positive attitudes, changing perceived social norms, and increasing the self-efficacy of the recipients. More specifically, the script emphasizes the severity and susceptibility of HPV infection, the benefit of the HPV vaccine, the ease of obtaining the HPV vaccine (self-efficacy), and the social norm of getting HPV vaccines (see Fig. 3 for the theoretical constructs used in the script.) We adapted the language used to be more casual and conversational for the younger audience. Like the first script, the second script was generated as a network representation using a graphical notational tool.

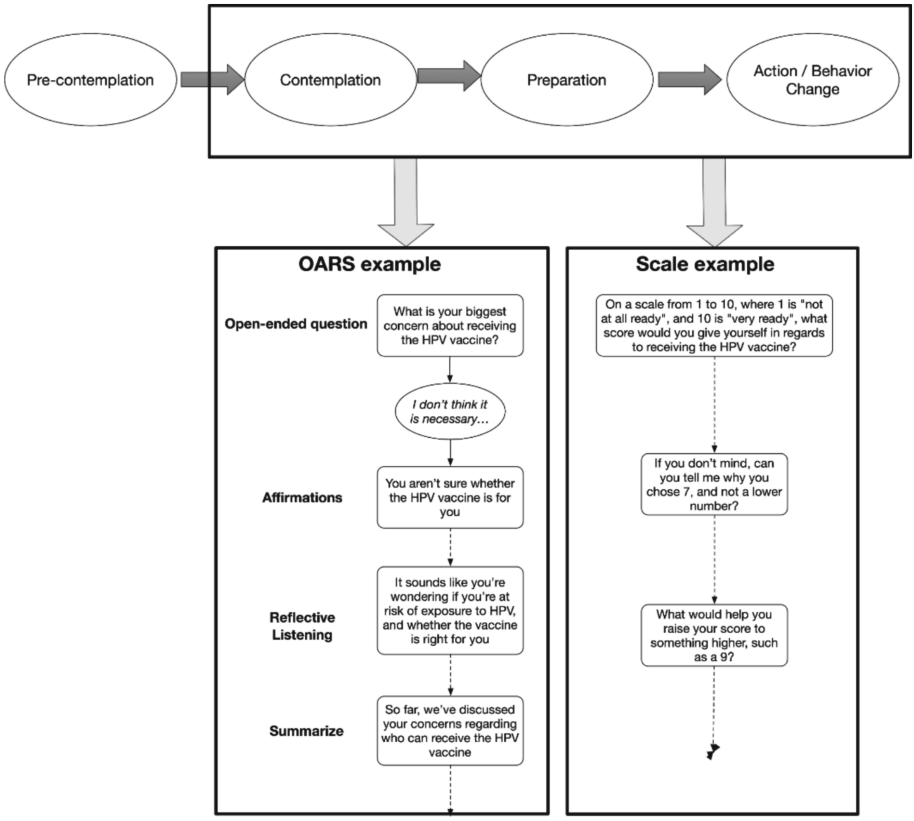


Fig. 2. Framework describing the theories integrated into developing a motivational interviewing script. Examples of OARS and readiness scale are shown from the script.

2.3 Setting and Population

Participants were recruited from several large lecture classes in social sciences at a large public university in a Southern state in the United States. They need to be at least 18 years of age to be eligible. In the end, 14 participants completed Study 1, and 10 completed Study 2. Each participant was given a gift card ($15) for their participation.

2.4 Survey Design and Measurement

2.4.1 Study Design

Study 1 and Study 2 used slightly different designs. Study 1 followed a post-test-only design where participants interacted with the speech bot first and then completed a paper-and-pencil survey measuring perceived usability, HPV vaccine attitudes and beliefs, and demographic characteristics including HPV vaccination status.

Study 2 employed a pre-and post-test design. Participants completed a brief pretest survey on Qualtrics the day before their scheduled lab interaction with the speech bot. The

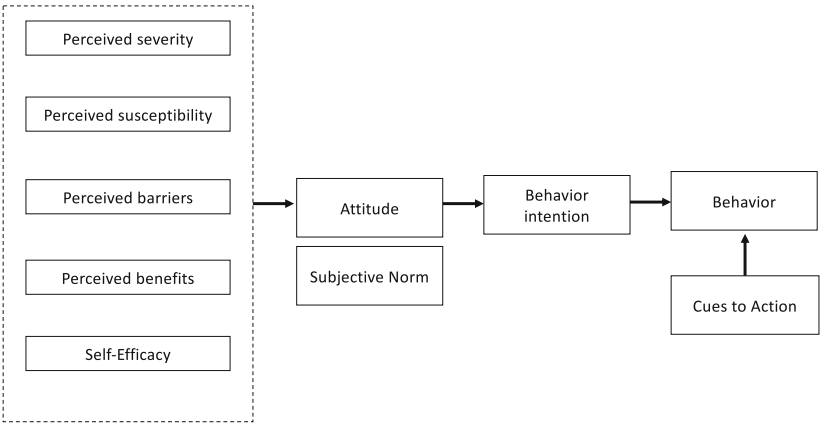


Fig. 3. Framework of the psychological variables incorporated into the development of the Theory-Based Script.

pretest measured their HPV vaccination status and vaccine attitudes. After interacting with the speech bot in the lab, participants filled in a paper-and-pencil questionnaire measuring the perceived usability of the bot, HPV vaccine beliefs and attitudes, and demographic information. This design allowed the assessment of the bot's effectiveness in changing users' vaccine beliefs and attitudes. Below are the measurements used in both studies.

2.4.2 Measurement Scales

The System Usability Scale (SUS) survey was used to measure the usability of the speech bot. SUS is one of the most widely used tools for assessing the perceived usability of a system or a product. It includes 10 five-point Likert scale questions and has been validated as a reliable measure of users' experience with various systems and products [21, 22]. The final score is calculated based on adjusting odd and even-numbered items to reach 100 points [23]. SUS is often used to compare the usability of systems. Sometimes, researchers use a modified version of SUS called SUS-MOD [24], which is the SUS scale minus the first question ("Would they use [Speech bot] frequently") when the users are unlikely to use the technology for multiple times. If a real-world speech bot was deployed in a clinical setting, removing this question would be essential to attain an accurate usability assessment, as the bot may only be used infrequently by the patient.

The Speech User Interface and Service Quality (SUISQ) was also used to measure the speech bot's usability [25, 26], focusing specifically its voice interface. This 25-item scale measures the usability of speech-based user interfaces such as a speech bot. It measures the following four dimensions: *User Goal Orientation* (UGO), *Customer Service Behavior* (CSB), *Speech Characteristics* (SC), and *Verbosity* (Vr) [27]. *User Goal Orientation* assesses the system's efficiency, the user's trust in it, and the clarity of the interface. *Customer Service Behavior* measures attributes such as the system's friendliness, politeness, speaking pace, and the use of familiar language. *Speech Characteristics*

factor pertains to the naturalness and enthusiasm of the system's voice. Finally, *Verbosity* measures how talkative and repetitive the system is.

The *Carolina HPV Immunization Attitude and Belief Scale* (CHIAS) was used to measure HPV vaccine attitudes [28]. CHIAS is an established tool to measure the attitudes and beliefs regarding the HPV vaccine. It measures four dimensions: perceived benefit, perceived barriers, perceived harm, and uncertainty about vaccine effects. We utilized the version that is tailored for young adults [29].

Demographic information of the participants was also collected, including age, gender, race/ethnicity, year in college, family income, and HPV vaccination status.

3 Results

Two arms of the study were conducted in the Spring of 2024 with IRB approval from our respective review board (HSC-SPH-22-1039). For the motivational interviewing arm (Study 1), 14 female participants were recruited between the ages of 18 and 26. These participants were majority non-Hispanic white ($n = 9$) with slightly more individuals who indicated they had the HPV vaccine ($n = 6$), while the remainder were unsure ($n = 4$) or did not have the HPV vaccine ($n = 4$). The second arm involved the Theory-based Intervention Approach (Study 2) where we recruited 10 female participants between the ages of 19 and 22. Four out of the ten reported that they had the HPV vaccine, and half of the participants were non-Hispanic white with the rest Hispanic white. Table 1 displays the breakdown of the demographic for the two arms.

Table 1. Participants' Demographic Information ($N = 24$).

	Study 1	Study 2
Age	Mean: 20.4 ($SD = 1.00$), range: 18–26	Mean: 20.1 ($SD = 1.10$), range: 19–22
Gender	Female: 14	Female: 10
Race/ethnicity	Non-Hispanic White: 9 Hispanic White: 2 Asian: 3	Non-Hispanic White: 5 Hispanic White: 5
Parents income	Less than $25,000 (n = 1), $25,000–34,000 (n = 2), $50,000–$74,999 (n = 2), $75,000–$99,999 (n = 1), $100,1000–$149,999 (n = 4), $150,000 or more (n = 4)	$35,000–49,000 (n = 2), $50,000–$74,999 (n = 0), $75,000–$99,999 (n = 3), $100,1000–$149,999 (n = 2), $150,000 or more (n = 2), Unknown (n = 1)
Vaccination status	Vaccinated (n = 6) Not vaccinated (n = 4) Not sure (n = 4)	Vaccinated (n = 4) Not vaccinated (n = 2) Not sure (n = 4)

3.1 Overall Usability of Speech Bots

Overall, the two speech bots demonstrate moderate levels of usability (See Table 2). We ran two one-sample t-tests to compare the SUS (MOD) of the two speech bots to the average usability of most interactive voice interfaces of 72 and found the differences to be statistically insignificant (Bot 1: $t(13) = -1.23$, $p = .24$; Bot 2: $t(9) = -1.82$, $p = .10$.)

Table 2. Descriptive statistics of the SUS and SUISQ of the two speech bots.

Usability measurement	Study 1	Study 2
SUS (Overall)	$M = 58.93$, $SD = 6.47$	$M = 57.25$, $SD = 25.62$
SUS (mod)	$M = 63.50$, $SD = 6.92$	$M = 49.75$, $SD = 9.31$
Q1: I think that I would like to use this [Speech bot] frequently	$M = 1.71$, $SD = 1.14$	$M = 1.40$, $SD = 1.26$
Q2: I found [Speech bot] unnecessarily complex	$M = 1.64$, $SD = 1.50$	$M = 1.70$, $SD = 1.34$
Q3: I thought [Speech bot] was easy to use	$M = 2.71$, $SD = 1.33$	$M = 2.40$, $SD = 1.51$
Q4: I think that I would need the support of a technical person to be able to use [Speech bot]	$M = 1.29$, $SD = 1.73$	$M = 1.70$, $SD = 1.49$
Q5: I found the various functions in [Speech bot] were well integrated	$M = 2.71$, $SD = 1.20$	$M = 2.30$, $SD = 1.06$
Q6: I thought there was too much inconsistency in [Speech bot]	$M = 1.29$, $SD = .99$	$M = 1.40$, $SD = 1.35$
Q7: I would imagine that most people would learn to use [Speech bot] very quickly	$M = 2.43$, $SD = 1.70$	$M = 2.80$, $SD = 1.23$
Q8: I found [Speech bot] very cumbersome to use	$M = 2.14$, $SD = 1.03$	$M = 2.10$, $SD = .99$
Q9: I felt very confident using [Speech bot]	$M = 1.93$, $SD = 1.33$	$M = 2.10$, $SD = 1.37$
Q10: I needed to learn a lot of things before I could get going with [Speech bot]	$M = 1.57$, $SD = 1.70$	$M = 1.20$, $SD = 1.23$
SUISQ (Overall)	$M = 4.66$, $SD = .70$	$M = 3.78$, $SD = .92$
UGO	$M = 4.68$, $SD = 1.29$	$M = 3.70$, $SD = 1.29$
CSB	$M = 6.33$, $SD = .62$	$M = 5.40$, $SD = .85$
SC	$M = 3.91$, $SD = 1.26$	$M = 2.98$, $SD = .86$
Vr	$M = 3.73$, $SD = 1.23$	$M = 3.03$, $SD = 1.10$

3.2 Comparison of Two Speech Bots

An independent samples t-test for the System Usability Scale (SUS) showed no significant difference in SUS scores between the speech bot using the motivational interview approach ($M = 58.93$, SD = 24.19) and the speech bot using the theory-based approach ($M = 57.25$, $SD = 25.62$), $t(22) = .16$, $p = .87$.

For the Speech User Interface and Service Quality (SUISQ) scale, an independent samples t-test revealed that the speech bot using the motivational interview approach (bot 1, $M = 4.66$, $SD = 0.70$), compared to the theory-based approach (bot 2, $M = 3.78$, $SD = 0.92$), scored significantly higher, $t(22) = 2.68$, $p = .014$. We further compared the two bots in terms of the four dimensions of SUISQ through four independent-sample t-tests and found speech Bot 1($M = 6.33$, $SD = .62$) was evaluated as significantly better than Bot 2 ($M = 5.40$, $SD = .85$) in terms of CSB score, $t(22) = 3.12$, $p = .005$. Furthermore, there was a near statistical significance difference in SC scores between Bot 1 ($M = 3.91$, $SD = 1.269$) and Bot 2 ($M = 2.98$, $SD = .86$), $t(22) = 2.02$, $p = .05$. Finally, while Bot 1 also had higher UGO ($M = 4.68$, $SD = 1.29$ vs. $M = 3.70$, $SD = 1.29$) and Vr scores ($M = 3.73$, $SD = 1.23$ vs. $M = 3.03$, $SD = 1.10$) than Bot 2, the differences were not statistically significant (URO: $t(22) = 1.83$, $p = .08$; Vr: $t(22) = 1.45$, $p = .16$.)

3.3 Effectiveness in Changing Vaccine Beliefs and Attitudes

Pre-exposure data was only collected in Study 2. So RQ3 was answered based on the data of Study 2 only. A paired samples t-test was conducted to compare CHAIS scores before and after participants talked with speech bot 2. It showed that conversation with the chatbot can significantly improve users' HPV vaccine beliefs and attitude (CHAIS) from ($M = 2.332$, $SD = .33$) to ($M = 2.10$, $SD = .53$) conditions, $t(9) = 2.77$, $p = .02$. (A smaller number in CHAIS indicates a more positive HPV vaccine attitude).

3.4 Observational Data

While we did not record and analyze the chat history, a research assistant observed participants' interactions with the chatbot in both studies and noticed several differences. Participants in Study 1, despite encountering more technology difficulties in the system connection, provided a higher overall evaluation of the conversation with the speech bot. They exhibited greater engagement and interest in the conversation with the speech bot, more curiosity, and willingness to access to further information about the HPV vaccine, more positive emotions and expressions, and more satisfaction with the interaction experience. Notably, when they were asked if they had questions at the end of the conversation, participants in Study 1 were more likely to respond with inquires, while participants in Study 2 tended to answer "no" with noticeable hesitation. Additionally, during the post-survey interviews, participants in Study 1 had more positive feedback about the speech bot (e.g., "It's cool" and "I feel it's easy to talk"). In contrast, participants in Study 2 commented that the speech bot sounded like a "robot" rather than a real human, which might have adversely affected their overall evaluation of the

speech bot. Additionally, participants in Study 2 looked more uncertain about answering the questions from the speech bot and appeared more tired and confused during the conversation.

4 Discussion

4.1 User Response to the Two Speech Bots

Our preliminary findings show that (1) the speech bot using the motivational interview approach outperforms the theory-based bot primarily in terms of Customer Service Behavior (CSB) and (2) despite its lower usability, the speech bot using the Theory-based approach is still effective in improving users' HPV vaccine-related beliefs and attitudes. CSB encompasses factors such as the friendliness and politeness of the system, its speaking pace, and its use of familiar terms. Coupled with our observation data, this difference can be attributed to the fact that the MI script is more interactive and less didactic, as it encourages the users to lead the discussion and ask questions that they are interested in. Additionally, both bots were also evaluated as relatively low in terms of Speech Characteristics (SC) and Verbosity (Vr). SC reflects the naturalness and enthusiasm of the bots' language, while Vr refers to the system's repetitiveness. Our findings show that to increase the perceived usability of the speech bot, in the future, we need to (1) use everyday language that users are familiar with, (2) incorporate more interaction led by users instead of the bots, and (3) personalize messaging by detecting users' preferences, level of knowledge and emotional states.

4.2 Creation of a New Script Combining the Two Approaches

The result of this work provides valuable insights for improving the interaction design of our speech bot. One notable strength observed in the MI experiment was the receptiveness of the user experience. The MI script actively engaged user by encouraging them to share their perception of the HPV vaccine and helping them explore pathways toward making an informed decision about HPV vaccination. However, we noted that the MI script might be more effective for individuals who have not received the HPV vaccine, as the MI strategy focuses on facilitating a shift in a person's mindset or behavior. All in all, the purpose of experimenting with different counseling strategies is to identify potential strengths and opportunities to enhance health outcomes. These findings highlight the importance of tailoring interaction designs to address the specific needs and states of the target audience.

Our future efforts in designing natural and engaging counseling interactions for speech bots will focus on combining features of both conversation models. Specifically, our preliminary design integrates the MI and theory-based model tested in the earlier experiments. Figure 4 shows the meta-structure of the model where features of MI (like the depiction of OARS technique) and HBM and TPB from the theory-based strategy are integrated. This hybrid model is designed to accommodates patients who have had the HPV vaccine and those who have not. It also seeks to educate vaccinated young adults as they might not be the primary decision-makers for their own vaccinations, and they might

need to make such decisions for their future children. For unvaccinated individuals, the model employs MI and theory-based counseling techniques, while for those who are vaccinated or uninterested, the interaction transitions into an informative discussion or concludes the counseling session. Our next step involves testing the interaction and engagement of the new conversation model and compare it with our previous attempts.

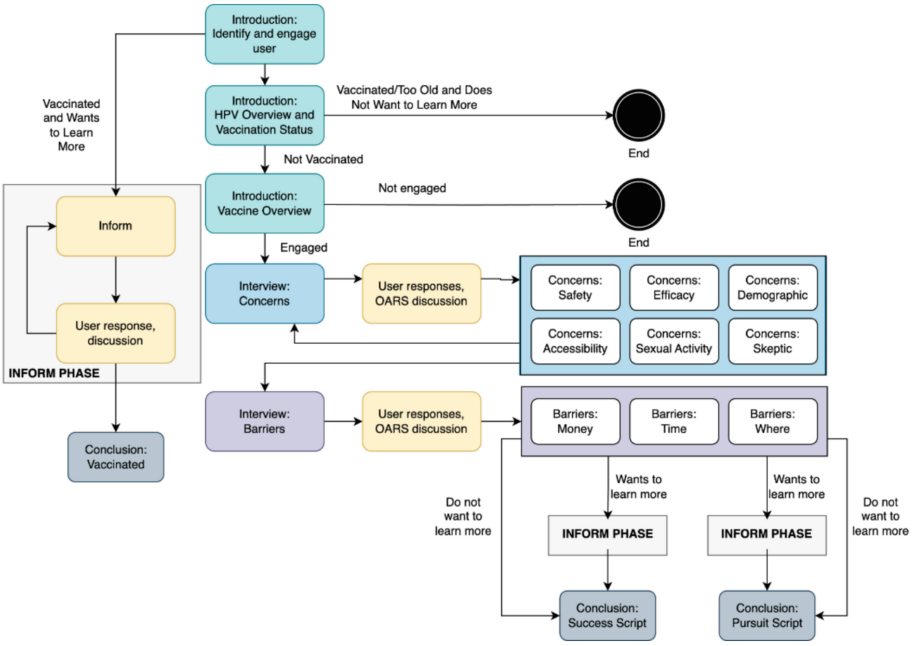


Fig. 4. Description of a new variation of the script that incorporates features from previously tested conversational strategies from the motivational interviewing and Theory-Based strategies.

4.3 Limitations

Given the exploratory nature of this study as a part of the efforts to develop effective conversational strategies for promoting HPV vaccination among young adults, the two experiments reported here relied on small sample sizes. In addition, all participants were women. Even though we did not exclude men in our recruitment, and a few men did sign up for the study, none showed up during the data collection for either study. Moving forward, we will strive to recruit participants of both genders to ensure a more representative sample in future studies.

5 Conclusion

This study demonstrates the feasibility of using speech bots as a conversational tool to promote HPV vaccination among young adults. We adopted a "conversation first" design approach to develop and focus on usable conversational interaction and the implementation of evidence-based theories. Our study identifies the features of conversation associated with higher perceived usability and provides guidelines for the development of an updated speech bot combining motivational interviewing and behavioral theories. The long-term goal is to transform these conversational scripts into computable formats (i.e., computational ontology) to create a fully automated speech bot to supplement clinician-patient communication about HPV and HPV vaccines.

Acknowledgments. This research was supported by the Cancer Prevention and Research Institute of Texas under award #RP220244, UTHealth-Houston-CPRIT Innovation for Cancer Prevention and Research Training Program Summer Undergraduate Fellowship (Cancer Prevention and Research Institute of Texas Grant #RP210042) and the National Institutes of Health under award # U24AI171008, # R21DK134815.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Centers for Disease Control and Prevention: HPV Vaccination Recommendations, HPV Vaccination Recommendations (2021)
2. U.S. Department of Health and Human Services: Increase the proportion of adolescents who get recommended doses of the HPV vaccine—IID-08. https://health.gov/healthypeople/objectives-and-data/browse-objectives/vaccination/increase-proportion-adolescents-who-get-recommended-doses-hpv-vaccine-iid-08. Accessed 1 Aug 2023
3. Pingali, C., Yankey, D., Elam-Evans, L.D., Markowitz, L.E., Valier, M.R., Fredua, B., Crowe, S.J., DeSisto, C.L., Stokley, S., Singleton, J.A.: Vaccination Coverage Among Adolescents Aged 13–17 Years — National Immunization Survey–Teen, United States, 2022. MMWR Morb. Mortal. Wkly Rep. **72**, 912–919 (2023). https://doi.org/10.15585/mmwr.mm7234a3
4. Sonawane, K., Garg, A., Meissner, E.G., Damgacioglu, H., Hill, E., Nyitray, A.G., Deshmukh, A.A.: Human papillomavirus vaccination among young adults before and during the COVID-19 pandemic. JAMA Netw. Open. **7**, e2356875 (2024). https://doi.org/10.1001/jamanetworkopen.2023.56875
5. Zheng, L., Wu, J., Zheng, M.: Barriers to and facilitators of human papillomavirus vaccination among people aged 9 to 26 years: a systematic review. Sex. Transm. Dis. **48**, e255–e262 (2021). https://doi.org/10.1097/OLQ.0000000000001407
6. Bratic, J.S., Seyferth, E.R., Bocchini, J.A.: Update on barriers to human papillomavirus vaccination and effective strategies to promote vaccine acceptance. Curr. Opin. Pediatr. **28**, 407–412 (2016). https://doi.org/10.1097/MOP.0000000000000353
7. Moore, R.J., Arar, R.: Conversational UX design: an introduction. In: Moore, R.J., Szymanski, M.H., Arar, R., Ren, G.-J. (eds.) Studies in Conversational UX Design, pp. 1–16. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-95579-7_1
8. Moore, R.J., Arar, R.: Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework. Morgan & Claypool (2019)
9. Allen, J.: Natural Language Understanding. Benjamin-Cummings Publishing (1995)
10. Jurafsky, D., Martin, J.H.: Speech and Language Processing (2009)

11. Miller, W.R., Rollnick, S.: Motivational Interviewing: Helping People Change. Guilford Press, New York (2012)
12. Lockhart, S., Dempsey, A.F., Pyrzanowski, J., O'Leary, S.T., Barnard, J.G.: Provider and parent perspectives on enhanced communication tools for human papillomavirus vaccine–hesitant parents. Acad. Pediatr. **18**, 776–782 (2018). https://doi.org/10.1016/j.acap.2018.05.012
13. Reno, J.E., O'Leary, S., Garrett, K., Pyrzanowski, J., Lockhart, S., Campagna, E., Barnard, J., Dempsey, A.F.: Improving provider communication about HPV vaccines for vaccine-hesitant parents through the use of motivational interviewing. J. Health Commun. **23**, 313–320 (2018). https://doi.org/10.1080/10810730.2018.1442530
14. Xiao, X., Lee, D.K.L., Wong, R.M., Borah, P.: The impact of theory in HPV vaccination promotion research: a systematic review and meta-analysis. Am. J. Health Promot. **35**, 1002–1014 (2021). https://doi.org/10.1177/08901171211012524
15. Becker, M.H.: The Health Belief Model and Personal Health Behavior, vol. 2, pp. 409–419. Slack (1974)
16. Janz, N.K., Becker, M.H.: The health belief model: a decade later. Health Educ. Q. **11**, 1–47 (1984)
17. Li, L., Wood, C.E., Kostkova, P.: Vaccine hesitancy and behavior change theory-based social media interventions: a systematic review. Transl. Behav. Med. **12**, 243–272 (2022). https://doi.org/10.1093/tbm/ibab148
18. Fraser, N.M., Gilbert, G.N.: Simulating speech systems. Comput. Speech Lang. **5**, 81–99 (1991)
19. Prochaska, J.O., Velicer, W.F.: The transtheoretical model of health behavior change. Am. J. Health Promot. **12**, 38–48 (1997)
20. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991). https://doi.org/10.1016/0749-5978(91)90020-T
21. Brooke, J.: SUS: a retrospective. J. Usabil. Stud. **8**, 29–40 (2013)
22. Brooke, J.: SUS-A quick and dirty usability scale. In: Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)
23. Lewis, J.R.: The system usability scale: past, present, and future. Int. J. Hum. Comput. Interact. **34**, 577–590 (2018). https://doi.org/10.1080/10447318.2018.1455307
24. Lewis, J.R., Sauro, J.: Can I leave this one out? The effect of dropping an item from the SUS. J. Usabil. Stud. **13**, 9 (2017)
25. Polkosky, M.D.: Machines as mediators: the challenge of technology for interpersonal communication theory and research. In: Mediated Interpersonal Communication, pp. 48–71. Routledge, London (2008)
26. Polkosky, M.D.: Toward a Social-Cognitive Psychology of Speech Technology: Affective Responses to Speech-Based E-Service, p. 161 (2011)
27. Lewis, J.R., Hardzinski, M.L.: Investigating the psychometric properties of the speech user Interface Service quality questionnaire. Int. J. Speech Technol. **18**, 479–487 (2015). https://doi.org/10.1007/s10772-015-9289-1
28. McRee, A.-L., Brewer, N.T., Reiter, P.L., Gottlieb, S.L., Smith, J.S.: The Carolina HPV immunization attitudes and beliefs scale (CHIAS): scale development and associations with intentions to vaccinate. Sex. Transm. Dis. **37**, 234–239 (2010)
29. Dempsey, A.F., Fuhrel-Forbis, A., Konrath, S.: Use of the Carolina HPV immunization attitudes and beliefs scale (CHIAS) in young adult women. PLoS One. **9**, e100193 (2014). https://doi.org/10.1371/journal.pone.0100193
30. Moore, N., et al.: Translating motivational interviewing for the HPV vaccine into a computable ontology model for automated AI conversational interaction. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–12. ACM (2024). https://doi.org/10.1145/3613905.3651051



Research on User Experience Design of Internet Healthcare and Intelligent Triage Systems

You Yang Wu, Zi Xin Zhou, Lin Guo Jin, Zi Hong Ye, and Yong Yan Guo(✉)

East China University of Science and Technology, Shanghai, China
g_gale@163.com

Abstract. Purpose. To explore the interface design of intelligent triage systems from the perspective of user cognitive behavior, optimizing page layout and interaction to enhance user satisfaction.

Method. A thorough investigation of existing triage systems both domestically and internationally was conducted to identify pain points in current products; questionnaires and statistical methods were used to present user needs and analyze user experience metrics. The interface design elements were distilled from three aspects: page browsing interaction, page functionality, and doctor interaction, aiming to improve the app's usability and fluency.

Conclusion. This study proposes design strategies for the interface of intelligent triage systems from three aspects: "user-friendly" page browsing, "rational" functionality layout, and "diverse" doctor interaction, to further enhance the user experience of the system and provide references for related design practices.

Keywords: intelligent triage system · cognitive behavior · interface design · user experience · design strategy

1 Introduction

In recent years, with the development of information technology, internet healthcare has gradually entered people's lives, generating online services such as online consultations, family doctors, and AI-assisted diagnosis and treatment, as well as helping users with online appointment scheduling and prescription services, saving users' time in offline medical visits.

As time goes by, many issues arising during user interactions urgently needs to be addressed, such as excessive pop-up ads, crowded interface content layout, long search paths, low recognition accuracy of triage systems, and chaotic icon colors. Many scholars have researched these issues; Liu Jihong proposed an adaptive design model for digital intelligent products, supported by various data analyses [1]; Ren Yanhui suggested that the application of data visualization design in medical apps can enhance the efficiency of information acquisition for users, meeting the emotional needs of medical platform users [2]; Li Xiaojing and Cao Sheng sheng proposed that page design should be humanized and metaphorical, focusing on user emotions to enrich the user experience [3]. In internet healthcare, intelligent triage systems, as the first stage of service for patients

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 85–96, 2025.
https://doi.org/10.1007/978-3-031-93502-2_6

seeking medical advice, significantly impact user experience and medical efficiency. Therefore, optimizing the interface design of internet healthcare triage systems is of great significance for improving patient medical efficiency.

1.1 Content of the Research

Based on the purpose of our study, i.e., to understand users' feelings, needs and expectations when using this kind of application, 'the questionnaire was used to analyze users' interface design preferences, functional layout and importance ranking, and evaluation of interaction experience. By quantifying the user experience indexes, we consider the design requirements comprehensively and provide a basis for the research of the intelligent diagnosis system.

1.2 Research Theories

User Experience (UE/UX) theory emphasizes the concept of user-centered design, which focuses on the integrated experience of sensory, emotional, behavioral and psychological dimensions of the user in the process of using a product or service. According to Norman, the user experience covers all interaction behaviors and their triggered emotional and psychological states, and it is a complex and subjective process [4]; Garrett proposed five levels of UX: strategic, scoping, structural, framework and presentation [5]. The goal of user experience is to achieve the unity of functional and emotional values and enhance satisfaction by satisfying user needs and reducing barriers to use.

In the field of Internet healthcare, user experience optimization is directly related to the accessibility and efficiency of healthcare services. In addition, based on the Technology Integration Model (UTAUT), a study by He Huixian and Zhang Ruichiu et al. pointed out that variables such as perceived risk, service pricing, and individual innovativeness are particularly important in the context of Internet healthcare, and that the ease of use of platforms needs to be improved by simplifying the operation, clarifying the classification of functions, and providing personalized services [6]. Therefore, we propose the following hypotheses for the user experience of the internet healthcare guidance system.

1.3 Research Hypothesis

1. Optimizing the page browsing experience (e.g., improving visual effects and simplifying the operation process) can significantly improve the smoothness and pleasure of user interaction. The overall page style of the consultation is crucial to the first impression of the user and the establishment of a sense of trust, a unified and simple page design can optimize the efficiency of information transfer, reduce the cognitive burden on the user and optimize the user experience.
2. Rationalization of the functional layout (such as clear functional classification and optimization of navigation design) can effectively help users to quickly locate the required services and improve operational efficiency. The logic of the functional layout of the guidance system needs to be supported by the coordination and consistency

of the overall page style, which are complementary and mutually reinforcing. Confusing layout or mismatch with the overall style will weaken the visual guidance effect of the page, affecting the user experience.

3. Enhance the interactive functions of doctors (such as intelligent guidance, real-time communication and personalized services) can improve the quality of interaction between patients and doctors and meet the diverse needs of users. The good and smooth experience of interacting with doctors is built on efficient visual guidance and friendly interface operation. Simple and clear page style and functional layout can enhance the usability of interactive functions (Fig. 1).

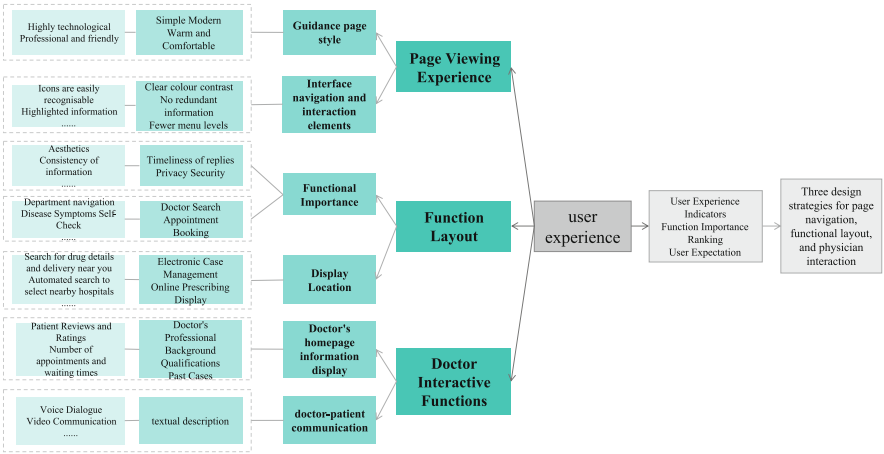


Fig. 1. Research model diagram of user experience based on questionnaire study.

2 Analysis of Results

2.1 Data Screening and Testing

2.1.1 Data Outlier Screening

In the process of analyzing the outliers of individual questions, the study found that there were some data points in the age option that were outside the whisker range of the box plot. After in-depth research, these outliers were attributed to the relatively small sample size of children and elderly people in the questionnaire survey. Based on this, it was decided not to remove this part of the questionnaire data as outliers. Subsequent analyses will focus on the 18–30 age group.

The time taken to complete the questionnaire was used as the main criterion to determine the outliers in the whole questionnaire. When the completion time of the questionnaire is too short or too long, there may be data quality problems. For this type of questionnaire, this study will further analyze the responses to the reverse questions before and after. If there is a significant difference, the questionnaire is judged to be less

credible, and the results of such questionnaires need to be deleted in order to improve the reliability of data analysis.

2.1.2 Data Reliability and Validity Analysis

Some of the questions of the questionnaire were subjected to the Cronbach Alpha test, and the Alpha values of 0.824 for the fifth question and 0.833 for the sixth question were greater than 0.5, which can be related to the reliability of the questionnaire results. Using the chi-square test and Bartlett's test to determine the validity of the data, from the perspective of the basic information, gender KMO = 0.050, age KMO < 0.01, are less than or equal to 0.050; from the perspective of the other options of the questionnaire, to the display of information on the doctor's page, for example, the doctor's professional background and qualifications, the number of patients' evaluations and scores of the number of appointments and the waiting time and so on, asymptotic significance value of the options are <0.010, there is a significant difference, and therefore have statistical significance. There is a significant difference, thus statistically significant.

2.1.3 Analysis of Information Related to Experimenters

In this survey, the total number of validly filled out 163 people, 69 men and 94 women. Among them, 101 people aged 18–30 participated in the survey, accounting for the highest proportion; 43 people aged 31–45, accounting for the second proportion; 9 people aged 46–60, accounting for the third proportion. Of the 163 people surveyed, 85 said they used Internet APPs every day; 35 said they used them a few times a week; 25 said they used them a few times a month; 18 said they seldom used that type of APP; and the remaining 14 said they had never used them.

2.2 Users' Perceptions and Preferences of Smart Guidance Page Style and Interaction Elements

Among the 163 experimenters, 95 of them ranked simple and modern in the style of the guide page, which has the highest percentage in the first-place share; and for the second-place share, there is a small difference between the four of simple and modern, warm and comfortable, technological and professional friendly. Combining the total number of the first and second places, the user's preferred style is inclined to simple and modern and warm and comfortable (Table 1).

Table 1. Style statistics.

	First		Second		Third		Fourth		Fifth	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Simple and modern	95	58.3	25	15.3	12	7.4	14	8.6	17	10.4
Warm and comfortable	38	23.3	32	19.6	26	16	14	8.6	53	32.5
highly technological	19	11.7	31	19.0	23	14.1	27	16.6	63	38.7
professional and friendly	11	6.7	34	20.9	35	21.5	24	14.7	59	36.2

In the importance of elements in the intelligent diagnostic interface browsing, the statistical average can be concluded that the key information is prominent, the color contrast is clear, there is no redundant information, the icon is easy to identify, the icon is easy to identify, the arrangement is concise, the menu hierarchy is less mean score are greater than 3, so for the six elements of the user is presented a satisfactory attitude. The user data reflects that the satisfaction degree of layout simplicity is 3.89 > highlighting key information 3.84 > easy to identify icons 3.80 > clear color contrast 3.73 > no redundant information 3.73 > fewer menu hierarchies 3.58. In the design of the intelligent diagnosis page, it is necessary to make trade-offs and innovations according to the order of the importance of the elements.

To improve the user satisfaction of the intelligent diagnosis system of Internet hospitals, not only should we start from the design of the page elements, but also the interactive experience of the user in the process of using the system is an important aspect of improving the user satisfaction. Users present a satisfactory attitude towards the seven elements of page aesthetics, communication convenience, response timeliness, information accuracy, privacy security, information coherence, and service friendliness. And the user data reflected the satisfaction of information accuracy 3.88 > communication convenience 3.85 = service friendliness 3.85 > information coherence 3.83 > page aesthetics 3.82 > response timeliness 3.76 > privacy security 3.75 (Table 2).

Table 2. Differences in the importance of browsing and interacting elements of the smart guide interface.

	Average value	Standard deviation	Variance (statistics)
Highlighted information	3.84	.955	.913
Clear color contrast	3.73	.839	.704
No redundant information	3.67	1.019	1.038
Icons are easily recognizable	3.80	.942	.887
Simple layout	3.89	.916	.840
Few menu levels	3.58	.974	.949
Aesthetics of the page	3.82	.870	.756
Ease of communication	3.85	.907	.822
Timeliness of reply	3.76	.974	.949
Accuracy of information	3.88	.887	.787
Privacy and security	3.75	.969	.939
Consistency of information	3.83	.870	.756
Service friendliness	3.85	.904	.818

2.3 Users' Perceptions and Preferences of the Functional Display Elements of Smart Guidance Pages

There are many functions in the interface of intelligent medical guidance, in order to improve the efficiency of users, we have collected and screened the function categories of this kind of app at home and abroad, and investigated the users' perception of their importance through questionnaires. For the design of the functions in the diagnostic page, we will investigate from two aspects: the importance of the functions and the display position of the functions.

The research on the importance of the display position of the function is to allow users to quickly find relevant information when entering the guidance system, improve the accuracy of the search information, and to a certain extent, optimise the design of the information hierarchy of the existing guidance system. From the analysis of the frequency of options, the user feedback for the booking of registration accounted for 69.9% > information query accounted for 57.7% > department navigation accounted for 54% > disease symptoms accounted for 52.8% > online consultation with doctors accounted for 43.6% (Table 3). Appointment booking involves the cooperation of multiple information platforms, and there are certain requirements for the interoperability and extensiveness of intelligent guidance.

Table 3. Frequency analysis of the importance of functional presentation.

	Be selected	
	Frequency	Percentage
Doctor Search	94	57.7
Departments	88	54.0
Appointment	114	69.9
Disease Symptoms Self-Check	86	52.8
Online Consultation	71	43.6

According to the binary order summary table statistics, the importance of the functions of the user feedback is the convenient management and enquiry of e-case 32.5% > online prescription display 27.6% > enquiry of drug details and nearby drug delivery 26.4% > automatic searching and selection of nearby hospitals 32.5% > professional health education push and lectures 28.2% > community communication and patient mutual support platform 35.6% > Family-based health record management and disease prediction reminder 42.9% (Table 4).

2.4 A Study of Users' Perceptions and Preferences of Doctor Presentation Elements

Liu Jusheng said that getting good service and more information from the doctor is the goal of the patient, and under the paid service, the interaction and information feedback

Table 4. Ranking of functional importance.

Functions	1st place		2nd place		3rd place		4th place		5th place		6th place		7th place	
	Case	%	Case	%	Case	%	Case	%	Case	%	Case	%	Case	%
Convenient case management and enquiry	53	32.5%	21	12.9%	23	14.1%	21	12.9%	15	9.2%	7	4.3%	23	14.1%
Online prescription display	41	25.2%	45	27.6%	24	14.7%	11	6.7%	21	12.9%	12	7.4%	9	5.5%
Enquiry on drug details and nearby drug delivery	28	17.2%	27	16.6%	43	26.4%	31	19.0%	18	11.0%	10	6.1%	6	3.7%
Auto search for nearby hospitals	17	10.4%	22	13.5%	22	13.5%	53	32.5%	24	14.7%	17	10.4%	8	4.9%
Professional health education and seminars	8	4.9%	20	12.3%	19	11.7%	17	10.4%	46	28.2%	32	19.6%	21	12.9%
Community support platform	5	3.1%	21	12.9%	19	11.7%	13	8.0%	21	12.9%	58	35.6%	26	16.0%
Family health records and disease reminders	11	6.7%	7	4.3%	13	8.0%	17	10.4%	18	11.0%	27	16.6%	70	42.9%

during the online service is very important to the patient [7]. Therefore, the use of the interaction between patients and doctors in the intelligent guidance page is a key factor in improving user satisfaction, but the existing intelligent guidance system is generally incomplete information about doctors, slow feedback from doctors, communication barriers and other problems make the user trust greatly reduced.

The interaction between the doctor and the user is mainly divided into three links: 'Recommendation - Communication - Evaluation'. Zhou Siming proposed that in the doctor recommendation of the medical guidance system, it is necessary to meet the functional requirements of doctor recommendation, appointment registration, and the query of the patient's medical history data [8]. Patients 'knowledge of doctors' information helps to promote the consultation process efficiently and improve the quality of medical care.

From the feedback of the questionnaire, in the interaction of the understanding session, users are more willing to prioritize information on professional qualifications, such as professional background and details of consultation cases. Therefore, in the doctor recommendation, doctors with better ability to match the disease need to be recommended first, and patient evaluation scores, number of appointments and waiting time are summarized as secondary labels for display, reducing patients' ineffective browsing time. In the communication with the doctor, taking into account personal privacy and security issues, the user prefers to give priority to the use of traditional text descriptions, more complex communication can be carried out in order of voice dialogue, video communication and graphic information supplement. In order to ensure the effectiveness of communication and the accuracy of information, Hu Song proposed in his research to use slot entity pairs and entity supplementary description information in communication, introduced the Copy mechanism, and proposed a natural language generation model based on entity guidance and Copy mechanism [9]. And in special cases, such as facing patients with low literacy level, it is necessary to recommend more personalized communication to the user after self-testing of textual descriptions. (See Tables 5 and 6).

Table 5. Users' ranking of importance of doctor's information presentation.

	1st place		2nd place		3rd place		4th place	
Type of information	case	%	case	%	case	%	case	%
Doctor's Professional Background and Qualifications	97	59.5%	25	15.3%	24	14.7%	17	10.4%
Past Cases	34	20.9%	48	29.4%	50	30.7%	31	19.0%
Testimonials and Ratings	18	11.0%	54	33.1%	54	33.1%	37	22.7%
Number of appointments and waiting time	14	8.6%	36	22.1%	35	21.5%	78	47.9%

Table 6. Ranking the importance of the way users communicate with their doctors.

	1st place		2nd place		3rd place		4th place		5th place	
Communication	case	%	case	%	case	%	case	%	case	%
Text description	75	46.0%	25	15.3%	29	17.8%	34	20.9%	1	9.1%
Voice dialogue	39	23.9%	58	35.6%	38	23.3%	28	17.2%	10	90.9%
Video communication	34	20.9%	38	23.3%	50	30.7%	41	25.2%	0	0
Picture message supplement	15	9.2%	42	25.8%	45	27.6%	60	36.8%	0	0
Other	0	0	0	0	1	0.6%	0	0	0	0

3 Design Strategies

3.1 Intelligent Diagnostic Page Browsing and Interaction Elements Design Strategy

Simplify the diagnostic interface to create a simple, modern, warm and comfortable atmosphere. Clearly divide the layers and streamline the hierarchy, adopt a clean, comfortable and non-irritating color scheme, reduce the saturation of the background color, and strengthen the visual contrast between the background and the main panels. Remove redundant information, reduce repetitive search entrances, optimize the interface layout, and minimize repetitive paths during user operation. Avoid using high brightness text colors. Reduce the display space of recommended and popular science content, present it in the form of information flow, and maintain appropriate white space on the page. (2) Unify the style of page elements to enhance the visibility and legibility of the page. Enhance color contrast, unify logo styles, distinguish similar logos and improve recognition. Reasonably control the number of logos to avoid page clutter. (3) Enhance privacy and security. Set access rights to users' personal information, cases, consultation videos and other data. When the user's private information is collected, used or shared, it is clearly notified through the page reminder message or pop-up window. Display privacy protection slogans or security lock icons on the privacy information page to remind users of the status of their private information. (4) Provide customization functions to enable users to adjust the order, position and number of different sections according to their personal needs, and set up hidden functions to simplify the page and hide unnecessary functions.

3.1.1 Intelligent Medical Guidance System Page Design Strategy

(1) Logical hierarchical design. Logical hierarchy is carried out according to the importance of functions. Key functions such as appointment registration, doctor information query and department navigation are placed in the first or second level to highlight their importance; information such as disease symptoms and online consultation with doctors are arranged in the third or fourth level so as to improve the logic and usability of

information on the page and facilitate the operation of users. (2) Convenient information enquiry system. Electronic case management, family health record management and disease prediction and reminder functions are the core functions used by users with high frequency, so these information query functions should be placed in the first or second level, which is easy for users to access quickly to meet the user's needs for case and family health record management. (3) Community Communication and Patient Support Platform. According to the results of the questionnaire survey, the community communication and patient mutual assistance platform is a function that users attach great importance to. The feedback and reminder function of platform information should be added in the page design to help users improve the timeliness and effectiveness of communication and mutual assistance. (4) Provide hospital search service. In the hospital selection function, based on the user's geographic location and consultation needs to provide recommended hospital programs, which is convenient for users to refer to and choose. This not only saves users' time in selecting hospitals, but also improves the efficiency of the navigation system.

3.2 Design Strategies for Intelligent Guidance to Facilitate Communication with Doctors

3.2.1 Optimize the Display of Doctor Information

Comprehensive presentation of doctor information: Improve the doctor's information database, which, in addition to basic professional background and details of consultation cases, should also include information on the disease segments that the doctor is good at treating and the professional awards he or she has won, to ensure that patients have a comprehensive understanding of the doctor's qualifications and to enhance the degree of trust. Intelligent Recommendation of Doctors: We build intelligent algorithms to prioritise the recommendation of doctors with strong disease-matching capabilities based on the symptoms of the disease entered by the patient. At the same time, patient evaluation scores, number of appointments and waiting time are clearly displayed in secondary labels to help patients quickly screen and reduce ineffective browsing.

3.2.2 Improve Communication Design

Provide diversified communication methods: text description is used as the main communication method by default, which meets users' needs for privacy and security. As the complexity of communication increases, voice dialogue, video communication and supplementary options for graphic information are provided in turn. In addition, obvious switching buttons are set in the communication interface to facilitate users to adjust the communication mode at any time. Precise communication assistance: Introducing a natural language generation model (Copy mechanism) based on slot entity pairs and entity supplementary descriptive information, it intelligently analyses and understands the text entered by the user, assisting the doctor in accurately obtaining information and improving the effectiveness of communication. Personalized communication recommendation: for patients with low literacy or special needs, the system automatically performs semantic analysis after the user inputs text, and if it finds that it is difficult to

understand, it will promptly recommend more understandable and personalized communication methods for the user, such as simplifying the textual expression, providing voice guidance, and so on.

Establish an Efficient Feedback Mechanism. Shorten the feedback time for doctors: set up a reminder mechanism on the doctor's side to notify the doctor in time when the patient initiates a consultation and ensure that the doctor replies as soon as possible. At the same time, the feedback time of doctors is counted and publicized to urge doctors to improve their response speed. Patient feedback channel: Setting up a special patient feedback portal in the guidance system, it is convenient for patients to put forward their opinions and suggestions on the doctor's service, system functions, etc., so as to improve and optimize them in a timely manner.

4 Conclusion

Internet healthcare is an emerging field in the medical industry and a trend in medical development. Intelligent guidance system is the entrance and infusion hub of Internet healthcare, which plays a key role in the efficient operation of the whole healthcare service and user experience. Based on the detailed market research and user research, the article quantifies the user experience indexes and proposes the design optimization strategy of the intelligent guidance system, which makes the logic of the interface design more friendly, rational and diversified. According to the research, during the browsing process, users prefer simple, neat and modern pages with clear and easy-to-recognize information; the feedback data on the importance of the interface's functional layer shows that users are inclined to make result-oriented choices for efficient consultation. This means that the whole interface design should be based on the line of booking—selecting a competent doctor—precise information communication—diagnosing and treating the disease, and the efficiency of consultation and user satisfaction should be improved by dividing the importance of multiple types of information. The user experience model of this study can provide certain guidance for the technical research of the medical guidance system, and provide reference and reference value for the design and deepening of the intelligent medical guidance system, which promotes the development of this system.

However, this study has certain limitations. Due to geographical restrictions, different users may have significant differences in interface design and functional requirements, and the lack of a comprehensive research sample may cause the conclusions to deviate from the reality. Moreover, this study is an idealized concept without technical practice and implementation, and lacks further optimization. In conclusion, in the study of user satisfaction of intelligent diagnosis system, the page design of intelligent diagnosis is an important factor to improve user satisfaction, and it is necessary to consider various elements after the research to conduct a detailed study of the user in various aspects, and to formulate reasonable metrics.

References

1. Li, J.: Development of Digital Intelligent Product Design Methods and Technologies. Beijing University of Aeronautics and Astronautics, Beijing (2023)
2. Ren, Y.: Application of Information Visualization Design in Internet Medical Platform. North University of Technology, Beijing (2022)
3. Liu, X., Chen, S.: Research on the Influence of Anthropomorphism and Metaphorization of Interaction Default Interface on User's Emotion. Ningbo University, Ningbo (2023)
4. Norman, D.A.: Design Psychology. Mei Qiong, Translation. CITIC Press, Beijing (2003)
5. Garrett, J.J.: Elements of User Experience. Machinery Industry Press, Beijing (2007)
6. Huang, H., Zhang, R., Song, W., Hu, Y.: Analysis and research on factors affecting patient satisfaction in internet medical care. Packaging Eng. **22**, 218–225 (2021). https://doi.org/10.19554/j.cnki.1001-3563.2021.22.026
7. Li, J.: Research on Doctor-Patient Interaction Behavior in Internet Medical Environment. Shanghai University of Finance and Economics, Shanghai (2022)
8. Zhou, S., He, J.: Design of doctor recommendation guidance system based on ATERDE and AHP data gravity classification algorithm. Heilongjiang Sci. **14**(24), 65–68 (2023)
9. Song, H.: Research on Chinese Medicine Consultation Model for Task-Based Dialogue. University of Electronic Science and Technology, Chengdu (2024)

How Social Cues in Depression Intervention AI Chatbots Affect Human-Computer Trust

Shuo Xu[1(✉)] and Tiancong Ma[2]

[1] School of Design, Guangzhou Higher Education Mega Center, South China University of Technology, Panyu District, Guangzhou 510006, China
sdxushuo@mail.scut.edu.cn

[2] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong, China
23044932g@connect.polyu.hk

Abstract. With the widespread application of artificial intelligence agents in daily life, social cues have received increasing attention as an important factor affecting human-machine trust. However, its specific role in depression treatment chatbots still needs further study. This study aimed to investigate how social cues influence users' trust in a depression intervention chatbot. In the experiment, we compared two groups of chatbots: one therapeutic chatbot with high levels of social cues, using avatar animation to express emotional tone, facial expressions, and body movements, and the other chatbot with low levels of social cues. A bot that uses humane and empathetic text conversations. We also collected participants' skin conductance and heart rate data to further explore the relationship between these physiological indicators and trust. The results of the study showed that in the self-help intervention for depression, therapeutic chatbots with higher levels of social cues significantly improved users' trust levels and showed clear advantages in forming therapeutic alliance and user satisfaction. Specifically, the high social cue agent group scored significantly higher on the dimensions of perceived reliability ($p < 0.05$), perceived technical competence ($p < 0.01$), and belief ($p < 0.01$) compared to the low social cue agent group higher. In addition, the high social cue agent group had significantly higher heart rate ($d = 0.90$) with a large effect size, and skin conductance ($d = 0.71$) had a medium effect size. Significant changes were observed in both heart rate ($F = 48.63$, $p < 0.01$) and skin conductance ($F = 26.25$, $p < 0.01$). In summary, the findings of this study not only reveal the key role of social cues in depression intervention chatbots, but also provide valuable practical guidance for improving human-computer interaction design in the future.

Keywords: Social cues · Human-computer trust · Agency · Chatbot · mHealth

1 Introduction

Across the world, depression is among the most prevalent mental diseases [1] and is associated with suicide attempts [2]. College students frequently experience depression [3, 4]. Depression negatively impacts academic achievement [5, 6], interpersonal

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 97–117, 2025.
https://doi.org/10.1007/978-3-031-93502-2_7

relationships [7], and quality of life [8]. To address these issues, researchers have proposed Internet-based psychological intervention (IPI) solutions, including chatbots [9, 10]. The past 10 years have seen a fast advancement in artificial intelligence (AI) technology, which has resulted in a growing body of research highlighting the benefits of chatbot-based interventions. Studies have shown that chatbot interventions have significant therapeutic effects because users can conduct personal self-help interventions at any time and place. Compared with minimal bibliotherapy, studies have found that chatbot interventions are superior in reducing depressive symptoms and enhancing user adherence [11]. These research results suggest that AI chatbots are potentially effective mediators of fully automated self-help psychological interventions [12, 13].

With the advancement of AI technology, trust between humans and computers has become a critical issue [14, 15], which is also important for depression treatment chatbots. AI agents, including chatbots, are becoming increasingly common in daily life. Interactions with communicative AI are shaping people's attitudes toward more specific AI agent [16]. People's views and behaviors on communication technology are further influenced by their interpretation of the communication technology [17]. Thus, it's critical to investigate how people understand the capabilities and applications of AI agents in order to comprehend how people assess agents in general [18] and the importance of trust in human-computer interactions is growing [14, 19, 20]. The research describes trust as " Trust in human-robot interaction is the human's belief or expectation that a robotic teammate will act in a way that protects the interests and welfare of every individual on the team." [21]. Agents here include computers and AI agents.

As robots become more and more anthropomorphic, humans may begin to view them as more social interaction objects rather than simple tools. Investigating how social cues affect people's trust in artificial intelligence agents is therefore very important. Unlike the abstract artificial intelligence agents used in previous studies, this study used a depression intervention chatbot with a certain user base.

In this study, the research team developed two chatbots that combined social cues to varying degrees and conducted experimental tests on 149 Chinese college students. The main purpose of this study is to explore how social cues affect users' trust in a depression intervention chatbot. Specifically, we aim to analyze the relationship between social cues and five dimensions of human-computer trust, including reliability, technical ability, understandability, belief, and personal attachment. In addition, this study hopes to further supplement the understanding of the trust-building mechanism by collecting physiological indicators such as heart rate and skin conductivity.

The scope of the study is mainly focused on a specific population of college students, considering their high incidence of depression, and using chatbots with different degrees of social cue integration. Through this setting, we hope to verify how the presence of social cues affects users' trust, thereby providing empirical support for the field of human-computer interaction and providing theoretical guidance for future chatbot design and psychological intervention measures. Overall, this study is not only an in-depth discussion of the theory of human-computer trust, but also an empirical exploration of the application of artificial intelligence in the field of mental health.

This study is based on the argument of the research group of Hamid et al., emphasizing that artificial intelligence should be used as a supplement to human labor, rather

than a substitute, to promote efficient and balanced coexistence beyond the traditional working environment [22]. By demonstrating how AI can augment human capabilities in complex decision-making environments, this study further reinforces the view that collaborative human-machine interactions are superior to purely autonomous or individualistic approaches [23]. Furthermore, this study extends the insights of Hamid et al. by showing that integrating machines with humanistic characteristics into human workspaces can drive ultra-flexible smart manufacturing environments while promoting human well-being and productivity [24]. Thus, this study contributes to the broader field by providing empirical evidence that, if automation is implemented strategically, a symbiotic relationship can be established in which the role of humans is not reduced but enhanced, a finding that complements Hamid's hypothesis of mutually beneficial human-machine collaboration [25].

2 Literature Review

2.1 Human-Computer Trust and Social Cues

This paper adopts Madsen et al.'s five-dimensional human-computer trust model [20], emphasizing the importance of cognitive and emotional dimensions of human-computer trust in interaction, which affect individuals' acceptance and trust in technology. In the cognitive dimension, the reliability of technology is a basic consideration. Individuals need to evaluate the stability and consistency of technology in different situations to judge its credibility [26]. In addition, individuals' understanding of the technology's capabilities is also crucial. Only when they are sure that the technology can perform tasks effectively will users be willing to rely on the technology. Ease of understanding is also critical. Intuitive interfaces and simple operations can help users master technology more quickly and enhance their sense of trust [27]. In the emotional dimension, trust and attachment constitute the emotional connection between individuals and technology. Trust reflects the degree of dependence of users on technology, based on their cognitive assessment of the reliability and capabilities of the technology. After establishing trust, users are more inclined to use the technology and believe in its performance [28]. Attachment refers to the user's emotional connection with technology. This emotional dependence affects the user's attitude and behavior, further affecting the formation of trust [29]. In summary, the cognitive and emotional dimensions of human-computer trust are intertwined and together constitute the basis for individuals' trust in technology. A deep understanding of these dimensions will provide an important reference for exploring the trust mechanism in human-computer interaction, and thus help design more trustworthy and humane technical systems.

Social cues refer to non-verbal and verbal elements that convey information and emotions in communication, and have a profound impact on interpersonal interaction and trust. In the context of human-computer interaction, Lobato et al. defined social cues as features that "serve as channels for social information" [30]. According to the classification of Burgoon et al., social cues can be divided into eight categories: movement, acoustics, appearance, interpersonal space, touch, time, environment and objects, and smell [31]. In addition, Leathers divides the interpersonal communication system into

four subsystems: language, vision, hearing, and intangible, highlighting the interactive role of these elements in the transmission of meaning [32].

Numerous researches have examined the influence of social cues on human-computer trust in agents, as they are one of the key variables influencing human-computer trust. It has been demonstrated that social cues, including the robot's head movement and eye contact during conversation, boost task immersion [33, 34]. An anthropomorphic chatbot developed in the early stages of study was intended to improve exercise and demonstrated that a therapeutic alliance could be formed between the client and the chatbot [35]. Similarly, randomized controlled trials have verified that embodied conversational agents can provide effective psychotherapy [36, 37].

Early studies have focused on the extent to which social agents should portray social cues [38–40]. A few hypotheses, such the theory of the media equation [41], argue that fundamental social characteristics are enough to induce social reactions, and some early research have validated this theory [42, 43]. Similarly, the social cue hypothesis [44] states that adding social cues to robots, such as facial expressions, voices, and body language, can lead humans to perceive technology more positively. Several studies have supported this hypothesis [45–47]. In order to expand on their media equation theory, one study introduced the computers-are-social-actors (CASA) paradigm, specifically, "As computers display human-like characteristics including language, social roles, gender, race, and personality, people attribute a wide range of social traits to them because computers are inherently social." [48].

Researchers are also exploring the impact of social cues on human-computer trust. Researchers are interested in using robots as a technology to modify attitudes and behaviors [49, 50]. According to social agency theory [51], agents with more social cues will induce stronger social responses from people. Therefore, understanding people's attitudes and behaviors towards social cues of robots is crucial to enhancing the perceptibility of artificial agents and optimizing the human-computer interaction experience. Early research concentrated on how vocal signals convey social cues and affect human-computer trust. Studies have found that existing gender stereotypes apply equally to users' computer interactions with female and male voices [52, 53]. Other studies have applied CASA to trust research and found that subjects tend to trust computerized voices that match their own personalities, whether introverted or extroverted [54]. Some studies have analyzed the impact of agents with human facial features on human-computer trust. For example, in competitive gaming situations, people are more likely to believe computers that prominently display faces [55]. Other studies have shown that when presenting a computer agent's facial image and voice, consistency between the two is crucial to trust perception [56]. A storytelling robot that combines gaze and gesture as social cues is more compelling than a robot that employs just one of these cues, according to study on the influence of body language social cues on human-robot trust [57].

Therefore, it can be found that AI agents that combine social cues increase the chances of effectively establishing human-computer trust, which has important potential value for the application of chatbots in mental health intervention. Based on this, we studied the correlation between human-computer trust and social cues in different dimensions: first, the relationship between the cognitive dimension of trust and social cues, including the three dimensions of reliability, technical competence, and understandability; second,

the relationship between the emotional dimension of trust and social cues, including the two dimensions of faith and personal attachment [20].

Our hypothesis: H1. In the (a) cognitive dimension, the chatbot group with higher levels of social cues will have higher human-computer trust, while in the (b) affective dimension, the chatbot group with higher levels of social cues will have higher human-computer trust.

2.2 Human-Computer Trust and Physiological Indicators

At the same time, researchers also explored the manifestation of human-computer trust in physiological indicators, among which heart rate is considered an important trust indicator. In an experiment in which trust was induced by adjusting the accuracy of AI responses, a significant correlation was found between the level of trust and heart rate [58]. Other studies have similarly shown that increased heart rate is associated with greater trust, because increased heart rate is associated with increased expectations of trust [59]. In addition, some studies have asked participants to complete a series of cooperative tasks to predict their expectations of their partners through heart rate, and the results show that heart rate can reflect the process of trust building [60].

At the same time, skin conductivity is also used as a physiological indicator to study human-computer trust. For example, a study using the decision-making path of an autonomous car found that participants' skin conductance increased significantly before facing high-reward and high-risk decisions [61]. Therefore, when people perceive high levels of trust, they may show changes in emotional activation on these physiological measures. These two indicators have been shown to be valid and sensitive measures of emotional activation [62, 63]. Emotional activation makes users feel more active and focused when interacting with the agent, just like communicating with other people. Emotional activation plays a key role in helping users master complex depression treatment techniques, and its importance is widely recognized in cognitive behavioral therapy (CBT) for depression, especially in treatment-related learning. In the analysis of process and emotional effects [64]. A large body of neurobiological research supports the view that emotional activation facilitates memory [65]. It can be inferred that if highly social agents can establish high levels of human-computer trust, then this will also promote high levels of emotional activation, further verifying the association between social cues and human-computer trust. This high level of emotional activation helps users learn depression treatment techniques.

Based on H1, we further hypothesized that: H2. Participants in the chatbot group with higher levels of social cues will have higher levels of emotional arousal, which will be reflected in their physiological indicators, namely (a) heart rate and (b) skin conductance (SC).

3 Methods

3.1 Participants

College students were recruited from four different universities in Guangdong, China and Hong Kong for the experiment. The study made use of a laboratory, where participants were asked to interact with Neil, measure some of their physiological data, and

answer questionnaires provided by us after the experiment. Considering that the research topic was not related to clinical effects and that unnecessary pain should be avoided for vulnerable individuals, we chose to collect experimental data in a non-clinical sample.

In 163 attempts, data from 149 participants were finally used for data analysis. Initially, 14 data were deleted from the dataset because some of the subjects failed to wear the device correctly, resulting in data that could not be measured continuously and accurately, and some participants left the experiment midway due to personal circumstances, and their data could not be included in the study. Participants did not receive any form of medication or psychotherapy before and after the experiment. The experimental methods used in this investigation were authorized by the Institutional Review Board (IRB). Lastly, participants had to be able to communicate in Chinese and give written informed consent.

3.2 Experimental Setting

3.2.1 Chatbot-Delivered Intervention

In this study, we developed and deployed a psychotherapy chatbot designed based on CBT principles, named "Neil". This chatbot can run on various smart terminals, including Windows, macOS, Android, and iOS devices.

Neil's workflow is shown in Fig. 1, and its internal architecture is shown in Fig. 2. We invited experts to carefully review and improve the text content of Neil's dialogue to ensure that it meets the theoretical and practical standards of cognitive behavioral therapy. In addition to text input, Neil chatbot also supports users to interact through voice input. When users use voice input, the system will first use the open source "Paraformer" Alibaba Chinese speech recognition model to convert the voice content into corresponding text information. Next, these text contents will enter Neil's natural language understanding module for in-depth analysis. Natural language processing, intent categorization, and emotion identification are the three primary machine learning model-based functions included in this module. First, the system will extract key information from the text input by the user. Then, the model will classify the information and recognize emotions. The model will annotate and analyze the text content using our pre-set label system for intent classification and emotion recognition. Based on the results of the above natural language understanding, the dialogue management module will generate corresponding output content according to the pre-defined response template. Finally, the database-based natural language generation module will convert these output contents into natural and fluent text feedback. It is worth mentioning that the response guide design in the dialogue management module also strictly follows the relevant principles of CBT.

In the process of communicating with users, this chatbot will evaluate and analyze the user's psychological state. Next, it will take corresponding conversational intervention measures based on the principles of CBT to provide users with psychotherapy support. Specific psychological intervention content includes: helping users to gain a deep understanding of CBT's treatment philosophy, assisting in the assessment of existing problems, identifying automatic negative thinking patterns, questioning and correcting irrational beliefs, cultivating positive and healthy coping methods, and consolidating treatment

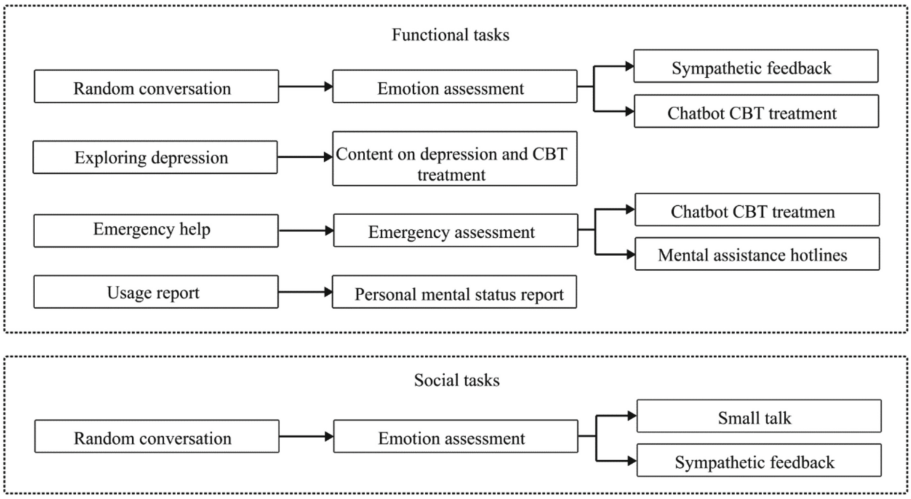


Fig. 1. The chatbot Neil's workflow.

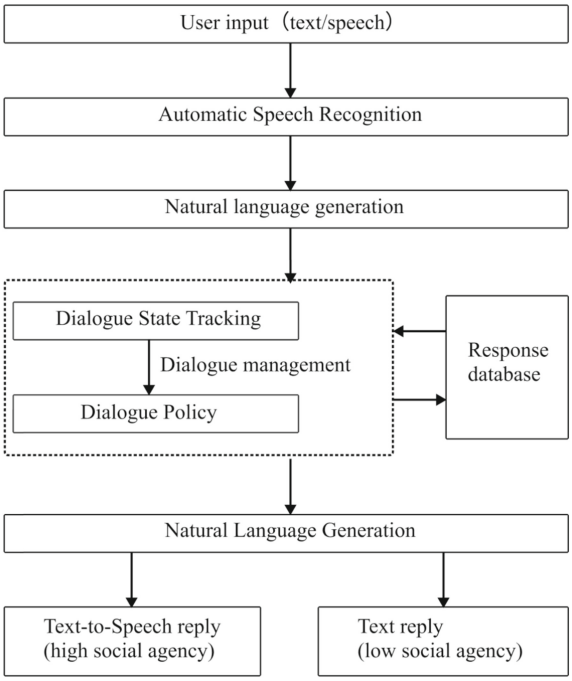


Fig. 2. The chatbot Neil's structure.

results. This systematic cognitive behavioral intervention can help patients improve their emotional state and behavioral performance. In addition to the above-mentioned professional psychological intervention functions, users can also have a more free and casual conversation with Neil, and the chatbot will also provide emotional feedback in an empathetic way. Neil allows individuals to log their own mental health issues and generate customized mental health reports at the same time.

3.2.2 Social Cues

We created two distinct chatbot designs, based on the quantity of social cues in order to investigate the influence of social cues on user trust. Low social agency group: Neil will use human-friendly and empathetic language to engage users in a socially cued textual dialogue, and users need to read the text content displayed on the mobile phone screen, as shown in Fig. 3. High social agency group: Likewise, Neil from this group will engage in socially-cued text conversations with users using humane and empathetic language. But going one step further, Neil in this group has several social cues, and can output responses through the mobile phone speaker playing emotionally toned synthetic speech, in addition to eye expressions (for example, the robot will squint to show seriousness, etc.), head and body movements (for example, nodding to show affirmation and waving to say hello). To implement these social functions, we used Variational Inference for Text-to-Speech (VITS) speech synthesis technology and Unreal Engine 5 (UE5) animation production tools, as shown in Fig. 4.

During the experiment, the feedback from the high social agency group was presented to the participants in the form of emotionally charged synthetic voices played through the mobile phone speakers, accompanied by vivid character animations. In contrast, the low social agency group's feedback required the participants to read the text that was on their phone screens. However, the two chatbots were based on the same core and gave exactly the same answers to the same questions from users.

3.2.3 Procedures

This study was conducted in a laboratory environment. According to the experimental materials and instruments provided, the specific steps for each group of subjects to participate in the experiment are as follows: (1) After understanding the experimental process, the subjects first sat quietly for 3 min, and then took the first physiological data measurement, recording the heart rate and SC before the measurement. (2) After that, the subjects interacted with the mental health chatbot "Neil" according to the tasks we gave. (3) Within 3 min after completing the interaction with Neil, the physiological data was measured again, and the skin conductance level and heart rate after the interaction was recorded. (4) The subjects then filled out the human-computer trust questionnaire. (5) Finally, user satisfaction was measured through the Client Satisfaction Questionnaire-8 (CSQ-8), and therapeutic alliance was measured through the Working Alliance Inventory-Short Revised (WAI-SR). Participants' feedback on their experience with high social agency was also collected.

In this study, the subjects were instructed to ask Neil to perform a series of pre-set tasks. Task characteristics are crucial in controlling the interactive relationship between

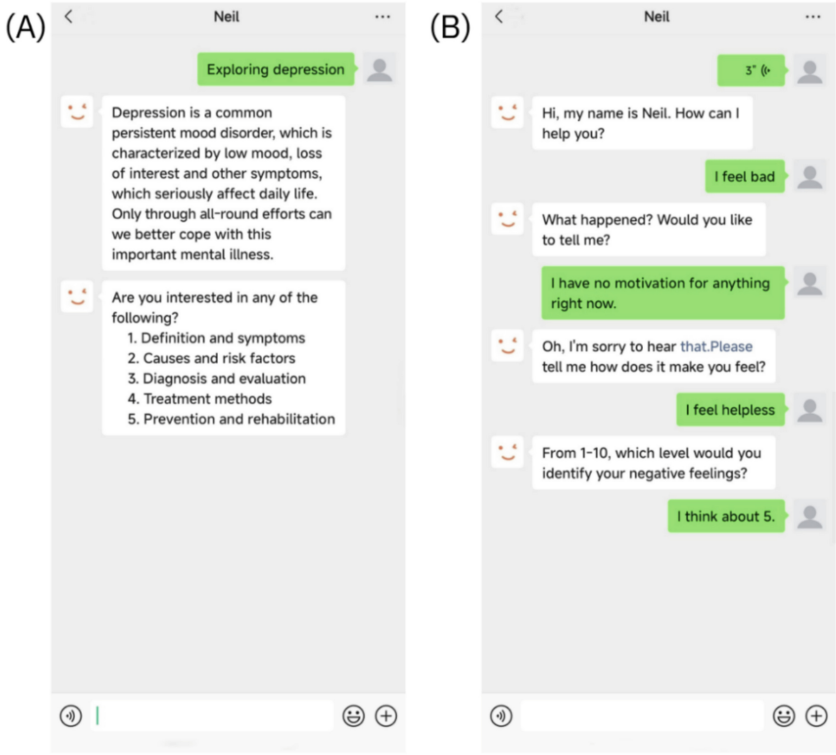


Fig. 3. Examples of low social agency. (**a**) A question-and-answer system for exploring depression is provided while helping users understand depression. (**b**) A demonstration of CBT-based treatment for depression. Supports text messages and voice messages.

users and artificial intelligence systems, despite being frequently overlooked [21, 66]. Based on previous research on the classification of tasks of artificial intelligence agents such as chatbots [67, 68], we divided the tasks that Neil can perform or assist humans to complete into two categories: functional tasks and social tasks. Tasks that include humorous conversations or providing emotional support fall into the category of social tasks, and we designed a series of social tasks accordingly. Mental health intervention tasks that require cognitive abilities can be classified as functional tasks [18]. In addition, we designed a series of functional tasks in detail based on Neil's topic tag system. This tag system is based on the classification of consulting problems and includes three dimensions: S1-trouble type, S2-psychological illness, and S3-SOS. Among them. It should be noted that some projects may require diagnosis and identification by clinical medical professionals, so we referred to professional guidance when designing labels and tried to control the accuracy of concept use. The tasks in these three dimensions cover a range of psychological problems from general troubles to emergency help. See the Appendix 1 for a comprehensive list of questions for every tasks.

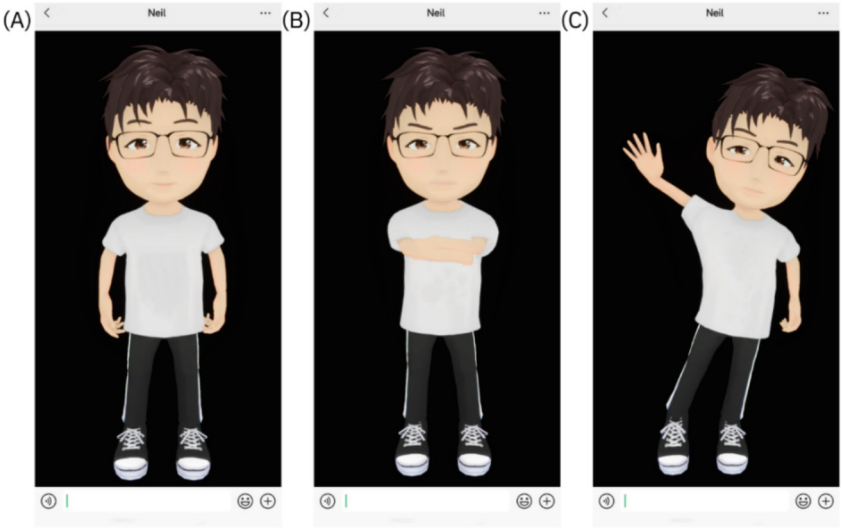


Fig. 4. Examples of high social agency. (**a**) A character in a resting state. (**b**) A serious animation with a frown and crossed arms. (**c**) An animation of a waving hand.

3.3 Measurement

3.3.1 Measurement of Human-Computer Trust

We employed a 25-item scale created by Madsen and Gregor in this investigation [20] to measure participants' trust in Neil. This scale contains the following five dimensions: perceived reliability, perceived technical competence, perceived understandability, faith and personal attachment. Five measuring items, each with a 5-point Likert scale (1 being absolutely disagree, and 5 being completely agree), make up each dimension. Confirmatory factor analysis (CFA) was conducted to assess the validity of the scale. The initial CFA model fit was χ^2 (265) = 360.44, $p < 0.01$, Bentler comparative fit index (CFI) = 0.95, Standardized root mean square residual (SRMR) = 0.06, and Steiger-Lind root mean square error of approximation (RMSEA) = 0.05 with 90% CI = [0.04, 0.06].Each of the above types of fit is good [69, 70].

We computed the average scores of the five trust scale dimensions based on the CFA outcomes of this investigation. The scale reliability index of each dimension was within the acceptable range: perceived reliability (M = 3.49, SD = 0.79, α = 0.84), perceived technical competence (M = 3.69, SD = 0.86, α = 0.84), perceived understandability (M = 3.90, SD = 0.81, α = 085), faith (M = 2.91, SD = 0.93, α = 0.89), and personal attachment (M = 3.47, SD = 0.78, α = 0.88).

3.3.2 Skin Galvanic Response and Heart Rate Measurement

Objective physiological indicators are used to quantitatively evaluate the differences in human-computer trust under different social cue conditions. In this experiment, we used the Grove Galvanic Skin Response (GSR) sensor to measure the SC level of participants and used an electronic heart rate monitor to monitor the heart rate changes of participants.

First, let the participants rest quietly for 3 min to stabilize the physiological indicators, and then start the measurement. Participants in the measurement were instructed to sit in a chair with a backrest, relax their feet naturally, wear a bracelet-type heart rate sensor on their wrists and tighten it moderately to ensure good contact. The GSR sensor is placed between two adjacent fingers. Each SC measurement lasts for 3 min, and 6 data points are recorded every 30 s. In order to eliminate the influence of the initial adaptation and end effect, the data from 30 s before and after are deleted, and the average SC level in the remaining 2 min is calculated as the skin conductance indicator of the participants in this state.

3.4 Randomization

Using the random number generator in the Statistical Package for the Social Science (SPSS) v.27 (IBM Corp., Armonk, NY), participants were assigned a random number, $0 < n \leq 1$. The high social agency test group was assigned to individuals who received the number $0 < n \leq 0.5$, whereas the control group was assigned to the remaining participants.

3.5 Statistical Methods

The data were analyzed using SPSS v.27 software, the significance threshold was established at $p = 0.05$. In order to determine if there were significant differences between the two groups at the baseline level, baseline factors (such as age, gender, education level, and clinical characteristics) were compared using analysis of variance (ANOVA) and chi-square analysis. The analysis of trust between humans and computers made use of an independent t-test. For the physiological indicator data, we used analysis of covariance (ANCOVA) to analyze, and we computed Cohen's d to determine the effect size. For CSQ-8 scores and WAI-SR scores, we used independent sample t-tests for inter-group comparisons. In addition, for the feedback data of high social agency group users, we also conducted word frequency analysis to supplement the results of quantitative analysis and explore the participants' usage experience and feedback more comprehensively.

4 Results

All participants were college students, including 113 undergraduates and 36 graduate students. There were more females ($n = 77$, 51.7%) than males, aged 18–28 years (mean = 23.17, SD = 1.02). All participants were able to speak and understand Chinese.

4.1 Baseline Data

We performed ANOVA and chi-square tests on age, gender, educational attainment, and relevant clinical factors in each group at T_1, as indicated in Table 1. The findings indicated that these baseline parameters did not differ statistically significantly between the two groups.

Table 1. Participant and variable demographics at baseline (T1).

	High social agency group[a]	Low social agency group[a]	χ^2/t	P
Age (years)	23.310(1.06)	23.026(0.95)	1.71	0.09
Gender				
Male	37(50.00)	40(53.33)	0.17	0.68
Female	37(50.00)	35(46.67)		
Education (years)	16.28(1.03)	16.04(0.93)	1.50	0.14
HR	83.31(8.21)	82.05(8.24)	0.93	0.36
SC(μS)	1.28(0.90)	1.35(0.91)	0.45	0.66

[a] The numbers are mean (standard deviation) or n (%)

4.2 Human-Computer Trust

Three of the five aspects of trust had significant differences, according to an independent t-test study that was conducted later. Compared with the low social agency group, the high social agency group reported significantly higher scores on the three dimensions of perceived reliability ($p < 0.05$), perceived technical competence ($p < 0.01$), and faith ($p < 0.01$). Both perceived reliability and perceived technical competence are considered to be cognitive aspects of human-computer trust, while faith is considered to be one of the emotional aspects of human-computer trust [20]. Therefore, these results support the association between high social agency and cognitive aspects of trust (H1a) as well as affective aspects (H1b). Figure 5 presents these results in the form of a bar graph.

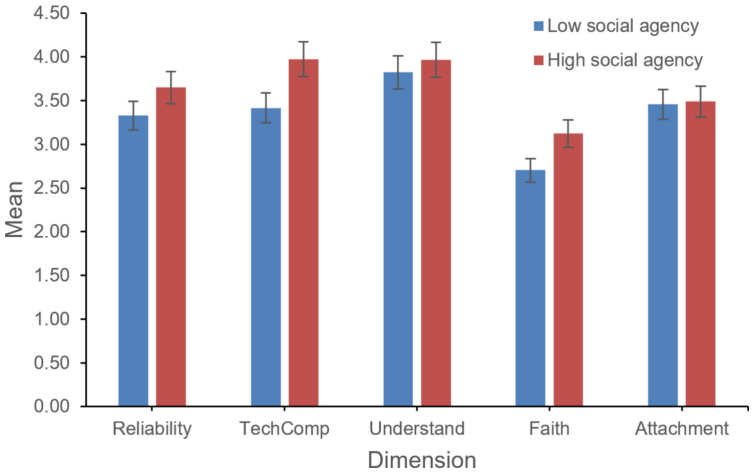


Fig. 5. Differences in human-computer trust of five dimensions by social cues.

4.3 Physiological Data

After adjusting for baseline factors, we performed an ANCOVA analysis on the outcome data gathered at T_2 and assessed the effect of experimental group membership on the result, as indicated in Table 2. The findings demonstrated that individuals in the high social agency group had significantly higher heart rates ($d = 0.90$) with a high effect size and SC ($d = 0.71$) with a medium effect size when compared to the low social agency group: heart rate ($F = 48.63$, $p < 0.01$) and SC ($F = 26.25$, $p < 0.01$) showed significant changes. This implies that high amounts of social cues can cause higher degrees of emotional arousal. These changes in physiological indicators show that participants in the high social agency group had higher levels of emotional arousal. Combined with the statistically significant finding that the high social agency group reported higher trust, this also confirms that changes in physiological indicators can effectively characterize human-computer trust. Overall, the associations between physiological indicators of emotional arousal, including heart rate (H2a) and skin conductance level (H2b), and social cues were supported.

Table 2. ITT analysis at T2.

	High social agency		Low social agency		F	P	d [c]
	T_2 [a]	95%CI [b]	T_2 [a]	95%CI [b]			
HR	93.65	92.66–94.65	88.69	87.97–89.69	48.63	<0.01*	0.90
SC(μS)	2.57	2.35–2.80	1.74	1.52–1.97	26.25	<0.01*	0.71

* The result is significant at the 0.01 level

[a] The numbers are pooled mean (standard error)

[b] 95% Confidence Interval

[c] Using means and standard errors, Cohen's d was demonstrated for between-subjects effects at T_2

4.4 Auxiliary Analysis

To examine the differences in client satisfaction and therapeutic connection between the two groups, we ran an independent t-test on the individuals who finished the trial at T_2. According to the WAI-SR scale ($t = 5.16$, $p < 0.001$), the high social agency trial group had a significantly better therapeutic alliance than the low social agency control group. At the same time, the high social agency trial group also scored significantly higher than the low social agency control group on the CSQ-8 client satisfaction scale ($t = 3.68$, $p < 0.001$).

According to the results of the frequency analysis of user feedback in Table 3, process factors have a slightly greater impact than content factors. It is worth noting that three participants said in the interview that although they trust Neil more than other similar products, they are still more willing to seek help from professional psychological counselors.

Table 3. Participant feedback results statistics[a].

The best experiences about using Neil	Process (48/67)[b]	Easy to access (31/67) Quick response (25/67)
	Content (36/67)	Empathetic feedback (26/67)
		Humanized wording (22/67)
		Professional (17/67)
		Turn-based dialogue (11/67)
The worst experiences about using Neil	Process (43/67)	Lack of personalization (20/67) Unnatural (14/67)
		Stipid (12/67)
	Content (40/67)	General (15/67)
		Irrelevant content (13/67)
		Repetitive dialogue (13/67)
		Expertise is difficult to understand (8/67)

[a] Several responses include more than one theme and were counted more than once
[b] Numbers are (counted number/total number of participants)

5 Discussion

This study experimentally examined how different levels of social cues affect participants' trust levels when interacting with a depression treatment chatbot. At the same time, the study also explored the relationship between social cues and emotional arousal using physiological indicators such as heart rate and skin conductance to further clarify the connection between social cues and trust. As artificial intelligence-based virtual agents become increasingly popular, how users interact with these agents and build trust has attracted increasing attention [71]. The results of this study not only strengthen the concept of human-computer trust, but also provide a new perspective on the relationship between social cues, trust, and emotional arousal. In addition, the results support the social agent hypothesis [72], which states that more social cues can activate richer human social interaction patterns. The following is an in-depth discussion of these findings.

The results showed that participants reported higher levels of trust when using a treatment chatbot with high social cues, especially in the cognitive dimensions of perceived reliability and technical ability, and the emotional dimension of belief. This suggests that participants believed that chatbots with high social cues were more reliable and capable, and developed a stronger sense of dependence on them. However, on the dimension of comprehensibility, participants' trust did not increase significantly, which may be due to their lack of knowledge of cognitive behavioral therapy (CBT). At the same time, although high social cues enhanced participants' emotional trust, no significant improvement was observed in the personal attachment dimension, which may reflect that people still regard emotional attachment as a unique trait of humans.

Analysis of physiological indicators showed that the emotional arousal level of participants in the high social agent group was significantly higher than that of the low social agent group. This finding provides quantitative evidence for the higher trust of participants in the low social agent group, thus proving the positive role of social cues in enhancing users' trust in therapeutic chatbots. High social agents may achieve more effective intervention by enhancing emotional arousal.

In addition, the results also showed that the high social cue group scored significantly higher than the low social cue group on the CSQ-8 user experience measure and the WAI-SR therapeutic alliance measure, indicating that therapeutic chatbots with high social cues have obvious advantages in establishing therapeutic alliances and providing high-quality user experience. This may be because strong social cues motivate users to have more trust in chatbots, thereby promoting the development of therapeutic alliances and improving user experience. This finding further verifies the importance of social cues in self-help intervention for depression and provides valuable inspiration for future system design.

5.1 Limitations and Future Directions

This study effectively controlled unknown influencing factors through random grouping experimental methods and explored the impact of social cues on a series of dependent variables. However, the limitations of the sample cannot be ignored: the participants were mainly college students. Although they are active users of high technology and have a certain representativeness, this also limits the generalization of the research results to a wider population. Although the use of non-clinical samples is in line with the safety principle, future studies still need to adopt random sampling, expand the sample range, and include a wider range of therapeutic chatbot users to enhance the representativeness and universality of the results.

In addition, the functional tasks of the therapeutic chatbot Neil used in this study are relatively fixed, mainly focusing on self-help depression intervention, so the results should not be simply generalized to other types of chatbots. Future research should consider different task settings, social agent designs, and user groups to obtain broader application insights.

In terms of the setting of social cues, this study has certain limitations. Although the research team tried their best to adjust the robot's movements, the difference between the chatbot's virtual image and the real human body may make it impossible for users to accurately identify its body movements, thereby affecting their trust in the chatbot. Therefore, the design of social cues needs to be further optimized, and future studies should clarify which specific social cues can effectively enhance trust in order to achieve precise optimization.

There are also limitations in emotional arousal. Although the results showed that high social agency did induce higher emotional arousal, participants' heart rates generally increased during the experiment. Although it can be inferred that the participants were in an aroused state, the lack of comparable research makes it challenging to determine the cause of this observation. In addition, this study used wearable devices to collect physiological data, but the collection time was short and the frequency was limited, which failed to fully analyze the participants' arousal moments. Future research should

consider increasing the time and frequency of data collection to analyze physiological data in more detail.

5.2 Design Implications

The results of this study can provide important guidance for the design of future therapeutic chatbots and other AI agents. Designers should consider giving systems richer social features to enhance users' trust. Future research should focus on users' preferences when interacting with AI agents, exploring their preferences for interactions or social features similar to human society and how these preferences affect the establishment of trust. In addition, when designing AI agents, allowing users to customize social cues will help provide a more personalized experience. Users can customize their AI agents by selecting social features according to their preferences, thereby improving user satisfaction and trust.

In the future, AI agents are expected to be able to measure users' physiological responses in real time and detect emotional arousal to assess the level of trust between humans and agents. This capability will make systems that integrate social cues more scalable and flexible. Because data is obtained instantly during the interaction, the personalized responses of AI agents will also be timelier and more appropriate, further enhancing user experience and trust. These design insights provide valuable directions for the development of future AI agents to better serve users.

6 Conclusion

This study explored therapeutic chatbots with different levels of social cues, focusing on the impact of social cues on participants' trust, while using heart rate and skin conductance as physiological indicators to assess emotional arousal. The results showed that high social cues were significantly correlated with cognitive dimensions (including perceived reliability and technical capabilities) and affective dimensions (beliefs) of human-computer trust and could induce higher emotional arousal. These findings not only expand the theoretical framework between social cues and human-computer trust, but also provide an important basis for establishing a more complete theoretical system of human-computer communication.

Acknowledgments. The authors thank all participants who participated in this study.

Disclosure of Interests. No potential conflict of interest was reported by the author(s).

Appendix 1

Questions for the Social Task Setting

Q 1. Hey, Nancy!
Q 2. Who is Neil?
Q 3. What is your job?

Q 4. Where are you from?
Q 5. How much money do you make?
Q 6. How old are you?
Q 7. Do you like to read book?
Q 8. What is your favorite animal?
Q 9. Which came first, the chicken or the egg?
Q 10. What is your favorite food?
Q 11. Tell me a joke.
Q 12. Can pigs fly?
Q 13. If you could eat egg tarts, how many could you eat?
Q 14. How many steps does it take to put an elephant in the refrigerator?
Q 15. Who is the best mental health smart assistant?

Questions for the Functional Task Setting

Q 1. I feel that studying is very difficult and I have no interest in studying. What should I do?
Q 2. My relationship with my colleagues is very tense and we had a quarrel today.
Q 3. Today I had a quarrel with my father because of charging money for the game. Do you have any suggestions?
Q 4. I like to drink very much, but I think I may be addicted to alcoholic beverages. What should I do?
Q 5. I couldn't sleep last night and fell asleep very late.
Q 6. I broke up with my girlfriend (boyfriend) and now I am very upset.
Q 7. I think my personality is a bit introverted and I am always very restrained when getting along with others.
Q 8. I think my accent is very strange, and sometimes I dare not speak loudly to others.
Q 9. My pet dog passed away, and I am so sad.
Q 10. I always feel tired and weak, often feel guilty and desperate, and try to avoid participating in social activities.
Q 11. I have a very important interview recently, but I always feel palpitations, sometimes I feel restless and have difficulty falling asleep.
Q 12. My mood has been fluctuating a lot recently. Sometimes I am manic, full of confidence, and then I fall into depression. What should I do?
Q 13. I was injured in an accident, and since then I have frequent nightmares, always feeling nervous and scared. What should I do?
Q 14. After a long period of dieting, I have no interest in food at all, and even have some anorexia.
Q 15. I always want to cut my arm with a knife. What should I do?

References

1. Auerbach, R.P., Alonso, J., Axinn, W.G., Cuijpers, P., Ebert, D.D., Green, J.G., et al.: Mental disorders among college students in the World Health Organization world mental health surveys. Psychol. Med. **46**(14), 2955–2970 (2016)
2. Hawton, K., Comabella, C.C., Haw, C., Saunders, K.: Risk factors for suicide in individuals with depression: a systematic review. J. Affect. Disord. **147**(1–3), 17–28 (2013)
3. Guo, L.P., Li, Z.H., Chen, T.L., Liu, G.H., Fan, H.Y., Yang, K.H.: The prevalence of mental distress and association with social changes among postgraduate students in China: a cross-temporal meta-analysis. Public Health. **186**, 178–184 (2020)
4. Sheldon, E., Simmonds-Buckley, M., Bone, C., Mascarenhas, T., Chan, N., Wincott, M., et al.: Prevalence and risk factors for mental health problems in university undergraduate students: a systematic review with meta-analysis. J. Affect. Disord. **287**, 282–292 (2021)
5. Andrews, B., Wilding, J.M.: The relation of depression and anxiety to life-stress and achievement in students. Br. J. Psychol. **95**(4), 509–521 (2004)
6. Hysenbegasi, A., Hass, S.L., Rowland, C.R.: The impact of depression on the academic productivity of university students. J. Ment. Health Policy Econ. **8**(3), 145–151 (2005)
7. Lee, T.S.-H., Wu, Y.-J., Chao, E., Chang, C.-W., Hwang, K.-S., Wu, W.-C.: Resilience as a mediator of interpersonal relationships and depressive symptoms amongst 10th to 12th grade students. J. Affect. Disord. **278**, 107–113 (2021)
8. Zhong, X., Liu, Y., Pu, J., Tian, L., Gui, S., Song, X., et al.: Depressive symptoms and quality of life among Chinese medical postgraduates: a national cross-sectional study. Psychol. Health Med. **24**(8), 1015–1027 (2019)
9. Andersson, G.: Internet-delivered psychological treatments. Annu. Rev. Clin. Psychol. **12**(1), 157–179 (2016)
10. Wang, K., Varma, D.S., Prosperi, M.: A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J. Psychiatr. Res. **107**, 73–78 (2018)
11. Liu, H., Peng, H., Song, X., Xu, C., Zhang, M.: Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv. **27**, 100495 (2022)
12. Gratzer, D., Goldbloom, D.: Therapy and e-therapy—preparing future psychiatrists in the era of apps and chatbots. Acad. Psychiatry. **44**, 231–234 (2020)
13. Vaidyam, A.N., Wisniewski, H., Halamka, J.D., Kashavan, M.S., Torous, J.B.: Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can. J. Psychiatry. **64**(7), 456–464 (2019)
14. Lee, J.-E.R., Nass, C.I.: Trust in computers: the computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In: Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives, pp. 1–15. IGI Global, Hershey, PA (2010)
15. Yagoda, R.E., Gillan, D.J.: You want me to trust a ROBOT? The development of a human–robot interaction trust scale. Int. J. Soc. Robot. **4**, 235–248 (2012)
16. Guzman, A.L.: Making AI Safe for Humans: A Conversation with Siri. Socialbots and Their Friends, pp. 85–101. Routledge, London (2016)
17. Sundar, S.S.: The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. MacArthur Foundation Digital Media and Learning Initiative, Cambridge, MA (2008)
18. Lee, S.K., Kavya, P., Lasser, S.C.: Social interactions and relationships with an intelligent virtual agent. Int. J. Hum. Comput. Stud. **150**, 102608 (2021)

19. Lee, E.J., Nass, C., Brave, S.: Can computer-generated speech have gender? An experimental test of gender stereotype, pp. 289–290 (2000)
20. Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems, pp. 6–8 (2000)
21. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors. **53**(5), 517–527 (2011)
22. Hamid, O.H., Smith, N.L., Barzanji, A.: Automation, Per Se, Is Not Job Elimination: How Artificial Intelligence Forwards Cooperative Human-Machine Coexistence, pp. 899–904. IEEE, Piscataway, NJ
23. Rothfuß, S., Wörner, M., Inga, J., Kiesel, A., Hohmann, S.: Human–machine cooperative decision making outperforms individualism and autonomy. IEEE Trans. Hum. Mach. Syst. **53**(4), 761–770 (2023)
24. Lu, M., Cui, T., Huang, Z., Zhao, H., Li, T., Wang, K.: A systematic review of questionnaire-based quantitative research on MOOCs. Int. Rev. Res. Open Dist. Learn. **22**(2), 285–313 (2021)
25. Smith, N.L., Hamid, O.H.: The Relative Importance of Perception, Embodiment, Metaphors, and Ethics for Cooperative Human-Machine Coexistence, pp. 429–435
26. Mayer, R.C.: An integrative model of organizational trust. Acad. Manage. Rev. **20**, 709 (1995)
27. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors. **46**(1), 50–80 (2004)
28. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors. **57**(3), 407–434 (2015)
29. Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. Hum. Factors. **58**(3), 377–400 (2016)
30. Lobato, E.J.C., Warta, S.F., Wiltshire, T.J., Fiore, S.M.: Varying social cue constellations results in different attributed social signals in a simulated surveillance task. In: Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference
31. Burgoon, J.K., Guerrero, L.K., Manusov, V.: Nonverbal signals. In: Handbook of Interpersonal Communication, pp. 239–280 (2011)
32. Leathers DG. Nonverbal Communication Systems. 1976
33. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Equilibrium theory revisited: mutual gaze and personal space in virtual environments. Presence Teleoper. Virtual Environ. **10**(6), 583–598 (2001)
34. Li, C.-Y.: Persuasive messages on information system acceptance: a theoretical extension of elaboration likelihood model and social influence theory. Comput. Hum. Behav. **29**(1), 264–275 (2013)
35. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Educ. Couns. **59**(1), 21–30 (2005)
36. Burton, C., Szentagotai Tatar, A., McKinstry, B., Matheson, C., Matu, S., Moldovan, R., et al.: Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J. Telemed. Telecare. **22**(6), 348–355 (2016)
37. Suganuma, S., Sakamoto, D., Shimoyama, H.: An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. JMIR Ment. Health. **5**(3), e10454 (2018)
38. Chetouani, M., Boucenna, S., Chaby, L., Plaza, M., Cohen, D., Chaby, L., et al.: Social Signal Processing and Socially Assistive Robotics in Developmental Disorders. Cambridge University Press, Cambridge (2017)

39. Choi, M., Kornfield, R., Takayama, L., Mutlu, B.: Movement matters: effects of motion and mimicry on perception of similarity and closeness in robot-mediated communication, pp. 325–335 (2017)
40. Thimmesch-Gill, Z., Harder, K.A., Koutstaal, W.: Perceiving emotions in robot body language: acute stress heightens sensitivity to negativity while attenuating sensitivity to arousal. Comput. Hum. Behav. **76**, 59–67 (2017)
41. Martin, C.D.: The media equation: how people treat computers, television and new media like real people and places [book review]. IEEE Spectr. **34**(3), 9–10 (1997)
42. Chidambaram, V., Chiang, Y.-H., Mutlu, B.: Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues, pp. 293–300 (2012)
43. Roubroeks, M., Midden, C., Ham, J.: Does It Make a Difference Who Tells You What to Do? Exploring the Effect of Social Agency on Psychological Reactance, pp. 1–6 (2009)
44. Louwerse, M.M., Graesser, A.C., Lu, S., Mitchell, H.H.: Social cues in animated conversational agents. Appl. Cognit. Psychol. **19**(6), 693–704 (2005)
45. Andrist, S., Spannan, E., Mutlu, B.: Rhetorical Robots: Making Robots More Effective Speakers Using Linguistic Cues of Expertise, pp. 341–348. IEEE, Piscataway, NJ (2013)
46. Cooney, S., Dignam, H., Brady, N.: Heads first: visual aftereffects reveal hierarchical integration of cues to social attention. PLoS One. **10**(9), e0135742 (2015)
47. Eyssel, F.A., Hegel, F.: (S)he's got the look: Gender stereotyping of robots. J. Appl. Soc. Psychol. **42**(9), 2213–2230 (2012)
48. Lee, K.M., Park, N., Song, H.: Can a robot be perceived as a developing creature? Effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. Hum. Commun. Res. **31**(4), 538–563 (2005)
49. Agrawal, S., Williams, M.-A.: Robot Authority and Human Obedience: A Study of Human Behaviour Using a Robot Security Guard, pp. 57–58 (2017)
50. Lopez, A., Ccasane, B., Paredes, R., Cuellar, F.: Effects of Using Indirect Language by a Robot to Change Human Attitudes, pp. 193–194 (2017)
51. Atkinson, R.K., Mayer, R.E., Merrill, M.M.: Fostering social agency in multimedia learning: examining the impact of an animated agent's voice. Contemp. Educ. Psychol. **30**(1), 117–139 (2005)
52. Lee, E.J., Nass, C., Brave, S.: Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype, pp. 289–290
53. Nass, C., Moon, Y., Green, N.: Are machines gender neutral? Gender-stereotypic responses to computers with voices. J. Appl. Soc. Psychol. **27**(10), 864–876 (1997)
54. Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J. Exp. Psychol. Appl. **7**(3), 171–181 (2001)
55. Brave, S., Nass, C., Hutchinson, K.: Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int. J. Hum. Comput. Stud. **62**(2), 161–178 (2005)
56. Gong, L., Nass, C.: When a talking-face computer agent is half-human and half-humanoid: human identity and consistency preference. Hum. Commun. Res. **33**(2), 163–193 (2007)
57. Ham, J., Cuijpers, R.H., Cabibihan, J.-J.: Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. Int. J. Soc. Robot. **7**, 479–487 (2015)
58. Bui, L., Pezzola, M., Bandara, D.: How Do AI Explanations Affect Human-AI Trust? pp. 175–183. Springer, Berlin
59. Merrill, N., Cheshire, C.: Trust Your Heart: Assessing Cooperation and Trust with Biosignals in Computer-Mediated Interactions, pp. 2–12

60. Mitkidis, P., McGraw, J.J., Roepstorff, A., Wallot, S.: Building trust: heart rate synchrony and arousal during joint action increased by public goods game. Physiol. Behav. **149**, 101–106 (2015)
61. Morris, D.M., Erno, J.M., Pilcher, J.J.: Electrodermal Response and Automation Trust During Simulated Self-Driving Car Use, 1st edn, pp. 1759–1762. SAGE, Los Angeles, CA
62. Christopoulos, G.I., Uy, M.A., Yap, W.J.: The body and the brain: measuring skin conductance responses to understand the emotional experience. Organ. Res. Methods. **22**(1), 394–420 (2019)
63. Electrophysiology TFotESoCtNASoP: Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation. **93**(5), 1043–1065 (1996)
64. Kuyken, W., Padesky, C.A., Dudley, R.: Collaborative Case Conceptualization: Working Effectively with Clients in Cognitive-Behavioral Therapy. Guilford Press, New York (2011)
65. McGaugh, J.L.: Emotional arousal regulation of memory consolidation. Curr. Opin. Behav. Sci. **19**, 55–60 (2018)
66. Rzepka C, Berger B. User Interaction with AI-Enabled Systems: A Systematic Review of IS Research. 2018
67. Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., Ivaldi, S.: Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. Comput. Hum. Behav. **61**, 633–655 (2016)
68. Hoffmann, L., Krämer, N.C.: Investigating the effects of physical and virtual embodiment in task-oriented and conversational contexts. Int. J. Hum. Comput. Stud. **71**(7–8), 763–774 (2013)
69. Kline, R.B.: Methodology in the social sciences. In: Principles and Practice of Structural Equation Modeling, p. 2 (2005)
70. Hu, L., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. Multidiscip. J. **6**(1), 1–55 (1999)
71. Guzman, A.L., Lewis, S.C.: Artificial intelligence and communication: a human–machine communication research agenda. New Media Soc. **22**(1), 70–86 (2020)
72. Mayer, R.E., Sobko, K., Mautone, P.D.: Social cues in multimedia learning: role of speaker's voice. J. Educ. Psychol. **95**(2), 419–425 (2003)

AI and Digital Human Modeling in Safety and Risk Management

Statistical 3D Body Shape Predictions for Standardisation of Digital Human Modelling Tools

Erik Brolin(✉), Estela Pérez Luque, and Aitor Iriondo Pascual

School of Engineering Science, University of Skövde, Skövde, Sweden
erik.brolin@his.se

Abstract. This paper presents the development of statistical 3D body shape prediction models and how these models can be shared to be used by other researchers, software developers or organisations. In Digital human modelling (DHM) tools it is important that the generated manikin models are accurate and representative for different body sizes and shapes. By using both 3D body scan data and one-dimensional data, provided in tabular format, a prediction model was created that based on a few input variables predicts both additional missing one-dimensional data and body shape data, in the form of landmark coordinates in XYZ point cloud format as well as a body shape mesh model in OBJ format. The data is handled in a sequential process where three different prediction functions were defined using similar implementations of a conditional regression model. The developed statistical 3D body shape prediction model described in this paper is based on body scan data from the CAESAR anthropometric survey. The statistical 3D body shape prediction models, consisting of MATLAB scripts have been prepared, packaged and shared in an online repository on GitHub under the MIT License. Since the model is shared under an open license, the idea and intention are that it can be further developed by other researchers and organizations. This first version only generates a static body shape whereas future versions could include joint center prediction models and deformation patterns in relation to joint angles, to enable accurate body deformation during different postures and motions.

Keywords: Anthropometry · body shape · statistical body model

1 Introduction

Contemporary product and production development are typically carried out with support of computer aided design and engineering (CAD/CAE) tools where design of products and workplaces are originated and evaluated within virtual environments. Ergonomics addresses factors important to consider in product and production development processes to ensure a good interaction between humans and items being designed, whether that is a product or a workplace [1].

Digital human modelling (DHM) tools are useful when evaluating human-machine interaction and enables consideration of anthropometric diversity by creating human

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 121–131, 2025.
https://doi.org/10.1007/978-3-031-93502-2_8

models, so called manikins, of different sizes and proportions. This functionality is a central feature when using DHM tools for product and production development to ensure that the design fits the intended proportion of the targeted population, from a physical perspective [2, 3]. Therefore, it is important that these generated manikin models are accurate and representative for different body sizes and shapes, for both women and men, as well as different age groups. DHM tools should include functionality that enables manikin measurement and data documentation to be compliant with standard documents [4, 5]. Still, in standard documents as well as other sources, anthropometric data are often presented either by mean and standard deviation values or with percentile values, e.g. 5th, 50th and 95th percentile. However, mean and standard deviation values, as well as percentile data, are univariate, i.e. one dimensional, and usually lack information about how different measurements correlate to each other as well as the shape of the body. One dimensional data is therefore not suitable to use when generating 3D human models in DHM tools. Since the first major 3D body scanning based anthropometric survey CAESAR [6], numerous additional surveys have been conducted around the world. Three-dimensional body scan data makes it possible to accurately recreate body shapes using statistical regression models where input data are one-dimensional anthropometric measurements, e.g. stature, body weight and sitting height [7]. Another benefit with statistical regression models is that the original raw data is no longer necessary to generate predictions when regression coefficients have been produced. Still, few available DHM tools [8] have implemented body scan based human models for simulation of product or workplace ergonomics. This could be due to a lack of standardised methodology and data for producing, as well as evaluating the accuracy of, statistical 3D body shape predictions. This paper will present how statistical regression models can be produced from 3D body scan data but also shared to be used by other researchers, software developers or organisations. The assumption is that if methodology and data are made available, the use of it will increase and enable more realistic digital human models in simulations of product and workplace ergonomics.

2 Method

By using both 3D body scan data and one-dimensional data, provided in tabular format, it is possible to create a prediction model that based on a few input variables predicts both additional missing one-dimensional data and body shape data. However, the data needs to be structured and handled in a sequential process for the prediction model to be both stable and capable of producing intended results (Fig. 1). Three different prediction functions were defined using similar implementations of a conditional regression model [9] which can take different and any number of input parameters to predict unknown variables. The conditional regression model examines the correlation between independent (predictive) and dependent (predicted) variables to determine the regression coefficients. These coefficients are then used to construct linear regression equations for each dependent variable. These equations enable the prediction of missing dependent variables. However, this approach assumes that the data used to generate the regression equation is sufficiently similar to the data of the target group. For instance, male and female data are analyzed separately, with distinct regression equations for each sex.

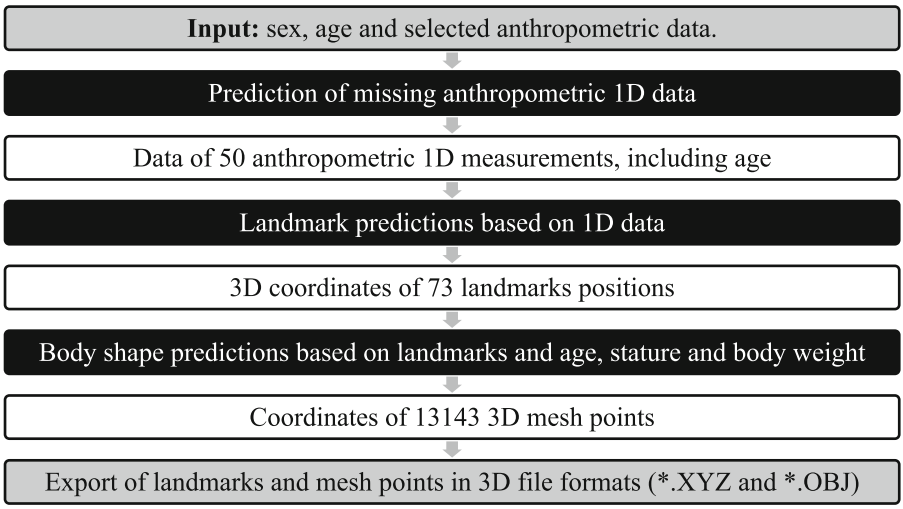


Fig. 1. Process steps of the statistical 3D landmark and body shape predictions starting with anthropometric input data and ending with export of landmarks and mesh points of body shape.

2.1 Anthropometric Body Scan Data

The statistical body shape predictions described in this paper are based on body scan data from the CAESAR anthropometric survey. In this first version, the North American data set is used since this data set has a large sample size and diversity considering the range of measurement of values for the anthropometric one-dimensional data, provided in tabular format. All prediction models developed in this first version were based on a sample from the North American CAESAR data set consisting of 1120 women and 1065 men. Apart from the body scan of participants, the data set consists of 50 measurements including age, where 30 of the measurements can be found in the standard document ISO 7250-1:2017 [10], and the other 19 measurements are defined in reports from the CAESAR survey [6]. 42 of the 50 measurements are reported in tabular format in the original CAESAR data documentation [6]. However, eight additional measurements have been added to the data through different approaches. Four measurements, Eye height, Elbow height, Cervicale height sitting, and Shoulder-elbow length are derived from other given measurements. Two measurements, Iliac spine height and Hand breadth at metacarpals are calculated as the vertical (Iliac spine height) and direct (Hand breadth) distance between landmarks. Two measurements, Waist and Calf circumference are estimated directly from the original body scans which have measuring data in horizontal layers every 2 mm. Both circumference measurement values were generated by calculating the cumulative distance between all scan points on specific heights (Fig. 2) as close as possible to the definitions in ISO 7250-1:2017 [10]. The height of the waist circumference was calculated as the midpoint between the Iliocristale height and the height of 10th rib Midspine (Fig. 2). The height of the calf circumference was found by identifying the maximum circumference of the calf (Fig. 2).

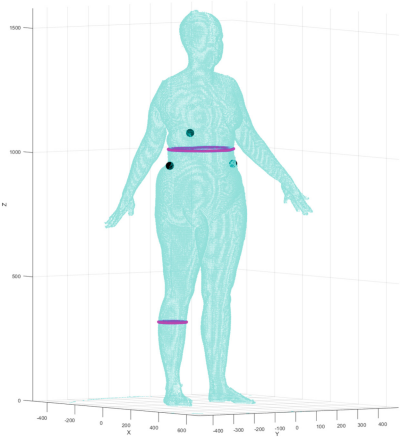


Fig. 2. Visualization of how waist and calf circumference (purple) was estimated for each body scan (turquoise) also showing the landmarks (black) for the left and right Iliocristale as well as the the 10th rib Midspine.

2.2 Prediction of Missing Anthropometric 1D Data

To be able to distribute a body shape predictions model, possible to use by other researchers and organizations, it needs to have some basic functionality for including user input data into the system. In this study, the conditional regression model [9] was set up so it automatically predicts all anthropometric measurements, that is not defined as independent predictive variables, of the 50 measurements necessary for the following body shape predictions. The implemented conditional regression model only needs to use descriptive statistics, i.e. mean and standard deviation values as well as the correlation matrix for all measurements to generate predictions. Therefore, the predictions can be adaptable and produce regression results without needing to have the original raw data from each participant which can be kept anonymized. The implemented conditional regression model also makes use of Principal Component Regression [9, 11] when there are several independent variables which otherwise could create issues of multicollinearity between the independent variables. The number of Principal Components (PC) used is in the current version automatically controlled by discarding all PCs with eigen values below 0.01, where the average eigen value is 1.

2.3 3D Landmark and Body Shape Predictions from Scan Data

To be able to generate statistical body shape predictions, the body scan data from the CAESAR study needed to be processed in three additional steps, also described in Brolin et al. [12]. The three steps were:

1. Rotate and translate landmark data together with body scan point cloud
2. Fit template mesh to body scan point cloud using homologous body modelling
3. Create statistical prediction of landmarks and mesh

In the first step a suitable rotation angle around the vertical axis could be calculated for each body scan using data of landmark pairs, found on both sides of the mid sagittal plane. A translation distance in three-dimensional space could be calculated using the known landmarks on the mid sagittal plane and height of the measuring platform for each subject. Through a MATLAB [13] script these rotation angles and translation distances were used to align all body scan point clouds in a similar position and orientation (Fig. 3).

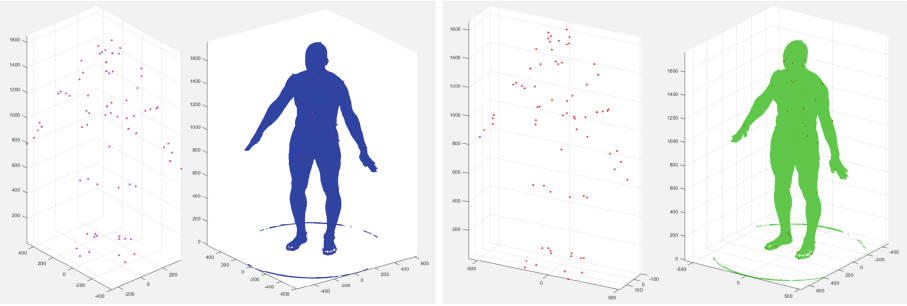


Fig. 3. Original landmarks and point cloud (left) rotated and translated (right).

In the second step, after the alignment had been done, a template mesh was fitted to all body scan point clouds using the software mHBM (Markerless Homologous Body Modelling) [14] (Fig. 4). The mHBM software deforms a mesh model such that the distance between the mesh and scan data is minimized, where software settings can allow for smaller or larger deformations. All fitted template meshes to body scans were visually inspected to look for unacceptable deformities, e.g. upper and lower limbs being misaligned to other parts of the body. Due to some unacceptable deformities during an initial template mesh fitting, two iterations were done where larger deformations were allowed for the torso and abdominal area, and smaller deformations were allowed for head and upper and lower limbs. Fitted mesh points from the two iterations could afterwards be put together to get a final fitted template mesh. The template mesh fitting was done on all 1120 women and 1065 men.

In the third step, the fitted template mesh and landmark data were used to generate statistical prediction models of both landmark points and mesh vertices coordinates, applying the principles of the conditional regression model but now predicting values for 3D coordinates. The statistical prediction, realized through MATLAB [13] scripts, was done in a two-step procedure where 73 landmarks were first predicted based on the 50 anthropometric measurements, including age, that had been generated in the previous step. Also, for the landmark predictions Principal component regression was used where the number of principal components used in the regression was 37 for both the female and male data. The coordinates of the landmarks were adjusted to make pairs symmetrical and mid points aligned to the mid sagittal plane. Then, all mesh vertices were predicted based on the predicted landmark coordinates as well as age, stature and weight. The number of principal components used in the last regression, predicting mesh points, were 116 for the male data and 121 for the female data. A challenge regarding the prediction of landmarks and body shape mesh 3D coordinates, compared to the prediction

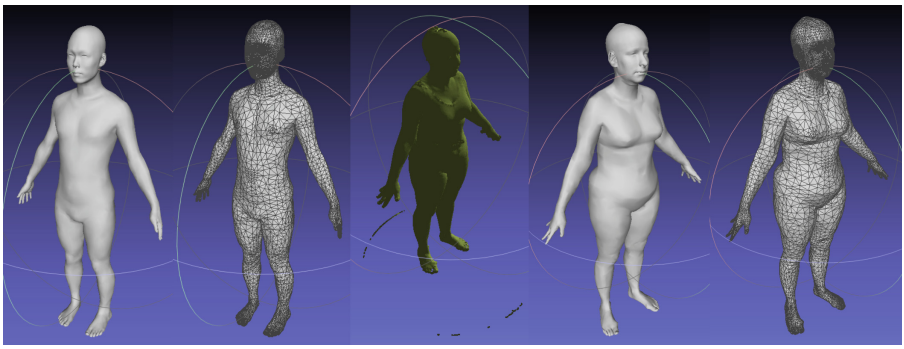


Fig. 4. Template mesh (left) fitted to (right) body scan point cloud (middle).

of missing anthropometric data, was that the number of predicted variables increased to 219 (73 × 3) for landmarks and 39429 (13143 × 3) for mesh vertices. However, the input variables will always be the same for these predictions and the final regression coefficients together with descriptive statistics, mean and standard deviation values, for both independent (predictive) and dependent (predicted) variables can saved and reused for each prediction. This makes the prediction process faster since the regression coefficients does not need to be recalculated each time.

2.4 Packaging and Publication of Statistical Body Shape Prediction Model

Both the prediction model for missing anthropometric data and the statistical body shape predictions consisting of MATLAB scripts were prepared and packaged to be shared and used by other researchers and organisations. Only the necessary and anonymised descriptive statistics and regression coefficients were included in the shared model. All necessary MATLAB scripts and additional files were included and published in a repository on GitHub [15] under the MIT License, which allows modification, distribution, and private use with attribution.

3 Results

The published repository is called *Statistical body shape predictions* and is publicly available on GitHub [16]. The first version of the repository contains license and readme files in addition to the main *Matlab* folder (Fig. 5). The idea is that other implementations through other programming languages could be added in future versions. In the *Matlab* folder a main MATLAB script mainInput.m can be found. Six additional scripts are placed in a folder and called by the script in *mainInput.m* where user input can be defined (Fig. 6). All descriptive statistics and regression coeffiecients are placed in the folder statBodyModel and called by the prediction scripts, *anthroRegression.m*, *landmarkPrediction.m*, and *bodyShapePrediction.m*. An additional output folder is also included as it is required for the scripts to produce results. The final results of predictions are in the form of landmark coordinates in XYZ point cloud format as well as a body shape mesh model in OBJ format. Both file formats can be visualised together, for example using Meshlab [17] (Fig. 7).

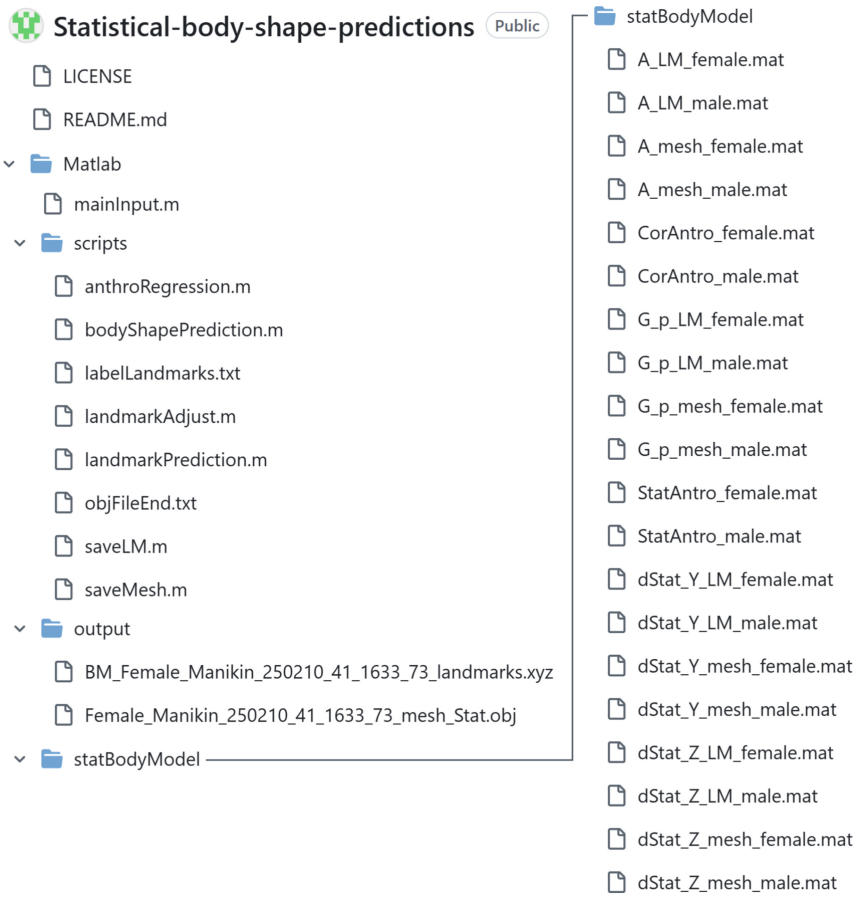


Fig. 5. Folder structure of the repository statistical body shape prediction model.

```
sex = 1; % Female = 1, Male = 0
% Predictive anthropometric variables, possible to change. See measurement list
below.
pred_id = [3 4 5]; % Age = 3, Weight = 4, Stature = 5

age = 41; % (year) Ex. 41
stature = 1633; % (mm) Ex. 1633
weight = 73; % (kg) Ex. 73

Z_test = [age weight stature]; % If predictive variables are changed, add them to
this Z_test variable.

addpath('./scripts/');
anthroMeasurements = anthroRegression(sex, pred_id, Z_test);
landmarkCoordinates = landmarkPrediction(sex, anthroMeasurements);
bodyShapePrediction(sex, anthroMeasurements, landmarkCoordinates);
```

Fig. 6. MATLAB code for user input in the file mainInput.m.

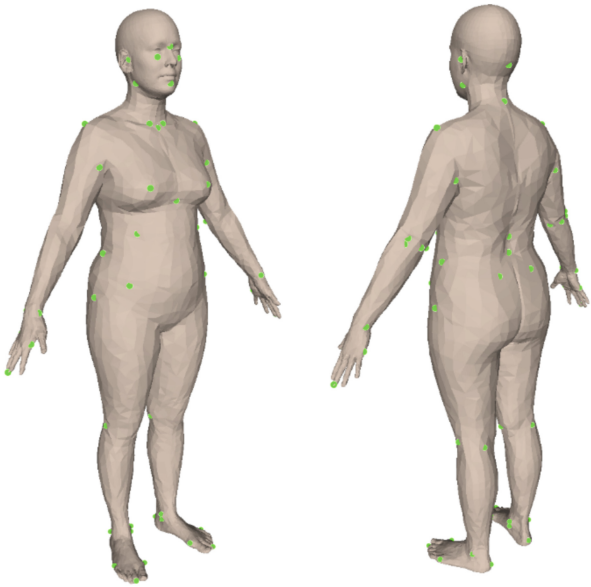


Fig. 7. Visualization of resulting body shape mesh together with predicted landmarks.

4 Discussion and Conclusions

The method and results presented in this paper demonstrates how statistical body shape prediction models can be produced from 3D body scan data but also shared through an online repository. By sharing the developed prediction model to be used by other researchers, software developers or organisations, the use of it might increase and enable further development of the model. The first version of the model could for example be extended by adding prediction models for other populations. The descriptive statistics and regression coefficients found in the folder *statBodyModel* are in this first version based on data from the CAESAR [6] study but could be changed to files based on another population with little to no changes in the MATLAB scripts. Since the data, in the form of descriptive statistics and regression coefficients, are separated from the methodology and process, which can be found in the script code, it can be argued that it would be possible to both share and standardize statistical body shape model data to a higher extent than to what is currently done. Future additions to the model could include additional anthropometric variables. The CAESAR study is, for example, not fully compliant with the standard document ISO 7250-1:2017 [10], where other data sets could have better coverage. Since the model is shared under an open license, the idea and intention are that it can be further developed by other researchers and organizations. This first version only generates a static body shape whereas future versions could include joint center prediction models and deformation patterns in relation to joint angles, to enable accurate body deformation during different postures and motions.

The current version of the statistical body shape predictions utilizes multivariate linear regression without considering the possibility of skewness in each variable. Instead, the regression model assumes that anthropometric measurements, and in this case also 3D points on the body, can be approximated with a normal distribution. This holds true in most cases when studying anthropometric measurements [18] but is not true for circumference and shape variables. However, the conditional regression model and principal component regression predict the expected value of the dependent variables with the smallest mean square error even if the normality assumption is not valid [19]. Still, additional functionality that handle skewness through transformation of data [20] is relevant to investigate in further research and development.

The accuracy of the statistically predicted meshes was in a previous study indicated to be relatively good even though differences could be seen [12]. These differences were mostly related to postural differences as the statistically predicted meshes have an average posture based on all included body scans and are centred and symmetrical on both sides of the mid-sagittal plane. Differences could also be seen regarding smaller areas with distinct shapes, e.g. muscles in the torso, which is difficult to recreate through statistical regression that predicts the most likely (average) body shape based on the input data. Still, methods for evaluating the accuracy of generated body shape models need to be further developed and standardised. To be compliant with standard documents, i.e. *ISO 15536-2:2007* [5] dimensions on the generated body shape models needs to be similar to the measurements of the real person that are used as input values. However, the question what should be considered as an acceptably accuracy of a 3D shape is difficult to answer [8], as standard documents only provide guidance of how the accuracy should be calculated and documented based on one-dimensional anthropometric measurements [5]. It could be argued that such accuracy analysis should also include distance analysis of three-dimensional data of body scan point clouds compared to the statistically predicted body shapes, which indicates that the standard documents might need to be updated and improved in some aspects. This need for standardisation of more complex 3D data could also be extended to other complex data sets such as motion data, generated from motion capture systems, 4D body scanners or simulation tools. However, if goals on the accuracy of predictions should be set, then both the given background information of a certain prediction (personal and environmental information) as well as possible additional diversity (physical, cognitive and behavioural) needs to be taken into consideration. If a software system has more background information on a certain situation, both regarding the person to be simulated and the environment and task, a better prediction of the body shape and motion can be achieved. Still, a motion or task could be achieved in infinite ways and that diversity needs to acknowledged, and if possible quantified, when the predicted shape and motion is assessed. The errors of a predicted body shape, posture or motion cannot be smaller than the human diversity that exist, and which cannot be explained by the given predictive variables. Further development of prediction models for body shapes, posture and motions as well as methods for assessing the accuracy of these predictions is believed to enable more realistic digital human models in simulations of product and workplace ergonomics.

Acknowledgments. This work has been made possible with support from the Knowledge Foundation and the associated INFINIT research environment at the University of Skövde (projects: Synergy Virtual Ergonomics and ADOPTIVE), and with support from Vinnova in the VIVA project, and SAFER—Vehicle and Traffic Safety Centre at Chalmers, Sweden, and by the participating organizations. This support is gratefully acknowledged.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. IEA homepage, What Is Ergonomics (HFE)? https://iea.cc/about/what-is-ergonomics/. Accessed 13 Feb 2025
2. Duffy, V.G. (ed.): Handbook of Digital Human Modeling. CRC Press, Boca Raton, FL (2009)
3. Scataglini, S., Paul, G. (eds.): DHM and Posturography. Academic, New York (2019)
4. ISO 15536-1: 2005 Ergonomics—Computer Manikins and Body Templates—Part 1: General Requirements. International Organization for Standardization (ISO), Geneva, (2005)
5. ISO 15536-2: 2007 Ergonomics—Computer Manikins and Body Templates—Part 2: Verification of Functions and Validation of Dimensions for Computer Manikin Systems. International Organization for Standardization (ISO), Geneva (2007)
6. Robinette, K.M., Blackwell, S., Daanen, H., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., Burnsides, D.: Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report. Air Force Research Laboratory, Wright-Patterson AFB, OH, and Society of Automotive Engineers International, Warrendale, PA (2002)
7. Allen, B., Curless, B., Popovic, Z.: The space of human body shapes: reconstruction and parameterization from range scans. In: Proceedings of the 2003 International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). San Diego, CA (2003)
8. Reed, M.P., Raschke, U., Tirumali, R., Parkinson, M.B.: Developing and implementing parametric human body shape models in ergonomics software. In: Proceedings of the 3rd International Digital Human Modeling Symposium, Tokyo (2014)
9. Brolin, E., Högberg, D., Hanson, L., Örtengren, R.: Adaptive regression model for prediction of anthropometric data. Int. J. Hum. Factors Model. Simul. **5**(4), 285–305 (2017)
10. ISO 7250-1: 2017 Basic Human Body Measurements for Technological Design—Part 1: Body measurement definitions and landmarks. International Organization for Standardization (ISO), Geneva (2007)
11. Jolliffe, I.: Principal Component Analysis. Springer, New York (2002)
12. Brolin, E., Delfs, N., Rebas, M., Karlsson, T., Hanson, L., Högberg, D.: Development of body shape data based digital human models for ergonomics simulations. In: Proceedings of the 7th International Digital Human Modeling Symposium. University of Iowa, Iowa City, IA (2022)
13. MATLAB: MATLAB Version: 9.13.0 (R2022b). The MathWorks, Natick, MA (2022)
14. Yamazaki, S., Kouchi, M., Mochimaru, M.: Markerless landmark localization on body shape scans by non-rigid model fitting. In: Proceedings of the 2nd Digital Human Modeling Symposium. Ann Arbor, USA (2013)
15. GitHub Homepage. https://github.com/. Accessed 13 Feb 2025
16. Statistical-body-shape-predictions. https://github.com/drBrolin/Statistical-body-shape-predictions. Accessed 13 Feb 2025
17. Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G.: Meshlab: an open-source mesh processing tool. In: Eurographics Italian Chapter Conference, vol. 2008, pp. 129–136 (2008)

18. Pheasant, S., Haslegrave, C.M.: Boyspace—Anthropometry, Ergonomics and the Design of Work. CRC Press, Boca Raton, FL (2006)
19. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs, NJ (1992)
20. Brolin, E., Högberg, D., Hanson, L.: Skewed boundary confidence ellipses for anthropometric data. In: Hanson, L., et al. (eds.) DHM2020—Proceedings of the 6th International Digital Human Modeling Symposium. Advances in Transdisciplinary Engineering, vol. 11. IOS Press, Amsterdam (2020)



Enhancing Virtual Restorative Environment with Generative AI: Personalized Immersive Stress-Relief Experiences

Kuei-Yuan Chang(✉) and Chang-Franw Lee

Graduate School of Design, National Yunlin University of Science and Technology, Yunlin, Taiwan
ken.chang.1011@gmail.com, leecf@yuntech.edu.tw

Abstract. In today's fast-paced world, stress and mental health challenges are becoming more common. Restorative environments help people relax and recover emotionally, and Virtual Reality Exposure Therapy (VRET) offers a way to experience these benefits beyond physical limitations. However, most VRET applications rely on pre-designed content, limiting their adaptability to individual needs. This study explores how Generative AI can enhance VRET by creating personalized, immersive environments that better match users' preferences and improve relaxation.

To evaluate the impact of AI-generated restorative environments, we combined EEG measurements with user interviews. Thirty university students participated in the study, experiencing two different modes: static mode and walking mode. The EEG results showed an increase in Theta (θ) and High Beta (β) brain waves, suggesting a state of deep immersion accompanied by heightened cognitive engagement and mental effort. While participants found the experience enjoyable and engaging, the AI-generated environments tended to create excitement and focus rather than conventional relaxation. These findings suggest that for AI-generated environments in VRET to be more effective for stress relief, future designs should reduce cognitive load while maintaining immersion. This study provides insights into how AI can enhance relaxation experiences and introduces a new perspective on personalized digital stress-relief solutions.

Keywords: Virtual Reality Exposure Therapy (VRET) · Artificial Intelligence Generated Content (AIGC) · Generative AI · Electroencephalography (EEG)

1 Introduction

1.1 The Growing Need for Mental Health Solutions

Modern life is stressful, and mental health challenges are becoming more common. The fast-paced urban environment often leads to accumulated stress, which can negatively impact emotional well-being. Being constantly exposed to high-pressure settings can cause exhaustion, making it harder to manage emotions and, over time, significantly affecting both mental and physical health [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 132–144, 2025.
https://doi.org/10.1007/978-3-031-93502-2_9

As a result, mental health has become a key concern for individuals and organizations alike. Many workplaces are now adopting different relaxation strategies, such as short breaks, mindfulness exercises, immersive environmental changes, or even quick gaming sessions, to help employees manage stress and stay productive. Studies have shown that these approaches can effectively reduce emotional strain and improve efficiency [2–4].

1.2 Restorative Environment Theory

Back in the late twentieth century, researchers [5–7] discovered that natural environments have a strong positive impact on emotional regulation and stress relief. This realization led to the development of the Restorative Environment Theory, which highlights the functional and interactive relationship between people and their surroundings (Fig. 1) The core idea is that well-designed environments can effectively reduce stress, ease negative emotions, and help people recover from mental fatigue.

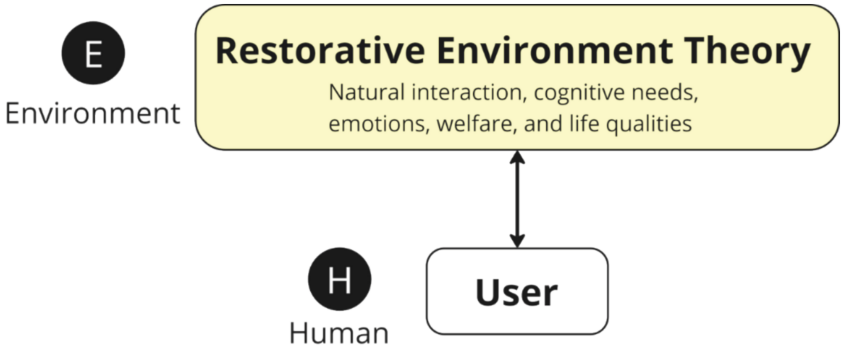


Fig. 1. The Interaction Between People and Restorative Environments [21].

A truly restorative environment provides a space where users can relax, boost their mood, and engage in self-healing. Researchers have identified four key factors that make an environment effective for mental recovery [1]:

- **The Principle of Distance:** This refers to the idea that stepping away from a familiar, stressful environment can help ease cognitive overload. By shifting attention to something new and engaging, people can experience relief from constant mental strain. This is one reason why immersive experiences, like interactive digital spaces, have become popular stress-relief tools.
- **The Principle of Richness:** A rich environment offers diverse and interconnected elements that create a sense of exploration and engagement. When an environment is designed to be immersive and cohesive, it provides enough stimulation for the mind to stay engaged while also allowing room for relaxation.
- **The Principle of Attractiveness**: Fascination occurs when an environment naturally captures attention without requiring conscious effort. This can range from highly engaging activities like sports or entertainment (strong fascination) to more passive, contemplative experiences like admiring a scenic landscape (soft fascination). Both types play a role in mental restoration.

- **The Principle of Compatibility**: A restorative environment should align with an individual's personal preferences and needs. The best restorative spaces are those that feel natural and inviting to the user, promoting relaxation and a sense of ease. Research suggests that people tend to recover best in environments that reflect their tastes and provide a sense of security [1].

1.3 Virtual Reality Exposure Therapy (VRET)

Although the benefits of restorative environments are widely acknowledged, many individuals who need stress relief the most—such as those with mobility limitations, patients requiring constant medical care, high-pressure professionals, and urban workers—may struggle to access them. To bridge this gap, Virtual Reality Exposure Therapy (VRET) has emerged as a promising solution (Fig. 2). By leveraging immersive technologies like virtual reality (VR) and spatial projection, VRET enables users to experience a therapeutic environment without the constraints of physical space and time.

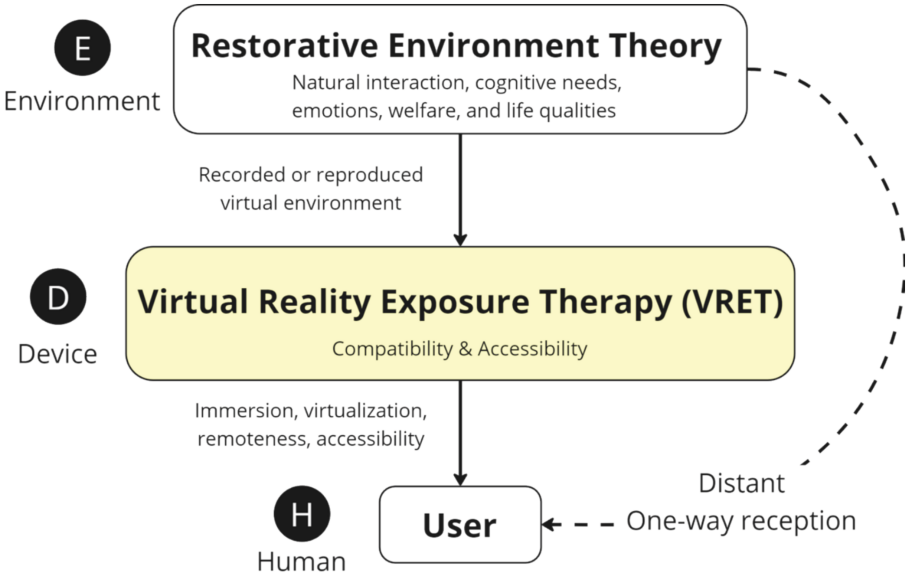


Fig. 2. Experiencing Restorative Environments Through Technological Devices [21].

VRET has been successfully applied to treat various psychological conditions, including phobias and post-traumatic stress disorder [8]. One of its biggest advantages is that therapists can control the environment, tailoring the experience to meet individual needs while ensuring accessibility within clinical settings.

Recent studies have explored the effectiveness of VRET in various settings, demonstrating its potential for stress relief and emotional recovery. A Swiss research team [9] developed a VR-based travel experience aimed at providing elderly dementia patients with safe and immersive journeys, allowing them to experience a sense of adventure within a controlled environment. Similarly, Liu and their research team [3] utilized VR

simulations of natural landscapes to assist professionals working in isolated environments, helping them recover from stress and maintain mental well-being. Misaki and their research team [10] investigated how immersive projection displays influence stress recovery, analyzing the psychological effects of different display conditions. Meanwhile, Zhang, Guan [3] and Shamri Zeevi [11] explored how integrating immersive technology into traditional art exhibitions and art therapy can enhance emotional healing, offering deeper therapeutic engagement for users. These findings illustrate how VRET can effectively extend the concept of restorative environments beyond physical constraints, providing accessible, adaptable, and engaging experiences that support mental health.

1.4 Human-AI Collaboration and Its Potential in Mental Health

While VRET has made significant strides in creating immersive restorative experiences, current implementations rely heavily on pre-recorded content and manually constructed virtual scenes. These limitations reduce adaptability and personalization, making it difficult to fully address individual emotional and cognitive needs. To enhance VRET's effectiveness, Generative AI has emerged as a potential game-changer by enabling the dynamic creation of personalized environments tailored to each user's preferences (Fig. 3).

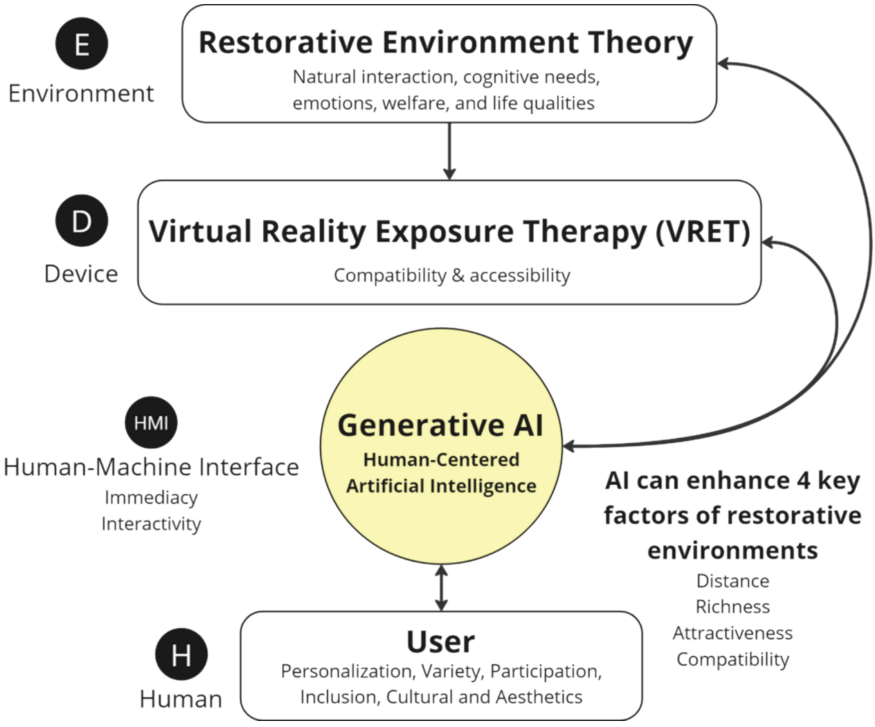


Fig. 3. Potential Opportunities of Generative AI in VRET [21].

Generative AI, including tools like ChatGPT for conversational interactions and image-generation models like Midjourney and Stable Diffusion, can create customized digital content in real-time. By leveraging user input and adaptive algorithms, AI-generated environments can provide a more interactive and personalized therapeutic experience. Research suggests that interactive technologies and AI-driven content creation fosters greater engagement, emotional resonance, and immersion, making it a valuable addition to mental health interventions [12–15].

Unlike traditional VR environments, which require extensive manual design, AI-powered restorative environments can be dynamically adjusted to fit an individual's mood, preferences, and stress levels. This human-AI collaboration shifts the role of technology from a passive tool to an active co-creator of meaningful, adaptive, and emotionally responsive experiences. In the context of mental health, AI can enhance emotional support, improve accessibility, and expand the scope of VRET beyond preset scenarios. The next stage in VRET evolution (Fig. 3) involves exploring how AI-generated restorative environments can enhance personalization, user engagement, and emotional regulation.

1.5 Assessing Emotional Responses in Virtual Environments

To get a more complete picture of how virtual restorative environments help relieve stress and affect emotions, this study uses a non-invasive EEG device, Emotiv EPOC X. Since surveys and interviews might not fully capture participants' emotional responses, EEG measurements provide additional insights. EEG devices are widely used in research on emotions and cognitive states, helping scientists track stress, anxiety, focus, and fatigue. They play a key role in psychology, human-computer interaction, and workplace health studies.

Research in cognitive neuroscience shows that the frontal lobe is responsible for planning, problem-solving, organization, and emotional regulation [16, 17]. Because of its role in higher-order thinking, this study focuses on EEG signals from this area. We analyze four important brainwave types—Theta (θ), Alpha (α), Low Beta (Low β), and High Beta (High β)—each associated with different states of relaxation and focus [18–20]:

- **Theta waves (4–8 Hz):** These slow, high-amplitude waves show up when people are deeply focused, relaxed, or retrieving memories. They're commonly seen during meditation, deep relaxation, or subconscious thought processes.
- **Alpha waves (8–13 Hz):** These fast, low-amplitude waves indicate a calm but alert state. They're associated with an ideal mental state for learning and thinking.
- **Beta waves (13–25 Hz):** These fast waves show up when the brain is actively thinking. More beta waves mean the brain is ready to react to the outside world. High Beta (16–25 Hz) also called stress-related beta waves, these are linked to high mental effort and emotional processing. Low Beta (13–16 Hz) reflects a relaxed but focused state, where the mind is engaged but not overwhelmed.

2 Research Methods

To explore the design considerations of using generative AI for creating virtual restorative environments, our research team previously conducted a small-scale pilot study [21]. As technological conditions have advanced, we have refined various research details and integrated new tools to further investigate the potential of generative AI in Virtual Restorative Environment Therapy (VRET) with greater depth.

2.1 Participant Selection and Study Scope

This study recruited participants primarily from university students who experience more academic stress. Prior to the experiment, participants completed an online consent form and a State-Trait Anxiety Inventory (STAI-T) questionnaire to assess their baseline anxiety levels. Those who met the study's inclusion criteria and provided informed consent were invited to participate. In total, 30 individuals (mean age: 24.6, SD: 1.91; 13 males, 17 females) fully completed the study.

2.2 Experimental Design and Procedure

As described earlier, participants completed the Trait Anxiety subscale (STAI-T) of the State-Trait Anxiety Inventory (STAI) via an online recruitment form before the experiment. This initial assessment was used to evaluate their recent stress levels. At the scheduled time, participants arrived at the designated experimental site to proceed with the second phase of the study.

The experiment was conducted in a 7 m × 5 m indoor space, lasting approximately 1 h. The process is illustrated in Fig. 4. Data was collected using computer recordings, paper-based questionnaires, and semi-structured interviews.

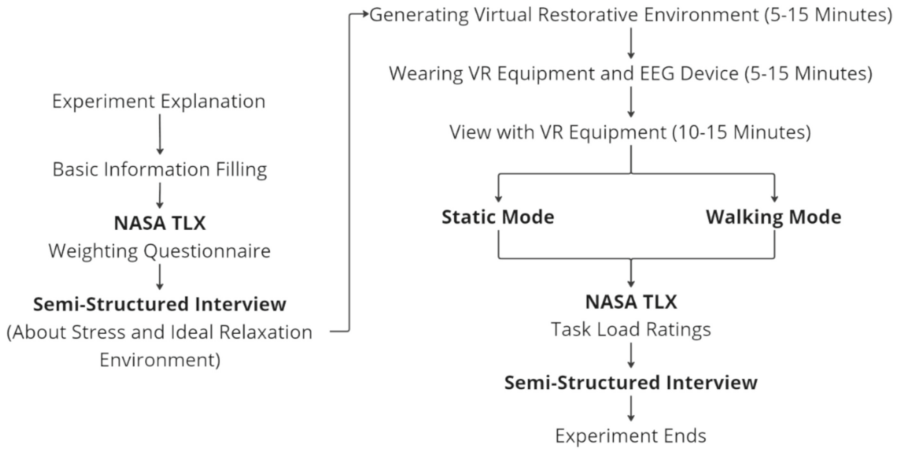


Fig. 4. Experimental Procedure.

Upon arrival at the testing site, participants were briefed on the experimental process and asked to complete a basic information form. They then filled out the NASA-TLX (Task Load Index) questionnaire, assigning weightings based on their expectations regarding the use of generative AI for creating virtual restorative environments. This was followed by a discussion on their daily stress-relief habits and their ideal restorative environment.

Participants were then guided to use Skybox AI, a generative AI tool, to generate virtual environment images based on text-based prompts (Figs. 5 and 6). Since many participants were not native English speakers, Google Translate was used to assist in formulating accurate AI prompts.



Fig. 5. Virtual Recovery Environment.



Fig. 6. Virtual Recovery Environment.

Once the ideal images were generated, they were downloaded and transferred to a VR device. The final immersive experience was conducted using Moon VR Home (now rebranded as Moon Portal). The visual content was simultaneously streamed onto a computer screen for real-time monitoring, allowing experimenters to observe participants' reactions and maintain communication throughout the experience. The complete VR experience setup is illustrated in Fig. 7.

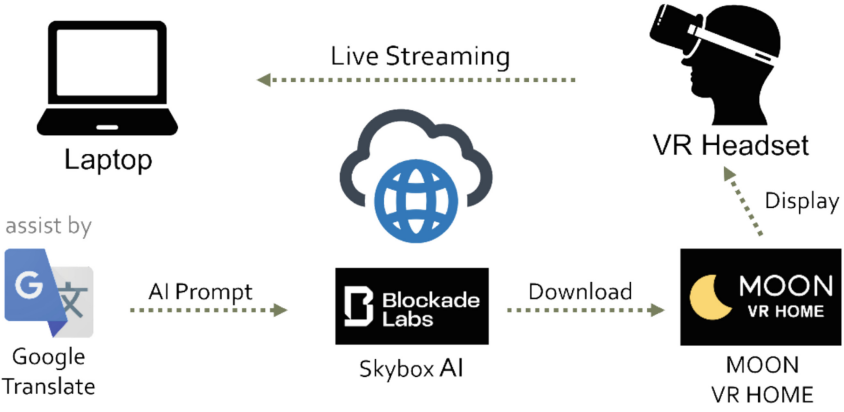


Fig. 7. Process of Generative AI and VR Experience Device Operation.

The experiment was carried out in two different VR experience modes (Fig. 8):

- **Walking Mode:** Participants could view the real-world environment and enter their custom-generated virtual environment through a virtual portal (Fig. 9), similar to Mixed Reality (MR) interactions.
- **Static Mode:** Participants remained seated in a swivel chair, experiencing a 360-degree VR display, similar to conventional VR experience.

To ensure accurate assessment of emotional responses beyond verbal feedback, this study utilized a non-invasive EEG device, Emotiv EPOC X, equipped with 14 electrodes to capture key brain signals. This device, widely used in cognitive psychology, user experience research, and human-computer interaction studies, offers a well-established and minimally intrusive method for EEG data collection (Fig. 10).



Fig. 8. Experimental site.



Fig. 9. Virtual perspective.



Fig. 10. Wearing the EEG Device EPOC X [22].

Since participants needed to wear both a VR headset (Meta Quest 3) and the EEG device (EPOC X) simultaneously, modifications were made to the VR headset for improved usability. The research team removed the default head strap of the Meta Quest 3 and used 3D-printed side handles, allowing participants to hold the headset while ensuring a stable fit for the EEG device (Figs. 11 and 12).

The EEG recording process was divided into four segments:

1. **Baseline Calibration (1 min):** Participants sat still while the EEG device automatically calibrated to establish baseline brain activity.
2. **Resting State Measurement (1 min):** Participants remained seated in a relaxed position, freely observing their surroundings.
3. **VR Static Mode (5 min):** Participants experienced their AI-generated restorative environment while seated.



Fig. 11. Wearing method (front view).



Fig. 12. Wearing method (side view).

4. **VR Walking Mode (5 min):** Participants actively explored their virtual restorative environment.

At the end of the experiment, participants completed a semi-structured interview and the NASA-TLX task load index, providing insights into stress recovery and user experience. The data collected was analyzed to determine the effectiveness of generative AI in virtual restorative environments, along with key design strategies for future implementations.

2.3 Ethical Considerations and Risk Assessment

Before conducting this study, ethical approval was obtained from the Human Research Ethics Committee at National Cheng Kung University (NCKU HREC). The study was reviewed to ensure compliance with ethical research standards and participant safety.

The approved case number for this study is NCKU HREC-E-112-789-2.

3 Research Results and Conclusions

3.1 Research Results

A total of 30 participants (13 males and 17 females) successfully completed the experiment. Based on the Trait Anxiety subscale (STAI-T) results, most participants exhibited moderate levels of anxiety, as shown in Fig. 13.

The NASA-TLX task load index results were divided into weighting calculations (Fig. 14) and actual perceived task load Ratings (Fig. 15).

The weighting scores (Fig. 14) represent participants' self-assessed priorities before engaging with AI-generated restorative environments. The results indicate that most participants were primarily concerned with Mental Demand and Performance, which reflect the cognitive load and the quality of the generated environment.

The actual perceived task load was recorded post-experiment, with participants rating their task load based on their experience. These scores were then multiplied by their pre-experiment weightings, yielding a maximum possible score of 100 per category (Fig. 15).

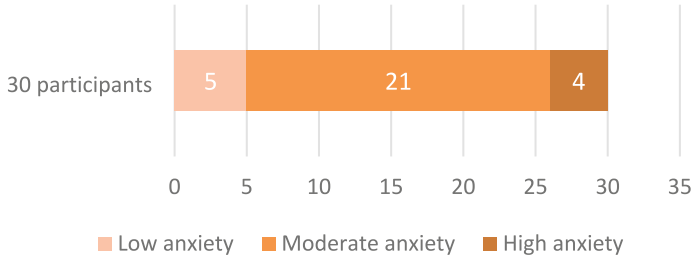


Fig. 13. Trait Anxiety (STAI-T) results.

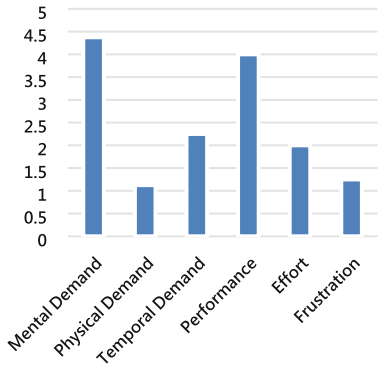


Fig. 14. Weighting Scores (Sum = 15).

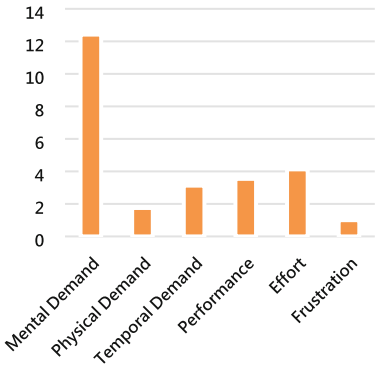


Fig. 15. TLX Perception Ratings (Max = 100).

The data suggest that overall workload during the AI-driven virtual restorative experience was minimal, with Mental Demand being the most prominent factor, significantly higher than other workload dimensions.

This study collected and processed EEG data from participants, removing three low-quality samples before conducting a frequency domain analysis on F3 and F4 electrode points in the frontal lobe (Group Band-Power Report).

Baseline (0 dB) was established through 1 min of static calibration before measuring three experimental conditions down below:

– Resting state (1 min)
– VR Seated Mode (5 min)
– VR Walking Mode (5 min)

The EEG data reveal a consistent increase in High Beta (β) and Theta (θ) waves across the three conditions, with the lowest levels in the resting state and the highest levels in the VR walking mode.

High Beta (16–25 Hz): Known as stress-related beta waves, these are associated with cognitive load and emotional processing. The increase suggests heightened mental engagement during the VR experience. Theta (4–8 Hz): Often linked to deep focus,

relaxation, and memory retrieval, the rise in theta waves indicates an immersive and mentally engaging experience.

This pattern suggests that while engaging with the AI-generated virtual restorative environments, participants experienced both heightened cognitive activity and deep immersion at the same time.

In addition to quantitative data, the study collected qualitative feedback through semi-structured interviews.

Most participants described the experience as novel and enjoyable, expressing excitement and curiosity about the AI-generated virtual environments. Many perceived the experience as more recreational than traditionally relaxing, with common descriptors such as "exciting and fun" rather than "calm and meditative". While the experience did not align with conventional relaxation methods, immersion in a personally curated virtual environment effectively helped distract participants from stress and worries.

When comparing the walking mode and the static mode, most participants felt that the static mode, where they sat and observed, was closer to their initial expectation of relaxation. The sense of security from sitting helped reduce tension and fear. However, the walking mode was considered more engaging and exciting, making it the most favored experience overall.

3.2 Conclusions and Recommendations

This study explores the application of generative AI in Virtual Restorative Environment Therapy (VRET) and compares two experience modes in the experiment. Based on participant feedback, we found that using generative AI to create personalized virtual restorative environments was widely recognized as having a relaxing effect. However, the experience did not induce the same sense of calmness typically associated with conventional meditation or relaxation therapies.

A review of previous literature suggests that the integration of generative AI better aligns with the four key factors of restorative environments identified by Zhang and Guan [1]: The Principles of Distance, Richness, Attractiveness, and Compatibility. Compared to existing VRET content, participants generally found AI-generated personalized experiences to be more engaging and immersive, although they also reported an increase in cognitive load. Moving forward, the greatest challenge for this promising application will be reducing the psychological and cognitive load experienced by users, such as optimizing AI communication processes and minimizing disruptive factors during immersion.

Acknowledgments. This study was funded by the National Science and Technology Council (NSTC), Taiwan. (Grant number 113-2410-H-224-008).

References

1. Zhang, R., Guan, X.: Research on the spiritual healing function of immersive art exhibition. In: Rauterberg, M. (ed.) Culture and Computing. HCII 2023. Lecture Notes in Computer Science, vol. 14035, pp. 206–214. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34732-0_15

2. Espinoza, A., Larraín, J., Gutierrez, F.J.: Biofeedback-controlled video games for emotional regulation. In: Fang, X. (ed.) HCI in Games. HCII 2023. Lecture Notes in Computer Science, vol. 14047, pp. 186–197. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35979-8_15
3. Liu, Y., Jiang, A., Dong, J.: Recovery effect of different virtual natural environments on stress in short-term isolation tasks. In: Chen, J.Y.C., Fragomeni, G. (eds.) Virtual, Augmented and Mixed Reality. HCII 2023. Lecture Notes in Computer Science, vol. 14027, pp. 634–648. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35634-6_46
4. Yu, C.-P., Lee, H.-Y., Luo, X.-Y.: The effect of virtual reality forest and urban environments on physiological and psychological responses. Urban For. Urban Green. **35**, 106–114 (2018). https://doi.org/10.1016/j.ufug.2018.08.013
5. Kaplan, S., Talbot, J.F.: Psychological benefits of a wilderness experience. In: Altman, I., Wohlwil, J.F. (eds.) Behavior and the Natural Environment, pp. 163–203. Plenum, New York (1983). https://doi.org/10.1007/978-1-4613-3539-9_6
6. Ulrich, R.S., Simons, R.F., Losito, B.D., Fiorito, E., Miles, M.A., Zelson, M.: Stress recovery during exposure to natural and urban environments. J. Environ. Psychol. **11**(3), 201–230 (1991). https://doi.org/10.1016/S0272-4944(05)80184-7
7. Kaplan, S.: The restorative benefits of nature: toward an integrative framework. J. Environ. Psychol. **15**(3), 169–182 (1995). https://doi.org/10.1016/0272-4944(95)90001-2
8. Rothbaum, B.O., Rizzo, A.S., McDaniel, D.D., Zanov, M.V.: Virtual reality exposure therapy. In: Friedman, H.S. (ed.) Encyclopedia of Mental Health, 2nd edn, pp. 370–374. Academic Press, New York (2016). https://doi.org/10.1016/B978-0-12-397045-9.00196-8
9. Jaquenod, S., Ducraux, D., Mugellini, E., Angelini, L.: Development of the train therapy in a nursing home in Switzerland. In: Gao, Q., Zhou, J. (eds.) Human Aspects of IT for the Aged Population. HCII 2023. Lecture Notes in Computer Science, vol. 14042, pp. 552–563. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34866-2_39
10. Misaki, T., Feng, C., Kawaguchi, A., Sugaya, M.: Measurement and evaluation of the healing effect of nature-themed video projection. In: Kurosu, M., Hashizume, A. (eds.) Human-Computer Interaction. HCII 2023. Lecture Notes in Computer Science, vol. 14012, pp. 264–278. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35599-8_17
11. Shamri Zeevi, L.: Making art therapy virtual: integrating virtual reality into art therapy with adolescents. Front. Psychol. **12**, 584943 (2021). https://doi.org/10.3389/fpsyg.2021.584943
12. Garibay, O.O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S.M., Garibay, I., Grieman, K., Havens, J.C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., McGregor, S., Salvendy, G., Shneiderman, B., Stephanidis, C., Strobel, C., Ten Holter, C., Xu, W.: Six human-centered artificial intelligence grand challenges. Int. J. Hum. Comput. Interact. **39**(3), 391–437 (2023). https://doi.org/10.1080/10447318.2022.2153320
13. Shneiderman, B.: Human-Centered AI. Oxford University Press, Oxford (2022)
14. Riva, G., Baños, R.M., Botella, C., Wiederhold, B.K., Gaggioli, A.: Positive technology: using interactive technologies to promote positive functioning. Cyberpsychol. Behav. Soc. Netw. **15**(2), 69–77 (2012). https://doi.org/10.1089/cyber.2011.0139
15. Calvo, R.A., Peters, D.: Positive Computing. MIT Press, Cambridge, MA (2014)
16. Davidson, R.J.: What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. Biol. Psychol. **67**(1–2), 219–234 (2004). https://doi.org/10.1016/j.biopsycho.2004.03.008
17. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. **24**, 167–202 (2001). https://doi.org/10.1146/annurev.neuro.24.1.167
18. Teplan, M.: Fundamentals of EEG measurement. Meas. Sci. Rev. **2**(2), 1–11 (2002) https://www.measurement.sk/2002/S2/Teplan.pdf

19. Başar, E., Başar-Eroğlu, C., Karakaş, S., Schürmann, M.: Gamma, alpha, delta, and theta oscillations govern cognitive processes. Int. J. Psychophysiol. **39**(2), 241–248 (2001). https://doi.org/10.1016/S0167-8760(00)00145-8
20. Kobe, M., Honma, I.: *図解脳波テキスト: 脳波の系統的読解のための新しい懇切なアプローチ*[Illustrated EEG Text: A New Systematic Approach for Understanding EEGs]. Bunkodo, Tokyo (1989)
21. Chang, K., Lee, C.: Designing Virtual Restorative Environment with Generative AI: Hardware and Design Considerations. In: The Asian Conference on Arts & Humanities 2024 Official Conference Proceedings, pp. 275–285 (2024). https://doi.org/10.22492/issn.2186-229X.2024.26
22. Emotiv: EPOC X User Manual: Regulatory Compliance (2024). https://emotiv.gitbook.io/epoc-x-user-manual/regulatory-compliance. Accessed 20 Jan 2025

The Visibility and Non-intrusiveness of Peripheral Vision Information During the Use of AR Devices: The Effects of Transparency Levels, Color, and Font Size

Ding Chen(✉), Kaige Wang, and Jinyun Cao

Beijing Institute of Technology, Beijing, China
cd040058ncepu@163.com

Abstract. The application of Augmented Reality (AR) devices is becoming increasingly widespread across various fields, including healthcare, rescue operations, military, and gaming. As a key component of AR, information display has been primarily studied in the context of target attention centers; however, research on peripheral visual information encoding remains underexplored. This study focuses on the encoding of peripheral visual information in AR devices, examining the effects of font size, color, and transparency levels. A total of 20 participants were recruited for the experiment. The results indicate that font size had no significant impact on the visibility and non-intrusiveness of peripheral visual information in AR devices. This suggests that, in designing peripheral visual information for AR devices, adjustments to color and transparency may take precedence over font size. A transparency level of T0.75 provided the optimal balance between visibility and non-intrusiveness for peripheral visual information. Additionally, the study partially identifies the optimal threshold range for transparency in peripheral encoding (T0.5-T0.75). The research also demonstrates the applicability of RAL 9016 Traffic White and RAL 5017 Traffic Blue colors. These findings expand the range of applicable visual areas for information encoding in the field of AR technology.

Keywords: Augmented Reality (AR) · Peripheral Vision · Transparency levels · Color · Font Size

1 Introduction

The application of Augmented Reality (AR) devices is increasingly widespread across various fields, including healthcare, rescue operations, military, and gaming [1–3]. Information display, as a core component of AR, has always been a research focus in the field of AR devices. By designing effective information displays, AR devices can offer users an immersive experience. During the augmented reality process, the information displayed by AR devices is typically concentrated in the user's peripheral vision to avoid obstructing the central visual area, thereby enhancing the user experience. The uniqueness of AR devices lies in this characteristic: unlike in the real world or with VR devices,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 145–159, 2025.
https://doi.org/10.1007/978-3-031-93502-2_10

where users can adjust their view by turning their head or performing other actions to shift peripheral visual information to the central visual area (similar to a driver adjusting the view of side mirrors while turning their head), information in the peripheral visual area of AR devices remains relatively fixed for the user and moves with the user's movements.

The visual information channel, as one of the primary sources through which humans receive information, accounts for 80% of all human sensory inputs. However, the distribution of visual information perception within the human visual field is not uniform [4]. As the viewing angle increases, the perceptibility of visual information shows a gradual decrease from the center towards the periphery [5]. Therefore, ensuring the visibility of information displayed in the peripheral vision area of AR devices is crucial. AR display refers to any visual element within an AR device that can guide the user while still allowing them to see the real environment. As a core goal of augmented reality, AR display design should not overly distract the user or induce symptoms of motion sickness during device use, thus ensuring the safety and comfort of tasks in the user's central visual area. Therefore, the non-intrusiveness of AR display design is equally important.

2 Literature Review

Color, transparency levels, and font size are key elements in AR display design, and numerous studies have demonstrated their significant role in determining the overall quality and usability of AR interfaces [6–8]. L. Zhang et al. highlighted in their study that color matching in augmented reality displays is critical, as it directly influences users' judgment of information and their overall experience [9]. C. Xue et al. noted that color coding in the interface can effectively reduce visual information overload and is commonly used in complex human-computer interfaces, such as those in VR and AR, to enhance interface quality and cognitive performance [10]. Michele Gattullo and colleagues, in their study of AR smart interactions in the context of Industry 4.0, also emphasized the importance of color coding in AR information design, with the combination of blue and white playing a crucial role in the legibility of information [10]. APK Kwok et al. explored the impact of text color on readability in head-mounted displays, finding that low contrast between text and background significantly reduced text readability [8]. Similarly, M. Fiorentino's research on the impact of text style and color on readability in transparent head-mounted displays revealed that color had a significant effect on error rates, especially in industrial settings [11]. While existing studies on color in AR information are relatively comprehensive, most do not consider the impact of increasing viewing angles on information perception.

Research has shown that font size significantly affects various aspects of user perception in augmented reality (AR interfaces), including text readability, recognition rate, and user performance. Appropriate font sizes can enhance reading speed and information retrieval efficiency, especially in visually complex environments. H. Renkewitz and colleagues investigated the impact of text size on readability and user performance in AR systems. Their study found that larger text sizes improve recognition rates and reduce cognitive load, particularly in augmented reality settings [12]. J. Grobelny and others discussed how font size affects user interaction and understanding in virtual reality (VR)

environments. They found that when text size is either too small or too large, user performance declines [13]. Johannes Novotny and colleagues evaluated the effects of different font sizes on reading speed and readability in VR scenes. Their research indicated that larger fonts help to improve reading speed and text readability, particularly in visually complex 3D environments [14]. ML Bernard and colleagues found that font size and formatting significantly impact text readability. For different fonts (such as Times New Roman and Arial), larger fonts enhance readability, especially in varying display environments [15]. In a study focusing on older adults, Y. Chen and others discovered that in virtual reality, the recognition ability of Chinese characters among seniors is significantly influenced by font size, stroke thickness, and background. Larger fonts and simpler backgrounds improve recognition rates and reading experiences for older adults [16]. From previous studies, it is evident that research on the impact of font size on information display in AR devices is still relatively limited. Most research has focused on virtual reality or target-based tasks. Therefore, further studies are needed to explore the effects of font size and its influence on information perception within peripheral vision.

Transparency, as an essential component in AR devices, can significantly affect text readability. Excessive transparency may reduce text visibility, especially when semi-transparent elements interact with complex backgrounds, causing text information to become unclear. Leykin and Tuceryan (2004) demonstrated that the transparency settings of text and background directly impact text readability in augmented reality systems. High transparency can significantly hinder reading performance [17]. Kruijff et al. (2018) further pointed out that in wide-field-of-view augmented reality displays, the use of semi-transparent elements reduces text visibility, impacting user efficiency [18]. Harrison and Vicente (1996) experimentally evaluated the use of transparent menus and found that when transparency exceeds 75%, text readability and user performance decline significantly, leading to higher error rates [19]. Additionally, Jankowski et al. (2010) explored the effects of text rendering styles and transparency on readability, suggesting that appropriate transparency enhances text clarity. However, excessive transparency negatively impacts task performance [20].

Based on the above research, it has been demonstrated that visual information encoding, such as color combinations, font size, and transparency, plays an important role in information perception within AR devices. However, the critical characteristic of peripheral vision in the context of information display in AR devices has been less frequently studied. Most research on information encoding in AR devices focuses on the user's central vision or task design, with relatively few studies addressing peripheral vision. A. Lewandowska and colleagues investigated the encoding of peripheral vision information on computer VDT devices, which represent real-world environments [21]. However, the scenarios in this study are still quite different from those of AR devices. Their results indicate that blue color combinations and moderate brightness contrast enhance peripheral vision information perception, though further validation is needed regarding the usability of peripheral vision information in AR devices. Therefore, this study primarily focuses on exploring the impact of color, font size, and transparency levels on the visibility and non-intrusiveness of peripheral vision information in AR interfaces. Visibility will be assessed by measuring the completion time of peripheral vision tasks,

while non-intrusiveness will be evaluated using the SSQ (Motion Sickness Symptoms Questionnaire) score [22].

3 Research Method

3.1 Participant

A total of 20 participants, including 11 males and 9 females, with ages ranging from 21 to 30 years (M = 25.7, SD = 3.81), voluntarily participated in this study. All participants had normal vision or corrected-to-normal vision, and none had color blindness or color vision deficiency. Prior to the study, participants received a briefing on the research objectives, potential risks and benefits, as well as their rights. They then provided informed consent by signing a consent form. None of the participants had previously participated in similar experiments. Based on their performance, each participant received a monetary reward ranging from 30 to 50 RMB.

3.2 Experiment Preparation

The experiment was conducted in a laboratory at the School of Design and Art, Beijing Institute of Technology, under normal lighting conditions (~300 lx). Task presentation and participant data collection were carried out using a custom experimental program written in Processing. The processor used was a 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz, with an NVIDIA GeForce RTX 4060 24G graphics processor, and the operating system was Windows 11. For the AR experience, a Pico 4 device was used, which included a head-mounted display and two wireless handheld controllers. The Pico device was set to the highest brightness level, with a brightness of 320 cd/m^2. The observation equipment during the experiment was a Dell U2723QX monitor, featuring a 4 K resolution and 100% sRGB color gamut coverage.

3.3 Experimental Materials Design

The experimental prototype was developed by drawing on previous AR research to ensure rigor in the AR study and minimize uncontrollable factors introduced by experimental materials. The prototype consists of nine square virtual buttons arranged in a 3 × 3 grid, with "Start" and "Return" buttons based on a 3 × 4 keyboard layout. [7] The button size corresponds to a 3° 50′ field of view (FOV), and targeting and interaction are facilitated through an AR-linked handheld controller. The target text is displayed on a black (RAL 9017 Traffic Black) background, while RAL 7021 Black Grey is used for the non-target buttons to reduce visual interference from other non-polar colors. The experimental prototype was positioned in the peripheral vision area, located 30° outside the central field of view of the human eye, and was randomly presented during the task.

This experiment involves three independent variables: (a) color (4 groups), (b) Transparency levels (T0.25, T0.5, T0.75,T1: fully opaque), and (c) text size (8 pt./10 pt./12 pt) [15]. A total of 48 different experimental conditions were created (4 colors × 4 Transparency levels × 3 text sizes). The experiment followed a within-subjects design. The

color selection was based on the CIEL*a*b* color space, which, unlike the RGB and CMYK color models, is closer to human visual perception and therefore exhibits perceptual similarity. In this study, four colors were experimentally compared using the CIEL*a*b* color space and the RAL color palette: green (RAL 6024 Traffic Green), orange (RAL 2009 Traffic Orange), white (RAL 9016 Traffic White), and blue (RAL 5017 Traffic Blue). These colors were chosen based on those most commonly used in existing AR-HUDs on the market. The opacity variable was controlled using the alpha channel of the color, with four conditions considered (0.25, 0.5, 0.75, and 1). An alpha value of 0 corresponds to complete transparency, while a value of 1 represents a solid color (opaque) [6].

3.4 Experimental Procedure

Using the Eriksen [23] experimental paradigm, the design was optimized to suit the research context. First, participants wore the AR device, read the experiment instructions, and clicked "Start" to begin the experiment. As shown in Fig. 1, the experimental flow starts with a fixed 1000-millisecond interval. Participants then saw a target number (ranging from 1 to 9) displayed in the center of the screen for 1000 milliseconds, which they were required to memorize. After this, the target number always appeared in the peripheral region of the AR display, with distractors positioned either above or below the target, forming a 3 × 3 grid. The experiment presented different combinations of independent variables (3 text sizes, 4 colors, 4 Transparency levels, totaling 48 combinations) in a random order to avoid sequence effects. Participants used the handheld controller provided with the Pico device to select the target, and once a selection was made, the experiment automatically proceeded to the next task.

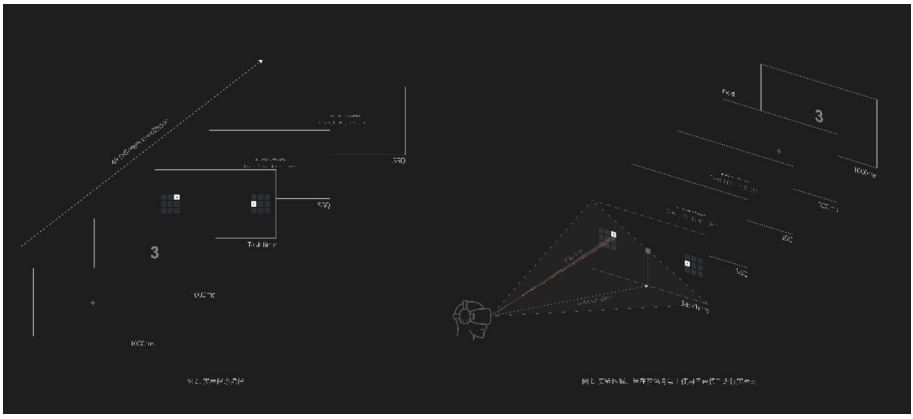


Fig. 1. Experimental Procedure Diagram.

3.5 Data Collection and Analysis

Data were analyzed using a one-way analysis of variance (ANOVA), with task time and Motion Sickness Symptoms scores as dependent variables. The three levels of text size, as well as the four levels of color and opacity, were considered as fixed factors. In the statistical model, individual differences between participants were treated as a random factor. Data were recorded using a custom experimental program written in Processing, and all analyses were performed using IBM SPSS Statistics 26 software, with statistical significance set at $p = 0.05$. One-way ANOVA was used to test for main effects and interaction effects across all treatment conditions, and post-hoc analysis was conducted using the Bonferroni test.

4 Results

Table 1 lists the average task time and motion sickness scores for each level of the independent variables. The experiment includes three independent variables: text size (3 levels), color (4 levels), and opacity (4 levels). Therefore, the total number of trials in this experiment is calculated by multiplying the number of levels for all independent variables by the total number of participants, i.e., $3 \times 4 \times 4 \times 20 = 960$. The number of trials corresponding to each factor was obtained by dividing the total number of trials by the number of levels for each respective variable.

Table 1. Average task time (in seconds) and Motion Sickness Symptoms scores, along with their standard deviations, for each level of the independent variables (Participants n = 20).

Variables	Number of trials	Task time		Dizziness	
		Mean	SD	Mean	SD
Font size	960				
8 pt	320	28.108	12.697	29.05	16.178
10 pt	320	21.838	12.002	26.95	12.996
12 pt	320	26.699	13.975	26.15	13.003
Colour	960				
RAL 6024 Traffic Green	240	26.871	9.974	20.00	10.662
RAL 2009 Traffic Orange	240	22.420	10.467	26.55	11.736
RAL 9016 Traffic White	240	16.563	9.784	19.85	9.74
RAL 5017 Traffic Blue	240	20.790	8.591	15.75	9.141
Transparency level	960				
T0.25	240	29.032	14.001	29.03	14.001
T0.50	240	21.529	9.129	21.53	9.129
T0.75	240	14.455	6.531	14.46	6.531
T1.00	240	21.630	9.387	21.63	9.387

Table 2 also summarizes the main effects results of the analysis of variance (ANOVA) for each corresponding independent variable.

Table 2. Descriptive analysis results, task time (in seconds) and Motion Sickness Symptoms scores for each level of the independent variable (Participants n = 20), results of the one-way ANOVA.

Variables	Number of trials	Task Time				
		Mean Square	df	F	p	η^2
Font size	768					
8 pt	256	216.474	2	1.30	0.281	0.044
10 pt	256	216.474				
12 pt	256	216.474				
Colour	768					
RAL6024 Traffic Green	192	363.115	3	3.837	0.013	0.132
RAL2009 Traffic Orange	192	363.115				
RAL9016 Traffic White	192	363.115				
RAL5017 Traffic Blue	192	363.115				
Transparency level	768					
T0.25	192	708.485	3	6.91	< 0.001**	0.214
T0.50	192	708.485				
T0.75	192	708.485				
T1.00	192	708.485				
Variables	**Number of trials**	**Motion Sickness Symptoms**				
		Mean square	**df**	**F**	**p**	**η^2**
Font size	768					
8 pt	256	44.867	2	0.224	0.8	0.008
10 pt	256	44.867				
12 pt	256	44.867				
Colour	768					
RAL6024 Traffic Green	192	398.879	3	3.712	0.015	0.128

(*continued*)

Table 2. (*continued*)

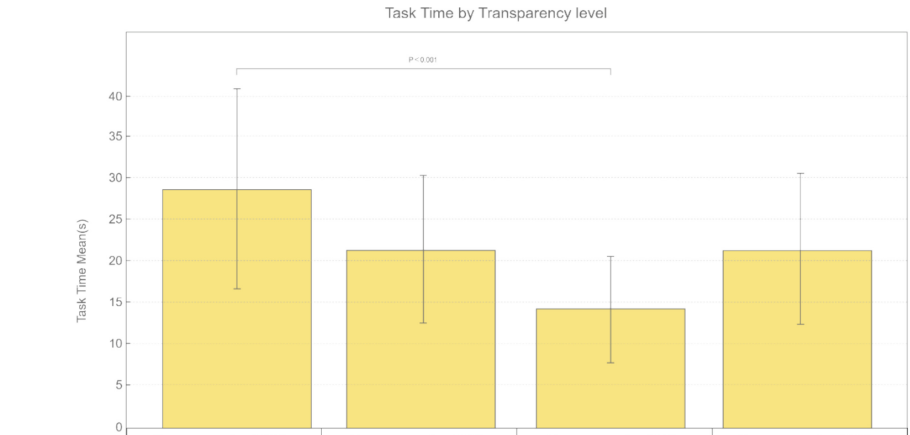
Variables	Number of trials	Task Time				
		Mean Square	df	F	p	η^2
RAL2009 Traffic Orange	192	398.879				
RAL9016 Traffic White	192	398.879				
RAL5017 Traffic Blue	192	398.879				
Transparency level	768					
T0.25	192	947.746	3	7.953	< 0.001**	0.239
T0.50	192	947.746				
T0.75	192	947.746				
T1.00	192	947.746				

4.1 Task Time

Table 2, along with Figs. 2, 3, and 4, presents the mean task completion times and statistical analysis results for each level of the independent variables. The ANOVA results indicate that both color ($F(3, 76) = 3.837$, $p = 0.013$, $\eta^2 = 0.132$) and transparency level ($F(3, 76) = 6.910$, $p < 0.001$, $\eta^2 = 0.214$) significantly affect task completion time, with a medium effect size for transparency level (as shown in Fig. 2). Regarding transparency level, T0.75 (14.454 seconds) significantly reduced task completion time compared to T0.25 (29.031 seconds), while no significant differences were observed between T0.50 (21.528 seconds) and the other levels. Post hoc Bonferroni tests revealed a significant difference between T0.75 and T0.25 ($p = 0.000$), while no other comparisons showed significant results.

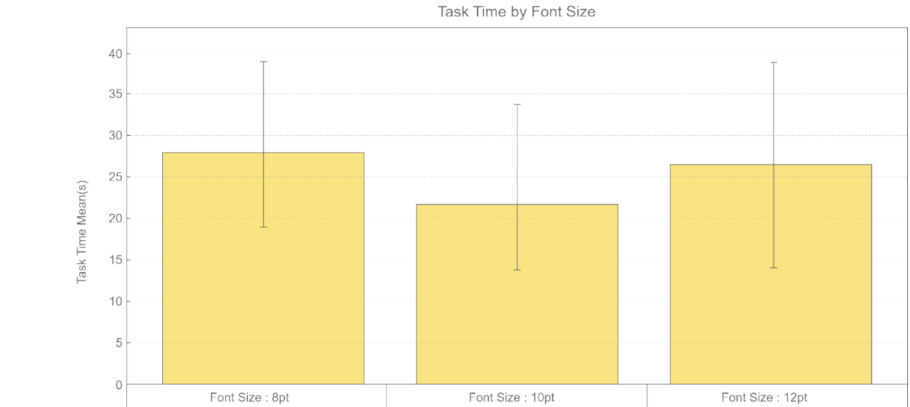
Although the task times vary between different font sizes as shown in the figure, the effect of font size did not reach a significant level ($F(2, 57) = 1.297$, $p = 0.281$, $\eta^2 = 0.044$) (see Fig. 3).

Meanwhile, the data results show that under different color conditions, the task completion time for RAL 9016 Traffic White (16.563 s) was significantly lower than that for RAL 6024 Traffic Green (26.871 s), while no significant differences were observed between the other color groups (see Fig. 4).



	Transparency level 0.25	Transparency level 0.50	Transparency level 0.75	Transparency level 1.00
Mean	29.032	21.529	14.455	21.630
Standard Deviation	14.001	9.129	6.531	9.387

Fig. 2. The average task time (SD) corresponding to the level of transparency.

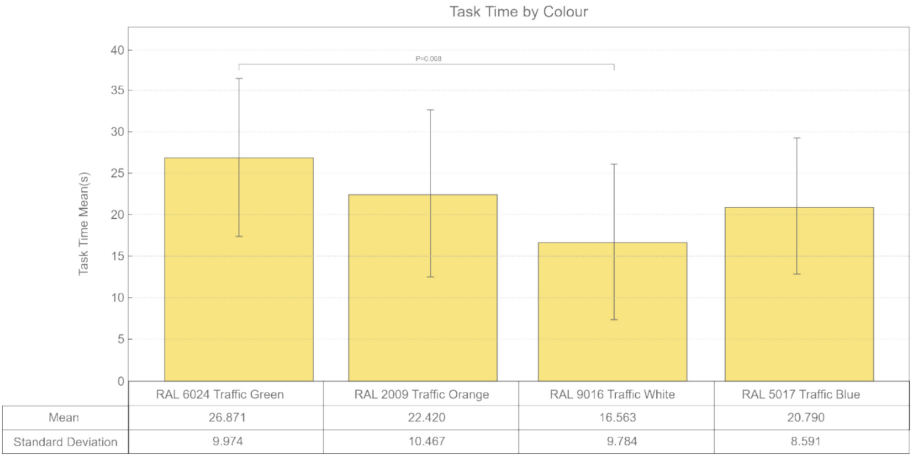


	Font Size : 8pt	Font Size : 10pt	Font Size : 12pt
Mean	28.108	21.838	26.699
Standard Deviation	12.697	12.002	13.975

Fig. 3. The average task time (SD) corresponding to the size of font.

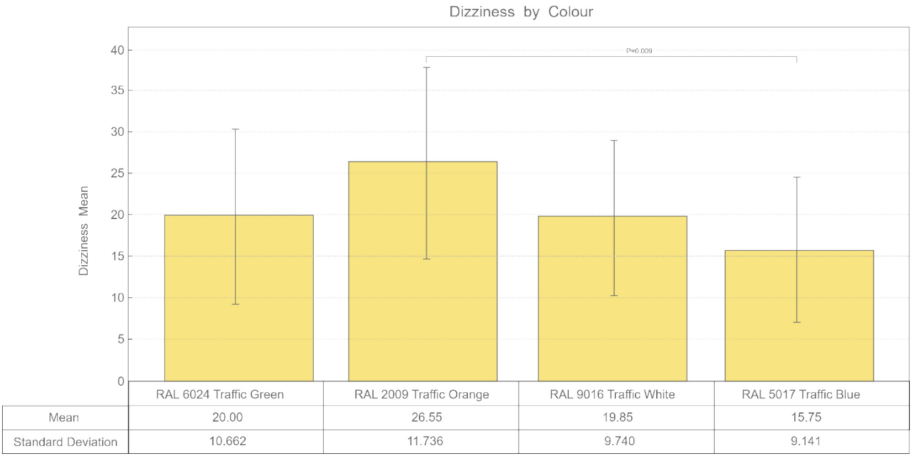
4.2 Motion Sickness Symptoms

Table 2 and Figs. 5, 6, and 7 show the mean scores of Motion Sickness Symptoms for each level of the independent variables, along with the statistical analysis results. The ANOVA results indicate that both color ($F(3, 76) = 3.712$, $p = 0.015$, $\eta^2 = 0.128$) and transparency level ($F(3, 76) = 7.953$, $p < 0.001$, $\eta^2 = 0.239$) significantly affected the Motion Sickness Symptoms scores, with a medium effect size for transparency level. The data show that under different color conditions, RAL 5017 Traffic Blue (15.75) produced the lowest Motion Sickness Symptoms score, which was significantly lower than RAL 2009 Traffic Orange (26.55) and RAL 6024 Traffic Green (20.00).



	RAL 6024 Traffic Green	RAL 2009 Traffic Orange	RAL 9016 Traffic White	RAL 5017 Traffic Blue
Mean	26.871	22.420	16.563	20.790
Standard Deviation	9.974	10.467	9.784	8.591

Fig. 4. The average task time (SD) corresponding to the color.

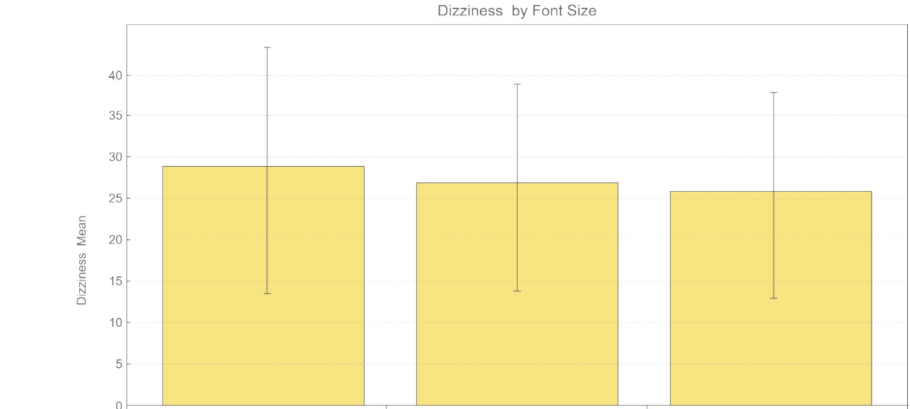


	RAL 6024 Traffic Green	RAL 2009 Traffic Orange	RAL 9016 Traffic White	RAL 5017 Traffic Blue
Mean	20.00	26.55	19.85	15.75
Standard Deviation	10.662	11.736	9.740	9.141

Fig. 5. Average (SD) scores of Motion Sickness Symptoms corresponding to colors.

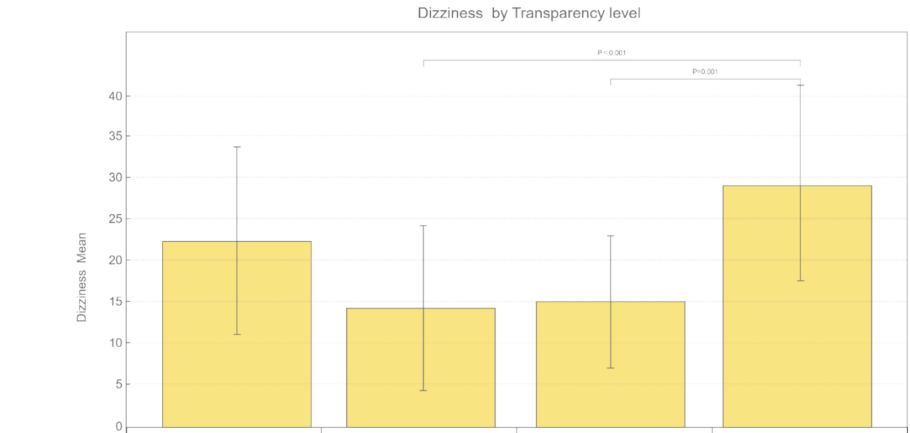
Although differences in Motion Sickness Symptoms scores between different font sizes are visible in the figures, font size did not have a significant effect on the Motion Sickness Symptoms scores ($F(2, 57) = 0.224$, $p = 0.800$, $\eta^2 = 0.008$).

Regarding transparency levels, T0.50 (14.65) and T0.75 (15.65) resulted in lower Motion Sickness Symptoms scores than T0.25 (22.35), while T1.00 (29.50) produced the highest Motion Sickness Symptoms score. Bonferroni post hoc tests showed significant differences between T0.50 and T1.00 ($p = 0.000$), and between T0.75 and T1.00 ($p = 0.001$), while no significant difference was found between T0.25 and T0.50.



	Font Size : 8pt	Font Size : 10pt	Font Size : 12pt
Mean	29.05	26.95	26.15
Standard Deviation	16.178	12.996	13.003

Fig. 6. Average (SD) scores of Motion Sickness Symptoms corresponding to transparency level.



	Transparency level 0.25	Transparency level 0.50	Transparency level 0.75	Transparency level 1.00
Mean	29.032	21.529	14.455	21.630
Standard Deviation	14.001	9.129	6.531	9.387

Fig. 7. Average (SD) scores of Motion Sickness Symptoms corresponding to font size.

5 Discussion

Through a peripheral vision recognition task in an AR device, this study investigates the effects of font size, color, and Transparency levels on peripheral vision information in AR environments. The collected data includes task completion time and Motion Sickness Symptoms ratings. Task time refers to the duration taken by users to perform the peripheral vision recognition task in the AR device, from the initiation of the task to its completion. A shorter task time indicates higher visibility of the peripheral vision information. The Motion Sickness Symptoms rating reflects users' subjective evaluation

of motion sickness symptoms while completing the peripheral vision recognition task. A higher rating indicates that the visual information, formed by the selected parameters, is more likely to induce negative experiences for the user, interfering with task performance and increasing the invasiveness of the information.

In this experiment, the effects of various factors are primarily seen in color and Transparency levels, both of which have a significant impact on the visibility and non-invasiveness of peripheral vision information in the AR scene. However, no significant effect of font size on peripheral vision visibility and non-invasiveness was found in this study. Generally speaking, RAL 5017 Traffic Blue and RAL 9016 Traffic White, combined with T0.75 opacity, performed the best overall, and this combination is the most recommended for encoding peripheral vision information in AR devices in this study.

5.1 Infuence of Font Size

This study explored the role of font size in the perception of peripheral vision information in AR devices. The results indicated that font size had no significant effect on task completion time or Motion Sickness Symptoms ratings. This finding contrasts with previous studies focused on central vision or virtual reality environments [14, 17]. Previous research generally suggests that larger fonts improve text readability, recognition rates, and user performance. However, it is important to note that these conclusions are often based on scenarios where users focus their attention on central vision or static text. In contrast, the experimental design of this study focused on the recognition of peripheral vision information. Therefore, the adjustment of font size may not have reached the critical threshold required to alter information processing efficiency in peripheral vision, leading to an insignificant impact.

In comparison, adjusting font size may be difficult to directly optimize this balance in peripheral vision. When the font size exceeds a certain threshold, it may increase invasiveness by occupying too much visual space. Conversely, smaller fonts reduce interference but also decrease readability due to the resolution limitations of peripheral vision. This nonlinear effect may have masked the potential influence of font size, preventing it from showing significant differences under the experimental conditions. Although this study did not find a significant effect of font size, the conclusion still provides valuable insights for AR interface design: in optimizing peripheral vision information, adjustments to color and opacity may take precedence over font size.

5.2 Infuence of Font Color

Color has a significant impact on the visibility and non-intrusiveness of peripheral visual information in AR interfaces. This result is similar to the findings of other studies on the role of color coding in AR/VR interface design. The current study also demonstrates a new characteristic of AR color coding, which remains effective in scenes where the viewing angle shifts from the central to the peripheral visual field. Consistent with the research of Saverio Debernardis et al., both white and blue exhibited high usability in AR information color coding, regardless of whether they were in the central or peripheral visual field [24]. Furthermore, this study found that the high visibility and low

intrusiveness of white (RAL 9016) may be related to its universal contrast with natural environments. In dynamic AR backgrounds, white can quickly attract users' unconscious attention (preattentive processing) by maximizing brightness contrast, such as when contrasted with dark backgrounds, thus reducing information recognition time. Ryun-Seok Oh et al. [26] also pointed out in their research that traffic white is often used in emergency signage due to its resistance to interference in complex environments, which aligns with the superior performance of white in dynamic AR backgrounds in this study. Additionally, the color coding of white and blue has been proven effective for encoding peripheral visual information in AR interfaces. This finding further extends the use of white and blue in AR environments and provides valuable insights into the recognizability of peripheral color coding in AR devices.

5.3 Infuence of Transparency Levels

The research results indicate that transparency level significantly affects both task completion time and Motion Sickness Symptoms (MSS) scores, with a moderate effect size. Specifically, when transparency was set to 75% (T0.75), task completion time was significantly shorter, while at 100% transparency (T1.00), MSS scores were significantly higher. These findings also support previous research suggesting that transparency impacts task completion time, particularly in augmented reality (AR), where excessively high transparency levels can affect text readability, thus reducing task efficiency. In this study, task completion time was shortest at T0.75, indicating that this transparency level achieves the best balance between the visibility and non-intrusiveness of AR information. In contrast, the longest task completion time was observed at T0.25, likely because the lower transparency accentuated the background color, interfering with information visibility and prolonging task completion time. The MSS scores aligned with previous studies, showing that lower transparency levels resulted in higher MSS ratings, especially at T1.00, where MSS scores were notably higher than at other transparency levels. This corresponds with the findings of Kruijff et al. [19]on the impact of transparent elements in wide-field AR displays, where they observed that semi-transparent elements reduced text visibility, affected user performance, and potentially increased discomfort. Similarly, Harrison and Vicente found that when transparency exceeded 75%, user performance significantly declined, error rates increased, and MSS scores rose. [25] This suggests that excessive transparency may cause visual disturbance, negatively affecting user comfort and task performance.

Interestingly, both the lowest transparency and full opacity were ineffective in ensuring good visibility and non-intrusiveness of peripheral AR information. In other words, there is an optimal transparency threshold for the peripheral visual field, whether transparency increases or decreases. This study partially reveals this optimal threshold range (T0.5–T0.75). The non-linear progression between transparency levels represents an innovative aspect of this study, as the discontinuous mapping suggests a critical interference threshold. Beyond this threshold, the gains from increased transparency diminish marginally.

6 Limitations and Future Work

This study provides insights for the design of peripheral vision information in augmented reality (AR); however, its application requires adaptive adjustments based on scene dynamics (e.g., motion speed, background complexity). Additionally, the font size gradient set in the experiment may not cover the sensitivity range of peripheral vision perception. Future research could explore more extreme size ranges or dynamic adaptive adjustment mechanisms, such as incorporating user head movements or fixation points. Moreover, the nature of the peripheral vision task (e.g., brief presentation, multitasking) may weaken the independent effect of font size, which should be further validated in relation to the complexity of the interactive scenario.

Finally, user adaptation to AR devices (e.g., training effects on peripheral vision information acquisition) may moderate the results, and this should be examined in long-term usage studies. Future research could integrate eye-tracking technology to quantify the impact of color on peripheral vision fixation shifts (saccadic shifts) and explore more optimized encoding combinations.

Acknowledgments. We express our thanks to all participants who took part in the survey, since their responses provided data supporting our study.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Huang, T.K., Yang, C.H., Hsieh, Y.H., Wang, J.C., Hung, C.C.: Augmented reality (AR) and virtual reality (VR) applied in dentistry. Kaohsiung J. Med. Sci. **34**(4), 243–248 (2018)
2. Qiao, X., Ren, P., Dustdar, S., Liu, L., Ma, H., Chen, J.: Web AR: a promising future for mobile augmented reality—state of the art, challenges, and insights. Proc. IEEE. **107**(4), 651–666 (2019)
3. Li, X., Yi, W., Chi, H.L., Wang, X., Chan, A.P.: A critical review of virtual and augmented reality (VR/AR) applications in construction safety. Autom. Constr. **86**, 150–162 (2018)
4. Vater, C., Wolfe, B., Rosenholtz, R.: Peripheral vision in real-world tasks: a systematic review. Psychon. Bull. Rev. **29**(5), 1531–1557 (2022)
5. Strasburger, H., Rentschler, I., Jüttner, M.: Peripheral vision and pattern recognition: a review. J. Vis. **11**(5), 13–13 (2011)
6. Hussain, M., Park, J.: Effect of transparency levels and real-world backgrounds on the user interface in augmented reality environments. Int. J. Hum. Comput. Interact. **40**(16), 4265–4274 (2024)
7. Yamin, P.A., Park, J., Kim, H.K., Hussain, M.: Effects of button colour and background on augmented reality interfaces. Behav. Inform. Technol. **43**(4), 663–676 (2024)
8. Kwok, A.P.K., Yan, M., Xu, Z.H., Lin, J.Y., Da Chen, R., Wen, S.L.: Effects of line length, number of lines, line spacing, and font size on reading performance of Chinese text in virtual reality environment. Int. J. Ind. Ergon. **96**, 103469 (2023)
9. Zhang, L., Albert, R., Kim, J., Luebke, D.: Develo** a peripheral color tolerance model for gaze-contingent rendering. J. Vis. **19**(10), 298c–298c (2019)
10. Xue, C., Li, J., Wang, H., Niu, Y.: Effects of target and distractor saturations on the cognitive performance of an integrated display interface. Chin. J. Mech. Eng. **28**(1), 208–216 (2015)

11. Fiorentino, M., Debernardis, S., Uva, A.E., Monno, G.: Augmented reality text style readability with see-through head-mounted displays in industrial context. Presence Teleop. Virt. **22**(2), 171–190 (2013)
12. Renkewitz, H., Kinder, V., Brandt, M., Alexander, T.: Optimal font size for head-mounted-displays in outdoor applications. In: 2008 12th International Conference Information Visualisation, pp. 503–508. IEEE, Piscataway, NJ (2008)
13. Grobelny, J., Michalski, R.: The role of background color, interletter spacing, and font size on preferences in the digital presentation of a product. Comput. Hum. Behav. **43**, 85–100 (2015)
14. Novotny, J., Laidlaw, D.H.: Evaluating text Reading speed in VR scenes and 3D particle visualizations. IEEE Trans. Vis. Comput. Graph. **30**, 2602–2612 (2024)
15. Bernard, M.L., Chaparro, B.S., Mills, M.M., Halcomb, C.G.: Comparing the effects of text size and format on the readibility of computer-displayed times new Roman and Arial text. Int. J. Hum. Comput. Stud. **59**(6), 823–835 (2003)
16. Chen, Y., Huang, G., Wang, K.: Effects of font size, stroke, and background on the legibility of Chinese characters in virtual reality for the elderly. Ergonomics, 1–11 (2024)
17. Leykin, A., Tuceryan, M.: Automatic determination of text readability over textured backgrounds for augmented reality systems. In: Third IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 224–230. IEEE, Piscataway, NJ (2004)
18. Kruijff, E., Orlosky, J., Kishishita, N., Trepkowski, C., Kiyokawa, K.: The influence of label design on search performance and noticeability in wide field of view augmented reality displays. IEEE Trans. Vis. Comput. Graph. **25**(9), 2821–2837 (2018)
19. Harrison, B.L., Vicente, K.J.: An experimental evaluation of transparent menu usage. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 391–398 (1996, April)
20. Jankowski, J., Samp, K., Irzynska, I., Jozwowicz, M., Decker, S.: Integrating text with video and 3d graphics: the effects of text drawing styles on text readability. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1321–1330 (2010)
21. Lewandowska, A., Dziśko, M., Jankowski, J.: Investigation the role of contrast on habituation and sensitisation effects in peripheral areas of graphical user interfaces. Sci. Rep. **12**(1), 15281 (2022)
22. Bruck, S., Watters, P.A.: Estimating cybersickness of simulated motion using the simulator sickness questionnaire (SSQ): a controlled study. In: 2009 Sixth International Conference on Computer Graphics, Imaging and Visualization, pp. 486–488. IEEE, Piscataway, NJ (2009)
23. Lavie, N.: Distracted and confused? Selective attention under load. Trends Cogn. Sci. **9**(2), 75–82 (2005)
24. Debernardis, S., Fiorentino, M., Gattullo, M.: Text readability in head-worn displays: color and style optimization. In: IEEE Transactions on Visualization and Computer Graphics, p. 125
25. Harrison, B.L., Vicente, K.J.: An experimental evaluation of transparent menu usage. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 391–398 (1996)
26. Oh, R.S., Kim, Y.C., Bae, Y.H., Choi, J.H.: Evaluation of the maximum cognitive distance per emergency exit sign colour in a smoke-filled environment simulated using a translucent eye patch. Fire Technol. **59**(6), 3445–3471 (2023)



Design of a Hierarchical Federated Generative Learning Based Smart Home System

Paulo Drefahl[1], Kevin Kostage[1], Sean Peppers[1], Wenbin Guo[2], Luca Mazzola[3], and Chengyi Qu[1(✉)]

[1] Florida Gulf Coast University, Fort Myers, FL, USA
{pdrefahl,kkostage,speppers,cqu}@fgcu.edu
[2] University of Florida, Gainesville, FL, USA
wenbin.guo@ufl.edu
[3] Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
luca.mazzola@hslu.ch

Abstract. The proliferation of Internet of Things (IoT) devices in smart home environments has significantly expanded the attack surface, making home networks increasingly susceptible to sophisticated cyber threats. Traditional cloud-based intrusion detection systems (IDS) often suffer from high latency and performance bottlenecks, limiting their effectiveness in real-time threat mitigation. This paper introduces a Hierarchical Federated Generative Learning (HFGL) framework for decentralized intrusion detection, leveraging a multi-tiered architecture composed of user routers, a local server, and edge nodes to facilitate real-time network monitoring and distributed anomaly detection. The system integrates OpenWRT-based router firmware with packet-level data capture, securely transmitting network traffic metadata to a local server via SSH-based Paramiko requests. A custom Electron.js-based desktop application provides an intuitive interface for homeowners, enabling seamless router configuration, real-time alert visualization, and security policy orchestration. At its core, our approach employs a federated deep learning pipeline augmented with Generative Adversarial Networks (GANs) to enhance intrusion detection capabilities while preserving data privacy. The GAN model not only identifies emerging threats but also generates adversarial attack simulations to improve model robustness. Experimental evaluations demonstrate significant improvements in detection accuracy and latency reduction compared to conventional IDS approaches, underscoring the potential of privacy-preserving, edge-driven cybersecurity solutions for modern smart home ecosystems.

Keywords: Cybersecurity · Intrusion Detection System · Federated Learning · Generative Adversarial Networks · IoT

1 Introduction

The rapid proliferation of Internet of Things (IoT) devices has significantly expanded the attack surface of modern home networks. From smart thermostats

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 160–177, 2025.
https://doi.org/10.1007/978-3-031-93502-2_11

and security cameras to connected lighting and appliances, these devices continuously exchange data over home networks, often relying on low-cost, resource-constrained embedded systems with limited security features [10]. As IoT adoption grows, so does the risk of cyber threats, with attackers exploiting vulnerabilities in home routers, insecure device configurations, and weak authentication mechanisms. Traditional cloud-based intrusion detection systems (IDS) struggle to provide low-latency, privacy-preserving solutions due to network congestion, bandwidth limitations, and centralized dependency. Consequently, an urgent need exists for decentralized, edge-based security mechanisms that can efficiently detect and mitigate emerging cyber threats in real time [7].

To address these challenges, we propose a Hierarchical Federated Generative Learning-Based Smart Home System for real-time intrusion detection at the router level. Our framework leverages a multi-tiered architecture integrating federated learning (FL) and Generative Adversarial Networks (GANs) to enhance threat detection capabilities while preserving user privacy. The system operates at three layers: local servers, edge servers, and a centralized high-performance computing (HPC) server, ensuring scalable, privacy-conscious network monitoring. At the local level, an OpenWRT-enabled home router captures network traffic using tcpdump and securely transmits packet data to a lightweight local server deployed on an Intel NUC. This server preprocesses the data and extracts network features before forwarding them to federated edge servers deployed on AWS virtual machines for collaborative model training. The edge servers participate in federated learning, aggregating intrusion detection models from multiple devices and refining security policies before relaying optimized models back to local networks.

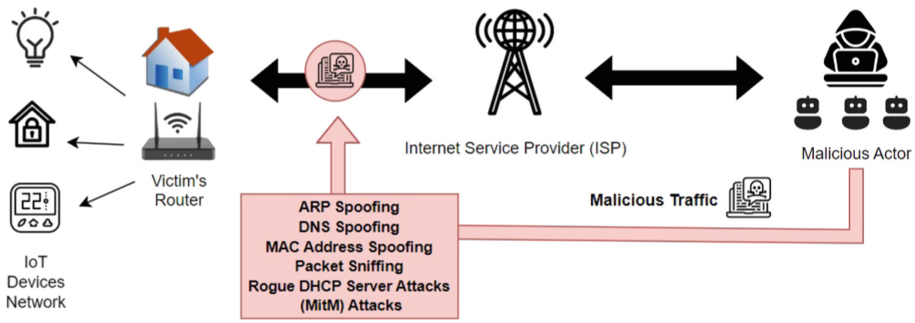


Fig. 1. Illustration of IoT Network Vulnerabilities: How Malicious Actors Exploit IoT Devices Through Victim's Router Using Spoofing, Packet Sniffing, and Man-in-the-Middle (MitM) Attacks.

The primary threats targeted by our system include spoofing attacks (ARP, DNS, MAC address), packet sniffing, rogue DHCP server attacks, and man-in-the-middle (MitM) intrusions. As illustrated in Fig. 1, malicious actors leverage the internet service provider (ISP) infrastructure to inject malicious traffic into

home networks, exploiting vulnerabilities in IoT devices and routers. Our system mitigates these threats by continuously monitoring network flows, detecting anomalies, and deploying adaptive mitigation measures.

To validate our approach, we conduct extensive experiments using three datasets: a real IoT traffic dataset, the CICIoT2023 benchmark dataset, and a synthetic adversarial dataset generated by our GAN model. The real dataset provides labeled normal and malicious traffic, capturing real-world attack patterns, while CICIoT2023 offers a diverse range of attack and benign traffic types for robust IDS evaluation. Additionally, our GAN model generates synthetic attack scenarios to enhance detection capabilities, training the discriminator to differentiate between legitimate and adversarial traffic while improving overall IDS robustness.

The experimental setup consists of a hierarchical deployment, where a local server (Intel NUC, 8GB RAM) handles preprocessing, and two AWS virtual machines (4 vCPUs, 8GB RAM each) serve as federated edge nodes for distributed learning. A Flask-based coordination server facilitates model exchanges between local and edge servers, ensuring real-time model updates while maintaining low-latency threat detection. The system also integrates an Electron.js-based desktop application, providing an intuitive user interface for security alerts, network monitoring, and automated response configurations.

Our results demonstrate that the proposed federated GAN-based intrusion detection system achieves high detection accuracy, reduced latency, and improved privacy protection compared to traditional cloud-based IDS solutions. However, several challenges remain, including computational overhead at the edge, limited OpenWRT compatibility across router models, and the need for further GAN stability improvements. Future work will focus on integrating a Mixture of Experts (MoE) model to optimize federated learning efficiency across heterogeneous network environments and enhancing GAN training methodologies to improve adversarial traffic simulations. Additionally, expanding the framework to enable cross-domain collaboration between home networks and ISP infrastructures will facilitate large-scale, privacy-preserving threat intelligence sharing.

2 Related Works

With the increasing prevalence of IoT devices in smart home environments, security challenges have become more significant. This has led to the development of complex intrusion detection systems (IDS) aimed at addressing evolving cyber threats. However, traditional IDS solutions suffer from limited network visibility, high maintenance costs, and an inability to adapt to novel attacks. Moreover, cloud-centric IDS models exacerbate privacy concerns by requiring the transmission of raw network data to centralized servers for analysis [3]. These limitations necessitate a shift toward near real-time, privacy-preserving IDS models. Our work addresses this challenge by introducing a hierarchical federated generative learning framework that enables real-time, decentralized intrusion detection at the router level while preserving user privacy.

Federated Learning (FL) has shown promise in privacy-preserving machine learning, allowing data to remain local while leveraging distributed training to produce robust, global models. This approach mitigates privacy risks associated with centralized learning while enabling continuous model updates [6]. The FedAvg algorithm demonstrated efficient communication in decentralized networks, leading to its adoption in applications such as animal monitoring methods [12]. However, applying FL to IoT environments presents additional challenges due to the need for real-time responsiveness and computational constraints. Our novel contribution lies in adapting FL for smart home IDS, ensuring that federated updates occur with minimal latency while maintaining threat detection accuracy across heterogeneous IoT devices.

Generative Adversarial Networks (GANs) have been increasingly applied in cybersecurity, particularly for generating synthetic attack patterns that mimic real-world cyber threats. This enhances IDS robustness by incorporating adversarial training [1]. GANs enable IDS models to learn from diverse attack scenarios, improving detection accuracy against previously unseen threats. However, GANs introduce risks such as training instability and susceptibility to data poisoning. Our work addresses these concerns by integrating stabilized GAN training techniques into a federated IDS framework, ensuring model reliability while maintaining adversarial robustness.

Recent studies have explored hybrid IDS models that incorporate both FL and GANs. For instance, prior research introduced GAN-enhanced IDS models to tackle class imbalance issues in intrusion detection datasets [5]. These studies analyzed multiple GAN architectures, including WGAN-GP, CGAN, CTGAN, and GWGAN-GP, comparing their performance against traditional sampling techniques. Their findings highlighted GANs' effectiveness in improving detection accuracy across multiple attack types. Our work builds upon this research by implementing GAN-augmented federated learning in a real-world deployment, demonstrating its feasibility and effectiveness in smart home network security.

Federated Learning has also been applied to dynamic network environments, such as drone swarms, where it has been used to enhance anomaly detection models for network security in autonomous systems [4]. Studies leveraging the AERPAW testbed demonstrated the benefits of differential privacy and adversarial training in mitigating data poisoning risks. Our work extends this concept to IoT networks, applying FL optimization strategies tailored for smart home environments. By doing so, we demonstrate FL's adaptability across diverse network structures, further enhancing its viability for resource-constrained IoT deployments.

While previous studies have independently explored FL and GAN-based IDS models, few have effectively combined these technologies to address security challenges in smart home environments. Smart home networks present unique constraints, including limited computational resources, device heterogeneity, and compatibility issues. Our proposed hierarchical federated generative learning framework bridges these gaps by leveraging FL to maintain user privacy and GANs to simulate diverse attack scenarios, improving detection accuracy, latency

reduction, and adaptability to zero-day threats [9]. By integrating these methods, our system offers a scalable, real-time, and privacy-preserving IDS designed for the next generation of smart home cybersecurity.

As we see larger IoT systems within homes continue to grow and evolve the challenge of ensuring usability and security will remain. Most IDS systems are implemented with out user interfaces (UI) which makes them less available to non-experts to manage and configure. As proposed, a user-centric framework for smart homes [11] would ensure accessibility and ease of use for complex technologies such as IDS based on federated learning. These systems could become quite cumbersome for our senior citizens, which is why usability should be prioritized [2] with our proposed design offering a user-friendly monitoring interface that allows easy access and management of security settings.

3 System Architecture and Design

In this section, we detail our multi-layered smart home security framework by providing a step-by-step workflow for the whole logic. In addition to that, we provide a detailed system architecture, offering a breakdown of each component within the front-end and back-end design.

3.1 System and Networks Architecture Overview

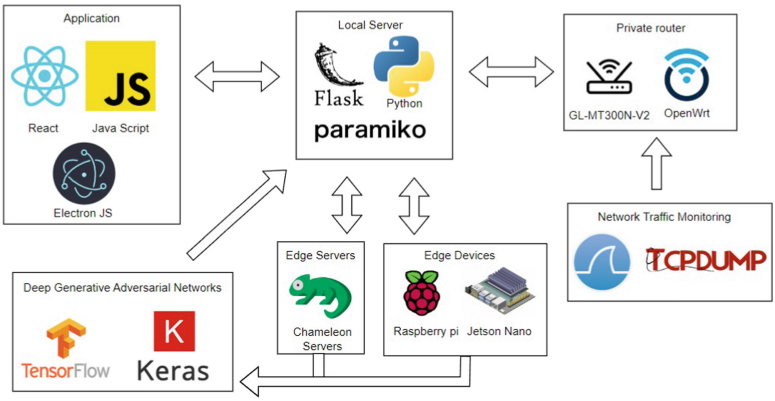


Fig. 2. Hierarchical Federated Generative Learning Framework for Smart Home Security.

Figure 2 presents the proposed system architecture from a network and computational perspective, highlighting a hierarchical, federated intrusion detection framework designed to enhance smart home security. At the foundation of this system, a smart home router running OpenWRT serves as the first line of defense,

continuously monitoring network traffic and capturing packet-level data using tools such as `tcpdump` and Wireshark. To minimize data exposure and preserve privacy, raw packet capture (PCAP) files remain on the router, ensuring that sensitive network information is not prematurely transmitted beyond the local network. A secure, key-based SSH connection, managed via Python scripts utilizing Paramiko, facilitates trusted and automated interactions between the router and the local server, enabling remote initiation or scheduling of packet capture sessions and secure retrieval of relevant traffic data.

Upon receiving the PCAP data, the local server performs an initial preprocessing and filtering stage, removing benign or routine traffic associated with whitelisted IP addresses and recognized services. This preprocessing step significantly reduces the volume of data requiring deeper analysis while concurrently flagging potentially anomalous patterns using lightweight heuristics. Instead of transmitting raw traffic logs, the local server extracts and anonymizes relevant network features, which are then securely relayed to designated edge servers. These edge nodes participate in a federated learning framework, allowing for decentralized model training without compromising user privacy.

At the edge computing layer, advanced Generative Adversarial Networks (GANs) enhance the intrusion detection process by generating adversarial traffic samples, enabling the classifier to identify both known and emerging cyber threats with greater accuracy. To further improve detection capabilities while ensuring privacy, a federated model aggregation protocol, such as FedAvg [6], periodically synchronizes model updates across multiple distributed nodes. This strategy enables the collaborative refinement of a global intrusion detection model without requiring direct access to individual network traffic data, thereby preserving privacy while improving detection robustness.

Detection outcomes and refined model parameters are continuously relayed back to the local server, facilitating real-time network monitoring and security updates. The user-facing desktop application, developed using Electron.js, React, and JavaScript, provides an interactive visualization dashboard, allowing homeowners to configure security settings, visualize alerts, and take proactive mitigation actions, such as blocking or quarantining malicious connections. By distributing computational workloads across the router, local server, and federated edge nodes, this architecture achieves low-latency, privacy-preserving intrusion detection, ensuring a robust and scalable cybersecurity solution for modern smart home environments.

The proposed system follows a structured workflow that enables privacy-preserving, real-time intrusion detection through a hierarchical federated learning approach. Initially, an OpenWRT-enabled router passively captures network traffic using tcpdump and Wireshark, storing raw PCAP files locally to minimize data exposure. A secure SSH-based communication channel, managed via Paramiko, allows the local server to retrieve these files for preprocessing, where benign traffic is filtered out, and initial anomaly detection is performed based on extracted network features. Instead of transmitting raw data, the local server forwards anonymized and summarized traffic features to federated edge

nodes, including Raspberry Pi, Jetson Nano, and Chameleon cloud servers, where advanced Generative Adversarial Networks (GANs) refine the detection process by synthesizing attack patterns and improving classifier robustness. Using federated averaging (FedAvg), model updates from distributed edge nodes are aggregated, enhancing global detection capabilities while preserving privacy. These refined models and detection outcomes are then relayed back to the local server, ensuring low-latency threat detection and adaptive security responses. A desktop application, built with Electron.js and React, provides users with an intuitive interface for real-time monitoring, security alerts, and router configuration, allowing homeowners to manage network policies while benefiting from a continuously evolving, collaborative cybersecurity framework.

3.2 Main User Interface Components

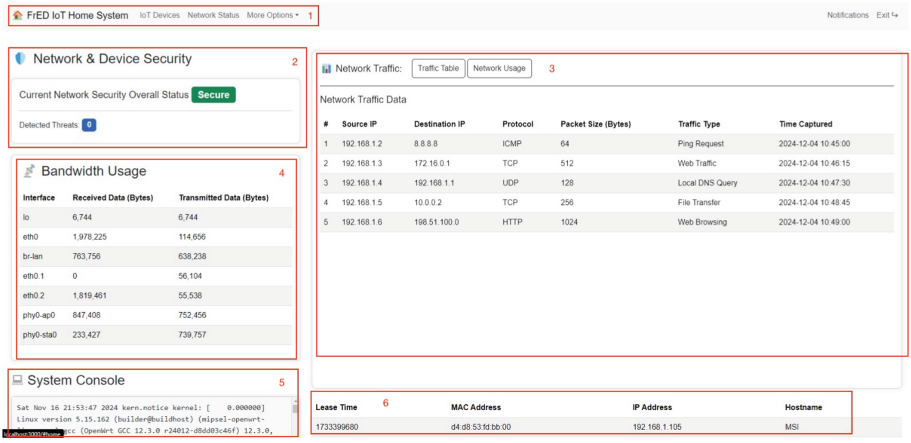


Fig. 3. Comprehensive Network Security Dashboard for Smart Home Intrusion Detection and Monitoring.

Figure 3 presents the main interface of the desktop application, which provides a comprehensive network security dashboard with functionalities such as real-time network monitoring, device activity tracking, bandwidth usage analysis, and network diagnostics. Each labeled section in the figure corresponds to a key component of the system, offering users an intuitive way to monitor and manage their home network security.

***Navigation and Menu* (Label 1)**. The top navigation bar offers quick access to core functionalities, allowing users to switch between different views, including "IoT Devices," "Network Status," and "More Options." Additionally, this section includes a notifications panel for security alerts and a clearly visible "Exit" button, ensuring seamless navigation and accessibility.

***Network & Device Security* (Label 2)**. This panel provides a high-level overview of the network's security status, offering a visual indicator (e.g., "Secure" or "At Risk") based on detected threats. The displayed threat count enables users to quickly assess potential risks and take immediate action if security anomalies are identified.

***Network Traffic* (Label 3)**. The network traffic section provides a detailed table of active network connections, displaying essential attributes such as source and destination IP addresses, communication protocols (e.g., TCP, UDP, ICMP), packet sizes, traffic types, and timestamps. Users can filter traffic logs based on protocol or IP range to detect suspicious activity and investigate potential security threats.

***Bandwidth Usage* (Label 4)**. This panel presents real-time bandwidth consumption across different network interfaces, showing both received (inbound) and transmitted (outbound) data volumes. Interfaces such as `eth0` and `br-lan` are monitored to help users diagnose network performance issues, identify bandwidth-heavy devices, and detect unusual data spikes, which may indicate malicious activities such as denial-of-service (DoS) attacks.

***System Console* (Label 5)**. The system console provides low-level diagnostic messages, kernel logs, and system status outputs from the OpenWRT-based router. This feature is particularly useful for troubleshooting, allowing advanced users to inspect real-time logs, validate network services, and debug connectivity issues.

***IP Leases and Device Information* (Label 6)**. This section displays active DHCP leases, listing key device attributes such as lease duration, MAC addresses, assigned IP addresses, and hostnames. By consolidating this information, users can easily identify unauthorized devices, track active hosts, and enforce customized security policies for different network endpoints.

Overall, this interface integrates multiple security and diagnostic tools, enabling proactive network monitoring and user-driven cybersecurity management in smart home environments.

3.3 Additional User Interface Components

Figure 4 presents key panels within the user application that facilitate firewall rule configuration, system performance monitoring, and network interface diagnostics. These components provide administrators with fine-grained control over security policies, network load assessment, and interface-level configurations to ensure optimal performance and protection.

***Firewall Rules* (Label 7)**. This panel displays and allows modification of the OpenWRT `/etc./config/firewall` file, enabling administrators to configure security policies, traffic filtering rules, and network protection mechanisms. Default settings define the handling of incoming, outgoing, and forwarded traffic, with options for enabling synflood protection and defining specific zones. As firewall rules take effect immediately, users can quickly adapt policies to mitigate threats, enforce segmentation, or optimize network flow based on real-time security needs.

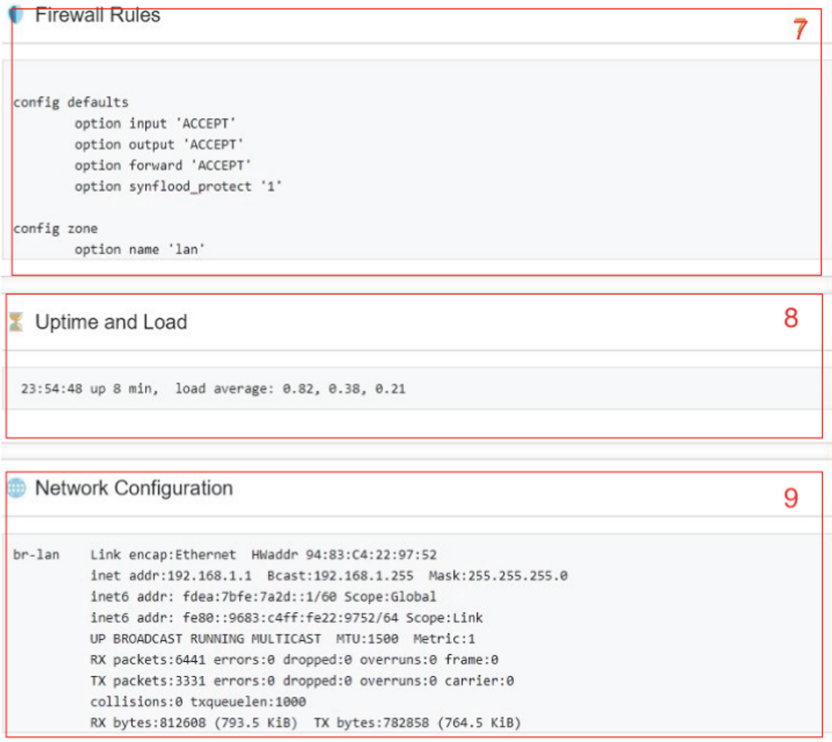


Fig. 4. Firewall Configuration, System Performance Monitoring, and Network Diagnostics Panel.

***Uptime and Load* (Label 8)**. This section provides real-time system performance metrics, including device uptime and CPU load averages over the past *1, 5,* and *15* minutes. Monitoring these indicators helps administrators detect excessive resource consumption, potential performance bottlenecks, or system instability. Elevated load values may indicate CPU-intensive tasks, while unexpectedly short uptime durations could suggest frequent reboots or underlying hardware/software failures requiring further investigation.

***Network Configuration* (Label 9)**. This panel presents detailed interface-level network diagnostics, displaying hardware addresses (MAC), assigned IPs, subnet masks, and traffic statistics. Users can monitor packet transmission and reception rates, detect packet drops or errors, and validate interface configurations to ensure proper routing and network segmentation. The information provided is crucial for troubleshooting connectivity issues, verifying correct bridge settings, and diagnosing potential hardware-related failures before they impact network stability.

These panels collectively enhance network security management by providing real-time insights, proactive diagnostics, and user-controlled firewall policy adjustments, ensuring a robust and resilient smart home security environment.

4 Federated Generative Learning Methodology

Building upon the previously discussed security and intrusion detection framework, this section introduces a hierarchical federated learning architecture designed to enhance distributed model training and collaborative threat detection.

4.1 Federated Learning Workflow

Figure 6 presents a hierarchical federated learning framework that enables distributed model training and adaptive cybersecurity solutions across multiple network layers. At the foundation of this system, edge devices within local networks independently train localized machine learning models based on the network traffic they monitor. These devices report to edge servers, which serve as intermediate nodes responsible for aggregating, refining, and distributing model updates. Each edge server collects multiple local models, integrates them, and generates an optimized local federated model, which is then transmitted to a remote high-performance computing (HPC) server for further refinement.

At the top tier of the hierarchy, the HPC server aggregates multiple local models from edge servers to produce a more generalized global model. This global model is then distributed back to edge servers, ensuring that each network benefits from collective intelligence while maintaining data privacy-since raw traffic data never leaves the local network. The bidirectional flow of local and global model updates fosters an adaptive intrusion detection system capable of identifying both known and emerging threats.

By organizing federated learning into multiple levels-from local edge devices to edge servers and finally to a centralized HPC system-this approach optimizes computational efficiency, reduces latency, and enhances detection accuracy. The depicted architecture highlights the potential for scalable, privacy-preserving cybersecurity solutions applicable to smart home environments, industrial IoT, and broader network security applications.

More specifically, federated learning facilitates decentralized model training while preserving data privacy by ensuring that raw data remains local, with only model updates being exchanged across network layers. The workflow is meticulously designed to uphold this principle, preventing the transmission of sensitive network traffic data while still enabling collaborative learning. The process begins with feature extraction, where local servers analyze PCAP flow data, extracting key network characteristics such as packet length, flow duration, and protocol distribution. To further safeguard privacy, sensitive identifiers like IP addresses and port numbers can be anonymized before further processing.

Following feature preparation, the edge training phase is initiated, where each edge device independently trains a local intrusion detection model using aggregated feature updates derived from observed network traffic. These localized models are periodically synchronized through federated updates, wherein edge devices transmit computed weight adjustments to a federated coordinator or central aggregation server. The coordinator then integrates these updates,

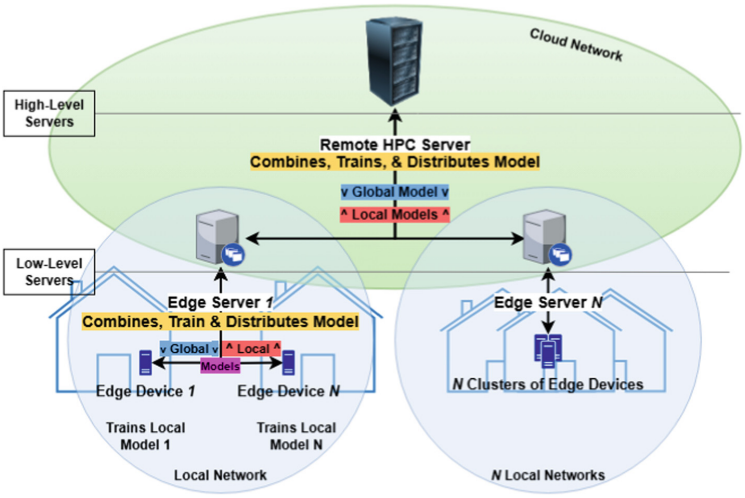


Fig. 5. Hierarchical Federated Learning Architecture for Distributed Intrusion Detection.

refines the global model, and redistributes the improved version back to the edge devices within the local network.

Finally, during the local inference phase, each edge device applies the latest model for real-time intrusion detection, leveraging the continuously evolving federated model to identify anomalies, detect cyber threats, and enhance network security. This hierarchical learning approach significantly reduces data exposure risks while enabling adaptive, privacy-preserving cybersecurity solutions within distributed smart home and IoT environments.

4.2 GAN and IDS Integration

Building on the federated learning framework, Generative Adversarial Networks (GANs) further enhance the effectiveness of intrusion detection by introducing an adversarial training paradigm. This approach leverages two competing models: a generator (G) and a discriminator (D), working in tandem to refine network intrusion detection capabilities. The generator (G) is responsible for synthesizing realistic malicious network traffic patterns, emulating attack behaviors such as DDoS floods, ARP spoofing, and other sophisticated cyber threats. By continuously generating deceptive attack flows, G challenges the discriminator (D), which is trained to differentiate between legitimate traffic and adversarial attack samples. Through this iterative adversarial training, D progressively improves its detection robustness, learning to recognize more nuanced and evasive attack strategies over time.

To ensure scalability and efficiency, the trained GAN model is deployed across distributed edge servers, enabling real-time, localized detection of malicious activities while reducing dependence on centralized infrastructure. This

distributed approach enhances both responsiveness and resilience, allowing the intrusion detection system (IDS) to adapt dynamically to evolving cyber threats.

The adversarial detection workflow begins with real-time scoring of incoming network traffic, where the discriminator module (D) evaluates extracted traffic features and assigns a probability score to determine whether the traffic is legitimate or adversarial. If the traffic is classified as malicious or artificially generated by G, it is immediately blocked, and an alert is sent to administrators, notifying them of a potential adversarial attack. The intrusion detection module then processes legitimate traffic, analyzing its characteristics to assess potential security risks. To further improve adaptability against emerging threats, the IDS incorporates federated deep learning, aggregating threat intelligence from multiple edge nodes while maintaining privacy. Additionally, GAN-generated synthetic attack simulations are integrated into the IDS training process, enabling it to identify complex and previously unseen cyber threats.

To achieve high detection precision while minimizing false positives, the system implements a thresholding mechanism, flagging network sessions as malicious only if their suspicion score exceeds a predefined threshold. This decision-making process is context-aware, taking into account factors such as time of day, typical user behavior, and device activity patterns. Once a threat is confirmed, real-time notifications are triggered via the Electron.js-based desktop application, providing administrators with detailed threat reports, including attack type (e.g., DDoS attempt), source IP address, and recommended mitigation actions. Designed for both technical and non-technical users, the notification system allows administrators to review logs, visualize network traffic trends, and configure security policies in an intuitive interface.

Upon detecting and confirming malicious network activity, the system enforces router-level mitigation measures, dynamically applying security policies such as throttling suspicious connections, dropping harmful packets, or updating firewall rules to neutralize ongoing threats. By integrating adversarial learning, adaptive intrusion detection, and automated threat response, this GAN-enhanced IDS framework strengthens smart home network security while ensuring efficient and precise real-time threat mitigation.

5 Evaluation

In this section, we first present a detailed overview of the experimental environment, including the design and deployment of both local and edge servers in a real-world setting. We then describe the dataset and evaluation metrics used for the federated learning (FL), Generative Adversarial Networks (GANs), and intrusion detection system (IDS) integration. This is followed by an analysis of experimental results, examining the system's performance across various simulated and real-world attack scenarios. Finally, we provide a discussion on the key findings, highlighting the effectiveness and limitations of the proposed framework.

5.1 Experiment Setup

Building on the system's design, the experimental setup ensures seamless integration between the router, local server, and federated edge nodes to facilitate real-time intrusion detection. The desktop application serves as the primary interface for network configuration, security monitoring, and response management, requiring users to provide essential router credentials, including IP address, SSH port, and administrative login details during the initial setup. These parameters are securely stored using encrypted configuration files, streamlining subsequent access without compromising security. Once configured, the application automatically initializes a local Python server, which establishes a secure SSH connection to the OpenWRT router via the Paramiko library.

With this secure channel in place, the local server retrieves real-time router diagnostics, such as CPU utilization, active network interfaces, and firewall configurations, ensuring continuous monitoring of system performance. It then initiates packet capture sessions using tcpdump or Wireshark, enabling network traffic analysis directly at the router level. Captured raw PCAP files are securely transferred to the local server via Paramiko, where they undergo preprocessing to filter out benign or whitelisted traffic. The extracted feature set is then transmitted to federated edge servers for distributed learning and collaborative threat detection, reducing computational load on the home router while preserving data privacy.

The local server, deployed on an Intel NUC with 8GB RAM, is responsible for handling initial data preprocessing and coordinating secure data transmission to federated nodes. The federated learning framework is implemented across two AWS virtual machines, each configured with 4 vCPUs and 8GB RAM, serving as edge computing nodes for distributed training. These edge servers aggregate intrusion detection models from multiple devices, refining security policies before relaying optimized models back to the local network for adaptive real-time protection.

A lightweight Flask-based service running on the local server coordinates the exchange of model updates and inference results between the federated learning nodes. Once the edge servers complete their analysis, the Flask server retrieves the refined threat intelligence and updated model parameters, integrating them into the local security framework. The desktop application then visualizes security insights, triggers real-time alerts, and enables user-configurable response mechanisms, providing a comprehensive and adaptive intrusion detection system within a single unified interface. This experimental setup ensures low-latency threat detection, decentralized model training, and privacy-preserving security monitoring for smart home environments. To clarify the hardware and software setup used in our experiments, Table 1 summarizes the specifications of each system component.

5.2 Dataset Discussion

The evaluation of our intrusion detection system (IDS) is conducted using three datasets: a real-world IoT traffic dataset, the [8] benchmark dataset, and a

Table 1. Experimental Setup Specifications

Component	Specifications	Function
Router	OpenWRT-based GL-MT300N-V2	Network traffic capture via tcpdump
Local Server	Intel NUC, 8GB RAM, Ubuntu 20.04	Data preprocessing, SSH, model updates
Edge Server 1/2	AWS EC2 VM, 4 vCPUs, 8GB RAM	Federated model training, threat analysis
Desktop App	Electron.js, React, Flask Backend	UI for alerts and network monitoring

synthetic traffic dataset generated by our GAN model. The real IoT traffic dataset comprises labeled normal and malicious traffic traces, capturing real-world adversarial behaviors such as port scans and distributed denial-of-service (DDoS) attacks. This dataset provides a ground truth reference for evaluating the system's ability to detect common cyber threats.

The CICIoT2023 traffic dataset, a widely recognized benchmark for IoT intrusion detection, includes a diverse mix of benign and attack traffic patterns, offering a comprehensive evaluation ground for IDS models. By incorporating multiple attack types, it allows for robust validation of the system's detection accuracy and generalization capabilities.

Additionally, our GAN model generates a synthetic traffic dataset designed to simulate adversarial attack patterns, improving the IDS's resilience against novel and evasive threats. While the discriminator in the GAN is trained on both real and synthetic network traffic to differentiate between legitimate and malicious activity, the generator does not interact directly with real network packets. Instead, it produces artificial adversarial samples that enhance the IDS's ability to detect previously unseen attack strategies. The IDS integrates all three datasets, leveraging real-time packet captures, labeled traffic from CICIoT2023, and synthetic adversarial patterns to refine its detection capabilities and adapt dynamically to evolving cyber threats.

5.3 Results

Figure 6 illustrates the comparative performance of the proposed Network Intrusion Detection System (NIDS) trained on both real and augmented datasets over five epochs. The evaluation focuses on key performance metrics, including accuracy, precision, recall, and loss, to assess how the system adapts and improves during training. The baseline performance, represented as "AVG Real Data," reflects the model's initial capabilities when trained solely on real-world traffic. At this stage, while accuracy is relatively high, both precision and recall exhibit moderate values, suggesting potential limitations in detecting minority-class attack patterns due to class imbalance. Loss values at this stage are also higher, indicating the model's initial struggle in distinguishing between normal and attack traffic.

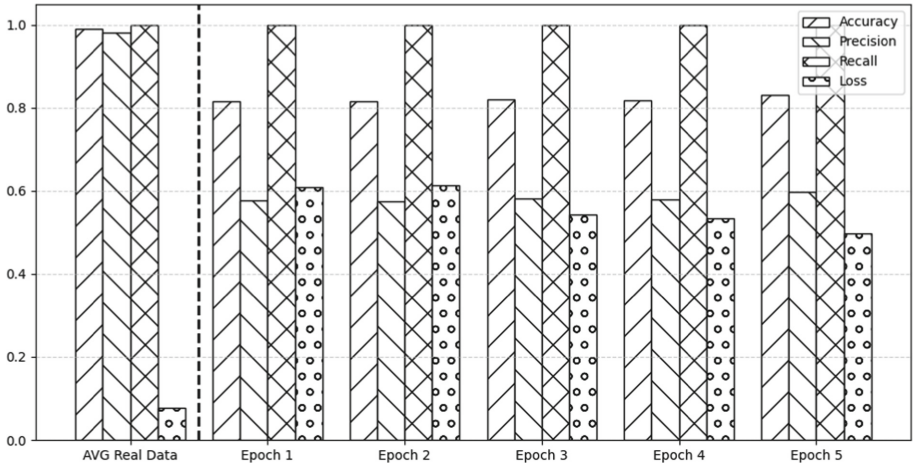


Fig. 6. Performance Comparison of NIDS Using Real Data vs. GAN-Augmented Data Over Five Training Epochs.

Table 2. Comparative Performance Metrics of NIDS on Real vs. GAN-Augmented Data Over Five Training Epochs

Data Type	**Accuracy**	**Precision**	**Recall**
Augmented	0.8206 ± 0.0066	0.5823 ± 0.0091	0.9996 ± 0.0001
Real	0.9908 ± 0.0005	0.9823 ± 0.0009	0.9996 ± 0.0001
Data Type	**AUC**	**Loss**	**LogCosh**
Augmented	0.9859 ± 0.0045	0.5596 ± 0.0499	0.0640 ± 0.0031
Real	0.9925 ± 0.0003	0.0781 ± 0.0024	0.0042 ± 0.0002

Table 2 further quantifies the impact of GAN-based augmentation by providing average model performance metrics across five training epochs. The results indicate that while models trained on real data achieve higher accuracy (0.9908) and precision (0.9823), they exhibit lower robustness against evolving adversarial threats. In contrast, models trained with augmented data demonstrate superior recall (0.9996), ensuring that fewer attack instances go undetected. However, the lower precision (0.5823) suggests an increased false positive rate, likely due to the diversity of synthetic attack samples generated by GANs. The AUC score remains high for both datasets, reaffirming the model's effectiveness in classifying network traffic. Notably, the loss value for augmented data (0.5596) is significantly higher than that of real data (0.0781), indicating that adversarial training introduces additional complexity in learning decision boundaries. Similarly, the LogCosh loss, which measures gradient smoothness, is noticeably higher for augmented data, reflecting the model's exposure to more diverse and complex attack patterns.

As training progresses, the model exhibits significant improvements across all key metrics. By Epoch 1, the integration of GAN-generated adversarial samples enhances the model's ability to recognize attack behaviors, evident from the steady increase in recall and precision. The recall metric continues to improve in subsequent epochs, demonstrating that the model becomes progressively more adept at detecting previously unseen attack patterns while reducing false negatives. The overall accuracy remains consistently high, validating the effectiveness of federated learning and GAN-based augmentation in reinforcing the model's decision boundaries. Additionally, the loss function shows a clear decreasing trend, indicating stable convergence and reduced classification errors. By Epoch 5, the model achieves its optimal performance, with near-maximal accuracy, precision, and recall, affirming the robustness of the proposed intrusion detection framework in recognizing sophisticated cyber threats with minimal performance degradation.

The quantitative results in Table 3 further support the efficacy of the proposed method in differentiating between attack and benign network traffic. The model successfully identifies 3,919 attack instances from a total of 4,000, demonstrating high sensitivity to malicious activities. The false positive rate remains exceptionally low, with only one benign instance being misclassified as an attack, minimizing unnecessary security alerts. Additionally, the system correctly classifies 39,999 benign network flows while only misidentifying 81 attack samples as normal traffic. This low number of false negatives highlights the model's effectiveness in mitigating the risk of undetected intrusions, a critical factor in real-world cybersecurity applications. The combination of federated learning and adversarial training ensures that the model remains adaptable to emerging attack strategies, improving its generalization across different IoT environments.

Table 3. Intrusion Detection Results in terms of Classification Performance on Attack and Benign Network Traffic

Traffic Type	True Positives	False Positive	True Negatives	False Negatives	Total Traffic
Attack	3919	1	39999	81	4000
Benign	39999	81	3919	1	40000

5.4 Results Discussion

Our experimental results validate the efficacy of a hierarchical federated generative learning approach in detecting advanced cyber threats in near-real time, particularly within resource-limited home networks. Traditional cloud-based security mechanisms often face challenges in balancing latency, privacy, and computational efficiency, whereas our distributed learning framework successfully mitigates these issues by enabling local inference and edge-based model updates. The

hierarchical model design not only enhances network security but also presents a viable commercial-grade ISP solution, leveraging edge computing for large-scale deployment. As IoT adoption continues to expand, this approach supports broader device-to-device collaboration, making it increasingly relevant in the modern cybersecurity landscape.

However, several limitations must be acknowledged. Deploying this system at scale would introduce high computational demands on edge devices, increasing operational costs for ISPs and potentially limiting widespread adoption. Many ISPs may be reluctant or unable to invest in high-performance edge computing infrastructures, thereby constraining the system's applicability. Another critical limitation is compatibility with OpenWRT. Although OpenWRT serves as a widely adopted open-source router firmware, not all router models support its full functionality due to hardware limitations. This incompatibility restricts the system's feasibility across diverse home networking environments, highlighting the need for broader hardware support or alternative integration strategies.

6 Conclusion and Future Work

This paper presented a Hierarchical Federated Generative Learning-Based Smart Home System for real-time intrusion detection at the router level. By integrating tcpdump-based local captures, a Paramiko-driven local server, and GAN-augmented federated learning at the edge, our approach enhances threat detection accuracy while preserving user privacy. Additionally, an Electron.js-based desktop application enables intuitive network monitoring and security management for non-expert users. Experimental results demonstrate reduced detection latency, improved accuracy, and strong privacy preservation, reinforcing the system's potential to secure modern home networks.

Future research will focus on optimizing computational efficiency by incorporating a Mixture of Experts (MoE) model, which can dynamically allocate resources for different cyber threats, improving scalability across resource-limited home routers. Additionally, improving GAN stability through adaptive loss functions and hyperparameter tuning will enhance attack simulations, ensuring more reliable adversarial learning. Finally, expanding the framework for cross-domain collaboration with ISP infrastructures could enable large-scale, privacy-preserving threat intelligence sharing, strengthening cybersecurity across interconnected networks.

Acknowledgments. This research was supported in part by the Florida Gulf Coast University (FGCU) Head-Start Funding and funding from DENDRITIC Human-Centered AI and Data Science Institute. We thank our collaborators at the University of Florida and Lucerne University of Applied Sciences and Arts for invaluable feedback.

Competing Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. de Araujo-Filho, P.F., Kaddoum, G., Campelo, D.R., Santos, A.G., Macêdo, D., Zanchettin, C.: Intrusion detection for cyber-physical systems using generative adversarial networks in fog environment. IEEE Internet Things J. **8**(8), 6247–6256 (2020)
2. Ashraf, A., Zhu, X., Liu, J., Rauf, Q., Firdaus, R.: Usability evaluation framework of smart home applications for senior citizens. In: 2022 12th International Conference on Software Technology and Engineering (ICSTE), pp. 29–39. IEEE (2022)
3. He, P., et al.: A survey of internet of medical things: technology, application and future directions. Digital Communications and Networks (2024)
4. Kostage, K., et al.: Federated learning-enabled network incident anomaly detection optimization for drone swarms. In: ICDCN '25: Proceedings of the 26th International Conference on Distributed Computing and Networking, pp. 104–114. ACM (2025). https://doi.org/10.1145/3700838.3700857, https://doi.org/10.1145/3700838.3700857
5. Kostage, K., West, D., Meinert, T., Qu, C., Calyam, P., Mazzola, L.: Enhancing autonomous intrusion detection system with generative adversarial networks. In: 2024 IEEE 20th International Conference on e-Science (e-Science), pp. 1–10 (2024). https://doi.org/10.1109/e-Science62913.2024.10678662
6. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data (2023). https://arxiv.org/abs/1602.05629
7. Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A.Y., Tari, Z.: Explainable intrusion detection for cyber defences in the internet of things: opportunities and solutions. IEEE Commun. Surv. Tutorials **25**(3), 1775–1807 (2023). https://doi.org/10.1109/COMST.2023.3280465
8. Neto, E., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., Ghorbani, A.A.: Ciciot 2023: a real-time dataset and benchmark for large-scale attacks in iot environment. Sensors **23**(13), 5941 (2023)
9. Pu, X., Lei, T., Wen, W., Feng, W., Wang, Z., Chen, Q., Jin, S.: Incentive mechanism and resource allocation for collaborative task offloading in energy-efficient mobile edge computing. IEEE Trans. Veh. Technol. **72**(10), 13775–13780 (2023)
10. Sadotra, P., Thakur, G., Chouksey, P., Nayak, M.H., Chopra, M.: Intrusion detection in smart homes: a comprehensive review. Environment **4**, 11 (2024)
11. Wickramasinghe, C.I.: Best-practice-based framework for user-centric privacy-preserving solutions in smart home environments. In: International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 101–120. Springer (2022)
12. Yu, Z., Xiangting, L., Yalin, S., Aidi, X., Yi, Z., Hailong, J., Weizheng, S.: Real-time monitoring method for cow rumination behavior based on edge computing and improved mobilenet v3. Smart Agriculture **6**(4), 29 (2024)



ExpertGen: A Comparative Analysis of User Performance, Cognitive Workload, and Trust in Domain-Tailored Generative AI

Nina Jiang[1], Wei Zhou[2], Sogand Hasanzadeh[2], and Vincent G. Duffy[1,3(✉)]

[1] Purdue University, Edwardson School of Industrial Engineering, West Lafayette, IN 47907, USA
duffy@purdue.edu
[2] Purdue University, Lyles School of Civil and Construction Engineering, West Lafayette, IN 47907, USA
[3] Purdue University, Agricultural & Biological Engineering, West Lafayette, IN 47907, USA

Abstract. Generative artificial intelligence has transformed multiple industries, including education, healthcare, and finance, by generating text, images, video, and audio content. Despite these advancements, existing generative AI models often fail to fulfill all expectations, generating factually incorrect or misleading responses, which impairs system accountability and user trust. To overcome this limitation, we designed ExpertGen, a generative AI system tailored for specific domains, designed to provide real-time, accurate, and context-specific insights. This study evaluates the effectiveness of ExpertGen on three major dimensions: learning performance, cognitive workload, and user trust. Experimental results show that ExpertGen significantly improves learning outcomes, with an average performance of 21.08 points higher than traditional e-books, and outperforming ChatGPT in domain-specific tasks. In addition, NASA-TLX cognitive workload analysis proves that ExpertGen significantly reduces the cognitive demands of e-books and its usability is comparable to ChatGPT. Trust analysis shows that users trust ExpertGen more than ChatGPT. These studies suggest that generative AI systems need to enhance reliability, transparency, and domain adaptability to improve user acceptance and educational outcomes.

Keywords: Generative AI · Trust · Cognitive Workload

1 Introduction

The development of Generative Artificial Intelligence (AI) revolutionized various fields, including education, healthcare, and business. Generative AI is an AI model that can generate content, including text, images, video, and audio

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 178–190, 2025.
https://doi.org/10.1007/978-3-031-93502-2_12

[4,9,18]. The generated content from the models, which are usually based on deep learning and transformer architectures, are based on learned patterns from training data [7,16]. The current generative system, while highly capable of generating human-like text, often suffers from hallucinations, where the generated content is factually incorrect or nonsensical, sometimes disconnected from the source content provided [10,19]. The seemingly plausible, but inaccurate and misleading responses undermine the system's accountability and hinder users' trust in the system [11,19].

To bridge this research gap, we develop a generative AI system, ExpertGen, which delivers real-time, domain-specific insights, enhancing learning experiences and decision-making processes across specialized fields [12]. Unlike conventional systems, such as Chat-GPT or static eBook formats, ExpertGen is optimized to meet the unique requirements of specialized domains. This tailored approach is expected to result in more efficient learning experiences, reduced mental demands, and increased user confidence in the information provided.

Therefore, to evaluate the effectiveness of ExpertGen, we conducted a comprehensive experimental study comparing it against both general-purpose Chat-GPT and traditional eBook methods. Our evaluation focused on three primary dimensions: performance improvement, cognitive workload as measured by the NASA Task Load Index (NASA-TLX), and user trust. The results from these experiments not only underscore ExpertGen's superior performance and reduced cognitive demands but also reveal a higher level of user trust, thereby validating our hypothesis that a domain-tailored generative AI system can significantly enhance the overall user experience.

In the following sections, we review the related work, introduce ExpertGen, describe our experimental methodology, present and discuss the results, and conclude with insights and future directions for research in domain-tailored generative AI systems.

2 Literature Review

To conduct the comprehensive literature review, we systematically collected bibliometrics data from the Web of Science database. The keywords "generative AI" and "generative artificial intelligence" were chosen to capture a broad range of literature on AI models that generate text, images, video, and audio, ensuring coverage of both technical advancements and practical applications in fields like education, healthcare, and business. Using both variations allows for a comprehensive search that includes discussions on foundational technologies, challenges like hallucinations and trustworthiness, and improvements in domain-specific AI systems such as ExpertGen.

The Fig. 1 shows the cumulative occurrence trend of multiple artificial intelligence (AI) related terms from 2018 to 2025. The cumulative growth of the term "Artificial Intelligence" is the most significant, especially after 2023, it shows an exponential increase, indicating that artificial intelligence technology has received widespread attention during this period and has become a core

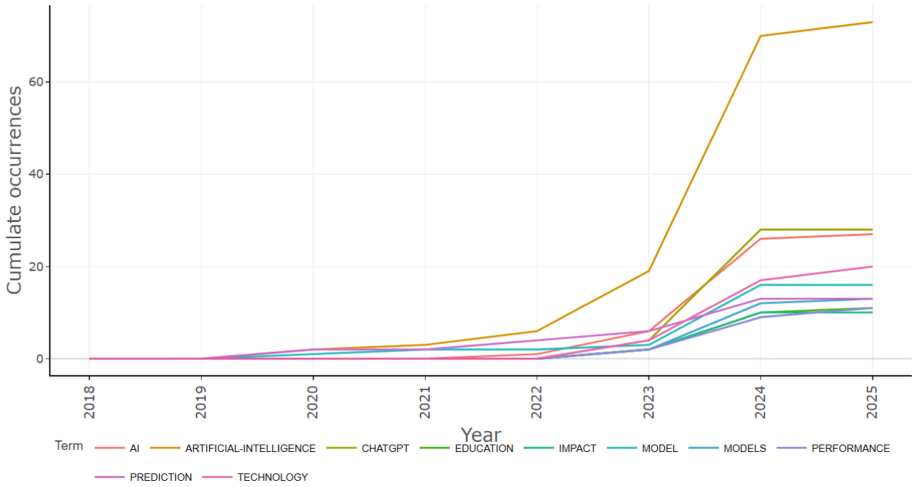


Fig. 1. Trend Analysis

issue both in academic research and industrial development [8]. At the same time, the frequency of occurrence of terms such as "AI" and "Technology" has also increased significantly after 2023, indicating that the technological evolution and application scenarios of artificial intelligence are constantly growing.

In addition, the cumulative growth of keywords such as "ChatGPT", "Education", and "Impact" shows that the application and influence of generative artificial intelligence (such as ChatGPT, DALL-E, Google Gemini, etc.) in the field of education have become important research directions [1,4]. The overall trend shows that artificial intelligence technology has entered a stage of rapid development in recent years and has shown far-reaching influence in many fields. It is expected to continue to grow in the future and promote interdisciplinary integration and innovation. The overall trend shows that generative artificial intelligence has entered a stage of rapid development in recent years and has demonstrated profound impacts in many fields [3,5,15]. It is expected to continue to grow in the future and promote interdisciplinary integration and innovation.

Trend analysis and keyword clustering analysis jointly reveal the path of technology evolution and the focus of application expansion. As illustrated in Fig. 2, this picture shows the keyword clustering network related to artificial intelligence, revealing the core of current research and the trend of technological evolution. From the overall framework, "artificial intelligence" is at the core of the network, indicating its extensive influence in multiple fields as a basic technology. Among them, the green cluster mainly centered around ChatGPT and its applications, wrapping keywords such as "large language models", "education", and "user acceptance", reflecting the deep penetration of generative artificial intelligence in knowledge generation, educational innovation, and human-computer interaction, especially with the focus of user experience and learning methods [6,7].

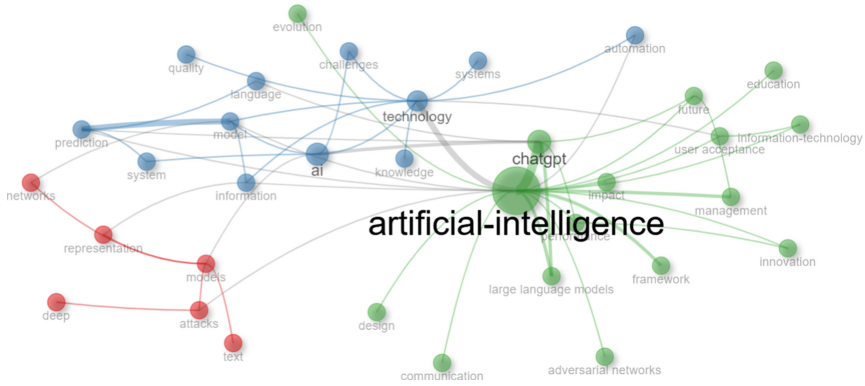


Fig. 2. Keyword.

Correspondingly, the blue clusters focus on the use of underlying technologies in the field of artificial intelligence, including "systems", "technology" and "language". This also illustrates that the development of generative artificial intelligence is inseparable from the continuous improvement of core supporting technologies such as computing architecture, algorithms, and natural language processing. Digging deeper into the technical challenges, which are in red, such as models, networks, and attacks, shows that generative AI is still facing severe problems in terms of security, robustness, and explainability. In particular, the challenge of transparency is highlighted as the research bottleneck in this field [2,11].

To better illustrate the state-of-the-art research of generative artificial intelligence and its application, Fig. 3 presents the thematic map to provide a scientific basis for academic research, industrial development, and policy-making by quantifying centrality and density, systematically reveals research hotspots, basic technologies, emerging trends, and marginalized topics in the field of AI [14]. This figure shows that topics such as Generative AI, Chatbots, and Bard have become the core driving force of current generative AI research, promoting the development of content generation, human-computer interaction, and intelligent decision-making. At the same time, machine learning, deep learning, and generative adversarial networks (GANs) are still the technical foundations of the development and progress of generative AI, while large language models (LLMs) are gradually moving from specific applications to core research and becoming a key direction in the future [13].

In addition, the results also indicate a shift in the trend of research directions. Diffusion models are gradually replacing GANs and becoming a more efficient content generation technology. It not only provides a systematic analysis framework for academic research but also effectively supports industry development.

To reveal the evolution of keyword trends, Fig. 4 presents a Three-Field Plot that shows the research topic changes in the field of Generative AI. From 1990 to 2020, Generative AI is still lack of development, early AI research mainly

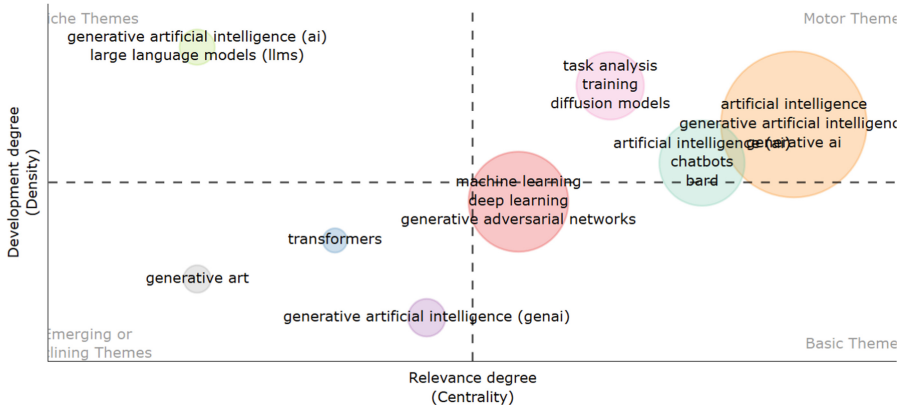


Fig. 3. Thematic Map.

focused on artificial intelligence and prediction, such as algorithm development and data-driven predictive models. Entering 2021 and 2022, artificial intelligence remains a core research direction, but the academic community has begun to pay attention to the limitations of AI systems, such as explainability and ethical issues.

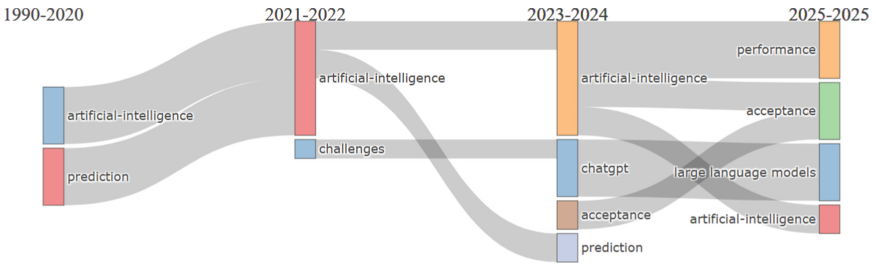


Fig. 4. Three Field Plot.

Between 2023 and 2024, the rapid development of generative AI has led to a shift in research topics, with ChatGPT and user acceptance becoming key topics, indicating that the research has expanded from underlying technology to user experience and social impact. In addition, prediction remains, indicating that the application of generative AI in decision support continues to receive attention. By 2025, the research focus further shift to large language models (LLMs), performance optimization, and user acceptance, showing that future AI research will not only focus on improving model capabilities but also on its social adaptability and practical application effects.

Overall, the three-domain diagram reveals the trend of AI research from early prediction models, to technical optimization challenges, to the integration

of generative AI and user acceptance, and finally to performance optimization and social impact. This evolution shows that AI technology is moving from pure algorithm-driven to application-driven and user experience optimization while emphasizing researchers' continued attention to the reliability, explainability, and social acceptance of generative AI.

Although generative AI has made significant progress, the above bibliometric analysis shows that challenges such as hallucination, credibility, and user acceptance remain critical challenges for the application of generative AI. From thematic analysis, research has gradually focused on the accuracy, performance, and social impact of generative AI, especially in the fields of education, healthcare, and decision-making. The rapid rise of large language models (LLMs) and their integration with numerous applications have highlighted the urgency of developing domain-specific generative AI systems that are more reliable and better aligned with user needs and expectations. Given these issues, there are still critical gaps in ensuring that AI-generated content is both accurate and contextually relevant, especially in high-risk areas where misinformation can have serious consequences. To overcome these limitations, our study introduces ExpertGen, which is a generative AI system that aims to provide real-time, domain-specific insights to improve user experience across various professional domains.

3 Method

ExpertGen is a domain-tailored generative AI system designed to overcome the limitations of general-purpose language models in specialized contexts [12]. By integrating domain-specific data and providing cited sources, ExpertGen delivers precise, context-aware responses that enhance both the reliability of information and user trust. The experiment aims to evaluate ExpertGen and compare it with general large language models, and eBook-based learning methods. In addition, it also assesses the impact of various learning methods on users' cognitive workload, system usability scale, and their trust in the system.

3.1 Hypothesis

Based on identified needs for comparative analysis of AI-assisted learning tools, this experiment aims to examine the following:

- Does ExpertGen lead to better learning outcomes?
- Is the cognitive workload reduced by using ExpertGen than Chat-GPT?
- Does the user trust more on the ExpertGen than Chat-GPT?

Participants need to review the consent form and complete the pre-test questionnaire. Then, the participants will start learning with the first assigned method (eBook, ExpertGen, or Chat-GPT) and different learning materials. The order of the assigned methods and materials is randomly assigned. Then, the participants will take two quizzes regarding the learning materials: an open-book quiz

and a closed-book quiz. Once the quiz is completed, participants will be asked to complete a post-test questionnaire, which includes a mental workload questionnaire (NASA-TLX) and a trust evaluation matrix. Then, the participants will learn the second and third randomly assigned methods with the same procedure as described above.

3.2 Experiment Procedure

As shown in Fig. 5, participants need to review the consent form and learn the objectives of the experiment along with the instructions on conducting the experiment. Then an IRB-approved consent form was provided, which they signed prior to beginning the experiments.

After confirming that the participants fully understand the upcoming procedures, they will be requested to complete a pre-test questionnaire. All questionnaires in this experiment are deidentified by making sure that all participants are using the assigned participant code provided by the researcher.

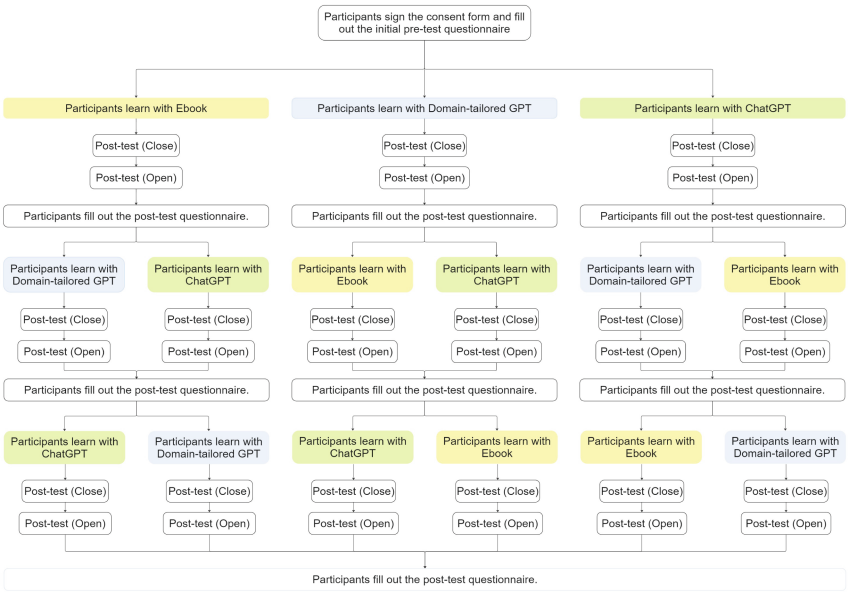


Fig. 5. Experiment Design.

After completing the pre-test questionnaire, participants will begin their learning assisted by the first randomly assigned method-either eBook, Domain-tailored GPT, or Chat-GPT-paired with one of the designated learning materials (human factors, statistics, or an introduction to machine learning).

Following this learning session, participants will take two quizzes based on the material: an open-book quiz and a closed-book quiz. These quizzes are designed

to assess performance improvements resulting from the interaction with the assigned learning method. Each quiz consists of 10 multiple-choice questions, randomly selected from a dedicated quiz pool, and participants will have up to 10 min to complete each one-though they may submit their answers as soon as they feel ready. Notably, during the open-book quiz, participants are permitted to interact with the assigned learning method.

Once the quiz is completed, participants will be asked to fill out a post-test questionnaire that includes a mental workload assessment (NASA-TLX) and a trust evaluation matrix. The same procedure will then be repeated for the second and third randomly assigned learning methods.

4 Results

4.1 Performance Improvement

The aim of the improved performance analysis is to evaluate our primary hypothesis: Does the domain-tailored GPT assistant cause better learning outcomes? To explore this, we collected quiz results from participants and determined the performance gain by calculating the difference between their open-book and closed-book quiz scores.

By utilizing the least significant difference (LSD) analysis, we assess the significance of the mean performance differences among the methods-eBook, Chat-GPT, and domain-tailored GPT. By applying LSD analysis to the calculated mean differences, we can identify which pairwise comparisons show statistically significant variations. This detailed analysis enables us to clearly illustrate the relative efficiency of each method compared to the others.

ExpertGen outperformed the eBook method by an average of 21.08 points, and this difference was statistically significant ($p < 0.001$). Figure 6 presents the visualization of the improvement in student performance. This analysis indicates that ExpertGen's performance improvements are significantly greater than those achieved by both the eBook and Chat-GPT methods ($p < 0.05$), thereby supporting our hypothesis that ExpertGen assistance leads to superior learning outcomes.

4.2 Cognitive Workload

To evaluate the cognitive workload associated with each method, all participants completed the NASA Task Load Index (NASA-TLX) as part of the post-test questionnaire [17]. Figure 6 presents the mean values for each NASA-TLX factor. The analysis reveals that the domain-tailored GPT method is the most effective at reducing cognitive workload, followed closely by Chat-GPT, with the eBook method being the least effective.

When comparing domain-tailored GPT with Chat-GPT, the mean difference was -0.53, which is not statistically significant ($p = 0.405$, exceeding the typical 0.05 threshold). This suggests that both methods have similar cognitive workloads, although domain-tailored GPT exhibits a slightly lower cognitive workload

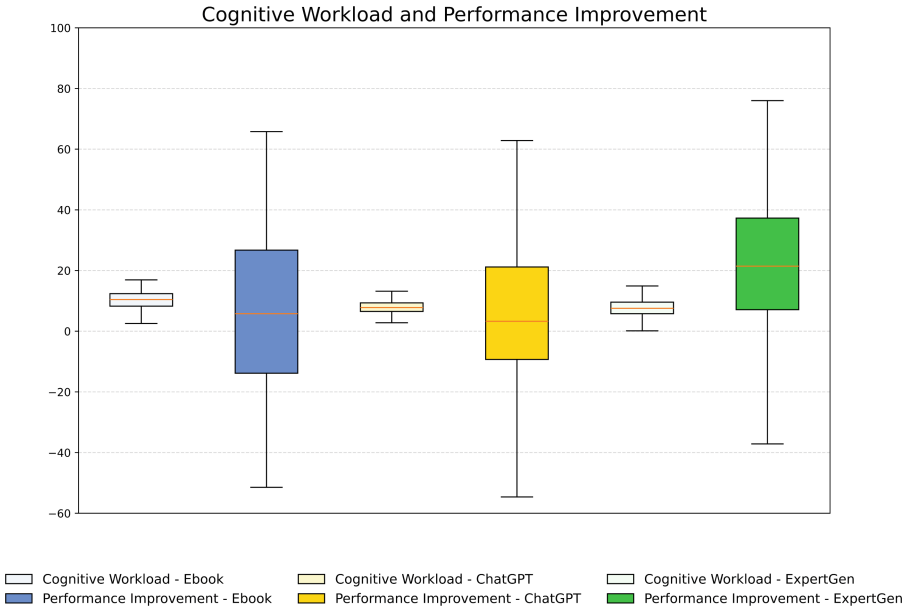


Fig. 6. Cognitive Workload and Performance Improvement

based on the mean values. In contrast, the comparison between domain-tailored GPT and the eBook method revealed a highly significant mean difference of -2.71 ($p < 0.001$), indicating that the cognitive workload for domain-tailored GPT is significantly lower than that for the eBook method.

The comparison between Chat-GPT and the eBook method yielded a statistically significant mean difference of -2.17 ($p < 0.001$), indicating that the cognitive workload for Chat-GPT is considerably lower than that of the eBook. Overall, both Chat-GPT and domain-tailored GPT are associated with a significantly reduced cognitive workload compared to the eBook method.

In this study, NASA-TLX analysis reveals significant differences in cognitive workload across the three methods. Domain-tailored GPT exhibits a lower cognitive workload compared to both Chat-GPT and the eBook method, with the difference between domain-tailored GPT and eBook being highly significant. Similarly, Chat-GPT demonstrates a reduced cognitive workload relative to the eBook method. These results indicate that both domain-tailored GPT and Chat-GPT facilitate more efficient management of cognitive demands, with domain-tailored GPT being the most effective in reducing user cognitive workload. Conversely, although Chat-GPT outperforms the eBook method in terms of cognitive workload, it is not as optimized as the domain-tailored GPT model. This discrepancy is likely attributable to the sophisticated capabilities of both the domain-tailored GPT and Chat-GPT models in comprehending questions and generating responses, thereby providing more efficient user interaction compared to the traditional, less interactive format of an eBook.

The eBook method exhibits the highest cognitive workload among the three methods, which may be attributed to its inherent characteristics of being less interactive and more static. Consequently, users may experience greater difficulty navigating the content and extracting relevant information, leading to increased mental demands.

4.3 Trust Towards the System

Establishing trust in AI systems is crucial for their acceptance and appropriate utilization, emphasizing the need for reliable trustworthiness measures. Accordingly, we employ Hamid's trust evaluation matrix [6] to assess the trust levels that participants attribute to the different systems. Figure 7 offers an intuitive visualization of users' trust across these various tools.

Table 1. Trust Analysis.

(I) Method	(J) Method	Mean Difference (I-J)	Std. Err	P-value	95% Confidence Interval	
					Lower Bound	Upper Bound
Domain-tailored GPT	Chat-GPT	0.40878*	0.15851	.011	0.0946	0.7230
	eBook	−0.43919*	0.15851	.007	−0.7534	−0.1250
Chat-GPT	Domain-tailored GPT	−0.40878*	0.15851	.011	−0.7230	−0.0946
	eBook	−0.84797*	0.15851	.000	−1.1622	−0.5338
eBook	Domain-tailored GPT	0.43919*	0.15851	.007	0.1250	0.7534
	Chat-GPT	0.84797*	0.15851	.000	0.5338	1.1622

The analysis indicates significant differences in trust across the various methods. Specifically, the eBook consistently achieves higher trust scores compared to both domain-tailored generative AI and Chat-GPT. While GPT is trusted more than Chat-GPT, it does not achieve the level of trust accorded to the eBook, establishing a clear trust hierarchy with the eBook at the top.

Table 1 summarizes the trust analysis by presenting the mean differences, standard errors, significance levels, and 95% confidence intervals for each pairwise method comparison. This table delineates the trust associated with each method and highlights where significant differences exist.

Therefore, key observations from the data include: The eBook method demonstrates significantly higher trust than both domain-tailored GPT and Chat-GPT, with the pairwise comparisons yielding p-values well below the 0.05 threshold.; ExpertGen is significantly more trusted than Chat-GPT, as indicated by a p-value of 0.011 ($p < 0.05$).

These findings clearly illustrate the trust hierarchy among the methods, with the eBook emerging as the most trusted tool.

This analysis reveals significant differences in user trust among the evaluated methods. Notably, the eBook method consistently exhibits higher trust scores compared to both the ExpertGen and Chat-GPT systems. Figure 7 offers an intuitive visualization of these trust differences. In contrast, while domain-tailored

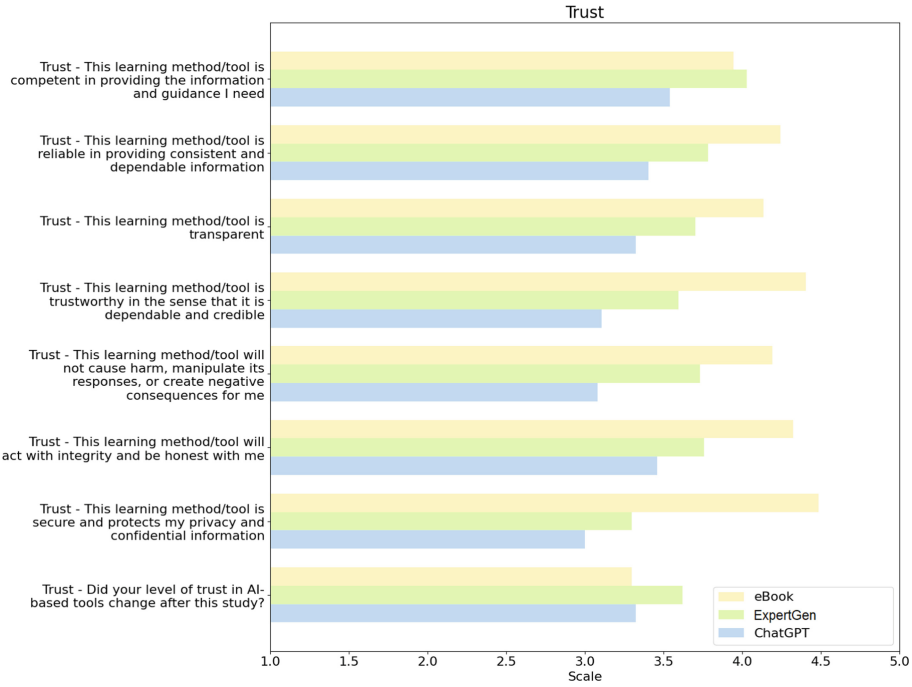


Fig. 7. Trust

GPT is trusted more than Chat-GPT, it still falls short of the trust level afforded to eBooks, suggesting a clear hierarchy with eBooks being the most trusted.

One plausible explanation for the higher trust in eBooks is their familiarity and long-established reputation as reliable sources of information. Unlike newer methods such as ExpertGen and Chat-GPT, eBooks have a well-documented history as a consistent and trusted medium for delivering educational content. This established credibility likely contributes to the elevated trust scores observed in our study. Moreover, the structured and comprehensive nature of eBooks may enhance users' feelings of security and confidence in the information provided.

Conversely, the increased confidence in the ExpertGen can be attributed to its specialized design. By integrating domain-specific data, the system is better equipped to address the unique needs and contextual nuances of a particular field, resulting in more relevant and context-aware responses. This specialization enables the system to deliver high-quality, concise, and easily understandable explanations, thereby fostering greater user trust.

5 Conclusion and Future Work

In this study, we addressed several critical gaps in applying generative AI models to customized domains by developing ExpertGen and conducting experimental

evaluations. The results demonstrated that ExpertGen significantly enhanced user performance while substantially reducing cognitive workload compared to both general-purpose models and eBooks. Moreover, the experiment demonstrated that users placed greater trust in ExpertGen than in the more general generative AI model. This increased trust can be attributed to ExpertGen's inclusion of cited sources and the heightened relevance of its responses based on the domain-tailored databases, which collectively strengthen users' confidence in the information provided. By delivering accurate and contextually consistent responses, ExpertGen further enhances credibility and reliability. Overall, these findings underscore the importance of developing domain-tailored generative AI systems to improve both system performance and the user experience.

References

1. Adetayo, A.J.: Reimagining learning through ai art: the promise of dall-e and midjourney for education and libraries. Library Hi Tech News, ahead-of-print (2024)
2. Bhanu, C., Chinju, J., Lekha, W., Kumaravelan, G.: Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. ACM Computing Surveys, 2 2025
3. Chen, Y., et al.: It3d: Improved text-to-3d generation with explicit view synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 1237–1244 (2024)
4. Cheong, M., et al.: Investigating gender and racial biases in dall-e mini images. ACM J. Responsible Comput. **1**, 1–20 (2024)
5. Chiu, T.K.F.: The impact of generative ai (genai) on practices, policies and research direction in education: a case of chatgpt and midjourney. Interactive Learning Environments, 11 2024
6. Choudhury, A., Shamszare, H.: The impact of performance expectancy, workload, risk, and satisfaction on trust in chatgpt: Cross-sectional survey analysis. JMIR Hum. Factors **11**, e55399 (2024)
7. Du, H.: The age of generative ai and ai-generated everything. IEEE Network (2024)
8. Giannakos, M., et al.: The promise and challenges of generative ai in education. Behaviour & Information Technology, 8 2024
9. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: large-scale pretraining for text-to-video generation via transformers. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
10. Lei, H.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**, 1–55 (2025)
11. Jiang, N., Duffy, V.G.: Use of artificial intelligence for training: a systematic review. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14711. LNCS, pp. 346–363 (2024)
12. Jiang, N., Hasanzadeh, S., Duffy, V.G.: Domain-tailored generative ai for personalized assistant. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 15376. LNCS, pp. 233–249 (2025)
13. Li, Y., Min, M.R., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 7065–7072 (2018)

14. Manduchi, L., et al.: On the challenges and opportunities in generative ai. Meta, February 2024
15. Noroozi, O., Soleimani, S., Farrokhnia, M., Banihashem, S.K.: Generative ai in education: pedagogical, theoretical, and methodological perspectives. Int. J. Technol. Educ. **7**, 373–385 (2024)
16. Rawte, V., Sheth, A., Das, A.: A survey of hallucination in large foundation models, September 2023
17. Rubio, S., Díaz, E., Martín, J., Puente, J.M.: Evaluation of subjective mental workload: a comparison of swat, nasa-tlx, and workload profile methods. Appl. Psychol. **53**, 61–86 (2004)
18. Singer, U., et al.: Make-a-video: text-to-video generation without text-video data. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
19. Towhidul Islam Tonmoy, S.M., et al.: A comprehensive survey of hallucination mitigation techniques in large language models, January 2024

Assessing Vision Obstruction and Safety in E-Scooter Accidents: An Integrated Approach Using Digital Human Modeling and Computer Vision

Gabrielle B. Joffe[1], Yitong Bu[1], and H. Onan Demirel[1,2(✉)]

[1] Oregon State University, Corvallis, OR 97331, USA
{joffeg,buyi,onan.demirel}@oregonstate.edu
[2] University of Strathclyde, Glasgow, UK

Abstract. The rising popularity of electric scooters (e-scooters) in the past few years has led to an increase in safety concerns. Accidents and mishaps involving e-scooters have resulted in injuries to riders as well as pedestrians. The most common injuries resulting from e-scooter accidents are head and face trauma. This paper focuses on vision obstruction-related risks involving the use of e-scooters. We propose a vision analysis methodology that integrates digital human modeling (DHM) and computer vision (CV) to quantify the visibility of e-scooter riders in a variety of traffic scenarios. The approach is presented through three case studies that analyze hypothetical scenarios involving e-scooters being operated in busy areas of Corvallis, Oregon. Each case study includes full-scale computer-aided design (CAD) models of local infrastructure and vehicles. Upon importing the models to the DHM software, Siemens Jack, dynamic footage of the traffic scene is generated for all case studies. Digital manikins of two anthropometric profiles (5^{th} percentile female and 95^{th} percentile male) are used to represent the e-scooter riders in each scenario. Then, a CV image segmentation process is used to quantify the visibility of the e-scooter rider at different timestamps of the videos. This approach is conducted under both clear and rainy weather conditions. The results from this study demonstrate how both anthropometry and weather impact the visibility of an e-scooter rider in situations where they are already at risk of being obstructed by other objects. Benefits of this methodology include saving time and generating more precise results than prior studies that aimed to quantify visibility.

Keywords: E-Scooters · Vision Obstruction · Digital Human Modeling · Computer Vision · Safety Engineering · Human Factors Engineering

1 Introduction

As electric scooters (e-scooters) have become a more common mode of transportation, the rate of accidents has increased [24,48,54]. These accidents have

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 191–209, 2025.
https://doi.org/10.1007/978-3-031-93502-2_13

often culminated in severe injuries to e-scooter riders [27], in addition to pedestrians [52]. A research study conducted by Trivedi et al. [62] revealed that 80% of 229 e-scooter riders sustained injury when falling off their scooter. At the same time, 8% were injured upon collision with a motor vehicle. In the latter case, a contributing factor was the e-scooter rider being in the driver's blind spot. [30]. These outcomes highlight the necessity of greater safety measures for e-scooters.

This paper aims to investigate the traffic safety concerns posed by the operation of e-scooters. We propose a rapid vision analysis approach based on digital human modeling (DHM) techniques and computer vision (CV) algorithms to explore specific accident scenarios involving e-scooter riders, pedestrians, and motor vehicle drivers. These accident scenarios particularly focus on vision obstruction-related events involving pedestrians, vehicles, and e-scooters.

DHM is a computational tool that enables designers to represent humans in a computer-aided design (CAD) environment to evaluate safety and ergonomics [14,21]. In a DHM design workflow, humans are depicted as digital manikins to examine the human interaction components of products, such as a driver's comfort or line of sight in a vehicle cab [40]. When DHM is employed proactively, designers can decrease the time and money spent on developing physical prototypes of their products [8,15,16]. In the same manner, another technique that has applications in safety is CV. CV is a branch of artificial intelligence (AI) in which computers are trained to recognize information from images and videos [60]. It has a wide variety of uses, ranging from object detection in autonomous vehicles [12] to cancer screening in the medical field [22].

Prior studies have demonstrated the use of DHM for analyzing vision obstruction related to vehicle design [36,55]. While these efforts, such as those written by Srinivasan and Demirel [17,56,57], showcased the advantages of using DHM to quantify vision obstruction with static simulations, this work augments the previous approaches by introducing the incorporation of dynamic simulations to capture visibility measurements in real time. More specifically, our proposed methodology injects CV to measure the percent visibility of e-scooters in a variety of traffic scenarios. Our approach shows how coupling DHM and CV provides a robust way to assist with identifying the most critical e-scooter safety concerns.

2 Literature Review

2.1 E-Scooters and Safety Concerns

Since the emergence of e-scooter sharing systems in 2017, e-scooters have drastically increased in popularity [7,39]. As of 2024, the New Urban Mobility Alliance (NUMO) recorded the presence of e-scooters in more than 1000 cities and 60 countries around the world [41]. Although e-scooters have become a widespread form of transportation due to being convenient and environmentally sustainable, there has been a rise in the number of accidents involving them [49]. These accidents are often serious, leading to injuries and casualties among riders as well as pedestrians [3].

In particular, one of the contributing factors for these accidents is the lack of consistent regulations in place for the use of e-scooters. For example, in the United States (U.S.), mandatory helmet laws vary widely, with many states only requiring helmet use for riders who are under 18 years of age [31,52]. Where not required, helmet use is uncommon [50]. For example, a study by Comer et al. revealed that less than 5% of riders in an Indiana city chose to wear helmets [11]. The articles written by Demir et al. [13] and Okudan and Dündar [44] examined e-scooter accidents in Turkey and found similarly low rates of helmet usage. As a result of not wearing helmets, one of the most common forms of injury found in e-scooter operators includes trauma to the head and face [25,52]. Specifically, a U.S. study that analyzed injuries to riders in e-scooter accidents found that 27.6% of emergency room admissions were due to head trauma [1].

Sidewalk riding also plays a role in e-scooter accidents, especially in crowded conditions [35]. For instance, the research work by Siman-Tov et al. studied injuries and fatalities involving e-scooter usage in Israel and found that injuries to riders and pedestrians have been influenced by inconsistent rules regarding sidewalk and bike lane usage [53]. Pedestrians who were struck by e-scooters have suffered brain injuries [52], and in some cases, trauma to their kidneys [5]. Another study suggested that the absence of dedicated paths for e-scooters leads to dangerous interactions between e-scooter riders and other travelers, such as cyclists and pedestrians [64]. Moreover, higher riding speeds have been linked to more severe injuries, and e-scooter speed limits differ greatly by location [9].

Visibility is another concern regarding e-scooters as their small profiles and smooth movements make them more prone to being obscured [30,33]. For example, in 2019, an e-scooter rider was killed in Denver upon collision with an oncoming motor vehicle. A contributing factor was a parked truck obstructing the driver's line of sight [10]. The research work by Karpinski et al. [29] indicated that inadequate lighting and poor weather conditions, such as rain and snow, played a role in greater than 70% of e-scooter accidents between 2018 and 2020. Efforts to improve the visibility of e-scooters have been recommended in the form of adding reflective lights that make their presence more obvious [33].

2.2 Digital Human Modeling for Addressing Vision Obstruction

DHM is a method of computationally representing the human body to conduct physical or cognitive ergonomic analyses [8,14,59]. This technique is especially beneficial in the early stages of product design before design teams must make decisions that would be costly to reverse. Demirel et al. [14–16] showcased the time- and money-saving potential DHM has in the manufacturing, aviation, and automotive domains.

In recent years, one of the ways DHM has been used is to address vision obstruction in different types of vehicles. For example, Demirel and Srinivasan [18] and Srinivasan and Demirel [56,57] examined quantifying vision obstruction in the early phases of automotive design. In these studies, the authors employed a ray-casting technique with the DHM software, Siemens Jack, coverage zone

analysis [51], to measure the the driver's forward field of view (FoV), as a percentage, with a variety of concept variants. Additionally, the research conducted by Joffe and Demirel explored vision obstruction relating to the design of airport ground support vehicles [28]. In a similar manner, the authors demonstrated the use of Siemens Jack's ray-casting procedure to enable designers to proactively identify components of the vehicle that caused obstructions to the driver's FoV. Likewise, Firouzi et al. investigated the automotive A-pillar vision obstruction problem from an optimization perspective [23]. In their paper, the researchers used Siemens Jack's coverage zone visibility results as an input to create surrogate models that would help evaluate visibility with different sizes of pillar cutouts.

While coverage zone analysis is a versatile tool that can be applied to many situations, it requires a substantial amount of computational power, making it time-consuming. As a result, it is important to note that alternate approaches to quantifying vision obstruction have been taken. As an example, Demirel et al. incorporated DHM with image processing to obtain visibility results quicker [17]. In this case, the authors used DHM software to simulate static scenes from a Formula One (F1) racing environment and then exported screenshots from the driver's forward FoV to an image processing software, where the visibility scores were calculated as a percentage. Another strategy for assessing vision obstruction with DHM is that of Summerskill et al. in which the authors measured a Heavy Goods Vehicle (HGV) driver's direct vision using the volume of space they were able to see [58].

2.3 Computer Vision and Traffic Safety

Research has found that traffic accidents are a significant source of injuries and fatalities each year. According to the Insurance Institute for Highway Safety (IIHS), more than 40,000 people in the U.S. were killed as a result of motor vehicle crashes in 2022 [26]. Moreover, the National Highway Traffic Safety Administration (NHTSA) reported that in the same year, greater than 2.3 million people were injured in U.S. traffic accidents [38]. These statistics underscore the importance of implementing systems that facilitate traffic safety.

CV is an area of AI that combines elements of image processing and machine learning (ML) to enable computers to understand information presented in videos and still images [45]. A couple of popular CV algorithms include You Only Look Once (YOLO) and the Meta Segment Anything Model (SAM). The YOLO series algorithms, originally developed by Redmon and Farhadi, comprise a one-stage object detection procedure [63]. Moreover, Meta SAM is a robust zero-shot performance image segmentation tool that was trained on 11 million images and 1 billion masks [32]. Because of the speed and automated nature of CV, it has become a solution implemented in roadway safety [43].

As advanced driver assistance systems (ADAS) become more prevalent in modern cars, object detection has emerged as a valuable application of CV in the traffic safety domain [46]. A few examples of object detection in practice include lane, traffic sign, and pedestrian detection [34,42]. These applications

are especially crucial in autonomous vehicles as this enables them to understand their environment and navigate safely [20]. Additionally, CV has been utilized for traffic surveillance, such as measuring congestion and recognizing the occurrence of accidents [6]. Lastly, street intersections are another area that benefit from CV. In particular, the study by Belgibaev et al. involved incorporating CV in traffic lights in major Kazakhstani cities and illustrated the potential to improve traffic efficiency [4]. These insights demonstrate how CV can assist humans by automating tasks that are visually complex.

3 Methodology

In light of the e-scooter safety concerns mentioned above, we propose a methodology that utilizes DHM, coupled with a CV approach, to assess the risks associated with the operation of e-scooters. In particular, we use these two strategies to measure the percent visibility of e-scooter riders in unique traffic conditions. By injecting CV into the methodology, the DHM environment becomes an input to the CV algorithms, enabling the collection of high-fidelity visibility data. In addition to greater fidelity, this CV approach requires significantly less time to run each vision simulation in comparison with alternate techniques, such as ray-casting. Specifically, the CV techniques we use in this paper are the YOLO algorithm, YOLOv8, and Meta SAM.

Our proposed methodology is presented through three case studies involving e-scooters, drivers, and pedestrians. The first case study examines a collision scenario involving an e-scooter and a full-size family car (sedan) at an intersection. Next, the second case study evaluates the visibility of an e-scooter when it is operated in an area heavily occupied by pedestrians. Finally, the third case study explores an accident scenario in which an e-scooter rider is obstructed by a parked car when riding in the traffic lane.

First, we begin the methodology by creating a full-scale CAD environment within the DHM software, Siemens Jack (abbreviated to "Jack"). Each environment is specific to each case study and is based on images of real infrastructure in Corvallis, Oregon. The virtual globe software, Google Earth, is used to extract the dimensions from each scene for CAD modeling purposes. Riders with two different anthropometric profiles (5^{th} percentile female and 95^{th} percentile male) are used from the Anthropometric Survey of U.S. Army Personnel (ANSUR II) database. This allows us to evaluate how a rider's physical stature can impact their visibility, according to a driver or pedestrian's forward FoV. After the creation of the DHM environment, Jack's Task Simulation Builder (TSB) tool is used to generate dynamic footage of the simulated scene, which is recorded from a driver or pedestrian's perspective. The videos obtained from TSB are then passed to the object detection model, YOLOv8, to identify e-scooter riders. YOLOv8 outputs bounding boxes around detected objects; however, since e-scooters are not included in the Microsoft Common Objects in Context (MS COCO) dataset, we adjusted the bounding box to include the e-scooter, mitigating the necessity to train the model to recognize e-scooters. Once the rider was

detected, we expanded the bounding box to encompass both the driver and the e-scooter, ensuring accurate identification of the entire subject. This adjusted bounding box was then passed to Meta SAM for instance segmentation, which allowed us to generate a segmentation mask and obtain a visibility percentage. This technique enables us to measure the percent visibility of an e-scooter rider by counting their visible and obstructed pixels. This is illustrated in Fig. 1.

We also incorporate rain as an adverse weather condition into each scenario to analyze the difference in visibility imposed by precipitation. Specifically, the inclusion of rain represents typical weather patterns in Corvallis. We used Corel Video Studio to apply a heavy rain effect to the video footage and simulate rainy conditions. Given that the rain effect is represented as white particles, we adjusted the appearance of the e-scooter rider to black to ensure the raindrops were properly visible against them. To quantify the impact of the rain, we calculated the visibility percentage under the assumption that only raindrops obscure the driver's FoV. Since we only consider the obstruction caused by the raindrops in the rainy weather simulations, while other obstructing objects are not included, we determined the actual visibility using the following equation:

$$\%V_{\text{rain,f}} = \%V_{\text{clear}} \times \%V_{\text{rain,r}} \tag{1}$$

where $\%V_{\text{rain,f}}$ represents the final scaled percent visibility in the rainy weather simulations, $\%V_{\text{clear}}$ denotes the percent visibility in the clear weather simulations, and $\%V_{\text{rain,r}}$ is the raw output from the algorithm in the rainy weather simulations.

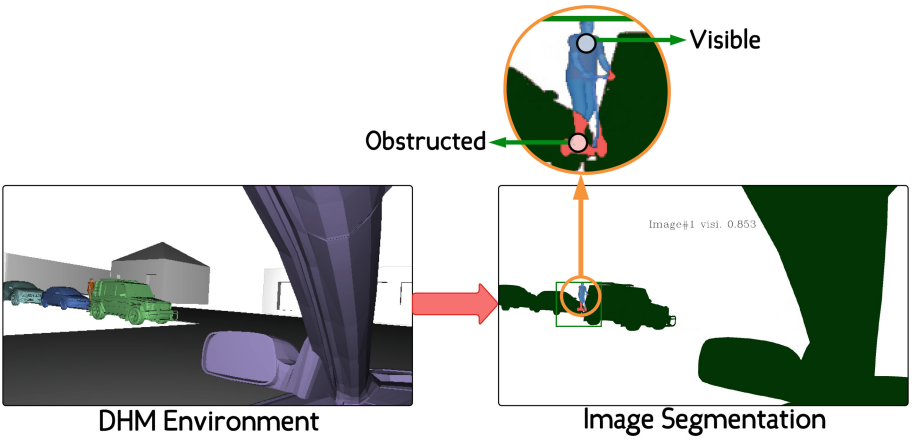


Fig. 1. The image segmentation portion of the methodology calculates the number of visible pixels of the target object. The red color indicates the obstructed area of the manikin and the blue color indicates the visible area. (Color figure online)

Overall, our approach shows how integrating CV with DHM saves time when analyzing vision obscuration-related scenarios, and highlights the dangers that

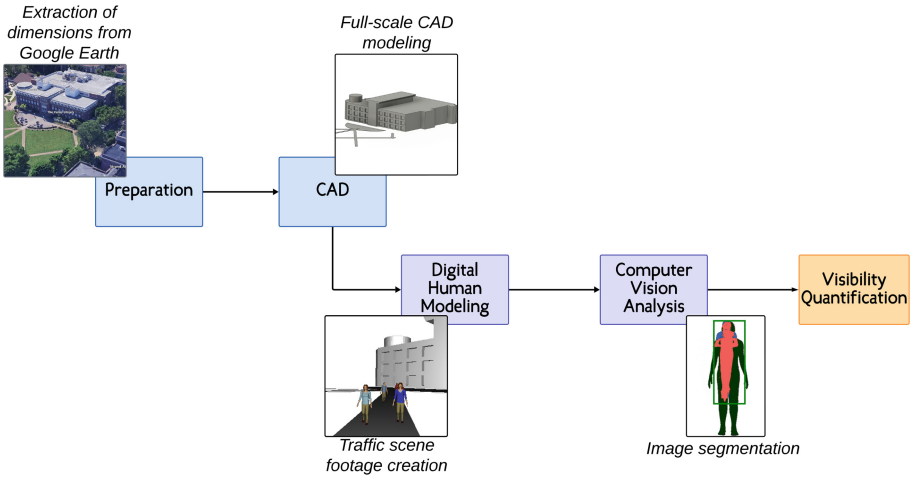


Fig. 2. The diagram shows the key elements of the methodology.

e-scooters can pose to riders and pedestrians. Figure 2 showcases a map of our proposed methodology.

4 Case Studies

The three case studies in this research serve to investigate the visibility of e-scooters in different traffic situations and weather conditions. This section of the paper explains the setup of the case studies and how the methodology applies to each one.

4.1 Case Study 1 - E-Scooter Obstructed by A-Pillar of a Sedan

In this case study, we investigated a scenario in which an e-scooter rider is obscured by the A-pillar of a sedan at an intersection. The situation takes place on one of the main streets near Oregon State University (OSU) in Corvallis, Oregon. While the scenario is not based on a known incident, it is inspired by traffic patterns in the area. Following the extraction of approximate dimensions of the region from Google Earth, we constructed low-fidelity full-scale CAD models to represent the scene, including the street, crosswalks, traffic light posts, and nearby buildings. The model we used for the traffic lights was retrieved from the CAD file repository, GrabCAD [61]. Additionally, we used a pre-existing CAD model of a sedan and also created a model of an e-scooter. Then, the completed models were placed into the Jack DHM software, where we set up the entire environment.

This scenario specifically involves an e-scooter rider using a crosswalk at 8 km/h (5 mph), perpendicular to where a sedan is starting a left turn. Due to

the distance between the two vehicles at their starting positions and the narrow profile of the e-scooter, the rider does not become visible to the driver until he has already begun to execute the turn. The driver was represented by a 50^{th} percentile male manikin, while manikins with of two anthropometric extremes (i.e., short and tall) were used to depict the e-scooter rider. The purpose of this was to assess potential differences in the visibility of e-scooter riders of varying heights from the driver's point of view. Utilizing the TSB module in Jack, we captured real-time, four-second recordings of the scenario from the driver's perspective for both of the manikins, producing two videos. These original two videos were made to represent the scenario under clear weather conditions and then were edited to simulate the driver's forward FoV when rain is introduced to the environment. Subsequently, the footage from the DHM scene was processed to the CV algorithms where the visibility of each e-scooter rider was quantified as a percentage at different timestamps of the video. To simplify the data, we elected to examine only the percentages at each second of the videos ($t = 0,\ 1,\ 2,\ 3,\ 4$). Figure 3 displays the traffic scene configuration, as well as the driver's FoV at a certain timestamp of the video.

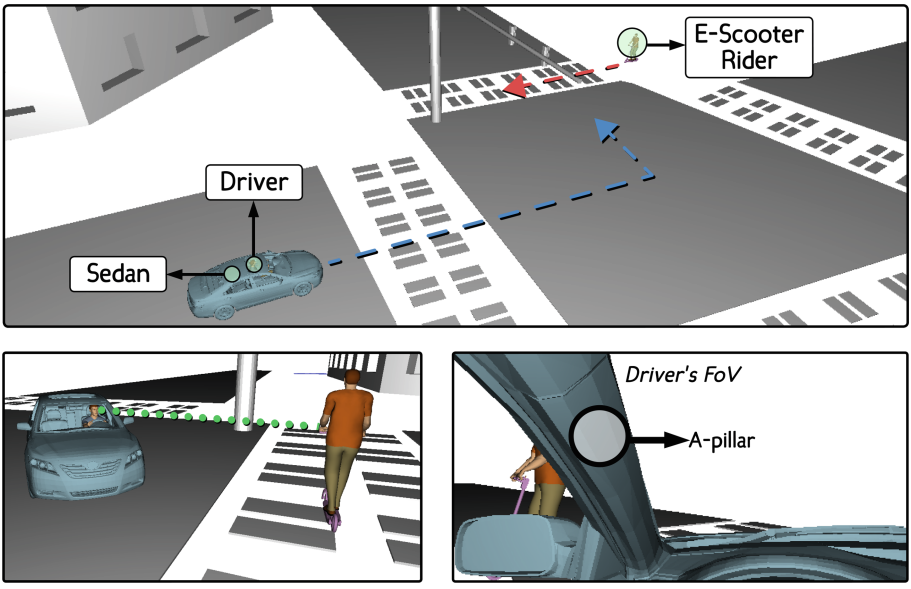


Fig. 3. Illustration of the DHM scene setup for the Case Study 1 video. The driver's FoV reveals the tall male e-scooter rider partially obscured by the A-pillar during the video.

4.2 Case Study 2 - Operating an E-Scooter Near Pedestrians

The second case study explores an instance of an e-scooter being operated in an area with major pedestrian activity. In particular, we assessed how an e-scooter rider can potentially create a hazard by blending into a crowd. This scenario is set at OSU's Library Quad, a location on campus with heavy foot traffic in addition to regular e-scooter usage. Since the setting is a crowded place, we set the speed of the e-scooter to be slower at 4 km/h (2.5 mph). After we built CAD models of the library building and main walkways, we inserted them into Jack. Then, we placed manikins of many different sizes on the walkways to act as pedestrians. The visibility aspect of this case study was examined by looking at the forward FoV of a 50th percentile male pedestrian who is walking towards the e-scooter rider while they are obscured behind another pedestrian (see Fig. 4). The same two anthropometric builds as in Case Study 1 were used to depict the riders.

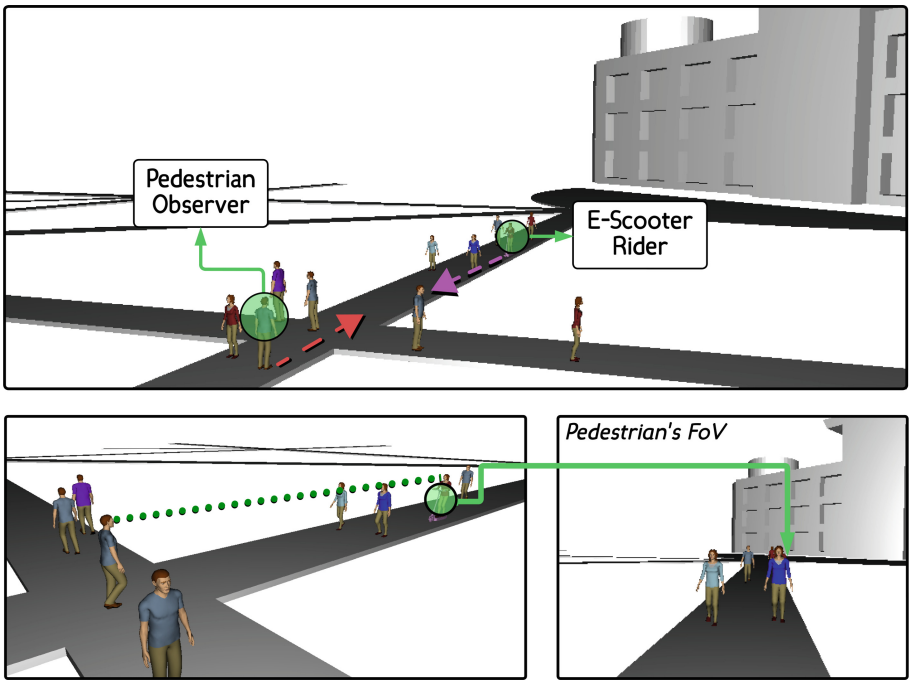


Fig. 4. Demonstration of the setup for Case Study 2, showing the directions of travel of the observing pedestrian and the e-scooter rider on the Library Quad walkway. From the observing pedestrian's FoV, the e-scooter rider is obscured behind another pedestrian.

In an identical manner to Case Study 1, the TSB functionality in Jack was used to animate the scene and create two four-second videos, from the pedes-

trian's perspective, of the pedestrians and e-scooter rider moving down the paths. We used the same techniques to edit these videos and create adverse weather variants with rain. Finally, the CV algorithms processed the videos to quantify the percent visibility of both of the two e-scooter riders at the same timestamps as the previous case study.

4.3 Case Study 3 - E-Scooter Obstructed by Parked Cars

Case Study 3 examines a situation where an e-scooter rider is obstructed by cars parked on the side of the road while they are riding in the traffic lane (see Fig. 5. It occurs on a different part of the same street that was used for Case Study 1. More specifically, the event is observed by a sedan driver who is scanning his surroundings as he prepares to make a right turn from a cross street. This brings attention to how easily an e-scooter rider can be obscured in traffic. In this scenario, the e-scooter rider is riding at 32 km/h (20 mph). This case study is not a recreation of a specific incident, but it is loosely inspired by the previously mentioned fatal accident that took place in Denver. In order to simulate a driver checking their environment, the manikin's eyes were set to track the e-scooter rider as they rode by.

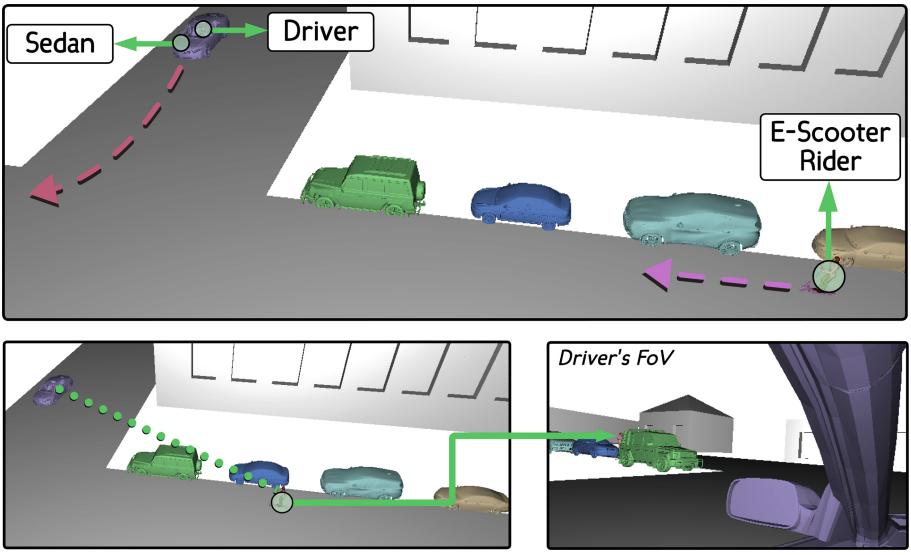


Fig. 5. The figure shows the scene setup for Case Study 3 with the sedan, e-scooter, rider, and parked cars visible at the intersection, as well as the driver's FoV at a selected timestamp of the video.

As in Case Study 1, the driver was depicted as a 50^{th} percentile male and the e-scooter riders were varied with the same two anthropometries. CAD models

of the street and buildings were produced in the same manner and the models of the parked cars were sourced from GrabCAD [2,19,37,47]. The process for creating the videos and analyzing them was the same as in the previous two case studies.

5 Results

The results of this study were collected from 60 CV-based vision obstruction simulations. The video footage of the traffic scenarios that was generated from the DHM environment functioned as the input to the CV simulations and the percent visibility scores were the output. More specifically, each of the three case studies was comprised of 20 simulations, including two manikins, two weather conditions (clear weather and rain), and five timestamps. The findings highlight how the visibility of an e-scooter rider changes with respect to anthropometry, as well as dynamically over time as the observer and rider progress along their respective paths.

5.1 Visibility Analysis for Case Study 1

The data collected from the Case Study 1 traffic scenario showed that neither of the two heights of e-scooter riders were visible at most of the timestamps in either of the weather conditions. The 0% visibility scores can be explained by the e-scooter rider being obscured by the sedan's A-pillar (see Fig 6) and also completely disappearing from the driver's FoV as he turned left. However, the riders were both 100% visible at $t = 1\,\text{s}$ under clear weather. It is important to note that the percentage visibility of both riders decreased by around 15% at $t = 1\,\text{s}$ when rain was added to the environment. Since there was not a significant difference between the visibility of the short and tall riders, this implies that anthropometry did not play a role in visibility for this traffic scenario. Figure 7 displays the results of this case study.

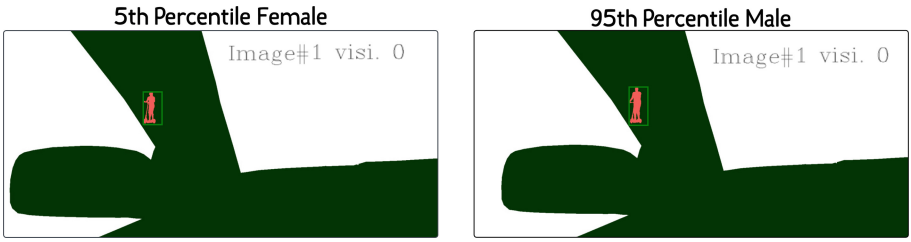


Fig. 6. Image segmentation output for Case Study 1 at $t = 0\,\text{s}$, showing the short and tall riders obstructed by the sedan's A-pillar.

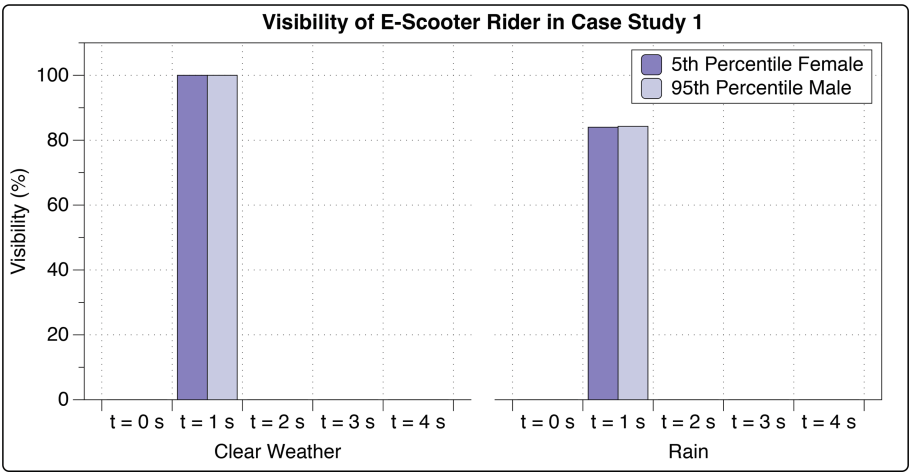


Fig. 7. The results from Case Study 1 show that rainy conditions adversely impact the visibility of the e-scooter rider. The rider also disappears from the driver's FoV during most of the timestamps, resulting in 0% visibility.

5.2 Visibility Analysis for Case Study 2

Observations from Case Study 2 show a drastic difference between the visibility of the 5^{th} percentile female and 95^{th} percentile male e-scooter riders. For example, at the $t = 2\,\text{s}$ timestamp, the small female rider was 16.7% visible while the tall male rider was 31.7% visible. The obscuration for both anthropometric profiles in this scenario was due to the pedestrian they were riding behind.

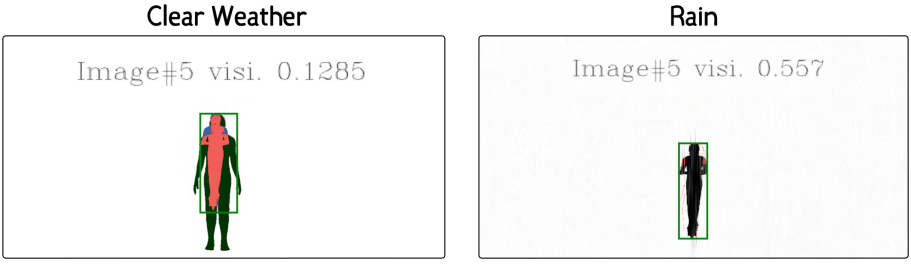


Fig. 8. Output from the $t = 4\,\text{s}$ CV simulations in Case Study 2, demonstrating the difference in visibility of the 5^{th} percentile female rider between clear weather and rainy conditions. The final visibility score for the rainy case (7.16%) was obtained using Eq. 1.

This fact emphasizes how difficult it can be for a person to identify that they are walking towards a moving e-scooter. Moreover, rainy weather was a contributing factor for poorer visibility. An example of this is at $t = 4\,\text{s}$ where the

visibility of the 5th percentile female rider decreased from 12.9% to 7.16% with the inclusion of rain (refer to Fig. 8). The data from Case Study 2 is demonstrated in Fig. 9.

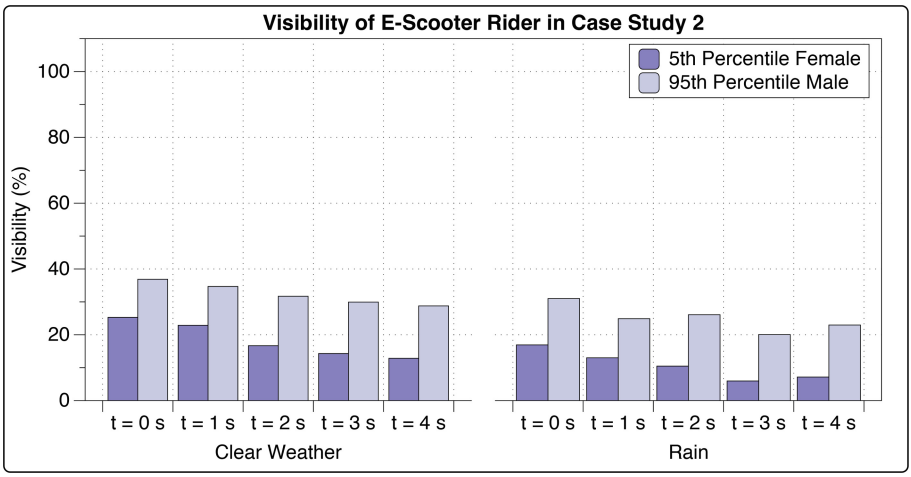


Fig. 9. Data from Case Study 2 demonstrates a marked difference between the visibility of the 5th percentile female and the 95th percentile male e-scooter riders behind the pedestrian.

5.3 Visibility Analysis for Case Study 3

This case study quantified the visibility of an e-scooter rider as they pass in front of a car that is preparing to turn right. Findings from Case Study 3 showed only a minimal difference in visibility between the short female and tall male e-scooter rider. At $t = 0\,\text{s}$ in clear weather, the riders were partially obstructed (around 80% visible) by the cars parked on the side of the street. Further, there was a negative correlation between rain and visibility (see Fig. 10). For instance, at $t = 1\,\text{s}$, the visibility of the 95th percentile male decreased from 100% in clear weather to 81% with rain. At $t = 3\,\text{s}$ and $t = 4\,\text{s}$, neither rider was visible to the driver as they exceeded the driver's line of sight. Since the e-scooter rider's speed was much faster in this scenario, it is crucial to note that their position changed quickly between each timestamp, leading to varied visibility measurements between the frames of the video (Fig. 11).

6 Discussion

This study aimed to assess visibility-related e-scooter safety hazards using an approach that integrates DHM with CV in the form of image segmentation. The

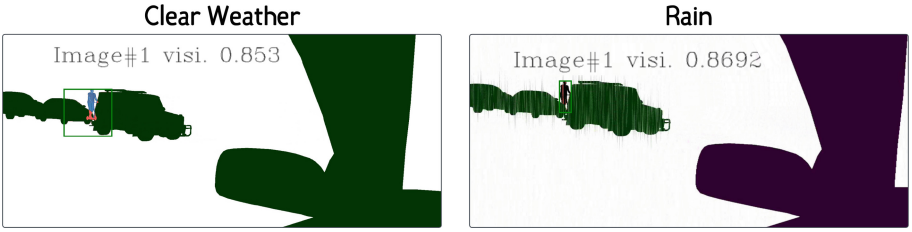


Fig. 10. The image segmentation output from Case Study 3 conveys the decrease in the visibility of the e-scooter rider when rain was added to the environment. The final visibility score for the rainy weather condition was calculated using Eq. 1.

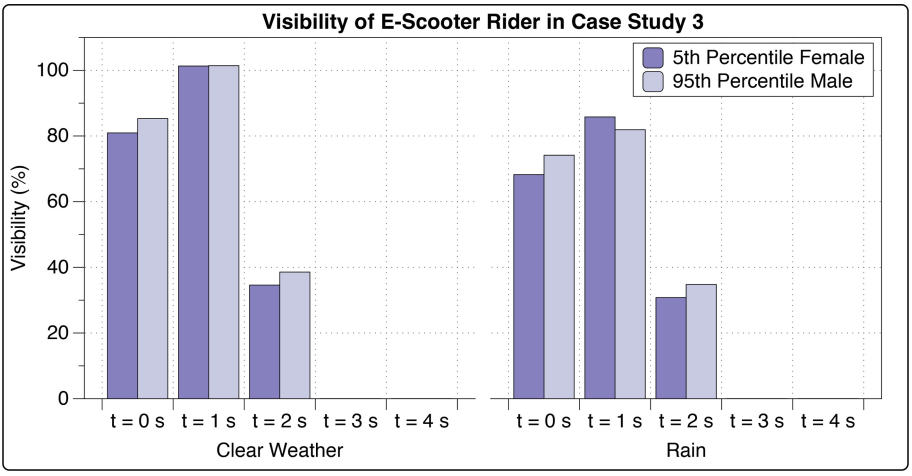


Fig. 11. The data from Case Study 3 shows that anthropometry did not have a significant impact on the visibility of the e-scooter rider. Rain decreased the visibility by up to 20% at the $t = 1\,\text{s}$ timestamp.

methodology we proposed assists with quantifying visibility in a consistent and rapid manner in comparison with other more cumbersome and time-consuming techniques, such as ray-casting within DHM software. While ray-casting vision obstruction analysis, such as the coverage zone method, provides insight into the visibility of objects, it is highly resource-intensive with each simulation ranging from a few seconds to several minutes. By transforming the DHM environment into an input and using CV for the simulations, the analysis is much faster, providing data in real time. Additionally, the image segmentation technique allows for the visible area of the target object to be calculated from its precise silhouette, rather than from a rectangular bounding box. This results in greater accuracy in the visibility scores. Overall, the findings from the three case studies offer valuable insights into the visibility concerns regarding e-scooters, such as their narrow profiles and adverse weather making it more difficult to see them.

While our CV approach promotes a rapid analysis of visibility, this study had a few limitations. In particular, although the videos from Jack's TSB module were generated based on real traffic conditions in Corvallis, the setup in the DHM environment was manual. This meant that we needed to make assumptions about the viewers' (i.e., drivers' and pedestrian's) and e-scooter riders' postures, as well as their trajectories. As a result, this manual approach to setting up the DHM environment can lead to biases. In the same vein, we did not have enough data available to us to recreate exact accident scenarios, so the setup process relied on intuition. Further, the observer in each scenario was a 50^{th} percentile male. Visibility scores would likely change if the observer anthropometrics were varied. Moreover, our data collection method, which only recorded visibility percentages at one-second intervals, increases the likelihood of missing momentary changes, especially in Case Studies 1 and 3, where the vehicles move at higher speeds and follow non-linear or perpendicular trajectories. This would explain the large number of timestamps where the visibility was measured as 0% when the video actually showed the e-scooter rider was partially visible in many frames. A higher frame sampling rate would improve the accuracy of the analysis of these scenarios. Lastly, our methodology does not include human cognition, which indicates that the percent visibility scores may not accurately reflect what a human observer would actually perceive in these scenarios.

7 Conclusion and Future Work

This research work presented a methodology that coupled DHM and CV to quantify the visibility of e-scooter riders in three unique traffic scenarios. Using this approach facilitated faster and more robust vision analyses than previous approaches. The results of this study illustrate that smaller e-scooter riders can be obstructed more easily by other objects or people, as well as how rain exacerbates visibility problems.

Future studies will augment the techniques used in this research by focusing on ways to improve it. This includes increasing the sampling rate to include more timestamps and ensure that rapid fluctuations in visibility are captured. It would also be valuable to examine how other weather conditions, such as fog and snow or varied intensities of rain, would impact visibility. Likewise, simulating various lighting conditions (e.g., nighttime and glare) would supply information about how lighting impacts an e-scooter rider's visibility. One last future research goal for this work is to use this image segmentation approach to quantify the visibility of a e-scooter rider from the perspective of drivers of a wider variety of vehicles, such as trucks and buses.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aizpuru, M., Farley, K.X., Rojas, J.C., Crawford, R.S., Moore, T.J., Jr., Wagner, E.R.: Motorized scooter injuries in the era of scooter-shares: a review of the national electronic surveillance system. Am. J. Emerg. Med. **37**(6), 1133–1138 (2019)
2. ALEKSEY TSYRIK: Hyundai Accent (Verna). https://grabcad.com/library/hyundai-accent-verna-1. Accessed 7 Jan 2025
3. Bascones, K., Méndez, T.M., Siller, F.Y.: [translated article] e-scooter accidents: A new epidemic. Revista española de cirugía ortopédica y traumatología **66**(2), T135–T142 (2022)
4. Belgibaev, B., Mansurova, M., Abdrakhim, S., Ormanbekova, A.: Smart traffic lights with video vision based on a control minicomputer in Kazakhstani megacities. Procedia Comput. Sci. **231**, 792–797 (2024)
5. Bianchi, A., Gallina, S., Cianflone, F., Tafuri, A., Cerruto, M.A., Antonelli, A.: E-scooter accidents: a rising cause of kidney injury. Urologia J. **89**(4), 506–510 (2022)
6. Buch, N., Velastin, S.A., Orwell, J.: A review of computer vision techniques for the analysis of urban traffic. IEEE Trans. Intell. Transp. Syst. **12**(3), 920–939 (2011)
7. Cao, Z., Zhang, X., Chua, K., Yu, H., Zhao, J.: E-scooter sharing to serve short-distance transit trips: a Singapore case. Transp. Res. Part A: Policy Practice **147**, 177–196 (2021)
8. Chaffin, D.B.: Digital human modeling for vehicle and workplace design. SAE International (2001)
9. Cicchino, J.B., Kulie, P.E., McCarthy, M.L.: Severity of e-scooter rider injuries associated with trip characteristics. J. Safety Res. **76**, 256–261 (2021)
10. Collaborative Sciences Center for Road Safety: e-scooter fatalities. https://www.roadsafety.unc.edu/wp-content/uploads/2021/03/escooter_fatalities_Mar_2021.pdf. Accessed 30 Dec 2024
11. Comer, A.R., et al.: Electric scooters (e-scooters): assessing the threat to public health and safety in setting policies: assessing e-scooter policies. Sustainability (2020). https://doi.org/10.18060/24194
12. Davies, E.R.: Computer vision: principles, algorithms, applications, learning. Academic Press (2017)
13. Demir, N., Dokur, M., Agdoğan, Ö., Koc, S., Karadağ, M., Dokur, İF.: Electric scooters as a silent source of danger in increasing use among young people: a single-center in-depth accident analysis. Turkish J. Trauma Emergency Surg. **29**(5), 596 (2023)
14. Demirel, H.O., Ahmed, S., Duffy, V.G.: Digital human modeling: a review and reappraisal of origins, present, and expected future methods for representing humans computationally. Int. J. Hum.-Comput. Interact. **38**(10), 897–937 (2022)
15. Demirel, H.O., Duffy, V.G.: Applications of digital human modeling in industry. In: Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1, pp. 824–832. Springer (2007)
16. Demirel, H.O., Duffy, V.G.: Digital human modeling for product lifecycle management. In: Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1, pp. 372–381. Springer (2007)
17. Demirel, H.O., Jennings, A., Srinivasan, S.: An early design method to quantify vision obstruction: formula one (f1) halo case study. In: International Conference on Human-Computer Interaction, pp. 32–44. Springer (2022)

18. Demirel, H.O., Srinivasan, S.: A proactive ergonomics framework to assess a-pillar vision obstruction. Digit. Hum. Model. Appl. Optim **46**, 16 (2022)
19. DinoKJW: Kia Sportage Hybrid 2022. https://grabcad.com/library/kia-sportage-hybrid-2022-1. Accessed 7 Jan 2025
20. Dong, X., Cappuccio, M.L.: Applications of computer vision in autonomous vehicles: Methods, challenges and future directions. arXiv preprint arXiv:2311.09093 (2023)
21. Duffy, V.G.: Human digital modeling in design. Handbook of Human Factors and Ergonomics, pp. 1016–1030 (2012)
22. Esteva, A., et al.: Deep learning-enabled medical computer vision. NPJ Digital Med. **4**(1), 5 (2021)
23. Firouzi, M.A., Bhaskaran, V., Demirel, H.O., Hoyle, C.: Metamodels and vision obstruction: a new lens on driver visibility. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. vol. 88346, p. V02AT02A056. American Society of Mechanical Engineers (2024). https://doi.org/10.1115/DETC2024-143657
24. Frank, S., et al.: Continuously increasing e-scooter accidents and their possible prevention in a large European city. Europ. J. Trauma Emerg. Surgery, 1–10 (2024)
25. Grill, F.D., et al.: E-scooter-related craniomaxillofacial injuries compared with bicycle-related injuries-a retrospective study. J. Cranio-Maxillofacial Surg. **50**(9), 738–744 (2022)
26. Insurance Institute of Highway Safety: Fatality Facts 2022. https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot. Accessed 1 Jan 2025
27. Iroz-Elardo, N., Currans, K.: Injury burden of introducing e-scooters: a review of e-scooter injury studies using retrospective review of emergency department records, 2015–2019. Transp. Res. Rec. **2675**(12), 1150–1159 (2021)
28. Joffe, G.B., Demirel, H.O.: The use of digital human modeling to assess vision obstruction in airport taxi operations. In: International Conference on Human-Computer Interaction, pp. 55–68. Springer (2024)
29. Karpinski, E., Bayles, E., Daigle, L., Mantine, D.: Characteristics of early shared e-scooter fatalities in the united states 2018–2020. Saf. Sci. **153**, 105811 (2022)
30. Karpinski, E., Bayles, E., Daigle, L., Mantine, D.: Comparison of motor-vehicle involved e-scooter fatalities with other traffic fatalities. J. Safety Res. **84**, 61–73 (2023)
31. Kazemzadeh, K., Haghani, M., Sprei, F.: Electric scooter safety: an integrative review of evidence from transport and medical research domains. Sustain. Urban Areas **89**, 104313 (2023)
32. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
33. Komova, G., Jallais, C., Ndiaye, D., Rogé, J.: The role of innovative visibility aids on e-scooter riders' conspicuity in daylight. Transport. Res. F: Traffic Psychol. Behav. **106**, 44–54 (2024)
34. Loce, R.P., Bernal, E.A., Wu, W., Bala, R.: Computer vision in roadway transportation systems: a survey. J. Electron. Imaging **22**(4), 041121–041121 (2013)
35. Maiti, A., Vinayaga-Sureshkanth, N., Jadliwala, M., Wijewickrama, R., Griffin, G.: Impact of e-scooters on pedestrian safety: a field study using pedestrian crowd-sensing. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 799–805. IEEE (2022)

36. Mellon, T.R.P.: Using Digital Human Modeling to Evaluate and Improve Car Pillar Design: A Proof of Concept and Design of Experiments. Master's thesis, Oregon State University, Corvallis, OR (2021)
37. Morteza Padarlar: TOYOTA CAMRY. https://grabcad.com/library/toyota-camry-1. Accessed 7 Jan 2025
38. National Highway Traffic Safety Administration: Overview of Motor Vehicle Traffic Crashes in 2022. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813560, Accessed 1 Jan 2025
39. National Operations Center of Excellence: New Urban Mobility alliance (NUMO). https://data.transportationops.org/new-urban-mobility-alliance-numo. Accessed 27 Dec 2024
40. Naumann, A., Rötting, M.: Digital human modeling for design and evaluation of human-machine systems. MMI-Interaktiv **12**, 27–35 (2007)
41. New Urban Mobility Alliance: New Mobility Atlas. https://www.numo.global/new-mobility-atlas#2/42.64/28.22. Accessed 27 Dec 2024
42. Nkuzo, L., Sibiya, M., Markus, E.: Computer vision-based applications in modern cars for safety purposes: a systematic literature review. In: 2023 Conference on Information Communications Technology and Society (ICTAS), pp. 1–9. IEEE (2023)
43. Ode, A.G., Samuel, O., Titus, A., Adejo, O.O.: Ai-based traffic controller using computer vision. Global J. Artif. Intell. Technol. Dev. **2**(1), 9–19 (2024)
44. Okudan, B., Dündar, S.: Analysis and comparison of electric scooter crashes: a review of crash characteristics and health outcomes in Türkiye. J. Transp. Logistics **9**(2), 353–364 (2024)
45. Prince, S.J.: Computer vision: models, learning, and inference. Cambridge University Press (2012)
46. Rezaei, M., Klette, R.: Computer vision for driver assistance. Springer International Publishing, Cham **45**, 16 (2017)
47. Shaylesh Roy: Mercedes Benz G Class G63 AMG 2020. https://grabcad.com/library/mercedes-benz-g-class-g63-amg-2020-1. Accessed 7 Jan 2025
48. Shichman, I., Shaked, O., Factor, S., Weiss-Meilik, A., Khoury, A.: Emergency department electric scooter injuries after the introduction of shared e-scooter services: a retrospective review of 3,331 cases. World J. Emerg. Med. **13**(1), 5 (2022)
49. Shmerling, R.H.: The popularity of e-bikes and e-scooters is soaring, but are they safe? https://www.health.harvard.edu/blog/the-popularity-of-e-bikes-and-e-scooters-is-soaring-but-are-they-safe-202409093069. Accessed 27 Dec 2024
50. Siebert, F.W., Ringhand, M., Englert, F., Hoffknecht, M., Edwards, T., Rötting, M.: Braking bad-ergonomic design and implications for the safe use of shared e-scooters. Saf. Sci. **140**, 105294 (2021)
51. Siemens PLM Software: Tecnomatix Jack. https://www.plm.automation.siemens.com/media/store/en_us/4917_tcm1023-4952_tcm29-1992.pdf. Accessed 1 Jan 2025
52. Sikka, N., Vila, C., Stratton, M., Ghassemi, M., Pourmand, A.: Sharing the sidewalk: A case of e-scooter related pedestrian injury. Am. J. Emerg. Med. **37**(9), 1807-e5 (2019)
53. Siman-Tov, M., Radomislensky, I., Group, I.T., Peleg, K., et al.: The casualties from electric bike and motorized scooter road accidents. Traffic Injury Prevention **18**(3), 318–323 (2017)
54. Singh, P., Jami, M., Geller, J., Granger, C., Geaney, L., Aiyer, A.: The impact of e-scooter injuries: a systematic review of 34 studies. Bone Joint Open **3**(9), 674–683 (2022)

55. Srinivasan, S.: Early Design Evaluation of See-Through Automotive A-pillar Concepts Using Digital Human Modeling and Mixed Reality Techniques. Master's thesis, Oregon State University, Corvallis, OR (2022)
56. Srinivasan, S., Demirel, H.O.: Quantifying vision obscuration of a-pillar concept variants using digital human modeling. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. 86212, p. V002T02A049. American Society of Mechanical Engineers (2022)
57. Srinivasan, S., Demirel, H.O.: Quantifying vision obstruction of formula one (f1) halo concept variants. In: Proceedings of the 7th International Digital Human Modeling Symposium, vol. 7. University of Iowa (2022)
58. Summerskill, S., Marshall, R., Paterson, A., Eland, A.: The use of dhm to quantify a measure of direct vision performance in trucks. In: Proceedings of the 7th International Digital Human Modeling Symposium, vol. 7. University of Iowa (2022)
59. Sundin, A., Örtengren, R.: Digital human modeling for CAE applications. Handbook of human factors and ergonomics, pp. 1053–1078 (2006)
60. Szeliski, R.: Computer vision: algorithms and applications. Springer Nature (2022)
61. thinkin3D: Traffic Signals. https://grabcad.com/library/traffic-signals-1. Accessed 7 Jan 2025
62. Trivedi, T.K., et al.: Injuries associated with standing electric scooter use. JAMA Network Open **2**(1), e187381–e187381 (2019). https://doi.org/jamanetworkopen.2018.7381
63. Varghese, R., M., S.: YOLOv8: a novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp. 1–6 (2024). https://doi.org/10.1109/ADICS58448.2024.10533619
64. Zuniga-Garcia, N., Juri, N.R., Perrine, K.A., Machemehl, R.B.: E-scooters in urban infrastructure: understanding sidewalk, bike lane, and roadway usage from trajectory data. Case Stud. Transport Policy **9**(3), 983–994 (2021)



The Application of Interactive Narrative in the Health Knowledge Visualization

Tian Lei, Yan Deng(✉), Rui Yang, and Zhaoxu Li

Huazhong University of Science and Technology, Wuhan, Hubei, China
1311061815@qq.com

Abstract. The popularization of health knowledge plays a crucial role in the transformation of the medical system from a treatment-oriented approach to a prevention-oriented approach that combines prevention and treatment. This study explores non-linear the common problems in the current popularization of health knowledge, such as high reading barriers, insufficient attractiveness, and difficulty in stimulating the public's willingness to actively learn about health knowledge. This study proposes a visualization design method for health knowledge based on interactive narrative, and applies and evaluates the method using health knowledge related to headaches as an example. The results indicate that the design method has a significant positive impact on the popularization of health knowledge. It not only enhances the attractiveness of health knowledge popularization and strengthens users' subjective learning willingness, but also effectively guides users to think deeply and better understand health knowledge.

Keywords: Health Knowledge · Knowledge Visualization · Interactive Narrative Design

1 Introduction

Currently, the global trend of healthcare system reform is gradually shifting from a treatment-centered approach to one focused on prevention. Therefore, enhancing the public's health literacy plays a crucial role [1]. However, the existing health knowledge content often lacks interactivity and attractiveness. As a result, the public tends to be repelled by and resistant to health knowledge, showing no interest in clicking on it and lacking the patience to finish reading. Nowadays, narrative forms are constantly being subverted and innovated. As a cutting-edge approach in the field of narrative, interactive narrative has been widely applied in many fields. Through interactive plot lines, interactive narrative can bring users a brand-new reading experience. Nevertheless, in the field of health knowledge, there are still few works applying the interactive narrative method. Therefore, how to apply the interactive narrative method to health knowledge so that it can better come into the public eye is what these paper aims to explore and study.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 210–226, 2025.
https://doi.org/10.1007/978-3-031-93502-2_14

2 Problems in Health Knowledge Popularization

Through the analysis of research literature on health science popularization from 1996 to 2015, it was found that China has formulated a series of policies and regulations to support the development of health science popularization, providing a strong guarantee for improving the public's health literacy. For example, in 2019, China issued the "Guideline to Implement the Healthy China Initiative", which is an important document used at the national level to guide disease prevention and health promotion in the next decade. The Healthy China Initiative Promotion Committee was established to increase the health awareness of Chinese residents and promote active participation of the whole population in health actions [2]. Subsequently, the Outline of the "Healthy China 2030" Plan was promulgated, which clearly defined the strategic goals and specific tasks for building a healthy China, and made detailed plans for promoting healthy lifestyles, improving health services, and building a healthy environment [3]. However, judging from the current practical effects, there are still certain problems and deficiencies in the popularization of health knowledge [4]. Text is the core of popular science's creation. If viewed from a functional perspective, texts used for health knowledge popularization can be divided into two categories: knowledge texts and narrative texts. Knowledge texts mainly undertake the task of directly describing medical knowledge, while narrative texts strive to connect medical knowledge in the form of stories. However, whether it is health knowledge tweets, illustrated medical posters, or medical experts' instructional videos and popular science animations, there are generally problems with their corresponding creative texts. Firstly, the difficulty of reading knowledge texts is high, which poses a challenge to public understanding. Many medical science popularization works are filled with a large number of professional terms and often have a mixture of multiple languages, without considering the cultural background and knowledge level of the public, which makes non-medical professionals rarely pay attention to them. Secondly, the attractiveness of narrative texts is insufficient. At present, most health knowledge works are mainly based on knowledge texts, such as expert lecture videos, medical illustration posters, etc. Only a few popular science works, such as some tweets and popular science animations, will add narrative texts to illustrate and polish the knowledge content with examples. Due to the lack of narrative text, popular science content often appears dull and uninteresting, making it difficult to stimulate and maintain public interest, thereby affecting the effectiveness of popular science. Therefore, it is particularly crucial to enhance the attractiveness of knowledge texts and increase the volume ratio of narrative texts in the design of health knowledge popularization. The method of interactive narrative can effectively address the above two challenges. Its narrative quality can increase the proportion of narrative text in the work, stimulate users' interest in health knowledge and establish emotional bonds, while interactivity promotes users' attention and deep understanding of knowledge texts.

3 Interactive Narrative and Health Knowledge Visualization

3.1 Interactive Narrative

Interactive narrative is the integration of "interaction" into traditional narrative. It allows the audience to interact with the story. This interactivity is the main difference between interactive storytelling and traditional storytelling, as it endows users with the ability to transition from passively receiving information to actively exploring stories. According to whether the interaction has an impact on the progress of the story, interactions in interactive storytelling can be divided into two categories: "instrumental interactions" that do not affect the storytelling and "structural interactions" that have an impact on the storytelling.

In interactive narrative, narrative structure is a very important concept. It concerns the organization and relationship between various parts in a narrative work, and is a macro layout and structure of the entire narrative text. The structure of interactive narrative determines the overall form and stability of the narrative. It is usually divided into two types: linear and non-linear [5]. Linear narrative structure refers to a structure that unfolds events in chronological order, usually including clear narrative stages such as beginning, middle, and end, advancing the story in a linear manner. The audience or readers can clearly follow the sequence of events to understand the entire story. In linear interactive narrative, interaction is usually the instrumental interactions. Non-linear structure refers to a narrative style that does not follow chronological order and causal relationships. It often tells stories flexibly and diversely by disrupting time and interweaving plot. In non-linear narrative structures, interactions are mostly structural interactions that change the direction and ending of the story [6].

If the narrative structure is compared to the route map of rail transit, then the trajectory routes shuttling between different stations are different storylines. The storyline is an important factor that constitutes the attractiveness and infectiousness of a story. It drives the development of the story in the narrative, allowing the audience to follow the plot to understand the full picture of the story, feel the growth and changes of the characters, and experience the emotions and conflicts in the story. In a linear narrative structure, there is usually only one storyline, as shown on the left side of Fig. 1, and there is only one ending to the story. In interactive non-linear narrative structures, there are many possible storylines for each type of narrative structure. This is precisely why interactive storytelling can address the problems in health knowledge popularization works. Figure 1 is a summary by Gonzalo Méndez et al. of linear and typical non-linear interactive narrative structures and possible storylines [7].

3.2 Interactive Narrative's Positive Effect on Health Knowledge Visualization

Research shows that applying interactive narrative techniques to the design of health popularization can not only stimulate users' intrinsic learning motivation and transform passive learning into active learning [8], but also promote users to engage in immersive learning and deeply grasp the essence of knowledge. For example, Sarasmita and her team adopted an interactive narrative approach to design a health reminder educational software, constructing elements such as virtual characters, story backgrounds, and target

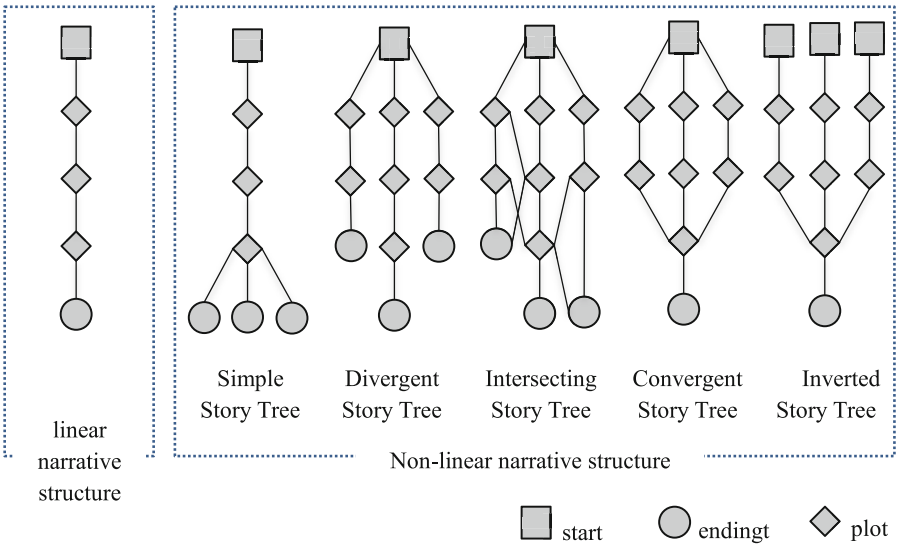


Fig. 1. Interactive narrative structures and possible story lines.

tasks for learners. The research found that driven by interactive narrative technology, learners demonstrated a stronger learning interest and higher learning enthusiasm compared with traditional learning methods [9]. Another example is that Ma Wenli et al. used interactive narrative technology to produce a popular science video about vaccine knowledge. Experiments have proven that interactive narrative is also conducive to users' understanding and memory of vaccine knowledge in the popular science design of vaccine knowledge [10].

In the context of the era where traffic is king, by applying interactive narrative techniques, we can further expand the creative forms of health knowledge works. This not only helps us master the traffic rules in the era of converged media, but also enables us to meet the needs of the emerging media environment [11]. The non-linear story structure of interactive narrative gives users the freedom to selectively watch certain parts of the story, which coincides with the characteristics of brevity and fragmentation of information in the short-video era. Therefore, health knowledge applications adopting interactive narrative are more adaptable to users' reading habits in the traffic-oriented era, thus achieving better communication effects.

In conclusion, the application of interactive narrative can not only stimulate users' learning interest, but also effectively promote the dissemination of health knowledge. Therefore, applying interactive narrative to the dissemination of health knowledge is undoubtedly an important direction for the development of popular science narrative in the future.

3.3 Design Principles of Health Knowledge Visualization Based on Interactive Narrative

When conducting the visualization design of health knowledge, if the method of interactive storytelling is adopted, three dimensions, namely text content, visual presentation, and interactive mechanism, need to be comprehensively considered, and the following design principles should be adhered to as much as possible.

1. **On the premise of ensuring the reliability and rigor of the knowledge content, it should be made as easy as possible for the audience to understand.**

In the popularization of health knowledge, the visualization of knowledge texts plays a vital role. Therefore, during the creative process of health knowledge visualization design, it is particularly crucial to ensure the authority and rigor of knowledge texts. When applying interactive narrative techniques, whether it is the story content that assists in the dissemination of knowledge texts, the visual representation transformed from knowledge texts, or the interactive means and forms of information acquisition and feedback, the principle of not distorting or misleading health knowledge texts must be ensured. Otherwise, it will lead to a deviation from the fundamental purpose, mislead the public, and thus fail to achieve the expected effect of health knowledge popularization.

2. **The narrative structure should be clear and interesting, conforming to the cognitive logic and framework of the audience.**

The core function of the narrative text is to expand and present the knowledge text, aiming to stimulate readers' interest in reading the knowledge content and promote their in-depth understanding of it. Therefore, the integrity and clarity of the narrative structure are of particular importance. Especially when applying interactive narrative techniques, if a non-linear interactive narrative structure is involved, a clear and complete story outline is crucial for sorting out the story logic. It can ensure that users do not feel a sense of fragmentation when reading each independent branch story line, thus providing a coherent and complete story experience. This experience not only helps users better understand the health knowledge content but also enhances their sense of immersion and empathy.

3. **The visual presentation should be intuitive and in line with the aesthetic characteristics of the audience.**

Visual presentation is the primary link for health knowledge to attract the audience. Only excellent visual effects can effectively capture the audience's attention and encourage them to watch the content in depth. In contemporary society with diversified aesthetic concepts, visual styles are diverse, and the audience has a relatively high acceptance of different styles, which provides creators with a broader creative space. However, the visual transformation of health knowledge is different from that of other entertainment products. Not all visual styles are suitable for expressing health knowledge content. When choosing the visual style of health knowledge, starting from the visual authenticity, it is necessary to comprehensively consider three dimensions: the difficulty and field of the health knowledge to be popularized, and the audience, and select an appropriate level of authenticity for visual transformation. Only by presenting beautiful and intuitive visual

effects of health knowledge can the effective dissemination of health knowledge be promoted.

4. **The interactive forms need to be consistent with the audience's interaction habits and be as diverse as possible.**

Interactivity is the core difference between interactive narrative and traditional narrative. Therefore, when applying interactive narrative to disseminate health knowledge, the choice of interactive forms is of particular importance. Common interactive forms include clicking, long-pressing, swiping, dragging, shaking, etc. By using a variety of interactive forms in an alternating manner, the freshness and interest of users' reading can be maintained. In addition, regarding the content of health knowledge, the rationality of the interactive forms also needs to be considered. For example, when introducing health knowledge with strong operability, adopting interactive forms that are closer to real-world operations helps users understand the operation steps and methods more accurately. A reasonable interactive design can also help users better integrate into the situation and enhance their sense of immersion.

4 Design Process of Health Knowledge Visualization Based on Interactive Narrative

Building upon the aforementioned design principles, this research delineates the interactive narrative design process for health knowledge into five stages: identifying audience needs, determining textual content, visual conversion of text, crafting the interactive prototype, and outputting along with feedback reception (see Table 1).

Table 1. Design process of health knowledge visualization based on interactive narrative.

Design process		Factors that need to be considered
Stage 1	Identifying audience needs	The age, knowledge background, and preferences of the audience
Stage 2	Determining textual content	Knowledge difficulty, knowledge domain, and narrative structure
Stage 3	Visual conversion of text	The degree of graphic verisimilitude and visual style
Stage 4	Crafting the interactive prototype	Interaction path, interaction habits, and user experience
Stage 5	Outputting and receiving feedback	Users' feedback

1. Identifying audience needs. As the cornerstone of the design activity, this stage requires clarifying the age, knowledge background, and demand preferences of the target audience, laying the foundation for determining the subsequent design direction.

2. Determining textual content. In this step, it is necessary to screen and refine the original knowledge content in combination with the age and knowledge background of the audience, so as to extract key knowledge texts that are easy for the audience to understand. Based on these knowledge texts, further write narrative texts, incorporate knowledge cases or stories to stimulate the audience's reading interest, optimize the narrative structure, and ensure the organic combination of knowledge texts and narrative texts, facilitating the audience's reading and understanding.
3. Visual conversion of text. When choosing the degree of verisimilitude, it is necessary to comprehensively consider the difficulty and field of health knowledge, as well as the age and knowledge background of the audience. For example, for health knowledge that is difficult and involves sensitive organs, if the audience is young or has a shallow medical background, a lower degree of visual verisimilitude is advisable. Conversely, if the medical difficulty is low and the audience has a certain knowledge background, a higher degree of verisimilitude can be selected. Given that not all audiences can accept pictures of real human tissues, and complex pictures may not clearly explain the pathology, therefore, comprehensively considering the difficulty and field of health knowledge as well as the characteristics of the audience to choose the degree of verisimilitude is a crucial step in the visual transformation process. After determining the degree of visual verisimilitude, use a unified visual style for expression to achieve a beautiful and clear picture.
4. Crafting the interactive prototype. In this stage, first, connect the content according to the narrative structure, and then enrich and adjust the interactive form in combination with interactive behaviors and user experience to ensure a smooth reading experience for users. At the same time, select an interactive form that is highly relevant to the text content to promote users' better integration into the text and understanding of the content. In addition, add necessary sounds and subtitles to enhance the user experience.
5. Outputting and receiving feedback. That is, invite some audiences to watch the results and make corresponding adjustments and modifications to the content according to their feedback.

5 A Case Design Based on Interactive Narrative

Building upon the aforementioned design process, this study adopts health knowledge about headache causes as its theme and embarks on a practical exploration of visualization design. Initially, the study identifies the target audience as middle-aged individuals aged between 40 and 50 with a limited foundation of health knowledge. This demographic is also a pivotal focus of China's strategy to promote the reform of the medical system, emphasizing prevention. If this group can amass a wealth of health knowledge and enhance their health literacy before various health issues begin to manifest as they enter the 50 to 60 age brackets, they will be better equipped to respond to and manage potential health challenges. Consequently, this study conducts in-depth interview surveys with this demographic, aiming to gain a comprehensive understanding of their health knowledge needs concerning common headache symptoms. Table 2 outlines the structure of this interview survey.

Table 2. Interview outline for users on the visualization design of headache-related health knowledge.

Stage number	Question theme	Questions
1	Introduction	Introduce yourself, briefly explain the purpose and approximate duration of this interview to the interviewee, emphasize the confidentiality and voluntariness of the interview, and obtain the interviewee's consent
2	Basic information	Basic information, including age, gender, occupation, and educational level, as well as daily living and work routines, such as working hours, sleep patterns, and exercise frequency, etc.
3	Experience of headaches	Do you often experience headaches? How frequently do they occur (several times a month, a few times a week, etc.)? Under what conditions do headaches typically arise (such as after exhaustion, during periods of high stress, in particular environments, etc.)? Where is the headache usually centered (forehead, temples, back of the head, etc.)? How would you characterize the intensity of the headache (mild, moderate, severe, and whether it impacts daily activities)?
4	Willingness to Learn about Headache Knowledge	Have you ever pondered the causes of headaches, such as those induced by mental stress or related diseases? Are you curious about the common triggers of headaches, including dietary, environmental, and psychological factors? Would you like to learn about the diagnostic methods for headaches, such as clinical and imaging examinations? Are you interested in the various treatment options for headaches, including medication, physical therapy, and traditional Chinese medicine? And do you wish to understand how to prevent headache symptoms through lifestyle adjustments and dietary precautions?

(*continued*)

Table 2. (*continued*)

Stage number	Question theme	Questions
5	Channels of information acquisition	Through which channels do you usually obtain health knowledge related to headaches (the Internet, doctors, books, relatives and friends, etc.)? Is the information provided by these channels accurate and comprehensive? What are the deficiencies?
6	Ending the Interview	Do you have any additional information or opinions to share? Thank you once again for taking the time to participate in the interview

In this interview survey, 10 interviewees aged between 40 and 50 were randomly selected. Their average age was 44, and the gender ratio was 1:1. All interviewees had received a high school education or above and had experienced headaches, thus meeting the interview qualifications. The interview results indicated that most middle-aged individuals in this age group were primarily concerned about diseases that may cause headaches, their potential risks, and methods of prevention and treatment. Building on this, we consulted Dr. Yang Rui from the Hospital of Huazhong University of Science and Technology. Under her professional guidance, we jointly developed the medical popular science material titled "Headaches Should Not Be Underestimated: Common Diseases Causing Headaches." Furthermore, based on this material, we created content that combines health knowledge with interactive narrative, which was ultimately presented in H5 format.

In the second phase of this study, we focused on refining the original knowledge materials and creating narrative texts. The original material is a popular science article with more than 2,700 words, covering various common diseases that may cause headaches, including hypertension, brain tumors, patent foramen ovale, and chronic angle-closure glaucoma. The article delves into the pathological mechanisms, causes, clinical manifestations, case analyses, as well as prevention and healthcare strategies of these diseases. Considering the potential lack of background knowledge among our target audience, complex pathological mechanisms could dampen their interest in reading. Additionally, some case analyses may not be suitable for all readers. Hence, we have opted to excise and simplify these elements to refine the educational content. In light of the reading preferences of middle-aged individuals, we have chosen the historical case of Cao Cao's headache from the Three Kingdoms era as the foundation for our narrative approach. Our goal is to tailor the text content to better match the reading tastes of our audience. Through this narrative technique, we aim to capture readers' interest and foster their comprehensive understanding of the scientific content. As for the narrative structure, we have implemented a straightforward story-tree framework with a non-linear narrative style. Centered around the story of Cao Cao's approach to his

headache, we guide users to virtually undergo the illness themselves. When confronted with the headache issue, users make various choices. Depending on these choices, we illustrate different life trajectories and health outcomes, and impart health knowledge pertaining to these possibilities. Therefore, the scientific knowledge each user acquires is personalized health information based on their individual decisions. Compared to the traditional linear narrative structure, this non-linear narrative method offers users the freedom to selectively access one or more pieces of popular science content on health knowledge based on their personal needs, thus avoiding information overload and reducing the reading burden on users. Furthermore, this method aligns more closely with the trend of fragmented reading times among Internet users (see Fig. 2).

Based on the aforementioned text, the third part involves the setting of the degree of graphic verisimilitude and visual style. Considering that the target audience has a limited level of knowledge about health, and the difficulty of the knowledge is relatively low, visual icons with a low degree of realism were selected to visually represent the disease symptoms, treatment plans, and preventive suggestions in the health knowledge text. Furthermore, by integrating the audience's preferences with the narrative content, this study adopted the visual style of the ancient-style game Canal Towns(江南百景图) as the visual theme for this H5 project. Meanwhile, the main characters and scenes from the popular science narrative were depicted, including Lady Bian, Cao Cao's wife, Cao Cao himself, and Hua Tuo, the miracle doctor. In the depiction of the characters' images and costumes, this study referenced the portrayals of Lady Bian and Cao Cao from TV dramas set in the Three Kingdoms period, as well as the general public's perception of the image of Hua Tuo, the miracle doctor. In this H5, Lady Bian primarily assumes the role of conveying disease information to Cao Cao; Cao Cao is tasked with making choices to progress the story; and Hua Tuo, the miracle doctor, is responsible for diagnosing Cao Cao's illness, disseminating health knowledge, suggesting treatment plans, and further driving the plot development. Considering the story is set during the Three Kingdoms period, the visual design of the scenes was inspired by the architectural styles, interior decorations, and traditional patterns of that era, such as cloud-head patterns, kui-phoenix patterns, and scroll-grass patterns. For more details, refer to Fig. 3.

During the creation of the interactive prototype, this study incorporated the brush of Hua Tuo, an esteemed ancient physician, as a design element for the loading interface. The fluidity of the ink imagery represents the completion of the loading process. Following this, the system introduces a door that signifies time travel. Users can open this door by long-pressing the page, allowing them to enter the historical time-space of 200 AD (the fifth year of the Jian'an era). In the design of the H5 option page, this study incorporated various emotional languages. The goal is to encourage users' clicking behavior, inspire them to try again, re-explore, and view more possible endings, thereby enabling them to absorb richer knowledge content. The medium has been designed with four decision-making nodes to ensure that users can fully experience the entire content after making four choices. The concluding segment, serving as the pivotal science-education component of this visual design, includes the following three primary aspects:

1. Test results, demonstrating to users their estimated lifespan during the Three Kingdoms period;

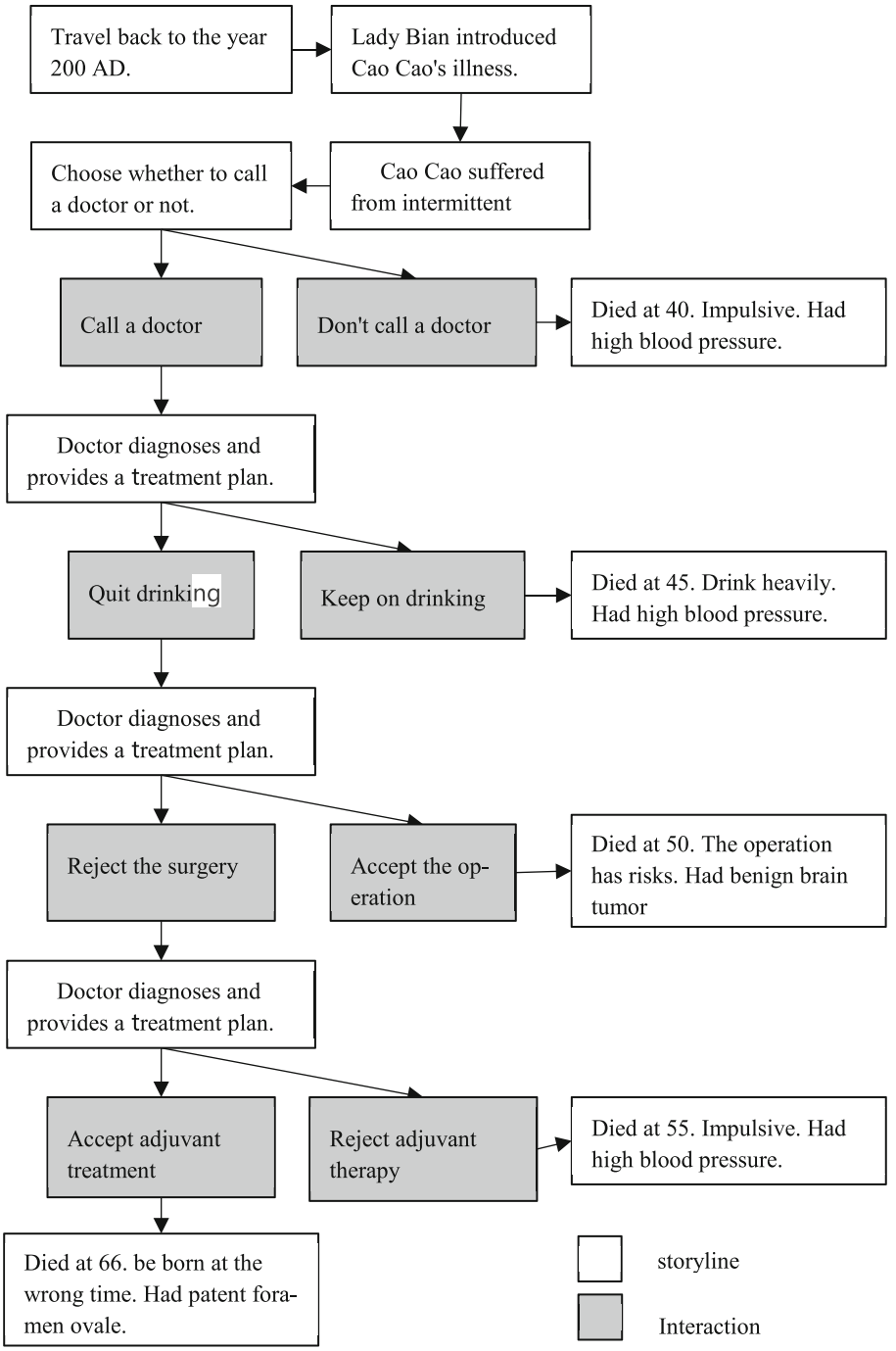


Fig. 2. Narrative structure of Headaches should not be underestimated: Common diseases causing headaches.



Fig. 3. Characters and scenes for Headaches should not be underestimated: Common diseases causing headaches.

2. The science-education aspect of the disease affecting the user and its causative factors. For instance, if a user has hypertension, it may be linked to factors such as extreme emotional fluctuations and excessive alcohol consumption;
3. The clinical manifestations, treatment strategies, and preventive measures associated with the disease.

In addition, classical-style sound effects have been added to this medium to enhance users' immersive interactive experience and emotional engagement (see Fig. 4).

To investigate the application effect of interactive narrative methods in the visualization design of health knowledge, this study conducted a user experiment using a



Fig. 4. Key pages of Headaches should not be underestimated: Common diseases causing headaches.

questionnaire survey. The questionnaire consists of 15 questions, structured across three dimensions. Initially, for the dimension of users' basic information, two non-scale questions are included. Subsequently, concerning the design's attractiveness, eight scale-type questions are formulated, encompassing users' willingness to click, willingness to read, overall evaluation of design attractiveness, and assessment of the appeal of design details. Finally, around the dimension of whether the design can guide users to think about and understand the popular science content, five scale-type questions are designed. For the specific questionnaire questions, please refer to Table 3.

The sample size of the subjects was calculated using G*Power. With a preset effect size, statistical power, and significance level, the calculation results indicated that a total of 34 subjects were required for the experiment. In this study, 82 questionnaires were collected. Given that the target user group was individuals aged 40–50, 7 questionnaires that did not meet the criteria for the target user group and 4 with excessively short answering times were excluded. Consequently, 71 valid questionnaires remained. The gender distribution among the respondents of the valid questionnaires was nearly balanced, at approximately 1:1. The average age was 46 years old. The respondents had at least a high school education, demonstrated good reading comprehension skills, and had a limited understanding of the experiment's knowledge theme - "health knowledge related to headache symptoms," which was in line with the experimental requirements. Subsequently, the data from the valid questionnaires were imported into SPSS, and reliability analyses were conducted on the questions within the two dimensions of attractiveness and understanding in the questionnaire. The results of the reliability analysis revealed that the Cronbach's alpha coefficient for the attractiveness dimension was 0.851 (see Table 4), and for the thinking and comprehension dimension, it was 0.822 (see Table 5). Clearly, the data from the questions in both dimensions of the questionnaire were relatively reliable, and the questionnaire data could be utilized.

Table 3. List of questions.

Dimension	Number	Questions
Basic information	1	What's your age?
	2	What's your gender?
Attractiveness	3	If you encounter this kind of H5 in life, would you be willing to click on it to view and browse?
	4	Does the design of this H5 enable you to patiently read through and understand the content described in it?
	5	Do you think the interactive narrative method in this H5 is useful for enhancing the attractiveness of health knowledge popularization content?
	6	How do you think the exquisite pictures in this H5 contribute to enhancing the attractiveness of health knowledge popularization?
	7	How do you think the gamified story arrangement in this H5 contributes to enhancing the attractiveness of health knowledge popularization?
	8	How do you think the interactive click - able method in this H5 contributes to enhancing the attractiveness of health knowledge popularization?
	9	How do you think the soothing music in this H5 contributes to enhancing the attractiveness of health knowledge popularization?
	10	How do you think the visualized popular science content in this H5 contributes to enhancing the attractiveness of health knowledge popularization?
Thinking and comprehension	11	Do you think the interactive narrative method in this H5 is helpful for guiding you to think about and understand the content of health knowledge?
	12	How do you think the visualized graphic design in this H5 contributes to guiding you to think about and understand the content of health knowledge?
	13	How do you think the interactive narrative method in this H5 contributes to guiding you to think about and understand the content of health knowledge?

(*continued*)

Table 3. (*continued*)

Dimension	Number	Questions
	14	How do you think the plain - language copy content in this H5 contributes to guiding you to think about and understand the content of health knowledge?
	15	How do you think the gamified approach in this H5 contributes to guiding you to think about and understand the content of health knowledge?

Table 4. Reliability analysis of the attractiveness dimension.

	Statistics for reliability	
Cronbach's alpha	Cronbach's alpha based on standardized items	N of items
.851	.855	8

Table 5. Reliability analysis of the thinking and comprehension dimension.

	Statistics for reliability	
Cronbach's alpha	Cronbach's alpha based on standardized items	N of items
.822	.825	5

Upon conducting an in-depth analysis of the questionnaire data, the following conclusions have been deduced:

1. Interactive narrative technology significantly enhances the appeal of health knowledge dissemination. A total of 96% of the participants concurred that health knowledge dissemination employing interactive narrative technology is more engaging than traditional graphic-text formats, leading to a greater propensity for them to click on and peruse the content.
2. The application of interactive narrative technology has demonstrated significant efficacy in enhancing users' comprehension of health-related information. A survey indicated that 65% of participants found that interactive narrative technology enabled them to engage more profoundly with the content and facilitated a more effective understanding of health knowledge, as opposed to traditional graphic-text dissemination methods.
3. The interactive narrative technology exhibits significant potential for future advancements in the dissemination of health knowledge and is highly anticipated by users. Within this experimental design for visualizing health knowledge, the interactive narrative technology not only highlights its distinctive advantages but also suggests extensive developmental opportunities. Furthermore, 73% of users anticipate that

interactive narrative technology will be widely utilized to develop content for future health knowledge dissemination efforts.

Upon conducting the aforementioned analysis, it becomes evident that interactive narrative technology exerts a favorable influence on the visual communication design associated with the dissemination of health knowledge. It notably augments the visual appeal of such knowledge, rendering users more disposed to engage proactively and peruse content related to health. Additionally, this technology possesses the capability to steer users towards profound contemplation, thereby augmenting their comprehension of health-related information. Consequently, the integration of interactive narrative technology within the visual communication design for health knowledge dissemination can efficaciously enhance the effectiveness of health knowledge transmission. This approach holds significant importance and possesses profound implications for bolstering the public's awareness of health preservation and alleviating the strain on medical resources.

6 Conclusion

Through case design and experimental research, this study has verified that the application of interactive narrative methods in the visual design of health knowledge has significant positive effects. This approach can remarkably enhance the attractiveness of health knowledge content. With its unique interactivity and storytelling nature, it subtly guides users to think deeply and comprehensively understand the health knowledge content. However, it must be pointed out that the scope of this study's tests is limited to people aged 40 to 50. Given the significant differences in cognitive levels, acceptance habits, and information needs among different age groups, whether the interactive narrative method also has positive effects on other age groups requires further in-depth research. In addition, this study mainly focuses on health knowledge popularization content. If the research field is expanded to other areas such as technology and culture, considering the differences in knowledge systems, audience characteristics, and communication requirements in each field, whether the interactive narrative method can still maintain its positive influence is also a topic worthy of extensive and in-depth discussion and research. This will provide a solid theoretical foundation for the more extensive and precise application of the interactive narrative method.

References

1. Li, H.: Research on Information Visualization Design in Medical Science Popularization. MA Thesis, Soochow University (2020). https://doi.org/10.27351/d.cnki.gszhu.2020.001987
2. Healthy China Initiative: Healthy China Initiative (2019–2030): general requirements, major actions and main indicators. Chin. Circ. J. **34**(9), 846–858 (2019)
3. The State Council of the People's Republic of China: The State Council's Opinions on Implementing the Healthy China Initiative. People's Daily, 7 (2019)
4. Wu, Y., et al.: Literature analysis of health science popularization research in China over the past 20 years. Sci. Popul. Res. **12**(03), 39–45 (2017). https://doi.org/10.19293/j.cnki.1673-8357.2017.03.006

5. Qiang, L., Bingjie, C., Haibo, Z.: Storytelling by the StoryCake visualization. Vis. Comput. **33**(10), 1241–1252 (2017)
6. Yu, S.: Several structural possibilities and application prospects of interactive narrative. J. Beijing Inst. Technol. **12**, 101–104 (2010)
7. Méndez, G., et al.: Dynamic emphatical narration for reduced authorial burden and increased user freedom in interactive storytelling. Connect. Sci. **31**(1), 33–59 (2019)
8. Wei, X., Wang, Y.: Research on the design strategy of educational games based on the interactive narrative mode. E-Educ. Res. **44**(09), 108–113+121 (2023). https://doi.org/10.13811/j.cnki.eer.2023.09.015
9. Ary, S.M., et al.: A computer-based winter active narrative and a serious game for children with asthma: development and content validity analysis. J. Med. Internet Res. **23**(9), e28796 (2021)
10. Tian, L., Wenli, M.: Application and research of interactive narrative in vaccine knowledge visualization. Res. Des. Arts. **12**(2), 104–109 (2022)
11. Yajuan, C.: Research on the cross-media narrative of popular science videos from the perspective of converged media. Sci. Popular. Res. **18**(06), 14–23+94–95 (2023). https://doi.org/10.19293/j.cnki.1673-8357.2023.06.002



Designing Automation for Enhancing Multitasking Performance

Jiaxin Li, Wenxin Xiao, and Ji-Eun Kim(✉)

University of Washington, Seattle, WA 98105, USA
jikim@uw.edu

Abstract. Partial automation enhances multitasking performance by reducing the overall attention demands of multitasking placed on human operators. However, there is a critical gap in researchers' understanding of how to design effective partial automation in multitasking, particularly when it comes to determining the types of tasks that should be automated. This study explores the impact of partial automation on multitasking performance under different sensory modalities and task prioritizations. Our study utilized the Multi-Attribute Task Battery II (MATB-II) as its multitasking environment and employed a 2×2 within-subjects design with two independent variables: sensory modality (cross-modality vs. intra-modality) and task priority (equal vs. unequal priority). The individual subtask and overall multitasking performances of 20 participants were measured. Results indicate that partial automation led to greater improvements in both types of performance under cross-modality than under intra-modality. These findings offer guidance for designing automation allocation strategies to enhance multitasking performance.

Keywords: Multitasking · Partial Automation · Sensory Modality

1 Introduction

1.1 Partial Automation in Multitasking

Partial automation has been increasingly integrated into daily activities, allowing systems to handle some tasks while leaving others under human control [1]. Partial automation has been recognized for its ability to enhance multitasking performance by reducing the workload on human operators [2]. For example, partial automation in aviation enhances pilots' multitasking performance by automating routine tasks, thereby reducing total task demands and pilot workload [3]. The autopilot system maintains altitude, heading, and speed, but pilots remain responsible for communicating with air traffic controllers and handling safety-critical situations, such as those that require the pilots to make evasive maneuvers to avoid collisions [4]. In addition, pilots are expected to continuously monitor the automated system to detect and mitigate automation failures [5].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 227–237, 2025.
https://doi.org/10.1007/978-3-031-93502-2_15

Compared to full automation, partial automation requires less investment in technology and reduces operator overreliance on automation by keeping humans in the loop [6]. Previous studies indicate that partial automation enables operators to intervene when automated systems encounter errors or human judgment is essential, thereby improving error mitigation and overall safety [1,7]. However, a key challenge in implementing partial automation is determining which tasks should be automated. When designing the automation strategy, system engineers must consider the attributes of each task and the task's impact on human operators, ensuring that the automation supports rather than hinders operators' ability to manage multiple tasks effectively [8].

1.2 Factors Impacting Multitasking

The effectiveness of partial automation depends on an understanding of the factors that influence operator performance during multitasking [6]. Two critical factors-sensory modality and task priority-influence operator multitasking performance and need to be carefully considered when designing automation interventions. Sensory modality refers to the channel through which individuals collect information from environmental stimuli. These stimuli are typically classified based on the sensory receptors they activate [9]. Visual and auditory modalities are predominant in most safety-critical settings [10]. According to multiple resource theory [11], multitasking interference is greater when tasks use the same sensory modality (e.g., two visual tasks) and lower when tasks use different modalities (e.g., one visual task and one auditory task).

Task priority, another crucial factor, affects multitasking by prompting operators to allocate more attentional resources to high-priority tasks. This prioritization enhances operator performance on these tasks but may lead to decreased performance on tasks of lower priority [12]. In a multitasking scenario involving rerouting, maintenance, and target detection of an unmanned aerial vehicle, participants exhibited higher accuracy and faster response times on the explicitly prioritized task while showing lower accuracy and slower responses on secondary tasks [13].

1.3 Research Question

While research has highlighted the importance of task-related factors in multitasking, key questions remain about the role these factors play in the presence of partial automation. Understanding their influence is essential for designing automation strategies that enhance the system and operator's combined multitasking performance. Specifically, it is unclear which tasks in a multitasking system should be automated given these factors [8]. In the following sections of this paper, we refer to the system and operator's combined multitasking performance as "overall multitasking performance". Existing research has only partially addressed this issue. For example, [14,15] investigated the impact of partial automation on operators' attention allocation and overall multitasking performance using the MATB-II. However, both studies preselected the automated

task and did not explore other possibilities for task selection. Meanwhile, [16] conducted a qualitative study examining how automation reliability and task-related factors affect operators' multitasking strategy but did not quantify their impact on overall multitasking performance. [17] analyzed how automating different tasks affects overall performance in air traffic control yet did not explain why the automation of certain tasks led to better performance. This study examines how to design partial automation in a multitasking environment, specifically focusing on how to select tasks for automation based on sensory modality and task priority in order to enhance overall multitasking performance.

2 Methodology

2.1 Participants

Twenty participants (14 women, 6 men) with a mean age of 25.55 years (SD = 4.14) from the University of Washington participated in this study. All participants had normal or corrected-to-normal vision. Individuals with visual or hearing impairments, color blindness, or photosensitive epilepsy, as well as those wearing medical devices that might have been affected by the equipment used in this study, were excluded.

2.2 Experimental Design and Procedure

We employed a 2 × 2 within-subjects design with the two independent variables, sensory modality (i.e., intra- and cross-modality) and task priority (i.e., equal and unequal priority). The level of sensory modality was defined by the combination of two tasks in the multitasking environment. The intra-modality condition involved a visual–visual task combination, while the cross-modality condition featured a visual–auditory task combination. Task priority was determined by the researcher, who before each session specified whether there would be a priority difference between the two tasks [18]. In this study, partial automation was implemented by automating one task while keeping the other task manual. The automation's reliability was set at 80%. Only the automated task would be prioritized if there was a priority difference.

Participants who met the inclusion criteria participated in the experiment by visiting the Human and Systems Laboratory at the University of Washington. Upon arrival, participants were asked to consent to participate in the study. Those who consented were provided with verbal instructions related to the experimental procedure. Participants then watched a short video about the experiment design and how to perform the tasks. Next, participants completed practice sessions to familiarize themselves with the keyboard controls. In the main experiment, each participant completed a total of six trials: two baselines without automation and priority differences, followed by four experimental trials, as shown in Fig. 1. After each session, the participants took a 1-minute break. The entire experiment lasted around 90 min.

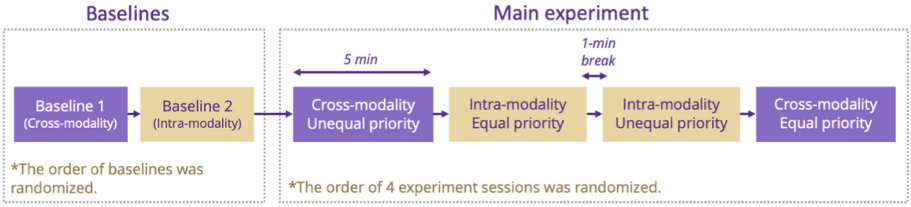


Fig. 1. Design of Experiment.

The dependent variables in this study include the individual performance measures for each subtask and a combined measure representing subtasks' combined performance. The description of each dependent variable can be found in Sect. 2.3. Table 1 lists all the independent and dependent variables.

Table 1. Independent and dependent variables

Category	Variable
Independent variable	Sensory modality (cross-modality, intra-modality)
	Task priority (equal, unequal)
Dependent variable	Inverse efficiency score (IES)
	Number of abnormal tanks
	Sum of errors

Multitasking Environment. MATB-II, a computer-based flight simulation software program developed by the National Aeronautics and Space Administration (NASA) [19], was used as the multitasking environment in this study. We utilized an open-source version of MATB-II, Open-MATB [20], so that we could create a customized user interface tailored to our research question. Unlike the original MATB-II, our customized setting using Open-MATB showed two subtasks displayed side by side. Our study had three subtasks: system monitoring (SYSMON), communications from the MATB-II user guide (COMM), and resource management (RESMAN). The left panel featured either the SYSMON or COMM task, depending on the experimental condition, while the right panel displayed the RESMAN task, as illustrated in Fig. 2. In the four experimental conditions, the task displayed on the left panel (SYSMON or COMM) was automated, while that on the right panel (RESMAN) remained manual.

The intra-modality (visual–visual) condition was created by activating SYSMON and RESMAN at the same time. SYSMON requires participants to respond when they detect abnormalities in any of the four gauges displayed on the interface. When an arrow on a gauge was offset, participants pressed the corresponding key (F1, F2, F3, or F4) on the keyboard to correct it. RESMAN

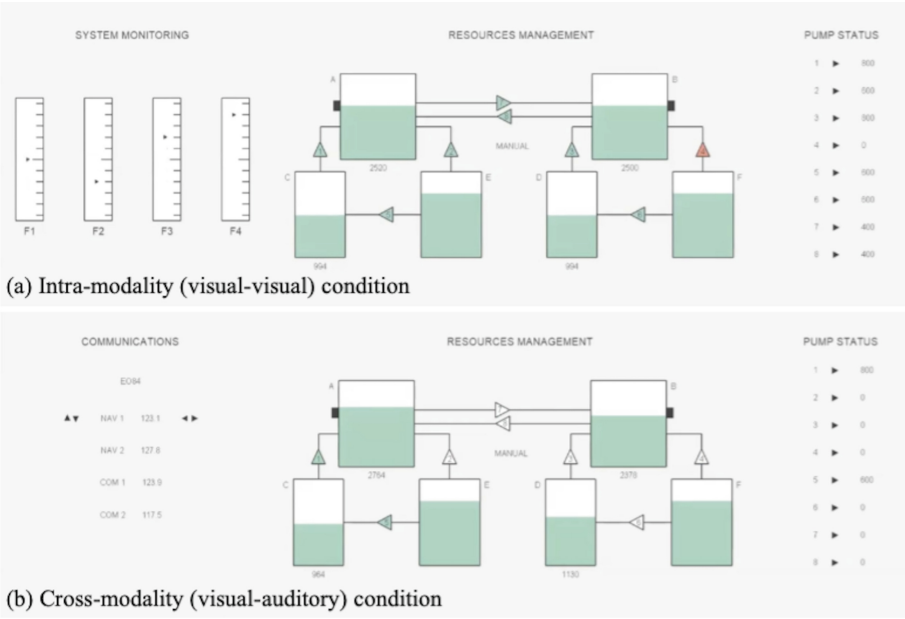


Fig. 2. User interface of multitasking environment: (a) The intra-modality condition included two visual tasks, system monitoring (SYSMON) and resource management (RESMAN). SYSMON was the automated task, while RESMAN was the manual task; (b) The cross-modality condition included an auditory task, communication (COMM), and a visual task, RESMAN. COMM was the automated task, while RESMAN was the manual task.

involves maintaining the fuel levels of two target tanks on the airplane within the desired range. Participants managed fuel by transferring it between the target tanks (Tank A and Tank B) and supplementary tanks using pumps, which they could activate or deactivate by pressing the corresponding number key on the numpad.

The cross-modality (visual–auditory) condition was created by pairing COMM and RESMAN. The COMM task requires participants to listen to radio messages and adjust the radio channel and frequency based on the messages. The participants used the up and down arrow keys to change the channel and the left and right arrow keys to adjust the frequency. They then confirmed their selection by pressing the return key.

To minimize any potential bias in performance incurred by differences in subtask difficulty, we balanced the difficulty of SYSMON and COMM by using the Baud rate [21], which quantifies the rate of information presentation to the participant, as measured in bits per second (bps). Equation 1 shows the Baud rate formula, where B represents the Baud rate of a subtask, N indicates the number of decision alternatives, and ΔT denotes the time interval between two events in the subtask [22]. We controlled the Baud rate for the pair of tasks

at 0.25 bps, with SYSMON presenting eight events per minute and COMM presenting three events per minute. The RESMAN task was designed with eight events per minute to balance the difficulty of the time-sharing tasks.

$$B = \frac{log_2(N)}{\Delta T} \tag{1}$$

2.3 Data Analysis

Participants' responses to the subtasks, along with their response times, were recorded by the Open-MATB software. The system's status, such as the fuel levels of Tank A and Tank B, was also logged for analysis. To evaluate overall multitasking performance, we employed a variety of behavioral responses.

- Inverse efficiency score (IES): We used IES, a performance measure that combines response time and the proportion of error responses [23], to generate a single metric by which to evaluate performance on the automated task (SYSMON or COMM). Errors in SYSMON included misses, defined as failures to correct an abnormal gauge, and false alarms, defined as incorrect adjustments to a normal gauge. Errors in COMM were categorized as missing a radio adjustment request, changing a radio channel or frequency without a corresponding request from the radio message, or adjusting the incorrect channel or frequency. Equation 2 indicates how to compute IES.

$$\text{IES} = \frac{\text{Response time}}{1 - \text{Error rate}} \tag{2}$$

- The number of abnormal tanks: We considered the number of tanks outside the desired fuel level (either above or below), referred to as "abnormal tanks", to evaluate performance on the manual task (RESMAN). The possible values for the number of abnormal tanks were 0, 1, or 2.

- Sum of errors: We used the combined sum of errors from the automated and manual tasks to evaluate overall multitasking performance.

To normalize the comparisons of behavioral responses across different conditions, we analyzed the relative change in each of the performance metrics from baseline, as shown in Eq. 3. $X_{\text{experimental}}$ represents the value of a behavioral response under the experimental conditions, and X_{baseline} represents its value under the baseline conditions. X_{relative} denotes the relative change in that behavioral response.

$$X_{\text{relative}} = \frac{(X_{\text{experimental}} - X_{\text{baseline}})}{X_{\text{baseline}}} \tag{3}$$

We conducted a non-parametric two-way ANOVA using the aligned rank transform (ART) to examine the effects of sensory modality and task priority on

the behavioral responses. ART is a robust method for analyzing factorial designs with non-normal data, allowing for the testing of main effects and interactions [24]. The extended ART method adjusts for the repeated measures in our within-subect design by accounting for individual variability across conditions [25].

3 Results

3.1 Descriptive Statistics

Table 2 displays the descriptive statistics of dependent variables for each experimental condition across all participants, including the mean and standard deviation for each variable. Based on the definitions of the three behavioral responses-IES, number of abnormal tanks, and sum of errors-lower values indicate better performance. Responses are shown in actual values.

Table 2. Descriptive statistics of dependent variables (mean and SD).

Dependent Variable	Task Priority		Sensory Modality	
	Equal	Unequal	Cross-modality	Intra-modality
IES	1258.349 (553.037)	1216.619 (553.713)	1031.025 (641.142)	1443.944 (338.882)
Number of abnormal tanks	0.400 (0.744)	0.600 (1.057)	0.350 (0.700)	0.650 (1.075)
Sum of errors	1.475 (1.552)	1.625 (1.720)	1.375 (1.444)	1.725 (1.797)

3.2 Effects of Sensory Modality and Task Priority on Behavioral Responses

Automated Task. Participants showed lower IES on the automated task during cross-modality multitasking compared to intra-modality multitasking ($F = 13.145$, $p < 0.001$). According to the definition of IES, the lower the ratio of the response time to (1 − Error Rate) is, the better the performance of the automated task. The negative values represent improvement in performance compared to baseline. There was no significant effect observed from priority or the interaction of modality and priority. Figure 3 depicts the interaction plot for IES across sensory modalities and task priority levels.

Manual Task. We observed significantly fewer abnormal tanks during cross-modality multitasking compared to intra-modality multitasking ($F = 9.737$, $p < 0.01$). This suggests better performance on the manual task. There was no significant effect observed from priority or the interaction of modality and priority. Figure 4 shows the interaction plot for the number of abnormal tanks across sensory modalities and task priority levels.

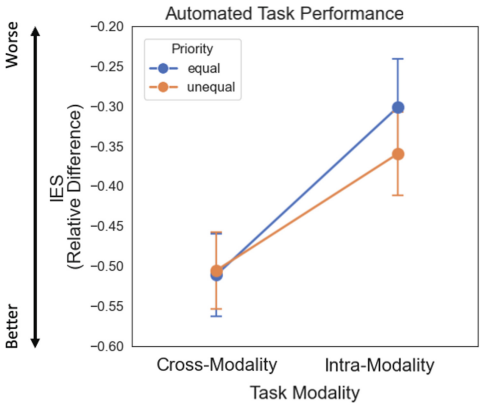


Fig. 3. Interaction plot for the inverse efficiency score (IES) in the automated task. The y-axis shows the relative difference in IES from baseline. Lower (more negative) values indicate better performance. Error bars represent standard errors.

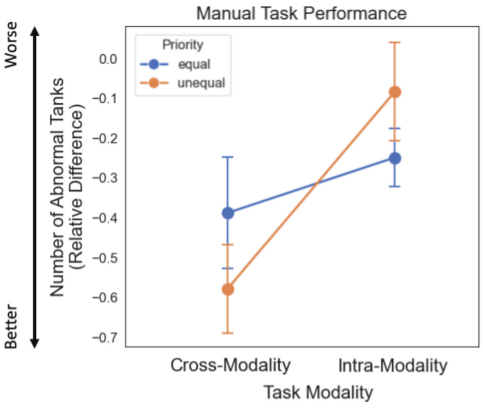


Fig. 4. The interaction plot for the number of abnormal tanks in the manual task. The y-axis represents the relative change of the number of abnormal tanks from baseline. Lower (more negative) values indicate better performance. Error bars denote standard errors.

Combined Multitasking Performance. With IES and the number of abnormal tanks, participants showed fewer sum of errors under the cross-modality condition than under the intra-modality condition, indicating a better overall multitasking performance under cross-modality ($F = 4.274$, $p < 0.05$). They also showed a marginal effect of task priority on sum of errors, with smaller sums of errors under the equal-priority condition compared to unequal-priority condition ($F = 2.982$, $p = 0.090$). There was no significant interaction effect of modality and priority. Figure 5 shows the interaction plot for the combined performance of the automated and manual tasks.

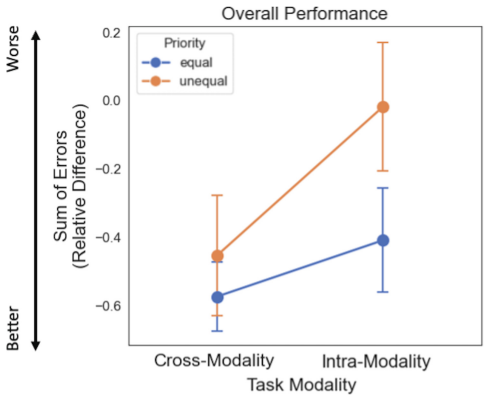


Fig. 5. The interaction plot for the sum of errors from the automated and manual tasks. The y-axis represents the relative change of the sum of errors from baseline. Lower values indicate better performance. Error bars denote standard errors.

4 Discussion

This study examined how sensory modality and task priority influence the effectiveness of partial automation in enhancing overall multitasking performance. Our results showed that cross-modality multitasking benefited more than intra-modality multitasking from the incorporation of partial automation, which highlights the importance of tailoring partial automation to account for sensory modalities in multitasking. No significant differences were observed between the equal- and unequal-priority conditions.

One explanation for the difference between sensory modalities is the impact of mapping auditory information onto spatially distributed visual elements on MATB-II during cross-modality multitasking. This process likely led to additional spatial transformation demands and increased cognitive effort, resulting in reduced performance [26,27]. In our study, participants had to map channel and frequency information provided in radio messages to visual panels using the MATB-II interface when performing an auditory task (COMM). To assess whether this could explain our results, we conducted a Wilcoxon signed-rank test comparing the cross-modality and intra-modality baselines. When there was no automation, participants in the intra-modality baseline exhibited better overall multitasking performance compared to those in the cross-modality baseline, as indicated by the former group's fewer sum of errors ($W = 41.5$, $p < 0.05$). Given that task difficulty between SYSMON and COMM was balanced, these findings suggest that the introduction of partial automation helped reduce the cognitive effort of performing spatial mapping in cross-modality conditions, leading to larger performance improvements.

This study has limitations that require further research. First, we focused on automation with 80% reliability, a value higher than the 70% used by other studies that have shown the efficacy of partial automation in enhancing performance

[14]. Future research could explore various levels of automation reliability to gain a more comprehensive understanding of how to design partial automation. Second, this study examined two levels of task priority, equal priority and unequal priority. The automated tasks, SYSMON and COMM, are both signal-detection tasks that are time-critical [28]. Considering the actual functionality of SYSMON and COMM in flight operations, responding to audio messages from air traffic control and detecting abnormal altitude gauges, are more safety-critical tasks, making them important for maintaining overall system functionality. In contrast, the manual task (RESMAN) involves planning and resource allocation [28], which is less time-critical and allows for more temporal flexibility. This difference is the primary reason we did not include a condition in which the manual task was prioritized. However, broadening the applicability of these findings would require exploring scenarios in which the manual task is prioritized while partial automation is active. Third, while MATB-II simulates the routine tasks of pilots during flights, exploring the effects of partial automation in real-world multitasking scenarios would provide valuable validation and enhance the generalizability of our findings beyond the simulator-based setting.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern.-Part A Syst. Hum 30(3), 286–297 (2000)
2. de Visser, E., Parasuraman, R.: Adaptive aiding of human-robot teaming: effects of imperfect automation on performance, trust, and workload. J. Cognitive Eng. Decision Making **5**(2), 209–231 (2011)
3. Wang, Y., Rongjin, H., Lin, S., Schultz, M., Delahaye, D.: The impact of automation on air traffic controller's behaviors. Aerospace **8**(9), 260 (2021)
4. Metzger, U., Parasuraman, R.: Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. In: Decision Making in Aviation, pp. 345–360. Routledge (2017)
5. Vu, K.-P.L., et al.: Pilot performance in trajectory-based operations under concepts of operation that vary separation responsibility across pilots, air traffic controllers, and automation. Int. J. Hum.-Comput. Interact. **28**(2), 107–118 (2012)
6. Endsley, M.R.: From here to autonomy: lessons learned from human-automation research. Hum. Factors **59**(1), 5–27 (2017)
7. Carsten, O., Martens, M.H.: How can humans understand their automated cars? hmi principles, problems and solutions. Cognition Technol. Work **21**(1), 3–20 (2019)
8. Calhoun, G.: Adaptable (not adaptive) automation: Forefront of human-automation teaming. Hum. Factors **64**(2), 269–277 (2022)
9. Auvray, M., Spence, C.: The multisensory perception of flavor. Conscious. Cogn. **17**(3), 1016–1031 (2008)
10. Hutmacher, F.: Why is there so much more research on vision than on any other sensory modality? Front. Psychol. **10**, 481030 (2019)

11. Wickens, C.D.: Multiple resources and performance prediction. Theoretical Issues Ergon. Sci. **3**(2), 159–177 (2002)
12. Barg-Walkow, L.H., Rogers, W.A.: Modeling task scheduling in complex healthcare environments: identifying relevant factors. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 61, pp. 772–775. SAGE Publications Sage CA: Los Angeles, CA (2017)
13. Liu, Y., Gao, Q.: Effects of secondary task eccentricity and visual salience on attention allocation in multitasking across screens. Int. J. Hum. Comput. Stud. **192**, 103363 (2024)
14. Onnasch, L.: Crossing the boundaries of automation-function allocation and reliability. Int. J. Hum. Comput. Stud. **76**, 12–21 (2015)
15. Karpinsky, N.D., Chancey, E.T., Palmer, D.B., Yamani, Y.: Automation trust and attention allocation in multitasking workspace. Appl. Ergon. **70**, 194–201 (2018)
16. Dan, C.S., Cullen, R.H., Rogers, W.A., Fisk, A.D.: Exploring strategy use in a multiple-task environment: Effects of automation reliability and task properties. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, pp. 2123–2127. SAGE Publications Sage CA: Los Angeles, CA (2012)
17. Strybel, T.Z., Vu, K.-P.L., Chiappe, D.L., Morgan, C.A., Morales, G., Battiste, V.: Effects of nextgen concepts of operation for separation assurance and interval management on air traffic controller situation awareness, workload, and performance. Int. J. Aviation Psychol. **26**(1-2), 1–14 (2016)
18. Jansen, R.J., van Egmond, R., de Ridder, H.: Task prioritization in dual-tasking: instructions versus preferences. PLoS ONE **11**(7), e0158511 (2016)
19. Santiago-Espada, Y., Myer, R.R., Latorella, K.A., Comstock Jr. J.R.: The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide. Technical Report L-20031, NASA (2011)
20. Cegarra, J., Valéry, B., Avril, E., Calmettes, C., Navarro, J.: Openmatb: a multi-attribute task battery promoting task customization, software extensibility and experiment replicability. Behav. Res. Methods **52**, 1980–1990 (2020)
21. Camden, A., Nickels, M., Fendley, M., Phillips, C.A.: A case for information theory-based modelling of human multitasking performance. Theoretical Issues Ergon. Sci. **18**(3), 266–278 (2017)
22. Liu, S., Nam, C.S.: Quantitative modeling of user performance in multitasking environments. Comput. Hum. Behav. **84**, 130–140 (2018)
23. Bruyer, R., Brysbaert, M.: Combining speed and accuracy in cognitive psychology: is the inverse efficiency score (ies) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)? Psychologica Belgica **51**(1), 5–13 (2011)
24. Elkin, L.A., Kay, M., Higgins, J.J., Wobbrock, J.O.: An aligned rank transform procedure for multifactor contrast tests. In: The 34th Annual ACM Symposium on user Interface Software and Technology, pp. 754–768 (2021)
25. Mark Beasley, T.: Multivariate aligned rank test for interactions in multiple group repeated measures designs. Multivariate Behav. Res. **37**(2), 197–226 (2002)
26. Stock, A.-K., Gohil, K., Huster, R.J., Beste, C.: On the effects of multimodal information integration in multitasking. Sci. Rep. **7**(1), 4927 (2017)
27. Tsang, S.N.H., Chan, A.H.S., Pan, X., Man, S.S.: Auditory versus visual spatial stimulus-response mappings in tracking and discrete dual task performance: implications for human-machine interface design. Ergonomics **64**(4), 485–501 (2021)
28. Pontiggia, A., et al.: Matb for assessing different mental workload levels. Front. Physiol. **15**, 1408242 (2024)

Establishing a Fire Risk Map Based on Planned Spatial Layouts and Environment in Rural Planning: A Case Study of a Stereotype Village in China

Changda Ma(✉), Yichao Shi, Chunlan Wang, and Patrick Kastner

Georgia Institute of Technology, Atlanta, GA 30332, USA
{cma326,yshi431,cwang932,patrick.kastner}@gatech.edu

Abstract. Among Chinese rural heritage, more than 3,000 ancient villages not only provide a rich historical narrative, but also face the challenges of slow economic development and inadequate fire risk management. Addressing these challenges requires more than analyzing the fire resistance of building materials and structures—it's necessary to study the spatial layout and wind environment to enhance fire prevention capabilities. In this research, Zhaoxing Dong Village, located in China is selected as the case study, with the village is divided into 100 units for fire assessment based on two aspects. First, the field of view area and evacuation characteristics of different routes to evaluate evacuation efficiency and to determine the fire risk coefficient of the spatial layout. Second, using Standard Fuel Models, slope analysis, the active and dead fuel moisture was estimated based on the average precipitation probability for the village in 2020. Wind adjustment factors (WAF) were calculated from average wind speeds at heights of 2.5 ft, 5 ft, and 22.5 ft. By combining these data, mid-flame wind speeds, flame lengths, and maximum surface fire spread rates are derived to access the impact of environmental factors on fire risk management. Ultimately, the fire risk for each unit is evaluated by integrating the fire risk factors associated with the spatial layout and environment, resulting in a fire risk map. This approach provides a comprehensive framework for understanding and mitigating fire risks, supporting fire risk management and reduction, shaping the future of health and safety through data-driven, people-focused solutions.

Keywords: Rural Fire Risk Assessment · Spatial Layout · Environments · Rural Planning · Fire Risk Map

1 Introduction

Fire, a major threat among natural disasters, is a sudden and rapidly widespread event that can cause severe damage and seriously impact buildings, urban and rural areas, as well as the ecological environment, thereby having a significant impact on people's health [1, 2]. Rural communities, particularly those residing in wooden structures, are typically at an increased risk of fire disasters [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 238–256, 2025.
https://doi.org/10.1007/978-3-031-93502-2_16

In Southwest China, there are frequent fires in local Dong Minority Rural Village due to overloaded electrical outlets and outdated wiring systems caused by indoor fire pits used for cooking, heating, lighting, and religious purposes, which has a serious impact on life safety [4, 5]. In mountain and highland areas, the *"Ganlan-Style House"* in Dong Minority Rural Village is built close together without fire breaks, with the aim of containing fires in a limited area. This lack of separation increases the sensitivity of large-scale hill fire incidents [3]. Moreover, the fire hazards are greatly increased by the flammability of the *"Ganlan-Style House"* which built by wood and the Dong people's shallow knowledge of fire risk.

Current research on fire risk management in villages is limited to the flammability of building materials [6], which poses a high fire hazard, isolated research on the spatial analysis of the village layout [7], or the single-sided study of urban microclimate on fire risk management [8]. However, there has been no comprehensive assessment of the relative importance of fire risk factors, such as the integration of spatial layouts, environmental conditions, and other contributing elements, to achieve effective fire control.

The spatial layout significantly impacts fire risk. Clear, spacious evacuation routes and proper building spacing prevent fire spread. Besides, effective area allocation enhances control and evacuation, reducing risks and losses. Therefore, A well-designed layout is crucial for safety and minimizes fire-related damage [9].

Environmental factors like temperature, humidity, precipitation, wind [10], and topography significantly influence fire behavior [11]. These elements affect fire spread, and control difficulty. Understanding their combined impact is also crucial for effective fire prevention and management strategies.

The main purpose of this study is to assess the impact of spatial layout and environmental factors on fire risk in Zhaoxing Dong Village in Guizhou Province, quantifying how these combined factors contribute to fire incidents, analyzing the spatial configuration of buildings and evacuation routes in terms of the area of visual field, length of evacuation routes, and zoning, as well as environmental influences such as moisture, wind speed, and topography, and visualizing a comprehensive fire risk assessment model(fire risk map). The model will not only deepen our understanding of fire dynamics in similar rural environments but also suggest targeted fire management strategies to protect such vulnerable communities from fire disasters.

2 Literature Review

- **Quantifying Fire Risk Assessments:** Fire risk assessment begins with risk identification, risk analysis, and risk management, and ultimately aims to address fire risks. First, risk identification is divided into three categories based on behavioral settings: fire risks caused by residents, automatic fire extinguishing system failure, and failure of fire departments to extinguish fires. Then, exploring fire risk analysis in depth by identifying risk types and estimating and calculating fire risk factors based on frequency data. In the final phase, emphasizing the frequency algorithm, using *"Pf"* to represent the multiplication of the frequency of ignition (P1), the frequency of system fire extinguishing failure (P2), and the probability of fire extinguishing failure by fire department personnel (P3) to assess fire risk [12].

- **Spatial Layouts on Accessing Fire Risk**: Regarding the spatial layout factors to assess the impact on fire risk, what is mainly evaluated are the characteristics of the spatial layout, such as the fire alarm system, fire protection system, and building specifications. Evacuation path [31], viewing area are variables [13], and frequency method is applied to process the data, or by comparing pairs to create a comparison matrix, to weigh the factors that affect fire risk. Finally, the fire risk is calculated from the characteristics of the spatial layout, and a fire risk map is created based on the analysis results [14].
- **Environment Factors on Accessing Fire Risk**: Wind environment in the analysis of accessing fire risk coefficients of microclimate has been extensively researched. The methodology includes introducing Wind Adjustment Factor(WAF) models, identifying WAF and midflame wind in fire modeling systems, and calculating fire risk factors like flame lengths and maximum surface fire spread rates [26], with concentrating on wind speed of flame inclination angle and radiation [15]. Another method assesses community vulnerability, ignition probability, and wildfire risk in wildland-urban interface (WUI), which highlights local wind patterns, showing how increased wind speeds intensify wildfire spread [16].

These literature reviews have mainly used different statistical methods to assess fire risk in terms of spatial layout or environmental factors alone, without integrating the effects of these two factors or their interactions with fire management. Estimating total fire frequency by integrating the effects of spatial layout and environmental factors holds significant research value for fire prevention and its impact on the health of village residents.

3 Method

This research focuses on Zhaoxing Dong Village in Guizhou, China, using the frequency estimation method. The study divides the rural planning into 100 units, analyzing spatial layout through the field of view areas and evacuation path lengths, which calculates spatial layout influential coefficients on fire risk. Environmental factors are analyzed using the average annual precipitation of Zhaoxing Dong Village in 2020 to estimate moisture and the simulation is conducted using *Grasshopper* with the "Butterfly" plug-in for computational fluid dynamics [17], to calculate the Wind Adjustment Factor (WAF), which represent the ratio of wind speed at different heights above ground. Slope analysis is conducted with "Ladybug" in *Grasshopper* [17]. Based on the standard fire model of *Behave Plus 5.0* [18], the mid-flame wind speeds, flame lengths, and maximum surface fire spread rates are obtained, which are key factors in determining environments influential coefficients on fire risk. Integrating spatial and environmental coefficients, the study evaluates Zhaoxing Dong Village's overall fire risk.

3.1 Case Description and Context

Zhaoxing Dong Village in Guizhou Province, China (Fig. 1), is one of the most famous villages of the Dong minority, characterized by a dense, disorganized spatial planning

based on the terrain, and the climate characterized by hot summers and cold winters is typical in Guizhou province. The region's climate combines the characteristics of a high plateau with a humid subtropical monsoon climate, which makes the local wind show unique characteristics. The characteristic spatial layout and environmental features make it a typical case to study these two factors to assess the local fire risk [19].

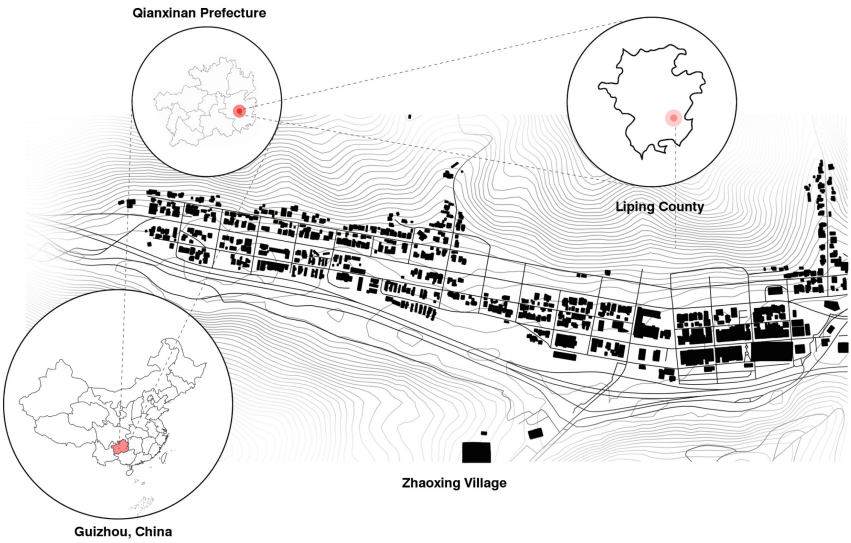


Fig. 1. Location Analysis of Zhaoxing Dong Village.

3.2 System Design

As for the fire risk management system (Fig. 2) design of this research, it consists of two sub-systems to evaluate, the first one is spatial layout fire risk management system and the other is environment fire risk management system.

- **Spatial Layout Fire Risk Management System:** This system selected 20 routes within the area and used the field of viewing area to see whether the evacuation path could be seen from surrounding buildings to evaluate the evacuation efficiency of each evacuation road. Each route is recorded five times at the same time and rate. Therefore, there will be 20 groups of the field viewing area to analyze their maximum and minimum values, which represents the fire evacuation efficiency of the routes. Besides, the length of each 20 evacuation routes also be a factor in evaluating the spatial layout fire risk. The equation is given in [12]:

$$FR = \sum_{i=1}^{n} P_{fi} C_{fi} \tag{1}$$

The spatial layout fire risk factor can be defined as a frequency function of evacuation efficiency and path length. P_{fi} can be the result of both evacuation efficiency and the length of route normalized from 0 to 1.

- **Environment Fire Risk Management System**: *BehavePlus 5.0* is a model-based fire behavior prediction and analysis system developed by the US Forest Service (USFS), which simulates fire behavior in a variety of natural and managed environments to help managers make better decisions [18]. Using *BehavePlus5.0*, managers can Predict the potential behavior and effects of fire, such as flame lengths, fire propagation speeds, and burn intensity [20]. In this research, based on *BehavePlus 5.0*, Fuel model, live moisture, dead moisture, WAF, 22.5 ft Wind speed, and slope degree can be calculated for mid-flame wind speeds, flame lengths, and maximum surface fire spread rates [21]. Moreover, normalizing these results from 0 to 1, The coefficient factor of environment fire risk is related to these results using the Eq. (1)

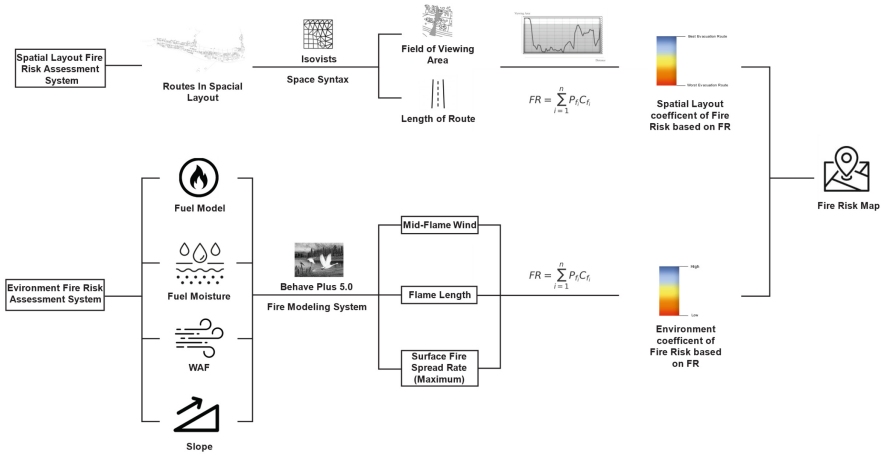


Fig. 2. System design of the whole fire risk management system

3.3 Fire Behavior and Environmental Factors Analysis

- **Fuel Model**: In *BehavePlus 5.0*, the fuel model is a key component that describes the vegetation type, structure, density, and combustibility of a specific area. "Anderson's 13 Fire Behavior Fuel Models", developed by Hal E. Anderson in 1982, this is one of the earliest systems of fuel models, categorized as follows: Grass Models (Models 1, 3, 4): Primarily composed of grass, with fire behavior greatly influenced by the moisture and density of the grass. Brush Models (Models 2, 5, 6): Contain a significant amount of brush, with fire behavior affected by the type of brush and its moisture content; Timber Models (Models 8, 9, 10): Mostly woodland, including litter and canopy layers. Slash Models (Models 7, 11, 12, 13): Primarily describe the debris

left after logging or natural disasters [22]. Since brushes are covered by most of Zhaoxing Dong villages area, this research will use Brush Models (Model 2).

- **Live and Dead Moisture**: Live Moisture is usually referred to as the water content of living plant parts, and dead moisture is pointed to the moisture content of dead parts of plants, such as dead branches and dry leaves. These indicators are influenced by various weather variables including precipitation, humidity, temperature, and more. Due to the many factors that influence it, there is no direct equation for using weather variables to measure live or dead moisture [23]. Several empirical models for measurements are widely used, as shown in Eq. (2), given in [23].

$$m = a_0 + \sum_{i=1}^{n} a_i X_i \tag{2}$$

where a_i are empirical constants and X_i are weather variables. In this research, the average monthly probability of precipitation in Zhaoxing Dong Village, Guizhou, China, for the year 2020 (Fig. 3) will be analyzed using data from the China Meteorological Administration.

In the process of calculating live moisture, a_0 and a_i are factors set based on plant type and region, due to the abundant rainfall in Guizhou Province and the high-water absorption capacity of vegetation, it can assume that precipitation has a greater effect on the live moisture content. It can be initially set between 0.6 and 0.8 for a_i, assuming 0.7; for a_0, the coefficient depends on the baseline of the average moisture content of the area, and it can be set between 5% and 15%, assuming 10% [24].

When calculating dead moisture, it's necessary to calculate relative humidity and the temperature, but humidity and temperature changes are small in Guizhou during the year [25], so only empirical constants need to be calculated. Guizhou's high humidity can be assumed that a_i is 0.4, temperature can be set to 0.2, and as for a_0, this coefficient stands for the baseline humidity, and it can be set to 15% [24]. The monthly live moisture and dead moisture of Zhaoxing Dong Village, Guizhou, China, 2020 (Fig. 4) are shown below.

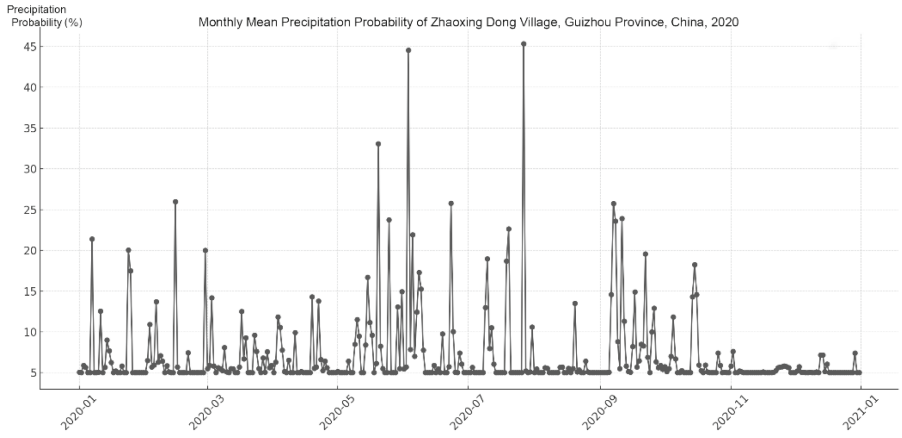


Fig. 3. Monthly Mean Precipitation Probability of Zhaoxing Dong Village, Guizhou, China, 2020.

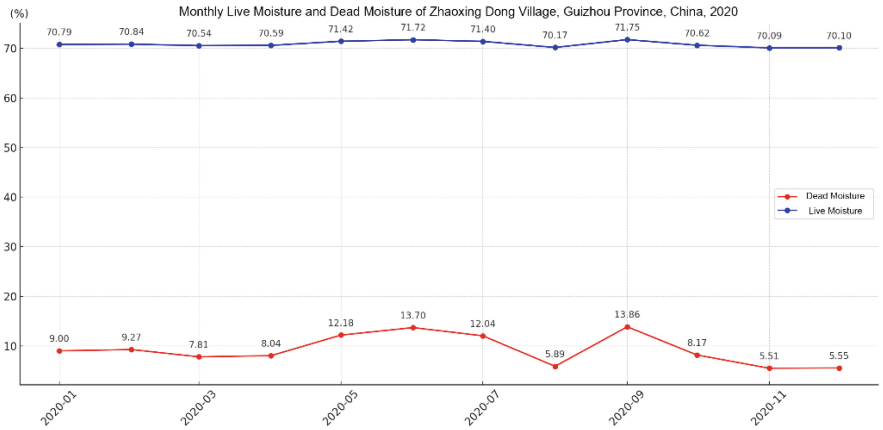


Fig. 4. Monthly live moisture and dead moisture of Zhaoxing Dong Village, Guizhou, China, 2020.

- **Wind Adjustment Factor (WAF)**: Wind Adjustment Factor (WAF) is a parameter used to describe the relationship between surface wind speeds and standard wind speeds measured at 20 ft elevation over open terrain [26]. Distinct models exist for wind affecting surface fuel that remains exposed to wind due to the absence of an overstay [27]. The WAF calculation is based on the wind speeds at different heights (Fig. 5). Since this research is more concerned with the relationship between spatial layout, and environment to manage fire risk, besides, the tree height in this area is nearly negligible, the "unsheltered" model has been adopted.

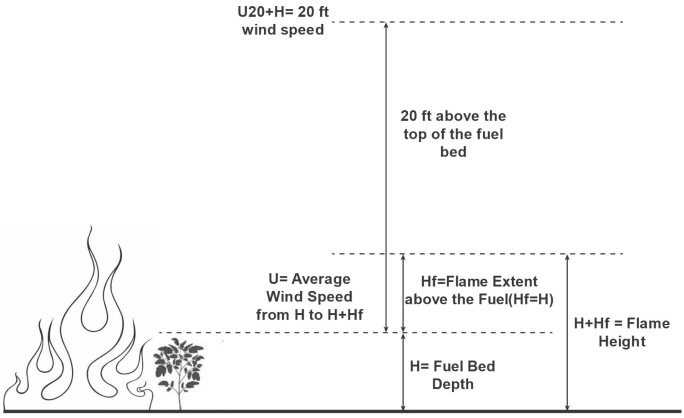


Fig. 5. Diagram of how “Unsheltered” WAF is calculated.

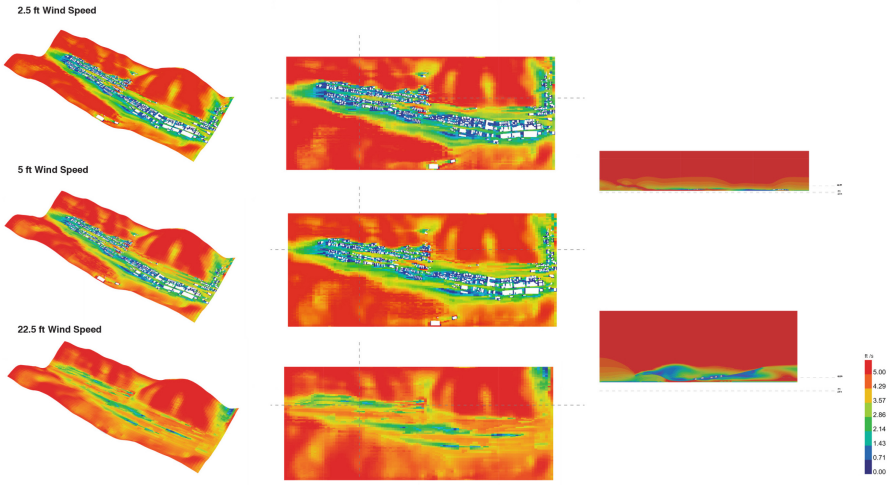


Fig. 6. 2.5 ft, 5 ft, and 22.5 ft average wind speed maps.

According to the local fuel bed depth(H) is 2.5 ft, it can be included that $H+H_f$ is 5 ft and $20+H$ is 22.5 ft. The WAF equation is given in [27], shown below:

$$WAF = \frac{U}{U_{20+H}} = \frac{1 + 0.36H/H_f}{\ln\left(\frac{20+0.36H}{0.13H}\right)}\left[\ln\left(\frac{H_f/H + 0.36}{0.13}\right) - 1\right] \tag{3}$$

U is average wind speed from to H and H_f. Based on the "Butterfly" plugin in *Grasshopper* [17], 2.5 ft, 5 ft, and 22.5 ft average wind speed maps were generated (Fig. 6).

Mid-flame wind speeds, Flame lengths, and Maximum surface fire spread rates: Mid-flame wind speeds refer to the wind experienced at the midpoint of a flame [28], shown in Eq. (4).

$$Mid\text{-}flame\,wind = WAF^*\left(20' + H\right) \tag{4}$$

Flame Length is the vertical distance from the base of the flame to the top of the flame. Besides, flame length is often used to estimate radiant heat and visibility effects in wildfire behavior prediction, the calculation method is given in [29], shown in Eq. (5), where I is the fire line intensity.

$$Flame\ Length = 0.45 * (I)^{\frac{2}{3}} \tag{5}$$

Maximum surface fire spread rates are the fastest rate at which a flame can spread over the surface, usually in the direction of the prevailing wind. A number of factors are involved in the Eq. (6), given in [30], including terrain, fuel type, fuel moisture, wind speed [30]. R is fire spread rate, ε is the propagation parameter, reflecting the heat transfer efficiency between the flames and the fuel. ρ_b is the apparent density of the fuel bed, and Q_{ig} is the

ignition heat.

$$R = \frac{I_r * \varepsilon}{\rho_b * Q_{ig}} \tag{6}$$

According to the Eq. (1), the coefficients of environmental factors for accessing fire risk are determined by the WAF, mid-flame wind, Flame length, and maximum surface fire spread rate, and all values are mapped to the 0–1 range to control the variables.

4 Result

4.1 Results for Spatial Layout Fire Risk Management System

By measuring the field of viewing area of the architecture to the evacuation route in the method, a data plot of 20 viewing areas (Fig. 7) is generated, the smaller the area means the greater the chance that the building will see the fire evacuation route [31], and the results of area (Fig. 8) of the 20 viewing area and length of 20 evacuation routes (Fig. 9) are shown as below.

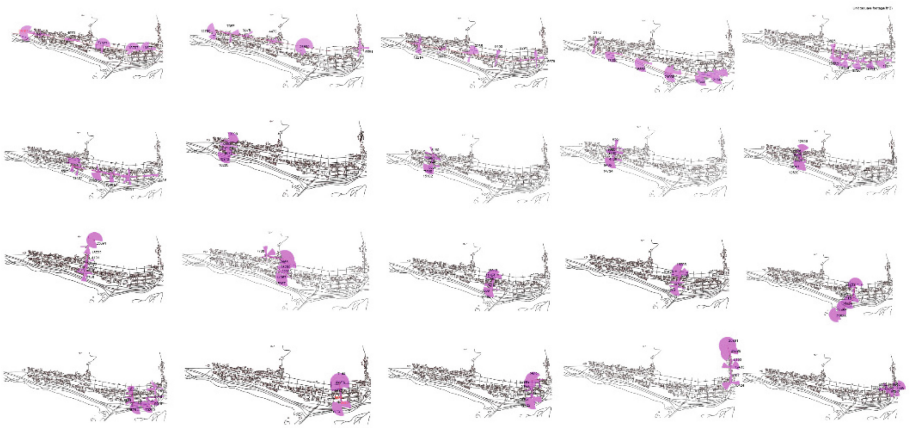


Fig. 7. The field of viewing area of 20 evacuation routes combined with spatial layout.

It is observed that longer evacuation routes with more turns exhibit greater fluctuations in viewing area, which in some cases can help avoid congestion during emergency evacuations, thus facilitating better evacuation. There are also cases where the route is long but the variations in viewing area are small, and overall values are low, indicating a route without turns, which is conducive to the evacuation of surrounding buildings, such as routes 2 and 3. For shorter evacuation routes, the smaller visible area suggests that more buildings will see this evacuation route when a fire occurs, leading to congestion and affecting evacuation efficiency, as seen with routes 7, 8, 9, and 10. It is necessary to use the frequency estimation method for further numerical analysis. Therefore, based on

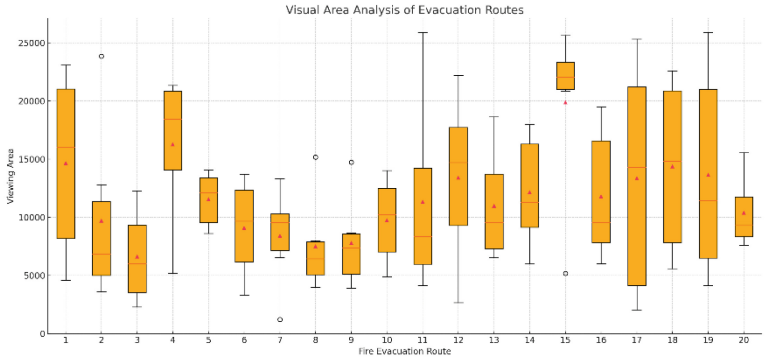


Fig. 8. The field of viewing area of 20 evacuation routes data analysis.

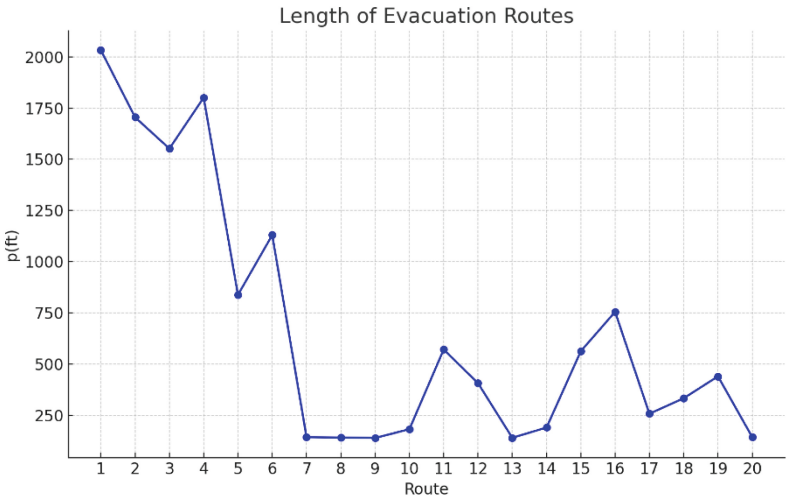


Fig. 9. Length of 20 evacuation routes.

the Eq. (1), the coefficient of Spatial Layout Fire Risk (f (Spatial Layout)) can be determined by the frequency of the field of viewing area (f (Viewing area) and the frequency of length of the evacuation route(f(p)), 1–5 routes calculation are examples, which are shown in Tables 1 and 2. The calculation of all routes is detailed in Appendix 1.

There is a positive correlation (Fig. 10) between the field of viewing area and the length of evacuation routes, with a factor of 0.81, which suggests that a smaller visual area or a shorter evacuation route increases the risk of exposure to the evacuation route and is prone to overcrowding in the event of fire risk. In addition, the length of evacuation routes was also positively related to the effect of f (spatial layout), with a correlation coefficient of 0.9. Simultaneously, the relationship between viewing area and the effect of f (spatial layout) had a positive value of 0.79, which showed a positive correlation between these factors.

Table 1. The coefficient of Spatial Layout Fire Risk Calculation of Route 1–5.

Route	Viewing area (ft^2)	f (Viewing area) (0–100)	p (ft)	f(p) (0–100)	f (Viewing area) (0–1)	f (Spatial layout) (0–1)
1	23101	88.30	2033	100.00	0.883	0.529
	6444	18.44			0.184	
	4573	10.59			0.106	
	21827	82.96			0.830	
	18707	69.87			0.699	
	13327	47.31			0.473	

Table 2. (Continued). The coefficient of Spatial Layout Fire Risk Calculation of Route 1–5.

Route	Viewing area (ft^2)	f (Viewing area) (0–100)	p (ft)	f (p) (0–100)	f (Viewing area) (0–1)	f (Spatial layout) (0–1)
2	12780	45.01	1705	82.86	0.373	0.267
	7057	21.01			0.174	
	3578	6.42			0.053	
	4478	10.19			0.084	
	23859	91.48			0.758	
	6581	19.01			0.158	
3	10214	34.25	1552	74.86	0.256	0.144
	2916	3.64			0.027	
	12265	42.85			0.321	
	5400	14.06			0.105	
	2291	1.02			0.008	
	6678	19.42			0.145	
4	5173	13.11	1801	74.86	0.098	0.447
	13206	46.80			0.350	
	16560	60.86			0.456	
	21006	79.51			0.595	
	20344	76.74			0.574	
	21344	80.93			0.606	

(*continued*)

Table 2. (*continued*)

Route	Viewing area (ft^2)	f (Viewing area) (0–100)	p (ft)	f (p) (0–100)	f (Viewing area) (0–1)	f (Spatial layout) (0–1)
5	8605	27.50	838	37.54	0.103	0.150
	14045	50.32			0.189	
	12322	43.09			0.162	
	13766	49.15			0.184	
	8726	28.01			0.105	
	11931	41.45			0.156	

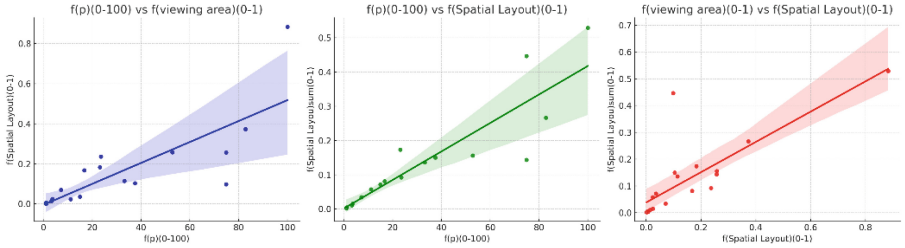


Fig. 10. Analysis of the relation between the field of viewing area, the length of evacuation routes and the coefficient of spatial layout on the accessing fire risk.

Therefore, related these three-correlation analysis shows, the longer the evacuation route, the lower the level of crowding and the better the evacuation in case of fire. However, increasing the visible area of buildings and evacuation routes can sometimes reduce the increased fire risk due to short evacuation routes. By visualizing the value of f (Spatial Layout) on the map to obtain the Spatial Layout Fire Risk map (Fig. 11).

4.2 Results for Environmental Layout Fire Risk Management System

Mid-flame wind speeds, Flame lengths, and Maximum surface fire spread rates were calculated, with data provided in Appendix 2. Utilizing *BehavePlus 5.0*, 20+H wind speed, slope, and the above variable, categorizing consistent WAF values are analyzed under specific fuel model, live moisture, and dead moisture conditions which are shown in Appendix 3. Besides, mid-flame wind speeds, Flame lengths, and Maximum surface fire spread rates are simulated under a constant WAF, with range of 0.8–0.85, for different slope with 20+H ft (Fig. 12). The Result indicates that with an increase in slope, both the maximum surface fire spread rates and flame lengths are escalated when the wind speed remains constant. However, mid-flame wind speeds remain unaffected by changes in the slope. When the slope is constant, higher wind speeds result in greater mid-flame wind speed, flame lengths, and maximum surface fire spread rates.

As for searching different relationships among each environmental factor, scattered linear relevance analysis plot between Mid-flame wind speeds, flame lengths, Maximum

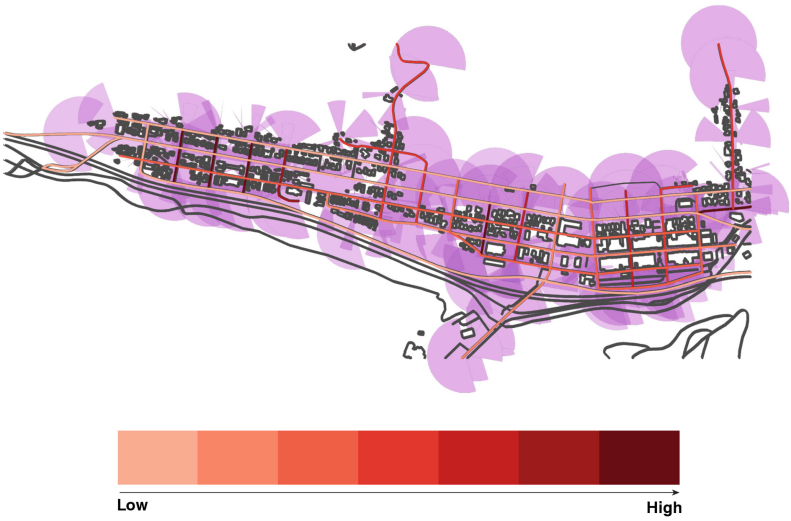


Fig. 11. Spatial Layout Fire Risk map.

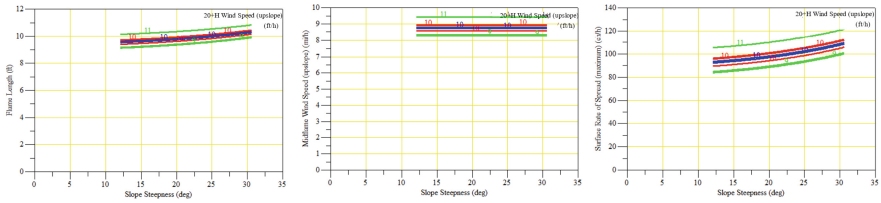


Fig. 12. Analysis of "Mid-flame wind speeds", "Flame lengths", and "Maximum surface fire spread rates" for "WAF" of 0.8-0.85.

surface fire spread rates and WAF (Fig. 13) are shown below. It can be concluded that the coefficients of fire assessed by environmental factors (f(Environment)sum) are most affected by maximum surface fire spread rates with a correlation factor of 0.924, and maximum surface fire spread rates are almost perfectly linearly related to Flame Length with a correlation factor of 0.998. However, the WAF and Mid-flame wind are positively associated with the environmental factor assessment of fire, but the relevance factors are 0.532 and 0.88, which suggests that the wind alone is not a single factor for the environmental factor assessment of fire.

Similarly, to the spatial layout fire risk map, the coefficients of fire assessed by environmental factors also can be visualized on planning maps (Fig. 14). It can be concluded that the high fire risk places are in the surrounding of the building complex, but the level of fire risk does not show a certain pattern because there are too many influencing factors, and it can only be said that the more thoroughly influencing factors are considered, the higher the accuracy will be.

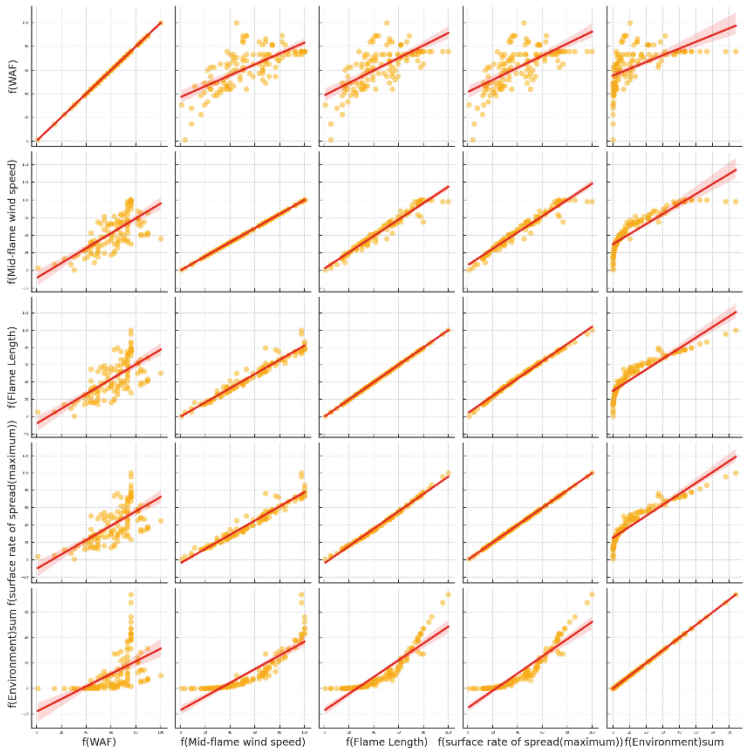


Fig. 13. A Scattered linear relevance analysis plot between Mid-flame wind speeds, Flame lengths, and Maximum surface fire spread rates and WAF.

4.3 Result for Overall Consideration

Based on the results of the environmental fire risk factors assessment, the serious fire risk area is located around the building complex. Combined with the arrangement of evacuation routes, the entire fire risk management system is evaluated through both coefficient of spatial layout and environment, with the calculation procedures detailed in Appendix 4. This diagram illustrates the correlation analysis between the spatial layout fire risk coefficient and the environment fire risk coefficient is presented below (Fig. 15), and the correlation between these is 0.18, demonstrating a negligible association.

The correlation between the entire fire risk system and spatial layout fire risk value or environmental fire risk figure are also generated (Fig. 15). It's concluded that the correlation between environmental factors and comprehensive fire risk management is 0.51, proving a positive relationship. Meanwhile, the correlation for spatial layout factors in comprehensive fire risk management is 0.87, suggesting that the influence of spatial layout on fire risk management exceeds that of environmental factors.

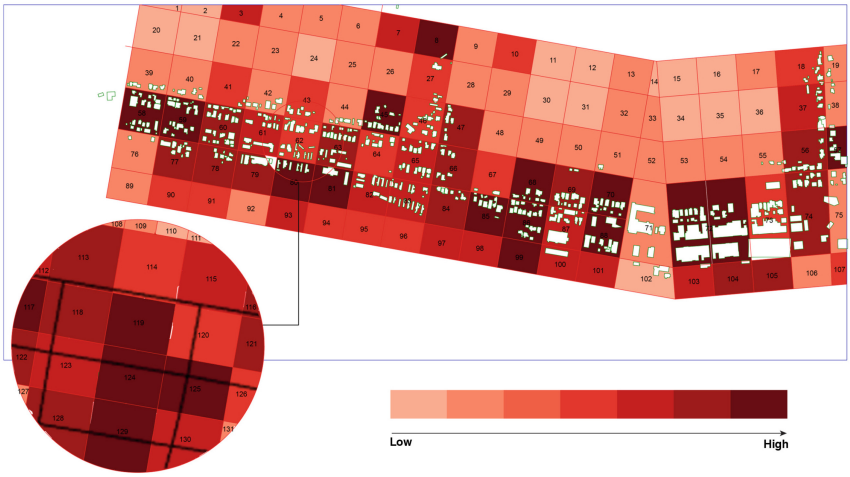


Fig. 14. Environment Fire Risk map.



Fig. 15. Analysis of the relationships between environment, spatial layout and entire fire risk management.

Finally, Visualizing the entire map of fire risk management (Fig. 16) is given by the frequency of entire fire risk management. Some of the units are assessed as high risk for fire considering only environmental factors, but when combined with the factors of spatial layout, the risk level decreases, such as units 39–83, but some of the risk level increases, for example units 27,45, and 46, which confirms that the spatial layout has a greater impact on the assessment of fire control than the environmental factors.

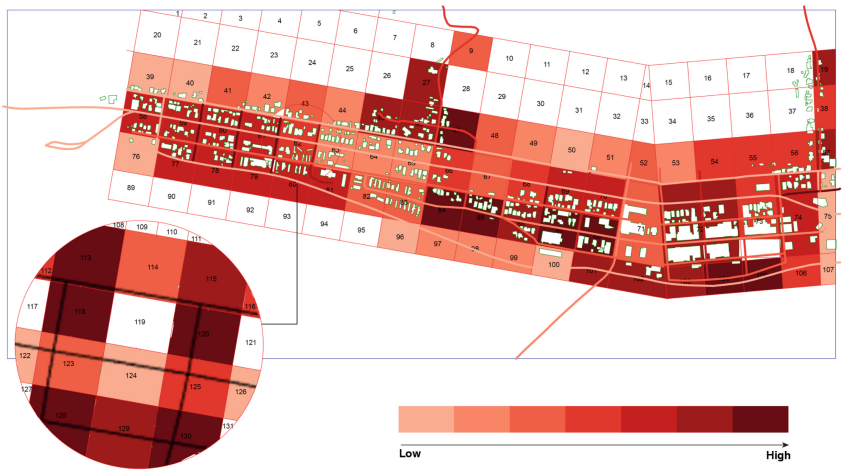


Fig. 16. Entire map for fire risk management.

5 Discussion

Our approach combines spatial layout and environmental factors to generate a fire risk management simulation model. As for the practical application of this simulation model, the technical departments of local governments should utilize GIS to integrate and optimize information such as urban data, environmental monitoring systems, and climatic conditions to improve the accuracy of simulation**.** Secondly, based on this model developed in this study, more refined fire prevention and building codes should be formulated, including clarifying the requirements for fire-resistant materials in high-risk fire areas, setting fire protection distances properly, and standardizing the standards for evacuation routes. In addition, the planning department should combine the fire risk management model to optimize the layout of functional areas, avoid setting up population-dense buildings in high-risk fire areas, and scientifically plan the distribution of fire stations and fire-fighting facilities to ensure fire resources cover the high-risk areas. Besides, more demanding fire protection design standards should be established, emergency shelters should be allocated, and ventilation systems should be optimized to effectively mitigate smoke hazards.

As for the limitation of this study, regarding to Zhaoxing Dong village itself, the village's spatial layout is deeply rooted in historical and cultural heritage, leaving limited room for substantial alterations. Thus, authorities must delve into detailed simulations to identify strategies that balance the preservation of historical elements with ensuring the safety and well-being of all villagers. When applying this model to other regions, there will be challenges in obtaining and integrating data. The completeness and accuracy of data vary across different regions, which may affect the accuracy of the simulation. Moreover, there are differences in policies and regulations in each region, and the

stakeholders' acceptance also varies, which requires coordination among multiple stakeholders concerning the results generated by the model and government recommendation. During this process, the rights and interests of developers, residents, and commercial stakeholders may be affected, which in turn increases the resistance to implementation. Finally, the operation of the model relies on high-performance computing equipment and professional technical personnel. For economically underdeveloped regions, the limitations of technology and resources may make it difficult to bear the high computing costs and professional technical support.

6 Conclusion

This study combines spatial layout as well as environmental factors to simulate a fire risk management model and is an attempt to break away from single factor assessment of fire risk management. Fire risk evaluation is a complex integration of multiple factors. While focusing on isolated factors is feasible and often necessary for detailed analysis, the interplay between various factors has a more profound impact, as these factors influence one another, potentially altering the overall fire risk model. Among them, the spatial layout has a greater impact on fire assessment than the environment factors. It would be more effective to take corresponding fire prevention measures in terms of spatial layout.

However, there is much approach that can be further researched and explored. Factors such as the density and placement of greenery, the adequacy of firefighting infrastructure, and even the influence of sunlight on fire prevention could be incorporated to enhance the model's accuracy. In addition, the advanced data collection tools and the statistical techniques will further improve the accuracy of the evaluation, to provide more precise and more informed decision-making support for fire risk management. Besides, employing machine learning techniques can further enhance the predictive accuracy and adaptability of fire risk assessment models. By utilizing large datasets, algorithms can identify complex patterns and interactions between environmental factors, historical fire incidents, and spatial characteristics. Techniques such as decision trees, neural networks, and ensemble learning methods could be optimized risk classification and predictive modeling. Ultimately, a multidimensional, data-driven approach will lead to more effective fire prevention strategies, ensuring safer and more resilient urban built environment.

Appendix

Appendix 1: Spatial layout assessment of the fire-related spatial layout factors calculation process. *(*.xlsx)*
Appendix 2: Environment assessment of the fire-related environment Factors calculation process. *(*.xlsx)*
Appendix 3: Diagrams of mid-flame wind speed, flame lengths and maximum fire spread surface rate based on *BehavePlus 5.0* calculation. *(*.pdf)*
Appendix 4. Calculations process of the effect of spatial layout combined with the environment on the entire fire assessment. *(*.xlsx)*

References

1. Wotton, M., Nock, C., Flannigan, M.: Forest fire occurrence and climate change in Canada. Int. J. Wildland Fire. **19**, 253–271 (2010). https://doi.org/10.1071/WF09002
2. Zhang, X., Zhang, Z., Zhang, Z., Xu, W., Luo, Q., Tao, H., Li, X.: Experimental investigation of compartment fires with circular opening: From the aspects of internal temperature and facade flame. Combust. Flame. **213**, 107–116 (2020). https://doi.org/10.1016/j.combustflame.2019.11.027
3. Chan, E.Y.Y., Lam, H.C.Y., Chung, P.P.W., Huang, Z., Yung, T.K.C., Ling, K.W.K., Chan, G.K.W., Chiu, C.P.: Risk perception and knowledge in fire risk reduction in a Dong Minority Rural Village in China: a Health-EDRM Education Intervention Study. Int. J. Disaster Risk Sci. **9**, 306–318 (2018). https://doi.org/10.1007/s13753-018-0181-x
4. Du, F., Okazaki, K.: Building improvement responses to multi-hazard risk in the historic Dali Dong Village, Guizhou, China. Int. J. Disaster Risk Reduct. **19**, 64–74 (2016). https://doi.org/10.1016/j.ijdrr.2016.08.014
5. Du, F., Okazaki, K., Ochiai, C.: Disaster coping capacity of a fire-prone historical dong village in China: a case study in Dali Village, Guizhou. Int. J. Disaster Risk Reduct. **21**, 85–98 (2017). https://doi.org/10.1016/j.ijdrr.2016.10.016
6. Renner, J.S., Mensah, R.A., Jiang, L., Xu, Q.: A critical assessment of the fire properties of different wood species and bark from small- and bench-scale fire experiments. J. Therm. Anal. Calorim. **148**, 1423–1434 (2023). https://doi.org/10.1007/s10973-022-11443-z
7. Qiao, Y., Lam, C.C., Wong, M.O.: A GIS-based framework of fire risk assessment and emergency route planning for heritage buildings. In: Li, D., Zou, P.X.W., Yuan, J., Wang, Q., Peng, Y. (eds.) Proceedings of the 28th International Symposium on Advancement of Construction Management and Real Estate, pp. 1721–1731. Springer Nature, Singapore (2024). https://doi.org/10.1007/978-981-97-1949-5_120
8. Cruz, M.G., Alexander, M.E.: The 10% wind speed rule of thumb for estimating a wildfire's forward rate of spread in forests and shrublands. Ann. For. Sci. **76**, 1–11 (2019). https://doi.org/10.1007/s13595-019-0829-8
9. Li, L., Li, N., Wu, X., Liu, B.: A method for evaluating the spatial layout of fire stations in chemical industrial parks. Appl. Sci. **14**, 2918 (2024). https://doi.org/10.3390/app14072918
10. Zhu, Z., Deng, X., Zhao, F., Li, S., Wang, L.: How environmental factors affect forest fire occurrence in Yunnan Forest Region. Forests. **13**, 1392 (2022). https://doi.org/10.3390/f13091392
11. Wildland Fire Behavior (U.S. National Park Service). https://www.nps.gov/articles/wildland-fire-behavior.htm. Accessed 29 Nov 2024
12. Xin, J., Huang, C.: Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management. Fire Saf. J. **62**, 72–78 (2013). https://doi.org/10.1016/j.firesaf.2013.09.022
13. Xia, G.-S., Datcu, M., Yang, W., Bai, X.: Information processing for unmanned aerial vehicles (UAVs) in surveying, mapping, and navigation. Geo-spatial Inf. Sci. **21**, 1–1 (2018). https://doi.org/10.1080/10095020.2017.1420510
14. Masoumi, Z., van Genderen, J., Maleki, J.: Fire risk assessment in dense urban areas using information fusion techniques. ISPRS Int. J. Geo Inf. **8**, 579 (2019). https://doi.org/10.3390/ijgi8120579
15. Gibson, L., Cicione, A., Stevens, S., Rush, D.: The influence of wind and the spatial layout of dwellings on fire spread in informal settlements in Cape Town. Comput. Environ. Urban Syst. **91**, 101734 (2022). https://doi.org/10.1016/j.compenvurbsys.2021.101734
16. Aksoy, E., Kocer, A., Yilmaz, I., Akcal, A.N., Akpinar, K.: Assessing fire risk in wildland-urban interface regions using a machine learning method and GIS data: the example of Istanbul's European side. Fire. **6**, 408 (2023). https://doi.org/10.3390/fire6100408

17. Willis, D., Braham, W.W., Muramoto, K., Barber, D.A.: Energy Accounts: Architectural Representations of Energy, Climate, and the Future. Routledge, London (2016)
18. Andrews, P.L.: BehavePlus fire modeling system, version 5.0: Variables. In: Gen. Tech. Rep. RMRS-GTR-213 Revised, p. 111. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO (2022). https://doi.org/10.2737/RMRS-GTR-213
19. Wang, S., Guo, Y., Zhang, H., Gao, M.: A life-cycle carbon emissions evaluation model for traditional residential houses: applying to traditional Dong Dwellings in Qandongnan, Guizhou Province, China. Sustainability. **15**, 13468 (2023). https://doi.org/10.3390/su151813468
20. Andrews, P.L.: BehavePlus fire modeling system: past, present, and future. In: Proceedings of 7th Symposium on Fire and Forest Meteorology; 23-25 October 2007, Bar Harbor, Maine, p. 13. American Meteorological Society, Boston, MA (2007)
21. Andrews, P.L.: Modeling Wind Adjustment Factor and Midflame Wind Speed for Rothermel's Surface Fire Spread Model, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ft. Collins, CO (2012). https://doi.org/10.2737/RMRS-GTR-266
22. Anderson, H.E.: Aids to determining fuel models for estimating fire behavior. U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT (1982). https://doi.org/10.2737/INT-GTR-122
23. Matthews, S.: Dead fuel moisture research: 1991–2012. Int. J. Wildland Fire. **23**, 78–92 (2013). https://doi.org/10.1071/WF13005
24. Luo, K., Quan, X., He, B., Yebra, M.: Effects of live fuel moisture content on wildfire occurrence in fire-prone regions over Southwest China. Forests. **10**, 887 (2019). https://doi.org/10.3390/f10100887
25. Xue, X., Hou, S., Meng, C.: Spatial–temporal variations in temperature and precipitation extremes during 1960–2019 in Guizhou Province, China. Atmos. **14**, 1162 (2023). https://doi.org/10.3390/atmos14071162
26. Ortega, M., Navarro, J.A., Molina, J.R.: Modeling Wind Adjustment Factor for a prescribed burn plan. An application to Mediterranean stands in Southern Europe. Agric. For. Meteorol. **342**, 109748 (2023). https://doi.org/10.1016/j.agrformet.2023.109748
27. Albini, F.A.: Estimating Windspeeds for Predicting Wildland Fire Behavior. Intermountain Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture (1979)
28. Rothermel, R.C.: A Mathematical Model for Predicting Fire Spread in Wildland Fuels. Intermountain Forest & Range Experiment Station, Forest Service, U.S. Department of Agriculture (1972)
29. Gao, Z., Ji, J., Wan, H., Li, K., Sun, J.: An investigation of the detailed flame shape and flame length under the ceiling of a channel. Proc. Combust. Inst. **35**, 2657–2664 (2015). https://doi.org/10.1016/j.proci.2014.06.078
30. Zhao, J., Huang, H., Jomaas, G., Zhong, M., Li, Y.: Spread and burning behavior of continuous spill fires. Fire Saf. J. **91**, 347–354 (2017). https://doi.org/10.1016/j.firesaf.2017.03.046
31. Guan, W., Hou, S., Yu, G., Gong, H., Guan, S., Zhao, J.: Dynamic evacuation path planning for multi-exit building fire: bi-objective model and algorithm. Fire Technol. **59**, 2853–2876 (2023). https://doi.org/10.1007/s10694-023-01448-x

Enhancing Human Detection in Post-disaster Scenarios Through Generative Adversarial Networks

Chengyi Qu[3], Xiaodong Hao[2], Feiyu Ma[2], Jizhuang Hui[2], Wenbin Guo[4], and Hong Wang[1(✉)]

[1] School of Artificial Intelligence, Hubei University, Wuhan 430062, China
hong.wang@hubu.edu.cn
[2] School of Construction Machinery, Chang'an University, Xi'an 710064, China
{xiaodong.hao,feiyu.ma,huijz6363}@chd.edu.cn
[3] Florida Gulf Coast University, For Myers, FL 33965, USA
cqu@fgcu.edu
[4] University of Florida, Gainesville, FL 32611, USA
wenbin.guo@ufl.edu

Abstract. Human body detection is critical in post-disaster response, especially in UAV-assisted search and rescue missions, where accurate localization of stranded individuals is essential for efficient emergency operations. However, current disaster scene image datasets are often scarce, poorly annotated, and of inconsistent quality, limiting the performance of deep learning models in complex disaster environments (e.g., ruins, fires, floods). These challenges are exacerbated when human targets are intertwined with complex backgrounds, making existing techniques inadequate for practical needs. To address these limitations, this paper presents an integrated approach combining image generation and target detection techniques to enhance the adaptability and accuracy of deep learning models. By constructing diverse, high-quality disaster scene datasets, the proposed method alleviates data scarcity and improves generalization under complex scenarios. Models trained on the expanded dataset achieve significant improvement, with generated images yielding an FID score of 45 and mAP50 scores exceeding 98% across three detection models. Additionally, this study explores the integration of computer vision and human-computer interaction in UAV search and rescue missions. The proposed framework provides an intelligent, efficient solution for post-disaster search and rescue, demonstrating the transformative potential of data augmentation and target detection technologies in disaster management. This work establishes a robust technological foundation for future post-disaster response and intelligent search and rescue operations.

Keywords: Post-Disaster response · Image generation · Target detection · Artificial intelligence · Generative Adversarial Networks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 257–272, 2025.
https://doi.org/10.1007/978-3-031-93502-2_17

1 Introduction

In the field of emergency response and disaster management, the rapid development of drone technology has marked the beginning of a new era. UAVs offer unique aerial perspectives and exceptional maneuverability, enabling efficient surveys of disaster areas and reducing the need for human responders in hazardous environments. By leveraging real-time monitoring and assessment, drones significantly enhance the efficiency of post-disaster operations, particularly in large-scale disasters where traditional search-and-rescue methods face considerable challenges. UAVs excel in rapidly deploying, covering vast areas, and transmitting real-time data, making them invaluable in scenarios such as earthquakes, fires, and floods, where complex terrain and environmental conditions can hinder manual rescue efforts. Through aerial photography, thermal imaging, and precise localization, drones provide intelligence to support rescue teams in identifying stranded individuals and assessing disaster zones effectively [1].

However, the practical application of UAVs in disaster scenarios is not without challenges, particularly in the area of target detection in complex environments. One major issue is the scarcity and ethical constraints associated with collecting and annotating disaster image datasets [2]. These challenges arise from the need to balance logistical feasibility with privacy concerns and the potential psychological impact on victims. Furthermore, existing datasets are often limited in scope and fail to represent the diverse and dynamic conditions found in real disaster environments, leading to misdetections and omissions in current models [3]. Addressing these issues requires innovative solutions to ensure ethical and effective use of UAVs in post-disaster operations.

Another significant challenge lies in the scarcity of data and the associated problem of overfitting in generative models. Generative adversarial networks, while capable of producing diverse images, often rely heavily on limited training data, leading to reduced representativeness and diversity in the generated images. Disaster scenarios frequently include complex backgrounds, occlusions, and varying environmental conditions, making it essential for generative models to learn accurate feature representations from constrained datasets [4]. This lack of diversity can hinder the adaptability and generalization of target detection models, especially when deployed in unfamiliar disaster environments. Improving the robustness of generative models to mitigate overfitting and produce high-quality, diverse datasets is a critical focus of this study.

Furthermore, traditional computer vision techniques and target detection algorithms often fall short in disaster scenarios [5]. These algorithms, trained on large-scale conventional datasets, are ill-equipped to handle the specificity and complexity of disaster environments. Factors such as occlusion by debris, smoke, or water, as well as visually constrained conditions, pose significant challenges to accurate human detection and localization. Existing models struggle with these complexities, underscoring the need for more specialized approaches to target detection in disaster management.

This study makes several important contributions to address these challenges. First, it provides a comprehensive overview of existing post-disaster search-and-rescue methods and datasets, analyzing current technological progress and identifying critical gaps. The research highlights the need for advancements in UAV-based search and rescue, particularly in addressing the limitations of existing datasets and detection algorithms. Second, the study emphasizes ethical considerations in data collection and processing.

By employing strict privacy protection measures and excluding sensitive real-world information, the study ensures that generated data adheres to ethical norms, thereby avoiding potential harm to disaster victims. Third, the study evaluates the performance of state-of-the-art target detection algorithms, such as YOLO11 and Faster RCNN, on newly generated disaster datasets. These evaluations reveal the strengths and weaknesses of different algorithms in human detection and provide recommendations for improving their robustness and generalization in complex scenarios. Finally, the study generates a high-quality, diverse disaster scenario dataset using generative models. This dataset simulates various disaster conditions, including earthquakes, fires, and floods, and incorporates diverse perspectives, occlusions, and complex backgrounds. By addressing the limitations of existing datasets, the new dataset enhances the training of target detection algorithms and improves the accuracy and reliability of UAV-based search-and-rescue missions.

Overall, this research provides innovative solutions to longstanding challenges in disaster management, advancing UAV technology and computer vision methods for post-disaster emergency response. It lays a robust foundation for future research and practical applications in intelligent disaster management and search-and-rescue operations.

2 Related Works

In recent years, with the development of post-disaster rescue and emergency response, rapid and effective assessment of post-disaster scene conditions, personnel localization and identification have become key research topics. Remote sensing technologies, such as drones and satellite images, have been widely used for rapid surveys of post-disaster scenes, but these methods usually rely on high-quality image data, as well as efficient and accurate target detection algorithms to identify affected people and important facilities. Therefore, how to generate high-quality disaster scene data and combine it with target detection algorithms for real-time analysis becomes the key to improve the post-disaster emergency response capability.

2.1 Ethical Issues of Disaster Scenarios

The application of generative AI in several fields poses significant ethical challenges, especially in terms of privacy protection and data security. Authors in [6] analyzed the application of generative AI and its ethical issues in several industries, with a special focus on the challenges in the healthcare field, investigated the Transformer-based generative AI system in software engineering and cyber-physical systems ethical implications and emphasizes the importance of rigorous ethical scrutiny and safeguards for emerging AI technologies. Authors in [7] further propose an ethical framework for generative AI that emphasizes transparency, accountability, and public participation, aiming to promote responsible development of the technology and to ensure that its applications truly benefit society and respect individual rights.

Ethical issues become more complex in the study of disaster scenarios, especially when it comes to vulnerable groups and those affected by disasters. Authors in [8] discussed the vulnerabilities of these groups, noting that they often suffer greater losses and

that studying them presents multiple methodological and ethical challenges. Ethical and methodological issues in disaster research were further analyzed in [9], with a focus on post-disaster studies of affected communities. This study emphasized the equal importance of ethics and methodology in disaster research, drawing insights from fieldwork conducted during the recovery process after Hurricane Katrina and Hurricane Rita. The framework of "ethical double-consideration" was proposed to support a paradigm shift in disaster research methodologies. Additionally, research in [10] investigated the use of generative AI in post-disaster emergency response, arguing that as its significance grows, ensuring fairness and transparency in its application becomes a key ethical challenge. The study highlighted the necessity of establishing clear ethical guidelines to assist emergency response teams in making responsible decisions when utilizing generative AI, ultimately promoting justice and transparency in post-disaster social recovery. The application of generative AI in numerous fields continues to raise significant ethical concerns, and as the technology evolves, balancing innovation with ethical responsibility will remain a key challenge for future development.

2.2 Generative Models

In machine learning, generative models are conceptually distinct from discriminative models. The current mainstream generative models include variational autoencoders, generative adversarial networks, flow-based models, autoregressive models, and diffusion models, each with unique characteristics and roles in different application scenarios. Authors in [11] proposed variational autoencoders based on autoencoders, achieving data generation by learning the distribution of latent variables in the latent space. Streaming models, though a niche category of generative models, are far less commonly used, with limited research available. Authors in [12] introduced a streaming model, which differs from other generative models in that it does not circumvent probability calculations but instead performs an invertible transformation between two distributions of consistent data dimensions using the Jacobian determinant transformation. Autoregressive models are a type of smooth time series model, primarily applied in natural language processing, though some studies have explored their use in image generation. Research in [13] utilized autoregressive connections to model images pixel by pixel, decomposing the joint image distribution into the product of conditional distributions. Diffusion models represent a cutting-edge approach in generative modeling, with rigorous mathematical derivations providing strong generative capabilities. Work in [14] proposed a denoising diffusion model that applies an iterative denoising process via a two-step Markov chain, allowing it to generate new data samples that align with the probability distribution of the training dataset.

In contrast, Generative Adversarial Networks (GANs) have rapidly emerged as one of the most influential techniques in deep learning since their introduction in 2014. The key innovation of GANs lies in their adversarial training framework, which pits a generator against a discriminator. With continuous advancements in GAN technology, numerous variants have been developed, expanding their application scope. Research in [15] optimized the objective function of GANs by introducing Wasserstein distance instead of Jensen-Shannon divergence, addressing the gradient vanishing problem. Further improvements were made in [16] with the introduction of Gradient Penalty (GP) in

WGAN, enhancing the enforcement of the 1-Lipschitz constraint and improving training stability and generation quality. StyleGAN [17] introduced the concept of "style," enabling the generated image to control both content features and multi-level visual attributes, making the outputs more diverse and detailed. StyleGAN2-ADA (Adaptive Data Augmentation) [4] further optimized this approach by incorporating adaptive data augmentation, significantly improving robustness in data-scarce scenarios. While StyleGAN2-ADA has demonstrated strong performance across various generative tasks, its application in disaster scene image generation remains relatively unexplored. The complexity of disaster scenes—such as ruins, fires, and floods, along with factors like human poses and occlusions—presents substantial challenges for image generation. Existing datasets often lack effective representations of these intricate conditions, making it difficult to apply generative models effectively in post-disaster scenarios.

2.3 UAV Target Detection

In UAV-based image target detection, detection methods face significant challenges due to the unique characteristics of UAV viewpoints. These challenges include the large number and high density of small targets in visible images, significant scale variations among targets, and complex background interference. To address these issues, authors in [1] proposed an autonomous UAV system based on thermal image processing. This method significantly improved UAV localization accuracy in low-bandwidth environments and successfully detected 38 stranded individuals across multiple experiments with an accuracy of 86%. Traditional image recognition techniques often experience performance degradation under adverse weather conditions such as low visibility and fog. To mitigate this issue, authors in [18] introduced BlendNet, a convolutional neural network that, when combined with illumination-aware image fusion techniques, effectively enhances detection accuracy in low-light and complex environments. Currently, UAV search operations in post-disaster search and rescue missions remain inefficient, and traditional methods struggle with the coordinated operation of multiple UAVs. To tackle this problem, work in [19] proposed a multi-UAV cooperative system based on YOLOv5, enabling real-time target detection and monocular localization. By allowing multiple UAVs to be freely combined and independently controlled, this system effectively reduces search and rescue time. Further improvements in small target detection were introduced by authors in [5] with PC-YOLO11s, an enhanced detection method specifically designed to address challenges such as low resolution, weak contrast, and background interference in UAV imagery. This method optimizes the YOLO11 network structure by incorporating a P2 layer and introducing a coordinate space attention mechanism, thereby improving detection accuracy and outperforming other YOLO-based models on the VisDrone2019 dataset.

Currently, UAVs have become an indispensable tool in SAR missions due to their excellent flexibility and portability [20]. In post-disaster emergency response, the combination of generating high-quality disaster images and target detection technology not only provides a solution to the problem of data scarcity in disaster scenarios, but also provides a strong support for improving the efficiency and accuracy of UAS in post-disaster SAR.

3 Methods

In this study, we propose a novel image generation and data expansion approach for disaster scenes that integrates generative adversarial networks (GANs) and target detection techniques. The objective is to address the scarcity of disaster image datasets and enhance the performance of target detection models in real-world disaster scenarios. Disaster datasets are typically constrained by limited accessibility, insufficient sample diversity, and high labeling workloads. These limitations pose challenges for the application of target detection models, especially in complex scenarios involving occlusions, multi-target detection, and intricate backgrounds.

The proposed workflow, as illustrated in Fig. 1, begins with the preprocessing of real disaster images, which are then used as input for advanced generative models, including GAN, DCGAN, WGAN, WGAN-GP, and StyleGAN. To improve the diversity and quality of the generated images, techniques such as adaptive discriminator augmentation and progressive training are employed. The generated images undergo a rigorous image quality assessment process to ensure they meet the required standards. These high-quality synthetic images, along with real disaster pictures, are then preprocessed to form an augmented dataset.

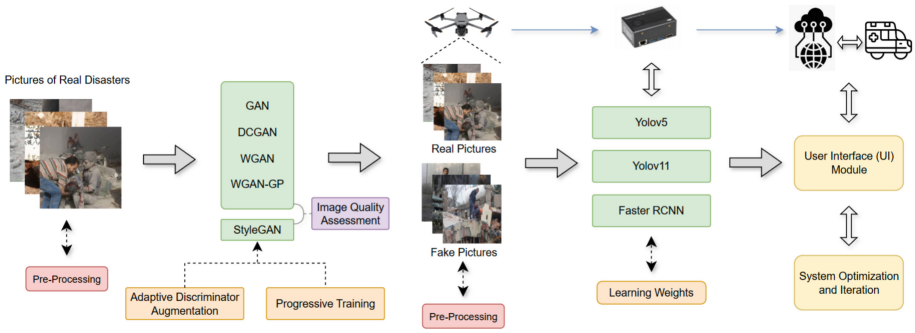


Fig. 1. Workflow of the Disaster Scene Image Generation and Target Detection Framework.

Using this expanded dataset, state-of-the-art target detection algorithms such as YOLOv5, YOLOv11, and Faster RCNN are trained, with optimized learning weights enhancing their accuracy and robustness in disaster scenarios. The detection models are integrated with a user interface (UI) module, facilitating system optimization and iterative improvements based on real-time feedback. This end-to-end framework provides a robust foundation for addressing data scarcity, improving target detection accuracy, and ultimately enhancing UAV-based emergency response in disaster environments.

The generative study primarily uses the Post-Disaster Dataset (PDD), comprising 1,144 disaster scene images with human targets and 155 test images [2]. The PDD dataset spans diverse post-disaster scenarios such as earthquakes, floods, and fires, with drone-captured images from various perspectives, including overhead, long-distance horizontal, and low-elevation views. These multi-view data capture scene variability and

environmental characteristics, providing a solid foundation for target detection model training.

To enhance dataset diversity and expand the image pool, publicly available datasets like the Comprehensive Disaster Dataset and the AIDER Dataset were incorporated. The AIDER Dataset focuses on identifying and classifying natural disaster scenarios, covering six major disaster types such as floods, earthquakes, and fires, along with additional scenarios like ruins and traffic accidents. After processing and filtering these sources, a comprehensive dataset containing 2409 images was created for disaster image generation.

All datasets were sourced from legally authorized public domains, and the generated images were based solely on the features and information from these datasets, ensuring no sensitive or personal data from real disasters was involved. The generated images, entirely synthetic, were used exclusively for academic research and technology development, adhering to ethical standards. Basic details of each dataset are summarized in the Table 1.

Table 1. Overview of datasets used for disaster image generation and target detection.

	PDD	CDD	AIDER	Ours data
Data set size	1144	12,834	6433	2409
Type of disaster	Earthquakes, floods, fires	Floods, earthquakes	Ruins, fires, floods, traffic accidents	Ruins
Data type	Images (with human targets)	Images (multiple target labeling)	Images (with human targets)	Images (without human targets)
Image annotation	Human target labeling	Unlabeled	Unlabeled	Unlabeled

To ensure the high quality and diversity of the training data, we employ advanced image segmentation techniques to extract the target regions during data preprocessing. Specifically, the human detection algorithm is utilized to generate a mask to accurately remove the human targets in the image from the original image. The mask generation related formula is as follows:

$$I_{final}(x, y) = M(x, y) \cdot I_{person}(x, y) + (1 - M(x, y)) \cdot I_{background}(x, y)$$

where $M(x,y)$ denotes the mask of the character, $I_{person}(x,y)$ and $I_{background}(x,y)$ are the character and background images respectively, and $I_{final}(x,y)$ is the background image after removing the character. By this method, we are able to preserve the complex background in the image, which further provides high-quality material for the image restoration and generation process. Subsequently, the background after character removal is automatically restored by a deep learning image restoration technique (Lama model), which is a self-supervised image restoration technique that generates natural and seamless image content based on the input loss regions, avoiding the common problem

of restoration defects. With this restoration technique, the original background after removing the characters is restored without obvious seams, ensuring the naturalness and authenticity of the background.

To expand the dataset, we utilized several generative adversarial network (GAN) models, including StyleGAN2-ADA [4], WGAN [15], DCGAN [21], and WGAN-GP [16]. Compared to other generative models like VAE, flow-based models, and diffusion models, GANs excel in generating structurally complex and detailed images. This capability is especially critical for disaster scenes, where intricate features like ruins, fires, and floods must be accurately depicted. While VAE offers stability and theoretical rigor, it often produces blurred or distorted images in complex scenarios. Similarly, flow-based models, despite their ability to model data distributions, have high computational demands and limited diversity in generated images. Diffusion models generate high-quality outputs but are slow in both training and image generation, making them less suitable for large-scale or real-time applications.

GANs, with their unique adversarial training mechanism, effectively capture complex structural features while maintaining image quality. They exhibit strong performance in data-scarce or complex scenarios, making them the preferred method for disaster image generation. Among these, StyleGAN2-ADA is central to this study due to its suitability for limited datasets. Its adaptive discriminator augmentation (ADA) dynamically adjusts data enhancement strategies during training, effectively mitigating overfitting. Traditional discriminators often overfit to training data in low-data settings, hindering the generator's ability to learn diverse features. ADA counters this by introducing random data augmentations—such as cropping, color perturbation, rotation, and flipping—when the discriminator becomes overly confident.

Additionally, StyleGAN2-ADA incorporates advanced regularization mechanisms, including path length regularization and a non-leakage discriminator design. Path length regularization enhances the generator's generalization by stabilizing gradient variations and reducing reliance on specific input noise. The non-leakage discriminator prevents gradients from overly influencing the generator, focusing instead on distinguishing overall distributional features. These multi-level optimizations allow StyleGAN2-ADA to generate high-quality and diverse disaster images, even in small-sample scenarios.

The loss function of StyleGAN2-ADA consists of the adversarial losses of the generator and the discriminator, respectively:

Generator loss function:

$$L_G = -E_{z\sim p_z(z)}\left[\log D(G(z))\right]$$

where $D(G(z))$ is the discriminator score of the image generated by the generator G, and $P_z(Z)$ is the latent spatial distribution of the generator input. The goal of the generator is to maximize the discriminator misclassification of the generated image, i.e., to make the discriminator think that the generated image is indistinguishable from the real image.

Discriminator Loss Function:

$$L_D = -E_{x\sim p_{data}(x)}\left[\log D(x)\right] - E_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

where X represents the real image, $G(z)$ is the generated image, and $P_{data}(X)$ is the data distribution. The goal of the discriminator is to maximize its ability to discriminate

between real and generated images, i.e., to keep the discriminant value of the real image as close to 1 as possible and the discriminant value of the generated image as close to 0 as possible. During the training process, the generator's goal is to minimize the loss between the output of the generated image and that of the discriminator so that the generated image becomes more and more difficult to discriminate, while the discriminator enhances its ability to discriminate between the real and the generated images by maximizing the loss between the real and generated images. Discriminator enhances its discriminative ability by maximizing the loss between the real image and the generated image. Through this adversarial training, the generator is able to produce increasingly realistic images, while the discriminator is able to more accurately distinguish between real and fake images.

To evaluate the quality of the generated images, we use the Fréchet Inception Distance (FID) metric, which quantifies the fidelity and diversity of the images by calculating the difference between the distributions of the generated images and the real images in the feature space.The FID is calculated using the following formula:

$$FID(p, q) = \left\|\mu_p - \mu_q\right\|_2^2 + Tr\left(\sum\nolimits_p + \sum\nolimits_q - 2\left(\sum\nolimits_p \sum\nolimits_q\right)^{1/2}\right)$$

where μ_p and μ_q are the mean values of the generated and real images, Σ_p and Σ_q are the covariance matrices of the corresponding distributions,and T_r denotes the trace of the matrix.The lower the value of FID, the higher the perceptual similarity between the generated image and the real image, and the better the generation effect. In this study, we use StyleGAN2-ADA for image generation in disaster scenarios to generate a variety of disaster background images. These generated images will be used for subsequent training dataset expansion to further improve the robustness and accuracy of the target detection model. In addition, in order to comprehensively evaluate the effectiveness of different generation models, we also compare four models, WGAN, WGAN-GP, DCGAN and traditional GAN, for generating FID scores. StyleGAN2-ADA generates higher-quality, more diverse images that better match the disaster scenario with limited samples, demonstrating its advantages in complex backgrounds.

After generating a large number of disaster background images, we fused character targets extracted from original images with the newly generated backgrounds. To achieve seamless integration, we employed an image mask-based synthesis technique, smoothing character edges to minimize artifacts and ensure realistic transitions. This process produced diverse disaster scene images, enriching the dataset and enabling simulations of various post-disaster search and rescue scenarios. These synthetic images enhanced dataset diversity, allowing the model to adapt more effectively to complex environments and multi-target detection tasks.

Once the dataset was expanded, we used the LabelImg tool [22] to annotate the images by creating bounding boxes for character targets and assigning corresponding category labels. To ensure labeling accuracy, we manually reviewed and corrected potential biases, considering both the image content and background context. The final dataset included 900 annotated images for model training.

For target detection, we trained three mainstream models—Faster R-CNN [23], YOLOv5 [19], and YOLOv11 [5]—using the labeled dataset. YOLOv5 and YOLOv11, as single-stage detection models, offered faster inference speeds compared to the two-stage Faster R-CNN. During training, the models learned features such as target location, morphology, and contextual information, progressively improving detection accuracy and robustness. Post-training, we conducted a rigorous performance evaluation using standard metrics like Precision and Recall to validate the models' effectiveness in disaster scenarios.

4 Results and Discussion

In this section, we discuss the ethical considerations and model performance evaluation in the process of generating disaster imagery. For image generation, we compare the performance of five generative models—WGAN, GAN, DCGAN, WGAN-GP, and StyleGAN2-ADA—in generating disaster scenarios. Additionally, we evaluate the performance of object detection models, including Faster R-CNN, YOLOv5, and YOLOv11, in identifying key elements within disaster images. By comparing their detection accuracy and robustness under different conditions, we assess their effectiveness in extracting meaningful information from both real and generated disaster scenarios. As discussed in detail later.

Ethical issues are always a concern in the process of generating imagery. Generated disaster imagery, while providing important support for disaster response and assessment, may also pose potential ethical risks, especially in terms of disinformation dissemination and data privacy protection. Therefore, when generating imagery, we have taken stringent measures to ensure that the use of these images is clear and that they are not misused or misleading to the public. In addition, during the image generation process, we ensure that we do not violate the privacy or likeness rights of any individual and take measures such as data encryption to ensure that the privacy of the affected individuals is protected. We compare the performance of five generative models—WGAN, GAN, DCGAN, WGAN-GP, and StyleGAN2-ADA—in the task of generating images of disaster scenarios, as shown in Figs. 2 and 3. By analyzing the loss curves and FID scores of each model, we evaluated their differences in terms of training stability and quality of generated images. The results show that DCGAN converges faster compared to GAN and shows better stability with less fluctuation in the loss curve; whereas the loss curve of GAN shows significant fluctuation during training, and the FID score also shows significant fluctuation and increases at 600 rounds. The loss curve of WGAN shows larger fluctuation, compared to the WGAN-GP that showed more stable performance, with less fluctuation in the loss curve and faster convergence, especially the FID score decreased more rapidly in the early stage of training. Through the comparison of FID scores (after 600 rounds of Epoch, the FID score of GAN is 1405, DCGAN is 930, WGAN is 813, and WGAN-GP is 854), it can be seen that there is a gap between the quality of the generated images of these models, especially the GAN model has a large fluctuation of FID scores in the later stage, indicating that there is an unstable quality of its generated images.

StyleGAN2-ADA outperforms all other models in disaster image generation, demonstrating more stable training and better quality, as seen in Fig. 4. After 1000 epochs, it

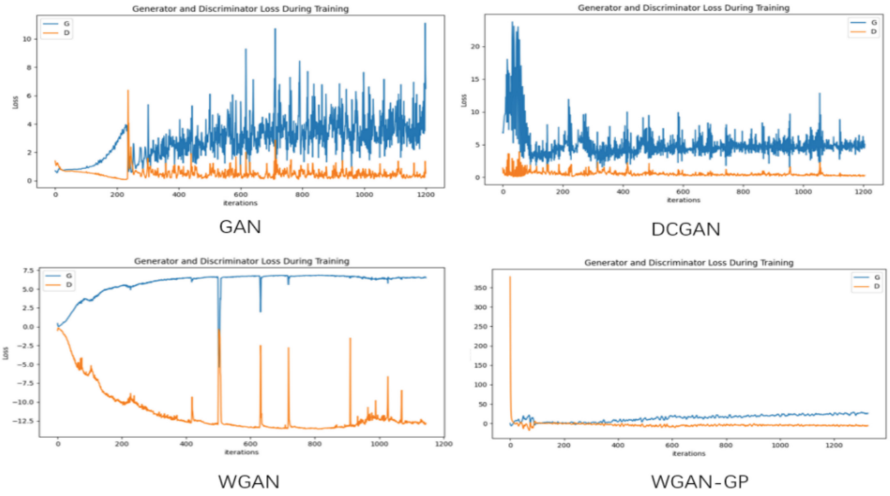


Fig. 2. Generator and Discriminator Loss Trends During Training Across GAN Variants (GAN, DCGAN, WGAN, and WGAN-GP).

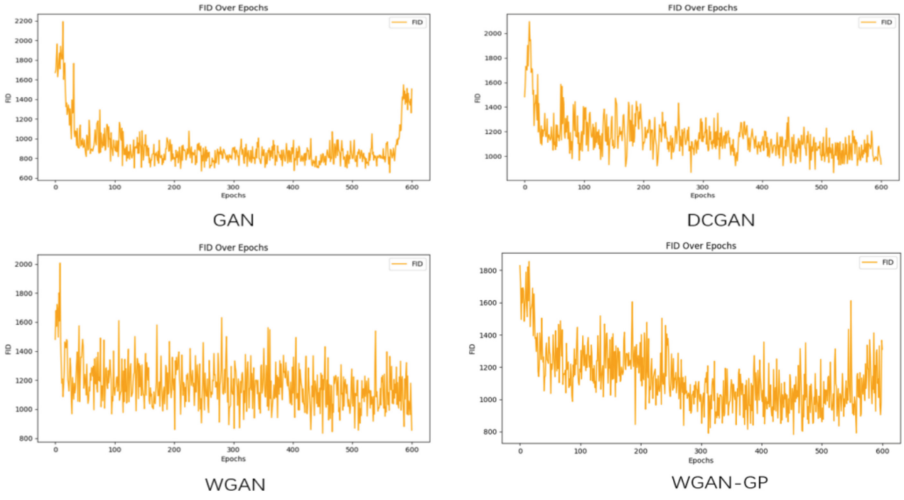


Fig. 3. FID Scores Across Epochs for GAN Variants (GAN, DCGAN, WGAN, and WGAN-GP)

achieves an FID score of 45.36, significantly lower than other models, indicating superior image quality. Its adaptive data augmentation and incremental training .allow it to handle data imbalance, generate rich scene details, and ensure seamless transitions even in complex environments. In contrast, models like GAN, DCGAN, WGAN, and WGAN-GP lack these advanced strategies, resulting in less stable training and lower image resolution. Limitations such as small dataset size (~2000 images) and suboptimal high-resolution performance also constrain these models. Figure 5 highlights the superior realism and detail of images generated by StyleGAN2-ADA.

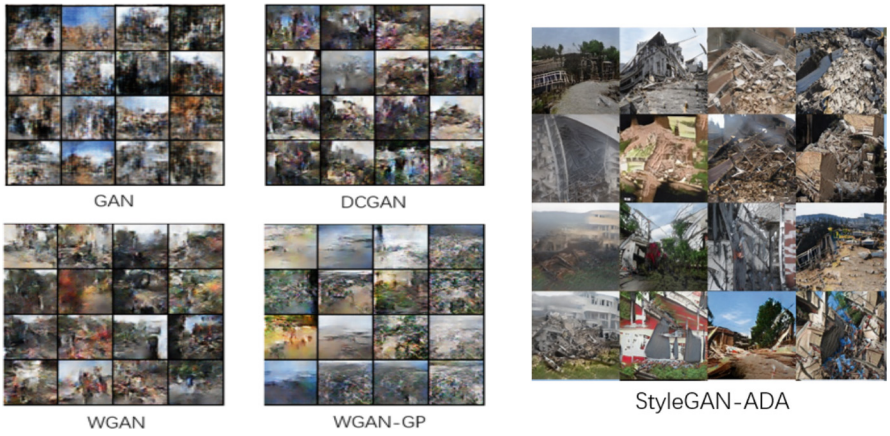


Fig. 4. Comparison of the quality of the generated images.

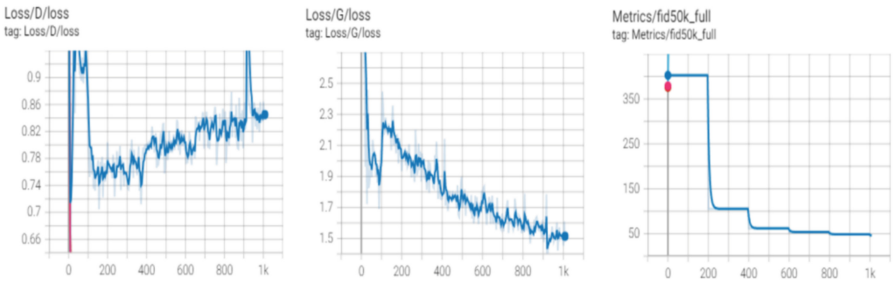


Fig. 5. Generator and Discriminator Loss Curves and FID Metric Convergence During Training.

The evaluation of target detection models in disaster scenarios highlights significant performance differences among YOLOv5, YOLOv11, and Faster R-CNN, as illustrated in Figs. 6 and 7. YOLOv5 consistently achieves superior precision, recall, and mAP scores, demonstrating robust and reliable performance in detecting human targets amidst complex backgrounds. Its mAP50 and mAP50–95 scores, 0.993 and 0.729 respectively, indicate its capability to handle multi-target detection tasks effectively, even under challenging conditions. YOLOv11, while achieving competitive recall scores, exhibits slight variations in precision and struggles with detailed detection tasks, resulting in marginally less stable performance in complex scenarios. Faster R-CNN, a two-stage detection model, demonstrates high precision and recall but lags behind in mAP50–95, reflecting limitations in detecting smaller or less distinct targets. These results position YOLOv5 as the most suitable model for disaster scenarios requiring high accuracy and adaptability.

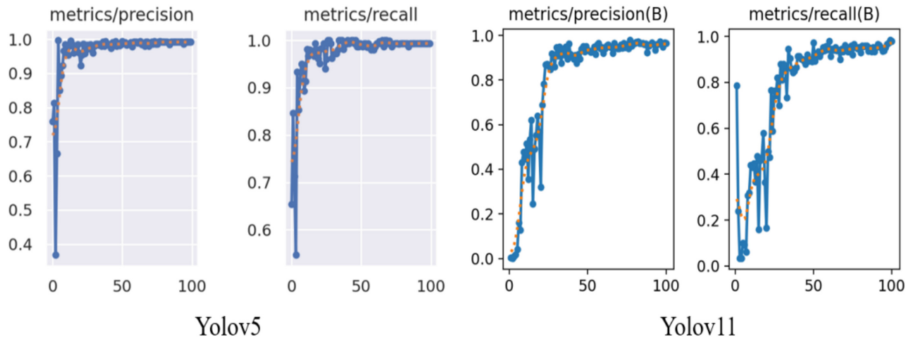


Fig. 6. Precision and recall performance curves for YOLOv5 and YOLOv11 during training, highlighting YOLOv5's consistent accuracy and stability compared to YOLOv11 in disaster target detection tasks.

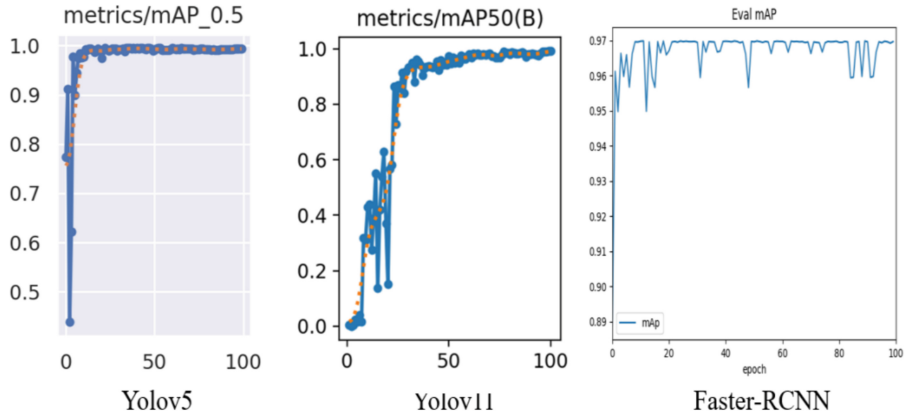


Fig. 7. Comparison of mAP metrics (mAP50 and mAP50–95) for YOLOv5, YOLOv11, and Faster R-CNN, demonstrating YOLOv5's superior performance in target detection accuracy and robustness across diverse disaster scenarios.

The UAV user interface, shown in Fig. 8, integrates advanced functionalities to enhance disaster response efficiency. The interface provides real-time video feeds with target detection overlays, allowing operators to monitor and identify individuals in disaster zones. Model selection options tailored to specific disaster types, such as floods, earthquakes, or fires, allow for dynamic adaptation to diverse scenarios. Essential features, including an SOS signaling button, emergency zone marking tools, and real-time flight parameter monitoring, ensure operational safety and effectiveness. The integration of manual and automated UAV control, combined with satellite mapping for precise navigation, further supports decision-making in time-critical rescue missions. This user-centric design streamlines UAV operation and target detection, providing a robust framework for post-disaster emergency response.



Fig. 8. UAV-based disaster rescue system interface, showcasing real-time video feeds with target detection, model selection options for specific disaster scenarios, emergency tools, and comprehensive UAV control features.

5 Conclusion and Future Work

This study presents a disaster scene image generation and analysis system combining generative adversarial networks (GAN) and target detection techniques to enhance disaster response and assessment. By generating realistic disaster images and performing accurate target detection, the system significantly improves the efficiency and accuracy of disaster scene analysis. Among the tested models, StyleGAN2-ADA achieved superior performance in image generation, offering stable training and high-quality outputs, while YOLOv5 demonstrated exceptional accuracy and robustness in target detection, particularly in complex scenarios. These advancements provide critical support for disaster management, offering decision-makers intuitive tools for analysis and planning.

While this system represents significant progress, some challenges remain. The fusion of characters with disaster backgrounds, especially in multi-target scenarios or complex environments, requires further optimization. Small target detection and performance in intricate backgrounds also present limitations. To address these issues, future work will focus on enhancing the fusion algorithm by improving the naturalness and consistency of synthesized images, refining size ratios, position selection, and multi-target handling.

Additionally, we aim to improve the robustness of the target detection model under diverse conditions and develop lightweight models for real-time target detection in UAV systems. These improvements will enable faster and more reliable disaster response while maintaining ethical considerations, such as privacy protection and accuracy in generated information. By addressing these challenges, this technology can further advance disaster management and human-computer interaction applications in related fields.

References

1. Schedl, D.C., Kurmi, I., Bimber, O.: An autonomous drone for search and rescue in forests using airborne optical sectioning. Sci. Robot. **6**(55), eabg1188 (2021)
2. Song, H., Song, W., Cheng, L., et al.: PDD: post-disaster dataset for human detection and performance evaluation. IEEE Trans. Instrum. Meas. **73**, 1–14 (2023)
3. Kostage, K., West, D., Meinert, T., et al.: Enhancing Autonomous Intrusion Detection System with Generative Adversarial Networks. In: 2024 IEEE 20th International Conference on e-Science (e-Science), pp. 1–10. IEEE, Piscataway, NJ (2024)
4. Karras, T., Aittala, M., Hellsten, J., et al.: Training generative adversarial networks with limited data[J]. Adv. Neural Inf. Process. Syst. **33**, 12104–12114 (2020)
5. Wang, Z., Su, Y., Kang, F., et al.: PC-YOLO11s: a lightweight and effective feature extraction method for small target image detection. Sensors. **25**(2), 348 (2025)
6. Kirova, V.D., Ku, C.S., Laracy, J.R., et al.: The ethics of artificial intelligence in the era of generative AI. J. Syst. Cybernet. Inform. **21**(4), 42–50 (2023)
7. Zlateva, P., Steshina, L., Petukhov, I., et al.: A conceptual framework for solving ethical issues in generative artificial intelligence. In: Electronics, Communications and Networks, pp. 110–119. IOS Press, Amsterdam (2024)
8. Rivera, R., Rolke, W.: Modeling excess deaths after a natural disaster with application to hurricane Maria. Stat. Med. **38**(23), 4545–4554 (2019)
9. Van Brown, B.L.: Disaster research "Methics": ethical and methodological considerations of researching disaster-affected populations. Am. Behav. Sci. **64**(8), 1050–1065 (2020)
10. Stellmacher, C.: Ethical Research Conduct Scholarly Recommendations: Exposure Research during and after Disasters. (2023)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Dinh, L., Krueger, D., Bengio, Y.: NICE: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
13. Van den Oord, A., Kalchbrenner, N., Espeholt, L., et al.: Conditional image generation with PixelCNN decoders. Adv. Neural Inf. Process. Syst., 29 (2016)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. **33**, 6840–6851 (2020)
15. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 (2017)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., et al.: Improved training of Wasserstein Gans. Adv. Neural Inf. Process. Syst. **30** (2017)
17. Karras, T., et al.: Analyzing and Improving the Image Quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
18. Golcarenarenji, G., Martinez-Alpiste, I., Wang, Q., Alcaraz-Calero, J.M.: Illumination-aware image fusion for around-the-clock human detection in adverse environments from unmanned aerial vehicle. Expert Syst. Appl. **204**, 117413 (2022)
19. Xing, L., et al.: Multi-UAV cooperative system for search and rescue based on YOLOv5. Int. J. Disaster Risk Reduct. **76**, 102972 (2022)
20. Qu, C., Drefahl, P., Guo, W., et al.: Autonomous Video Transmission and Air-to-Ground Coordination in UAV-Swarm-Aided Disaster Response Platform. In: International Conference on Human-Computer Interaction, pp. 339–355. Springer Nature Switzerland, Cham (2024)
21. Fang, W., Zhang, F., Sheng, V.S., et al.: A method for improving CNN-based image recognition using DCGAN. Comput. Mater. Contin. **57**(1), 167–178 (2018)

22. Yakovlev, A., Lisovychenko, O.: An approach for image annotation automatization for artificial intelligence models learning. Адаптивні системи автоматичного управління. **1**(36), 32–40 (2020)
23. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)

Biomechanics, Ergonomics, and Risk Mitigation

Research on Ergonomic Evaluation and Design Improvement of Grain Analyzer Based on Jack

Yu Chen and Meiling Zhou(✉)

School of Design, South China University of Technology, Guangzhou 510641, People's Republic of China
zhoumeiling1@163.com

Abstract. To improve the operating comfort of the Nuclear Magnetic Resonance (NMR) grain quality analyzer as well as the work efficiency of the inspectors, this study utilized the Hierarchical Task Analysis (HTA) method to sort out the operating process of the grain analyzer, and constructed 2 key postures, upright operation, and upright stooping posture, based on Jack's Digital Human Model (DHM). Jack was used to analyze the visual domain, reachable domain, and comfort of grain inspectors. The results showed that the upright posture when operating the grain analyzer was relatively better than the standing with bending posture, and the inspectors were less comfortable working in the upright posture for a long period. Based on the results of the analysis, an improved design of the operating area and the overall arrangement of the grain analyzer was made, and the improved solution was analyzed, it was found that the comfort level of the instrument was enhanced after the improvement. The research indicates that Jack's suggestions for improving the instrument and proposing new design forms from a human factors perspective can serve as valuable references for the design of future NMR grain analyzers.

Keywords: ergonomics · NMR grain analyzer · Jack software · operating posture · HTA

1 Introduction

China's huge grain production makes the inspection of grains before storage extremely important, if the storage conditions are not appropriate, it will affect the storage time of agricultural products or cause grain deterioration. The demand for storage inspection of grain and the pursuit of safety and high quality of grain both require high efficiency of grain inspection.

The grain quality analyzer can assist inspectors in analyzing and detecting grain in various indoor and outdoor settings, including laboratories, production workshops, warehouses, and field sites. Improper operating posture can affect not only the efficiency of grain detection but also the physical and mental health of the inspectors. Consequently, to reduce the risk of work-related musculoskeletal disorders caused by long-term muscle fatigue among inspectors, it is very necessary to conduct ergonomic studies on the testing operations of grain analyzers.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 275–288, 2025.
https://doi.org/10.1007/978-3-031-93502-2_18

Grain quality analyzers currently on the market mainly use Near Infra-Red (NIR) and NMR technology for detection. Among them, NMR technology has the advantages of being more stable as well as unaffected by the thickness, shape, color, and size of the sample compared to NIR technology, and it still has a lot of room to play in the field of agriculture [1]. Fewer grain analyzers on the market apply NMR technology and are only suitable for indoor scenarios, such as the mq-one seed analyzer from Bruker.

In addition, there is still a gap in ergonomic research on the inspection operation posture of grain analyzers, and the working comfort of inspectors has been neglected. However, there have been theoretical studies on the analysis of operating postures in other fields and the human-machine simulation of user operation tasks. For example, Zhou and Li et al. [2, 3] used Jack software to ergonomically analyze as well as design improvements to key postures in the operation process, which improved the personnel's operational comfort as well as their work efficiency; Zhang et al. [4] constructed and evaluated a digital human and welding torch model based on the Jack software, which improved the welder's operational posture; Emmatty et al. [5] used the Jack's Digital Human Model (DHM) to evaluate the static posture for waste sorting tasks, suggesting appropriate heights for sorting tasks and appropriate positions for sorted items; Banga et al. [6] used Jack software and Kinect interface to ergonomically evaluate the working posture of lawnmower operators, addressing the ergonomic issues of musculoskeletal disorders associated with operator whole body work. Their research validated the effectiveness of using Jack software for work posture assessment and improving workflow.

Therefore, to promote the wide application of the NMR grain analyzer, this study aims to task decompose the operation flow of the mq-one analyzer, extract typical operation postures, conduct ergonomic evaluation through the DHM in Jack, and optimize the design of the user's operation flow from the operation postures based on the results of the assessment, and then propose the improvement of the instrument from the perspective of ergonomic engineering and a new design form, providing reference for the design of subsequent NMR grain analyzers.

2 Methods

DHM, based on 3D models, enables designers to analyze and evaluate conceptual solutions in the early stages of design, aiding them in exploring the ergonomics of the proposed designs [7]. Therefore, to optimize the usage process of the NMR grain analyzer, a human factors engineering evaluation of the mq-one analyzer was conducted using DHM. Firstly, it is necessary to organize the operational process of the grain analyzer, extract key postures involved, and then utilize Jack simulation software to analyze and simulate the instrument. Before analysis, it is essential to construct an instrument model and establish a simulation environment. Subsequently, the analysis subject model should be imported into Jack to align with the virtual digital human of the constructed inspector, thereby creating a simulation scenario for grain analysis. After that, a comprehensive analysis of the inspector's visual field, reachable area, and comfort level will be conducted. Based on these analysis results, design improvements and instrument verification will be carried out. The detailed human-machine engineering simulation process for the grain analyzer is illustrated in Fig. 1.

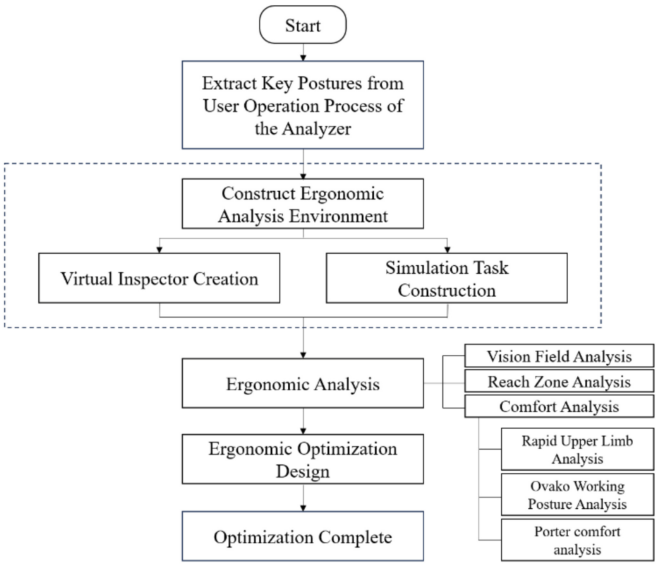


Fig. 1. Grain analyzer ergonomics simulation flow.

2.1 Key Pose Extraction

Hierarchical task analysis (HTA) is a task analysis method commonly used in human factors engineering, which represents tasks through a top-down hierarchical structure, decomposing tasks into main objectives and their sub-objectives [8, 9]. Hierarchical combing of the operation tasks of the grain analyzer through HTA can refine the tasks and help understand the key operations involved. The specific combing results of the instrument operation process are shown in Fig. 2.

In the process of grain analysis, since the tester needs to analyze different grain samples, the tester will operate cyclically in conducting the detection and data processing stages. After the data processing of the previous grain sample is completed, the sample tube will be replaced for a new round of detection and data processing analysis. Therefore, we primarily focus on analyzing the operational postures in the detection and data processing stages.

When placing a sample, the tester has to stand to operate the instrument due to its size and height. The instrument's detection usually takes one minute or less. Thus, the user typically chooses to stand and wait. Since the data processing time after the test is short, the user generally stands and bends over to operate the computer to facilitate the next-round testing. As a result, the operator mainly alternates between the upright and the bending postures. The mapping relationship between the specific target task and the operating posture is shown in Fig. 3.

2.2 Simulation Environment Construction

Since the grain analysis task imposes no specific requirements on the body size of operators, to accommodate the usage needs of the majority, in line with the GB/T 10000—1988

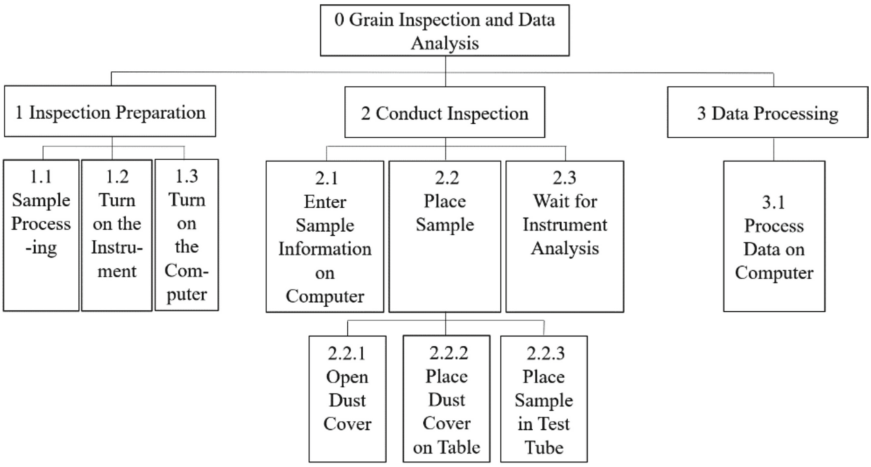


Fig. 2. Grain analyzer operation process hierarchy.

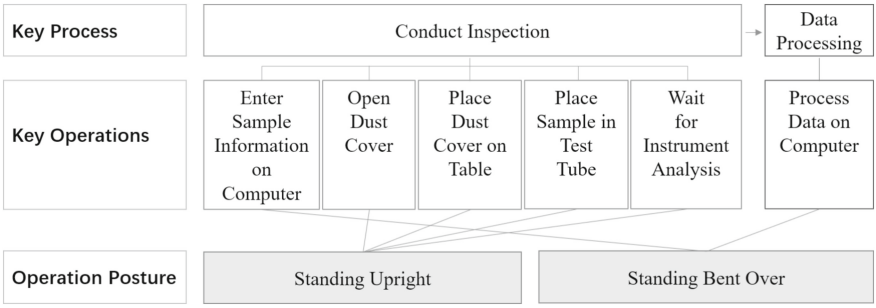


Fig. 3. Mapping of grain analyzer operation flow to operation posture.

standard "Human Dimensions of Chinese Adults", the 50th percentile digital human models of Chinese males and females were selected in the Jack simulation software to simulate grain inspectors, as depicted in Fig. 4.

In addition, based on the actual size, a simplified 3D model of the main body of the grain analysis equipment was established in Rhino. The model includes the grain analyzer, experimental table, computer, etc. According to the GB/T 14531—2017 standard "Office furniture – Tables, chairs and stools for reading", the height range of the desktop is 680 – 760 mm. For modeling and simulation, the average height value of the desktop, 720 mm, was taken. The Rhino model was saved in STL format and imported into the Jack software. Subsequently, according to the key operation postures derived from the above-mentioned analysis, the corresponding virtual human body model was constructed in Jack. It was then matched with the imported grain analysis environment model to construct the virtual simulation environment shown in Fig. 5.



Fig. 4. P50th percentile numerical mannequins for males and females.



Fig. 5. Simulation task construction.

2.3 Simulation Analysis and Evaluation

Visual Domain Analysis. When the inspectors conduct grain analysis, the visible range affects the accuracy and efficiency of the operation to a certain extent. The View Cones tool in Jack software was used to analyze two postures of the inspectors' operation. The viewing distance was set at 200 mm and the viewing cone angle was set at 40°, and then the visual fields of the inspectors were obtained, as shown in Fig. 6. The results indicate that the visibility of the inspectors is good in both the standing and bending postures.

Reach Zone Analysis. Reach zone analysis serves to determine whether the items to be manipulated are within the reach of the inspectors' hands. The Reach Zones tool in Jack software was used to simulate the reachable area of the inspectors' hands. Figure 7 presents the analysis results of the inspectors' reachable area in different postures.

Findings show that in the standing posture, the inspectors' reachable area fails to cover the desktop. This situation may impede the inspectors from placing the dust cover on the desktop, demonstrating relatively poor reachability. Redesigning the dust cover

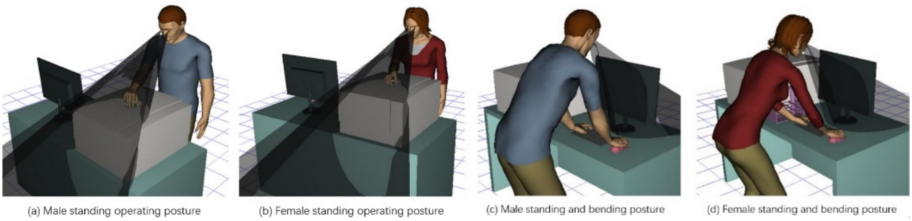


Fig. 6. Detection personnel's visual cone range under different operating postures.

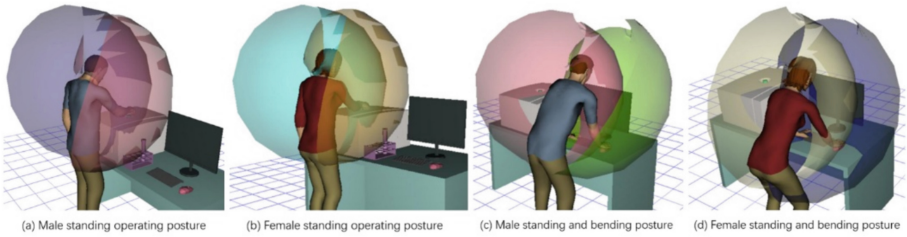


Fig. 7. Detection personnel's reach zone under different operating postures.

can enhance this operation. In contrast, in the bending posture, the reachability of the inspectors' hands is relatively good.

Comfort Analysis. The comfort of the inspectors' operating postures has an impact on both the efficiency of completing the grain analysis task and the fatigue level of the grain inspectors during work. Thus, to further explore the comfort during the operation process, the Rapid Upper Limb Assessment (RULA), the Ovako Working Posture Analysis (OWAS), and the Porter tool in Jack software were employed for analysis.

Rapid Upper Limb Assessment (RULA). RULA is capable of assessing how upper-limb movements affect the human body. The higher the score it generates, the more urgent improvement is required. Specifically, a score ranging from 1 to 2 indicates that the working posture is acceptable; a score of 3–4 suggests that the posture should be altered after an extended period; a score of 5–6 implies that immediate posture change is needed; and a score of 7 means that the posture demands instant investigation and modification [10].

The RULA was employed to simulate and analyze the standing and bending operating postures. The analysis results are presented in Fig. 8. Both male and female inspectors received a score of 3 for the standing posture during grain analysis, which is acceptable in the short term. In the bending posture for computer operation, the male digital human scored 6, and the female digital human scored 4, indicating a relatively low comfort level. Notably, the bending operating posture of male inspectors calls for prompt improvement.

Ovako Working Posture Analysis (OWAS). OWAS analysis can assist in identifying the working postures that need to be improved on a priority basis. Different levels can reflect the degree of potential harm that the operating postures may cause to the human body. Specifically, Level 1 represents a normal posture, for which no investigation and

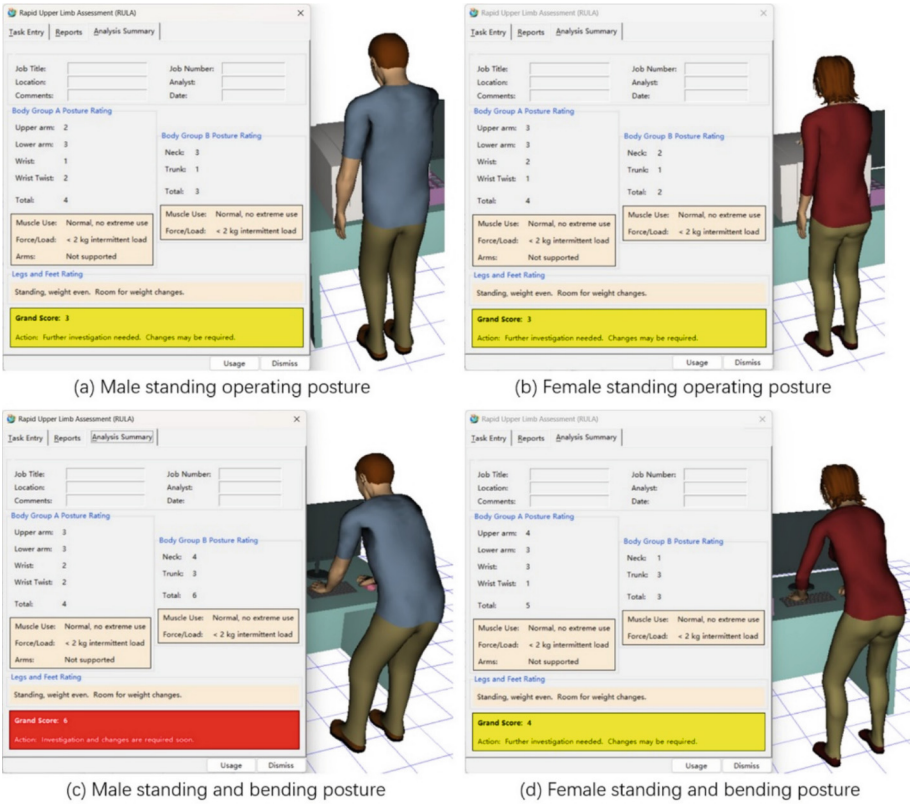


Fig. 8. Rapid upper limb assessment analysis results of testing personnel.

improvement is required; Level 2 indicates that the working posture may have harmful effects on the musculoskeletal system and further investigation and improvement can be carried out; Level 3 means that the working posture needs to be investigated and improved as soon as possible; and Level 4 indicates that the posture requires immediate improvement [11].

The OWAS tool was used to conduct a simulation analysis of the standing operating posture and the bending operating posture. The analysis results are shown in Fig. 9. The standing posture is rated as Level 1, and the bending posture is rated as Level 2. The results show that for the inspectors, standing to operate the instrument is acceptable, but bending over to operate the computer has a certain impact on the inspectors' comfort.

In the analysis results of RULA and OWAS, the standing posture when operating the instrument is relatively superior to the bending posture when operating the computer. Therefore, it is advisable to consider redesigning the module layout of the instrument starting from its internal structure, combining the computer module and the instrument module, so that the inspectors can complete the operation in a single standing posture.

Porter Comfort Analysis. Porter is commonly used to evaluate the comfort level of human joint bending angles, offering the comfortable angle range and reference values

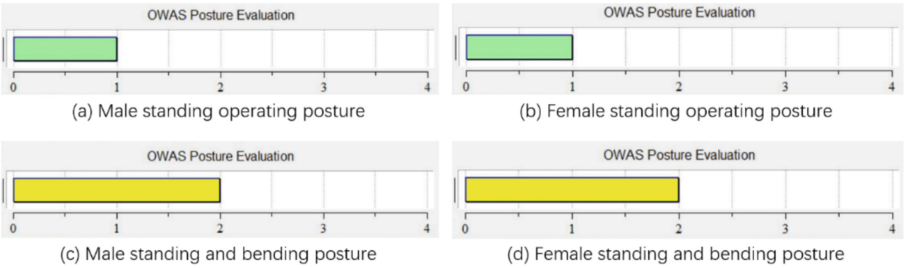


Fig. 9. Analysis results of the working posture of the testing personnel.

for joints. In the result presentation, green indicates that the joint bending is within the specified range, whereas yellow indicates it exceeds the specified interval. A smaller value signifies higher comfort and a larger one indicates lower comfort.

The Porter tool in Jack was used to analyze the inspectors' different operating postures, and the results in Fig. 10 indicate that, in both postures, the bending of the right and left thighs and inward bending of the right and left knees of both male and female inspectors are all marked in yellow. Moreover, in the standing-posture operation, the head-twisting angle and the left-bending angle of the upper arm of male inspectors, as well as the left-bending angle of the upper arm of female inspectors, exceed the specified ranges, suggesting relatively low comfort and the need for further improvement.

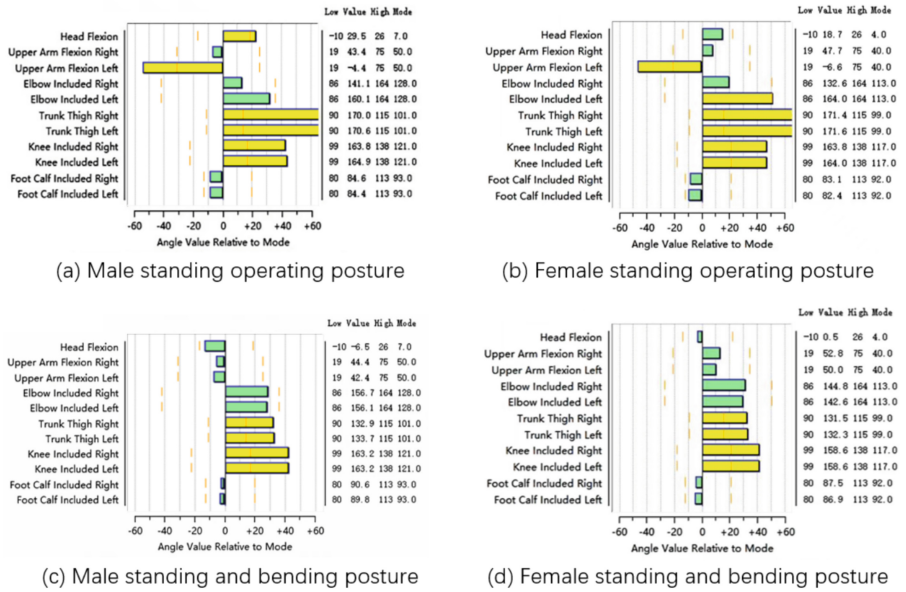


Fig. 10. Porter comfort analysis results of operating posture.

From the Porter analysis results, it can be seen that in the standing state, the bending angles of multiple joints of the inspectors exceed the specified values. If the grain analysis

instrument module and the computer module are simply combined and the operators are required to perform the operation in a standing posture, it will also have a significant impact on the work comfort of the grain inspectors. Therefore, the size and layout of the instrument can be further designed to enable the operators to work in a relatively comfortable posture.

3 Design Optimizer

3.1 Optimization Recommendations

The results of the human factors engineering simulation analysis of the grain analyzer show that the accessibility of the inspector and the comfort of the operating posture need to be further optimized in the grain inspection as well as the data analysis phase, and the design of the grain analyzer is improved mainly in the following aspects: (1) Redesign the dust cover and layout of the instrument's operating area to improve work efficiency. (2) By integrating the instrument detection module and the computer module, operators can avoid alternating between standing and bending postures and instead conduct grain detection in a single posture. (3) The layout and size of the instrument and computer are reconsidered to prevent the inspector from working in a standing position for a long period.

3.2 Improve Design

Overall Optimization of the Instrument. First, the combination of the instrument module and the computer module eliminates the need for inspectors to move back and forth between the instrument and computer operation areas and switch postures. Second, the internal structure of the instrument is rearranged to position it upright. This enables inspectors to conduct detections in a sitting position using laboratory chairs, thus avoiding long-term standing work. Figure 11 presents the pre- and post-optimization comparison and dimensions of the grain analyzer, with dimensions in millimeters. Third, the addition of pulleys and a pull rod to the instrument facilitates its movement by analysts in various indoor and outdoor detection scenarios.

Optimization of Instrument Operation Area. The grain sample placing area and computer operation area are arranged and designed as shown in Fig. 12. So that the two operation areas overlap, and the dust cover is connected with the instrument to reduce the operation of the inspector placing the dust cover and improve the efficiency of the inspector's work.

3.3 User Flow Improvement and Key Posture Extraction

In the test preparation stage, the original operation of turning on the instrument and turning on the computer has been combined into a one-button startup, which simplifies the operation of analysts. In the testing phase, the sample placement task is optimized from three steps of opening the dust cap, placing the dust cap, and putting the sample

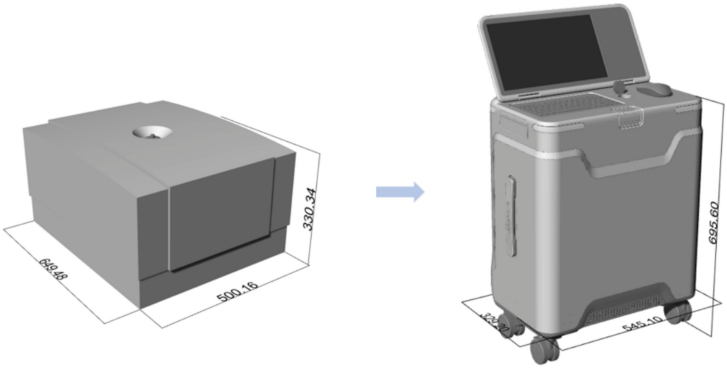


Fig. 11. Grain analyzer before and after optimization

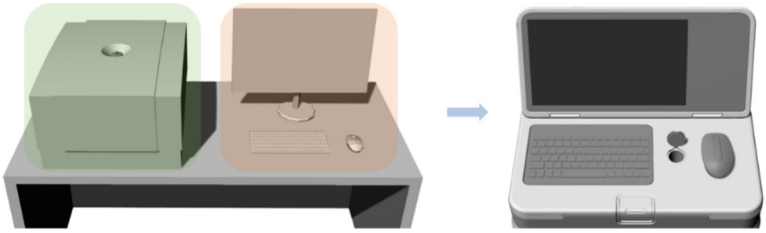


Fig. 12. Comparison of operating areas before and after instrument optimization.

tube to two steps of opening the dust cap and putting the sample tube, which optimizes the operation of the personnel and avoids the possibility of losing the dust cap. The hierarchical structure of the improved grain analyzer operation process is shown in Fig. 13.

In the two critical stages of testing and data processing, analysts can cooperate with laboratory chairs to use a sitting posture for grain testing. Therefore, the next step is to mainly simulate and verify the sitting posture during testing.

4 Result

The chairs used in the simulation analysis are the adjustable-height chairs commonly found in laboratories, with a height range typically between 38 and 51 cm. The male and female virtual persons use chairs with seat heights of 42 cm and 44 cm respectively for the simulation analysis. Additionally, to enhance the simulation efficiency, the optimized instruments are simplified and then imported into Jack for analysis.

4.1 Visibility Analysis of the Improved Instrument

After the instrument was improved, in the sitting posture, the results of the inspectors' viewing cone analysis show that their visual field can cover the display and the items to be manipulated, indicating good visibility. The specific results are shown in Fig. 14.

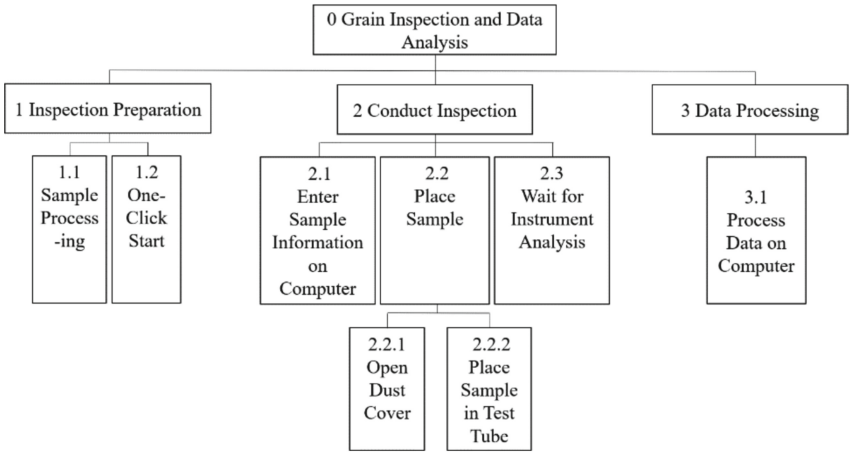


Fig. 13. Improved grain analyzer operation process hierarchy.

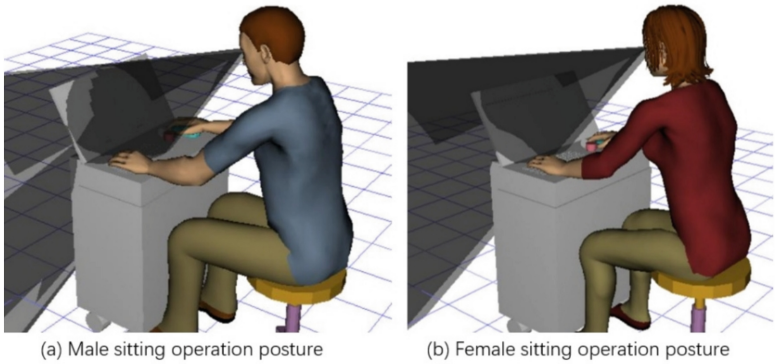


Fig. 14. Improved instrument sitting posture visual cone range.

4.2 Reach Zone Analysis of the Improved Instrument

After the instrument improvement, the reachable area of the inspectors covers the items to be manipulated, indicating good reachability. The results are shown in Fig. 15.

4.3 Comfort Analysis of the Improved Instrument

After the instrument was redesigned, a comparison of the parameters from the RULA results reveals that, compared with the pre-improvement standing and bending postures, the post-improvement sitting posture for detection effectively enhances the joint comfort of grain analysts. The comparison of the RULA results before and after the improvement is shown in Fig. 16. Additionally, the post-improvement OWAS analysis results, as shown in Fig. 17, are all rated as Level 1, indicating that the operating posture is normal.

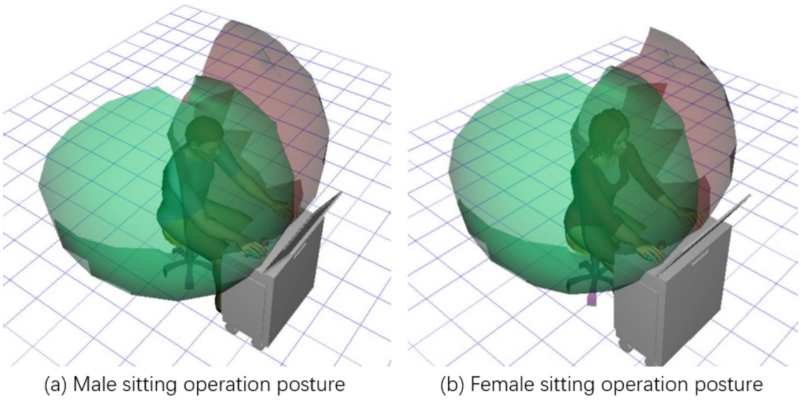


Fig. 15. Visual field of the improved instrument.

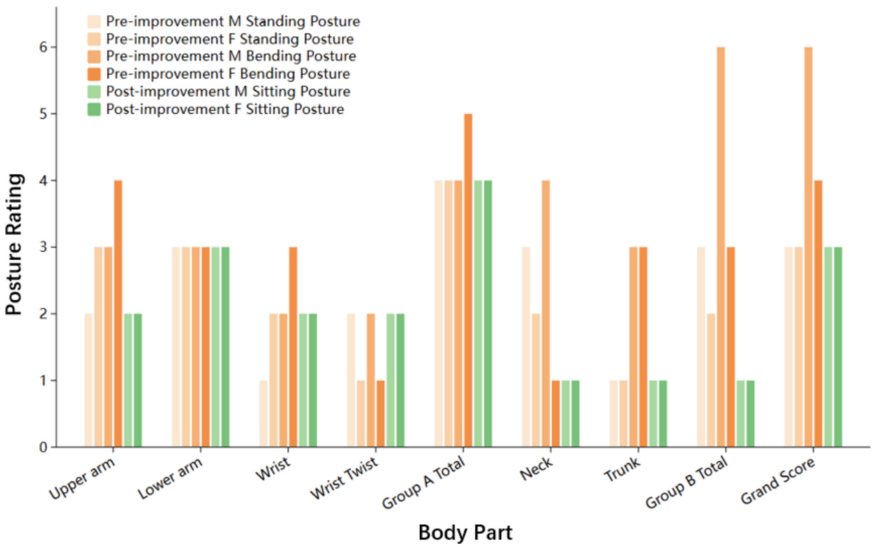


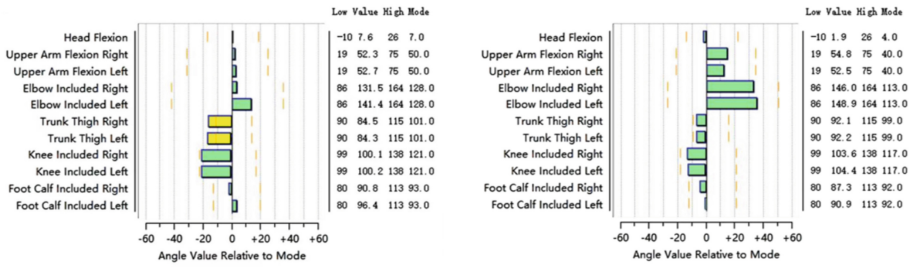
Fig. 16. Comparison of RULA results before and after improvement.



Fig. 17. Working posture analysis results of the improved instrument.

After the instrument improvement, the Porter analysis results of the inspectors are shown in Fig. 18. For female inspectors, the bending degrees of joints in the sitting-posture operation are all displayed in green, indicating a high level of comfort. Although

the bending degrees of the right and left thighs of male inspectors still exceed the specified range, the degree of exceedance is less than that before the improvement. Generally speaking, the comfort level when operating the improved instrument has been enhanced.



(a) Male sitting operation posture (b) Female sitting operation posture

Fig. 18. Porter comfort analysis result after Instrument Improvement.

5 Discussion and Conclusion

In this paper, we used the HTA method to decompose the inspectors' tasks and employed the DHM in Jack to evaluate the ergonomics of grain analyzers and improve the operation process. This approach enables designers to identify the challenges in user operation at the early design stage. Subsequently, they can optimize the instrument design based on the results of human-factors analysis. Moreover, the use of virtual digital humans significantly reduces the cost of later-stage testing and validation.

The study results indicate that, in comparison with the existing separate design of the grain analyzer and the computer, the integrated design can prevent analysts from frequently switching between standing and bending postures. Moreover, the results of the Porter comfort analysis reveal that in the standing position, joints such as the head, arms, and thighs exceed the limit ranges. Therefore, if the instrument with the integrated design is still placed on the desktop and used by inspectors in a standing position, it will significantly affect their comfort and work efficiency. The improved vertical grain analyzer allows inspectors to operate it in a sitting position. The analysis results demonstrate that the operating comfort and posture of the inspectors have been improved, which verifies the feasibility of this solution and can serve as a reference for the future design of grain analyzers.

This paper mainly conducts static simulations on the working postures of grain inspectors, lacking research and analysis on dynamic operations and working time. In the future, more in-depth research can be carried out in this regard to further evaluate the entire working process of inspectors. Moreover, human-factors assessments can be conducted on different heights of the vertical grain analyzer to explore a suitable height range. In addition, the improved grain analyzer can be moved and transported using a pull rod. However, it is still unknown whether the instrument's own weight will cause muscle damage or other safety risks to the hands, shoulders, and other parts of the

operator during the moving process. Further evaluation and analysis can be carried out in future research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Xi, Z.Y., Wang, F.H.: Research progress in grain quality non-destructive testing methods. Sci. Technol. Food Ind. **33**(15), 394–396+400 (2012)
2. Zhou, N., Yao, X., Luo, Q.L., et al.: Ergonomic simulation analysis and design improvement for aircraft tow tractor driving control console. J. Mach. Des. **41**(08), 209–214 (2024)
3. Li, X.Z., Tan, Y.Z.: Human-machine analysis and design improvement of main fire engine based on Jack simulation. J. Mach. Des. **41**(10), 178–183 (2024)
4. Zhang, Y., Wu, X., Gao, J., et al.: Simulation and ergonomic evaluation of welders' standing posture using Jack software. Int. J. Environ. Res. Public Health. **16**(22), 4354 (2019)
5. Emmatty, F.J., Panicker, V.V., Baradwaj, K.C.: Ergonomic evaluation of work table for waste sorting tasks using digital human modeling. Int. J. Ind. Ergon. **84**(7), 103146 (2021)
6. Kumar Banga, H., Kumar, R., Kalra, P.: Ergonomics evaluation of lawn mower operator's working posture using JACK software and kinect interface. Work. **72**(2), 497–510 (2022)
7. Ahmed, S., Irshad, L., Gawand, M.S., et al.: Integrating human factors early in the design process using digital human modelling and surrogate modelling. J. Eng. Des. **32**(4), 165–186 (2021)
8. Stanton, N.A.: Hierarchical task analysis: developments, applications, and extensions. Appl. Ergon. **37**(1), 55–79 (2006)
9. De Maya, B.N., Komianos, A., Wood, B., et al.: A practical application of the Hierarchical Task Analysis (HTA) and Human Error Assessment and Reduction Technique (HEART) to identify the major errors with mitigating actions taken after fire detection onboard passenger vessels. Ocean Eng. **253**, 111339 (2022)
10. McAtamney, L., Corlett, E.N.: RULA: a survey method for the investigation of work-related upper limb disorders. Appl. Ergon. **24**(2), 91–99 (1993)
11. Xue, H., Chen, J., Zhang, X.: A task simulation and ergonomics analysis method based on JACK. In: Duffy, V.G. (ed.) HCII 2020, LNCS, vol. 12429, pp. 584–594. Springer, Denmark (2020). https://doi.org/10.1007/978-3-030-59987-4_41



Evaluating Multi-User VR for Ergonomic Assessments Using a DHM Software

Soureesh De(✉), Henrik Söderlund, Jose Maria van der Ploeg Feriche, and Björn Johansson

Department of Industrial and Materials Science, Chalmers University of Technology, Chalmersplatsen 4, 412 96 Gothenburg, Sweden
soureesh@student.chalmers.se

Abstract. In the field of manufacturing assembly and workplace, ergonomics plays a critical role in safeguarding workers' health while ensuring workers efficiency. Musculoskeletal diseases (MSDs), caused by poor ergonomics in the workspace, are shown to have significant impact on productivity and absences, costing the industry billions yearly as well as human suffering.

Simulation of ergonomics through Digital Human Modelling (DHM) has become a key tool for addressing ergonomic challenges early in the design phase of workstations and factories since it enables ergonomic issues to be resolved before production, avoiding costly redesigns and delays. Furthermore, Virtual Reality (VR) as a tool to enhance ergonomic assessments has received increased attention in recent years. To reach its full potential the VR simulations will need to account for multi-stakeholder collaboration through what is referred to as multi-user VR.

This study examines the benefits and limitations of VR, particularly multi-user VR, in ergonomic assessments. Interviews with ten experts revealed key limitations as well as potential solutions and future directions for VR and multi-user VR in the field of DHM and ergonomic simulations.

A validation case study together with four ergonomics experts was conducted, comparing the ability to collaborate and assess manikin postures though a DHM desktop application, DHM single-user VR application, and a DHM multi-user VR application. The results from the interviews and case study are further discussed and future outlooks of multi-user VR in DHM simulations are presented.

Keywords: Virtual Reality · Multi-user · Ergonomics

1 Introduction

Today's manufacturing industry faces intense competition worldwide, necessitating the optimization of production costs and quality, alongside compliance with increasingly stringent environmental regulations [1]. Among the numerous aspects receiving heightened attention from companies striving to differentiate themselves in the market is the enhancement of their workstations and workspaces, or more specifically, the improvement of ergonomics within these spaces. This emphasis underscores the importance of creating work environments that are not only conducive to productivity but also prioritize comfort and user-friendliness, thereby fostering the well-being of employees [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 289–300, 2025.
https://doi.org/10.1007/978-3-031-93502-2_19

The increasing significance of ergonomics in the manufacturing industry in recent years is primarily attributed to the substantial economic losses associated with issues stemming from poor ergonomics. One of the main concerns revolves around the work-absences that workers need to take due to Musculoskeletal Disorders (MSDs). These disorders encompass injuries and/or conditions that affect the movement of the human body or its musculoskeletal system, thereby hindering their ability to perform their work effectively [3]. This issue results in millions of individuals worldwide having to leave their jobs for periods of time to recover from such injuries. According to the World Health Organization, approximately 1.71 billion people globally suffer from musculoskeletal conditions, making them the leading cause of disability [4]. Additionally, from the perspective of production and the quality of manufactured products, it has been demonstrated that poor ergonomics also have a negative impact [5], as they limit the capabilities and potential of workers in performing any task.

For this reason, production ergonomics has become a key discipline in optimizing worker interactions with industrial tasks. This branch of ergonomics is focused on creating safe and sustainable work environments that not only improve worker productivity and well-being but also the throughput of the company [6]. To achieve optimal improvements in the ergonomics of the workstation, it is essential to involve multiple stakeholders. This may include more than one ergonomist or a collaborative effort between an ergonomist and an engineer.

In this paper, a series of interviews conducted with experts from both academic and professional fields is presented to establish an initial understanding of the concept of multi-user digital environments for ergonomic simulations as well as validate our findings and demonstrate the potential of a multi-user VR ergonomic assessments through a case of multi-user VR in the IPS IMMA software. This paper thus addresses the following two research questions:

RQ1: Which are the current potential limitations of using VR for ergonomic assessments in workstation design today and how can they be solved?
RQ2: In assessing work station design, can the incorporation of multi-user VR demonstrate overall improvement in ergonomics assessment?

1.1 DHM—IPS IMMA

Identifying ergonomic issues in production lines once they are operational and the product is already on the market can escalate costs significantly. This is because rectifying the problem requires halting or relocating production during the replacement period, in addition to modifying or procuring new machinery to address the ergonomic issue. This is where new technologies come into play, as they can be immensely useful in proactively identifying problems. If during the design phase, work is conducted using a computer model enabling simulations, it becomes easier to make alterations if necessary [7].

Digital Human Modelling (DHM) aims to develop detailed models of the human body to accurately simulate various physical aspects, focusing on anthropometrics and biomechanics, primarily to assess ergonomic factors. Ergonomists can evaluate motions and postures of workers performing tasks, making necessary improvements to enhance comfort and safety, especially in production engineering where tasks involve significant risks [8]. Industrial Path Solutions (IPS) is a comprehensive software suite used by

leading manufacturing companies to streamline processes through advanced algorithms, such as evaluating assembly potential, designing flexible components, optimizing robotic operations, and simulating flow and surface treatments [9]. One notable module, IPS IMMA, uses efficient algorithms to evaluate assembly ergonomics, incorporating realistic biomechanical models, manikins, to ensure collision-free and ergonomic assembly [10].

1.2 Virtual Reality

In recent years, virtual reality (VR) has become a popular topic in information technology. With the advent of affordable, consumer-grade VR headsets designed for gaming and entertainment, VR is experiencing a resurgence. VR creates a simulated environment that immerses users so deeply they feel as though they are truly present in that space [11]. Beyond gaming and entertainment, VR is also very utilized in production and manufacturing. Engineers use VR to design and simulate production lines, enabling virtual presentations and modifications before physical setup. In the context of this paper, VR also facilitates ergonomic assessments, allowing for the optimization of workstations and identification of potential ergonomic problems [12].

VR technologies enable interactions with digital prototypes and virtual models of planned workstations right from the initial stages of planning, facilitating stakeholder involvement and feedback early in the design process [13]. Aiming for a more collaborative approach in the use of VR technology, the concept of "multi-user VR" has emerged, allowing multiple stakeholders to collaborate in the same VR environment.

2 Method

To address the proposed research questions, the study was organized into two distinct but interconnected phases: an interview study, and a validation case involving experimental trials. Each phase was aimed at exploring different aspects of the use of VR and multi-user VR in ergonomic assessments and at providing a comprehensive view of its potential benefits and the challenges.

2.1 Interview Study

The interview study was designed to address RQ1 and establish a foundation for approaching RQ2. This phase aim to deepen the understanding of the practical and theoretical challenges and opportunities associated with using VR in ergonomic assessments, with a particular focus on multi-user VR and its impact on the field.

To capture a wide and informed perspective, we selected a mix of academic researchers and industry practitioners known for their work in ergonomics and VR. Participants who have participated in the interview case study and the validation case study are shown in Table 1.

Using a semi-structured interview format allowed for flexibility and deepened discussions, ensuring that all relevant topics were covered thoroughly [14]. Participants

Table 1. Participants of the Interview Study and Validation Case Study.

No.	Title	Interview	Case study
1	PhD. Student—VR & Product Design	X	
2	Prof.—Human Factors Design & Ergonomics	X	
3	Prof.—Human Factors Design	X	
4	Prof.—Human Factors Engineering & Ergonomics	X	
5	Prof.—Ergonomic Assembly	X	X
6	Ergonomic Simulation Engineer	X	X
7	Ergonomist	X	X
8	Product owner—Multi-user VR platform	X	
9	PhD. Student—Geometry & Ergonomic Simulation	X	
10	Researcher—Human Factors Design	X	X

provided their perspectives on the current state of VR technology in ergonomic assessments, influenced by their respective backgrounds and expertise. They discussed the specific technologies they employ in ergonomics and identified any limitations they face in their current workflows. A significant focus was on how VR tools, particularly multi-user VR, fit into these processes.

2.2 Validation Case Study

The validation case study was designed to apply and test the insights gathered from the interviews, mainly aimed at answering RQ2. The primary aim was to explore how three different technology approaches: *desktop applications, single-user VR*, and *multi-user VR*, influence the ergonomic assessment process in a practical setting. Each case study required the collaboration of at least two ergonomic experts assessing three representative manufacturing tasks to identify and suggest improvements in the posture of a manikin. Following the case study an in-depth interview was held as well as a System Usability Scale (SUS) [15] survey was conducted. The SUS builds on a survey of ten liker-scale questions assessing the subjective understanding of the usability on a scale of 0–100. The case study was structured so that each task was assessed using three different platforms with various collaboration capabilities: *desktop application*, *single-user VR*, and *multi-user VR*.

Desktop Application Using the IPS IMMA 2023 R2 DHM software, the two participants viewed a manikin posture on a desktop screen. They navigated the scene using the mouse and keyboard, manipulating the 3D view through the 2D screen of the computer.

Single-User VR Application Using the IPS IMMA 2023 R2 software, one of the participants used the HTC Vive VR headset to view and assess a manikin posture while the other participant observed the same VR view on a desktop screen. Participants were given the flexibility to switch roles.

Multi-User VR Application Using a multi-user VR application developed by researchers at Chalmers University [16], both participants used the HTC Vive VR headsets to assess the posture of a manikin together. Throughout the assessments participants could collaborate in the VR environment through the representation of avatars in 3D including voice chat.

3 Results

The following sections represent findings and results from both the interviews study and validation case study. Every section offers a distinct perspective on the application of VR and multi-user VR in ergonomic evaluations, emphasizing the benefits, difficulties, and real-world uses of these technologies.

3.1 Limitations of VR in DHM

Participant 3 highlights a significant limitation in VR's inability to support effective collaboration among users, which is crucial for team-based ergonomic assessments. Both participant 1 and participant 3 note that body and hand tracking in VR often lacks accuracy and responsiveness, which can detract from the realism and effectiveness of the simulations. Additionally, participant 3 points out the limited Field of View (FoV) provided by many VR systems, which can restrict users' perceptions and interactions within the virtual environment. For instance, if the hand is positioned behind the body, the VR system might fail to detect it.

Comfort remains a critical concern with VR tools, particularly headsets, which participant 3 finds can be uncomfortable for extended use. Both participant 2 and participant 3 discuss the ease of use of VR tools, emphasizing that despite advances, significant usability barriers still exist that can delay the general adoption and effectiveness in ergonomic studies.

Participant 1 identifies usability friction specifically related to the multi-user VR experience, where the complexities increase due to the need for multiple functionalities that often are not compatible with each other. This lack of compatibility can significantly complicate the user experience as different systems struggle to integrate seamlessly within a shared virtual environment. The industrial experts further elaborated on software friction; particularly emphasizing issues related to cybersecurity. This includes concerns that to utilize some of the available VR systems effectively, companies would need to share sensitive private data. The unwillingness to compromise on data privacy leads to reluctance in adopting these systems, posing additional challenges for companies looking to integrate VR into their regular workflows. These frictions highlight significant hurdles in not only adopting VR technology but also in scaling its use across collaborative platforms that require stringent data security measures.

Participant 1, participant 2, participant 3, participant 4, participant 5, participant 6 and participant 7 all discussed the difficulty of simulating accurate weight sensation and the physical feeling of objects within VR, which is crucial for true-to-life ergonomic assessments. Issues such as proper simulation of lighting, heat, and other environmental factors also present challenges that affect the authenticity and utility of VR environments.

Participant 3 addresses additional limitations related to regulations or safety considerations when using VR tools. She highlights the current absence of specific regulations governing the use of VR technologies in work environments, which complicates the establishment of standardized safety practices. Furthermore, participant 3 notes potential issues with collisions within virtual environments, a problem exacerbated by the loss of a sense of reality that current VR technology often induces. This issue becomes particularly pronounced in multi-user VR environments, where the interaction between participants without accurate spatial awareness can lead to misunderstandings and accidents. These collision problems, if not properly managed or anticipated during the design of VR experiences, can pose significant risks to users and hinder the practical utility of VR for collaborative work settings.

The interviews with experts also noted some limitations within ergonomics and the simulation technologies used to do ergonomic assessments. One of the primary concerns discussed by participant 3 and participant 4 revolves around the fidelity of software-generated ergonomic scores. There is often uncertainty about which ergonomic assessment method should be applied in different scenarios, leading to potential inconsistencies in outcomes.

Further complicating ergonomic evaluations, participant 2 and participant 4 pointed out the issues with robotic and non-natural body movements within VR simulations. These unnatural movements can impact the accuracy of ergonomic assessments, as they may not accurately represent the physical dynamics of human motions in various work-related tasks.

Moreover, the consistency of ergonomic assessments is also a challenge, as highlighted by participant 1 and participant 3. Different ergonomists may produce varying scores for the same assessment scenario due to differing interpretations and applications of the observational ergonomic methods. Participant 3 additionally noted the issue of non-standardized ergonomic certification, which can lead to disparities in assessment results and methods employed.

Another related limitation seen in the interview study is the variation in ergonomic methods used across different companies. This lack of standardization in ergonomic assessments means that results can vary significantly depending on the company or the ergonomic expert conducting the evaluation. This diversity in approaches can make it difficult to compare results across different settings or to establish a unified standard for ergonomic safety in VR environments.

These challenges underscore the need for more standardized procedures and better simulation technologies in VR to enhance the reliability and accuracy of ergonomic assessments, ensuring they are effective in improving workplace safety and design.

3.2 Solutions and Future Outlook for VR in DHM

To end the interview study results, the participants revealed several innovative solutions addressing current limitations in VR ergonomic assessments and also, some insights of the future direction of the VR technologies in ergonomic assessments. Participant 1 discussed the development of multi-user VR systems specifically designed to overcome the inherent inability to collaborate effectively in traditional VR setups. To address the challenge of accurately simulating weight sensations, both participant 2 and participant

3 have explored biomechanical solutions that enhance the physical realism of virtual environments. Additionally, participant 9 introduced the concept of Smart Gloves and hand tracking designed to improve the simulation of hand and finger movements, offering a more intuitive and responsive way to interact with virtual objects, thereby increasing the accuracy of ergonomic assessments.

Looking towards the future, participant 1 anticipates the rise of asymmetric collaboration models in VR, which would allow different users to have varying levels of control and interaction based on their specific roles and tasks. Participant 2 and participant 9 see significant potential in AI-driven motion prediction, although they note that this technology will require extensive training to achieve high levels of accuracy. Participant 2 also emphasizes ongoing efforts to enhance the accuracy and responsiveness of VR technologies, which are critical for effective ergonomic evaluations. Furthermore, participant 5, participant 7 and participant 9 highlight the development of dynamic ergonomic scoring systems within software platforms, enabling real-time, adaptive assessments that can respond to the changing conditions of the work environment.

3.3 Validation Case Study

According to participant 5 and participant 7, the multi-user VR platform offers significant collaborative advantages and is technically similar to other VR systems. In contrast, desktop applications often display robotic behaviours and are time-consuming, primarily due to issues like misinterpreting the scale of CAD models. The difference in familiarity with these platforms was clearly evident between participant 10, who was new to all three platforms, and participant 6, who used them regularly. This discrepancy highlighted how prior experience with these platforms can significantly affect user engagement and effectiveness. All individuals agreed that the multi-user VR platform was especially user-friendly, facilitating easier navigation. The ability of VR to allow users to physically move around enhances the ergonomic assessment process, leading to more informed decision-making in a virtual environment.

The preference for ergonomic assessments distinctly favours the multi-user VR platform in case of participant 5 and participant 7, primarily because of its flexibility to support both single-user and multi-user operations. This adaptability is crucial as it enables users to mimic and physically experience the tasks performed by the manikins. Engaging directly with these simulations provides a more accurate ergonomic assessment, allowing for a thorough understanding of the physical demands and potential strain involved in specific tasks. However, according to participant 10 and participant 6, the choice of platform can vary depending on the scenario. If the assessment involves multiple stakeholders or needs collaborative input, then either single-user or multi-user VR platforms would be more suitable due to their interactive capabilities. On the other hand, if the assessment is to be conducted independently without the need for collaborative input, a desktop application might be enough. This option is generally adequate for solo work and can be effective in environments where simpler assessments are required.

According to participant 5 and participant 7, both VR environments and desktop applications hold their merits in the realm of ergonomic assessments. VR technology is particularly crucial for confirming and fully understanding scenarios that desktop applications can only outline. The immersive nature of VR allows for a deeper, more intuitive

understanding of ergonomic postures and work environments, which is less feasible with desktop applications. Knowing when to use each tool—taking advantage of VR's immersive capabilities for complex assessments and employing desktop applications for more straightforward tasks—is key to effective ergonomic evaluation. However, participant 10 and participant 6 note that while seeing things in 3D and experiencing a more immersive environment can significantly enhance the accuracy and confidence in ergonomic assessments, these elements are not strictly necessary. Desktop applications, although less immersive, still provide a valuable platform for assessing ergonomics. They can be particularly useful for initial evaluations or in situations where VR accessibility is limited. Thus, the choice between VR and desktop applications should be guided by the specific requirements and constraints of the assessment at hand.

Participant 5 and participant 7 emphasize that incorporating multiple stakeholders in a multi-user VR setting significantly enhances the ergonomics assessment and improvement processes for operators. While companies typically do not employ multiple ergonomists in day-to-day operations, the participation of additional experts is crucial for complex tasks. In environments such as automotive assembly, where tasks often require the simultaneous activity of two operators, the inclusion of diverse stakeholders—ergonomists, engineers, operators, and managers—becomes essential. Each brings a unique perspective that contributes to a more comprehensive evaluation of ergonomic needs. This multi-disciplinary approach, facilitated by VR, allows for more effective communication and collaboration, leading to better-informed decisions and more precise ergonomic enhancements. Participant 10 and participant 6 note that while a single ergonomist can assess specific tasks effectively, the collaborative nature of a multi-user VR environment fosters a deeper, more nuanced understanding of ergonomic challenges. This collective examination not only broadens the perspective but also enhances the decision-making process, ensuring that the ergonomic improvements implemented are both accurate and beneficial for all operators involved.

SUS Survey Results This section presents the findings from the SUS [16] conducted to evaluate the user-friendliness and effectiveness of three different platforms for ergonomic assessments. The survey's overall results are shown in Table 2.

Table 2. Results from the SUS survey with the validation case study participants.

Participant no.	Desktop	Single-user VR	Multi-user VR
5	82.5	85	80
6	32.5	45	45
7	80	52.5	52.5
10	72.5	65	60
Average	66.9	61.9	59.4
Std. deviation	23.31	17.49	15.05

The Desktop Application received the highest average usability score of 66.9, suggesting it is generally viewed as the most user-friendly of the three platforms assessed. This preference likely stems from participants' greater familiarity with desktop interfaces, as traditional computing environments have been in widespread use much longer than the comparatively newer VR technologies. However, the data also showed a high standard deviation of 23.31, indicating significant variability in user satisfaction and experiences. This variability is exemplified by the differing ratings from Participant 3 and Participant 2: Participant 3 rated the Desktop Application much higher than both VR platforms, indicating a strong preference for the conventional interface. In contrast, Participant 2 found both Single-user and Multi-user VR more effective than the Desktop Application, possibly reflecting a preference for more interactive and immersive environments. The broad range of experiences among users suggests that personal preferences and the non-immersive nature of the Desktop Application significantly influence usability ratings. Users who are accustomed to the interactive experiences offered by VR might find the Desktop Application less engaging, especially if their tasks benefit from the dynamic features that VR provides. Conversely, those more familiar with traditional computing might favour the well-known layout and controls of a Desktop Application, particularly if they have limited exposure to VR.

Single-user VR scored slightly lower than desktop application on average with 61.9 but exhibited more consistency among users, as indicated by a standard deviation of 17.49. Participant 1 found Single-user VR most beneficial for ergonomic assessments, potentially due to the combined benefits of having one participant in VR and one on the computer, effectively covering all bases without the risks associated with full immersion. Such risks might include collisions or other safety concerns, as VR users cannot see their physical surroundings. This setup might also explain why Multi-user VR did not emerge as their top choice, reflecting a preference for safety and effectiveness in an immersive environment where at least one person remains aware of the physical space.

Receiving the lowest average score of 59.4, Multi-user VR was nonetheless noted for the most consistent user experiences, shown by the smallest standard deviation of 15.05. This consistency suggests that despite a lower overall satisfaction, participants generally agreed on Multi-user VR performance level. The stable and predictable user experience, combined with the immersive nature of VR, makes Multi-user VR potentially more suitable for tasks requiring detailed environmental interaction and user engagement. This platform's ability to facilitate immersive assessments, where multiple users can interact within the same virtual environment, could be particularly advantageous for complex ergonomic evaluations.

Participant 2's equal score of both VR platforms above the Desktop Application provides a particularly insightful perspective. It suggests a strong preference for immersive experiences, which are crucial for effective ergonomic assessments. His perspective supports the notion that VR technologies, despite some usability challenges, provide significant advantages in settings that benefit from a high degree of interaction.

4 Discussion

The interview study provided valuable qualitative insights from a group of academicians and industry professionals in ergonomics and VR technologies, enhancing our understanding and informing our research and experimental design. However, the study's limitations include potential subjective biases inherent in qualitative interviews and the small number of participants, which may limit the extent to which the findings reflect the broader field. Additionally, the fact that all interviewees were from Sweden may have introduced biases related to national policies, corporate cultures, and educational systems, potentially limiting the relevance of the findings to other regions.

The validation case study empirically evaluated the effectiveness of VR technologies in ergonomic assessments through controlled experiments, where ergonomic experts assessed static posture across three tasks using Desktop Applications, Single-User VR, and Multi-User VR platforms. The goal was to compare the advantages and limitations of Multi-User VR versus other platforms to identify the most suitable for specific scenarios. However, the study has limitations, including its controlled environment, which may not fully reflect the complexities of real-world manufacturing settings, and the small sample size of only four experts, which could introduce bias.

The interview study reveals significant barriers hindering the widespread adoption of VR for ergonomic assessments, especially in collaborative contexts. Participants highlighted issues such as inaccuracies in body and hand tracking, limited field of view (FoV), and discomfort during prolonged use, all of which diminish the utility and user experience of VR. A key concern is the lack of standardization and reliability in ergonomic scoring across industries, making consistent assessments challenging. This underscores the need for unified frameworks and more responsive systems to simulate real-time human motion and environmental factors, such as weight and heat, to enhance realism. Promising solutions include advancements like smart gloves for improved hand simulation, hand tracking [17, 18] and better motion prediction [19] using AI, which could provide real-time, adaptive feedback in VR environments. These innovations, along with the development of asymmetric collaboration models, are expected to address many of the current usability and functionality limitations, improving VR's effectiveness for ergonomic applications.

The validation case-study highlights the effectiveness of desktop applications, single-user VR, and multi-user VR for ergonomic assessments, using qualitative and quantitative data from semi-structured interviews and empirical analysis. VR technologies, particularly multi-user VR, are preferred due to their immersive, collaborative features that enhance decision-making, though opinions vary on the necessity of these features for simpler tasks. While desktop applications are less immersive, they remain useful for simpler tasks due to their familiarity and ease of use. As VR technologies evolve, they promise to enhance ergonomic evaluations, enabling remote collaboration and broader accessibility.

To address limitations in future research, expanding the pool of interviewees to include experts from diverse regions, industries, and academic institutions would provide a broader and more representative understanding of VR applications in ergonomics. In validation studies, assigning multiple ergonomists to assess each platform—desktop,

single-user VR, or multi-user VR—would enable clearer comparisons and more reliable data. As technology advances, allowing ergonomists to evaluate entire simulations rather than static postures across all platforms would enhance the optimization of workplace designs. Future simulations of full factory operations could improve ergonomic conditions while monitoring productivity, offering deeper insights into the relationship between ergonomics and efficiency.

5 Conclusion

The possibilities and difficulties of adopting multi-user virtual reality (VR) for ergonomic assessments in manufacturing contexts are thoroughly examined in this research. The results, which were obtained through expert interviews and a validation case study, show that although VR technologies—especially multi-user VR—offer considerable advantages in terms of improved decision-making, collaboration, and immersion in ergonomic assessments, there are still obstacles standing in the way of their wider adoption, including issues with usability, standardization of ergonomic assessments, and tracking accuracy. Smart gloves and AI-powered motion prediction, among other VR innovations, present encouraging ways to address these issues. VR technologies have the potential to significantly improve ergonomic standards as they develop, especially for complex collaborative work. Future research should focus on expanding the scope of validation case studies, standardizing assessment frameworks, and exploring full-scale factory simulations to maximize the benefits of VR in optimizing both worker well-being and productivity in industrial settings.

References

1. De, S., Kar, A.K.: Exploring IOT applications in industry 4.0—insights from review of literature. Stud. Comput. Intell., 15–38 (2023)
2. Grajewski, D., Górski, F., Zawadzki, P., Hamrol, A.: Application of virtual reality techniques in design of ergonomic manufacturing workplaces. Procedia Comput. Sci. **25**, 289–301 (2013)
3. National Research Council (US) and Institute of Medicine (US) Panel on Musculoskeletal Disorders and the Workplace: Musculoskeletal Disorders and the Workplace Low Back and Upper Extremities. National Academy Press, Washington, DC (2001)
4. Musculoskeletal Health [Internet]. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions. Accessed 8 Jul 2024
5. Falck, A., Örtengren, R., Högberg, D.: The impact of poor assembly ergonomics on product quality: a cost–benefit analysis in car manufacturing. Hum. Factors Ergon. Manufac. Serv. Indus. **16**, 24–41 (2009)
6. Berlin, C., Adams, C.: Production Ergonomics: Designing Work Systems to Support Optimal Human Performance. Ubiquity Press, London (2017)
7. Da Silva, A.G., Mendes Gomes, M.V., Winkler, I.: Virtual reality and digital human modeling for ergonomic assessment in industrial product development: a patent and literature review. Appl. Sci. **12**(3), 1084 (2022)
8. Billing, E., Hanson, L., Lamb, M., Högerg, D.: Digital human modelling in action. In: Proceedings of the 15th SweCog Conference (2019)
9. IPS: Industrial path solutions [Internet]. https://industrialpathsolutions.se/. Accessed 14 Jul 2024

10. Ips Imma [Internet]. https://industrialpathsolutions.com/ips-imma/. Accessed 14 Jul 2024
11. Wohlgenannt, I., Simons, A., Stieglitz, S.: Virtual reality. Bus. Inf. Syst. Eng. **45–61**, 455–461 (2020)
12. Liao, T., She, J.: How does virtual reality (VR) facilitate design? A review of VR usage in early-stage engineering design. Proc. Des. Soc. **19**, 2115–2124 (2023)
13. Reinhard, R., Mårdberg, P., Francisco, R., et al.: The use and usage of virtual reality technologies in planning and implementing new workstations. In: Proceedings of the 6th International Digital Human Modeling Symposium, vol. 11, pp. 388–397 (2020)
14. Kallio, H., Pietilä, A., Johnson, M., Kangasniemi, M.: Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J. Adv. Nurs. **12**, 54–65 (2016)
15. Brooke, J.: Sus: a "Quick and Dirty" Usability Scale. In: Usability evaluation in industry. Taylor & Francis, London (1996)
16. Söderlund, H., Zamola, S., Boström, J., Li, D., Mugur, P., Cao, H., Johansson, B.: The creation of a multi-user virtual training environment for operator training in VR. Adv. Transdiscip. Eng. **52**, 173–184 (2024)
17. Meta [Internet]. https://www.meta.com/en-gb/help/quest/articles/headsets-and-accessories/controllers-and-hand-tracking/. Accessed 15 Jan 2025
18. VIVE [Internet]. https://www.vive.com/us/accessory/tracker3/. Accessed 15 Jan 2025
19. Mårdberg, P., Högberg, D., Carlson, J., Söderberg, R.: Towards enhanced functionality and usability of giving manikin task instructions in a DHM tool. Adv. Digit. Hum. Model. **744**, 44–51 (2023)

Exploring Physical Demand Metric: Slouching Score in Augmented Reality-Based Biomechanics Education

Will Mastrantuono[1], Jung Hyup Kim[1(✉)], Siddarth Mohanty[1], Sara Mostowfi[1], Yuanyuan Gu[2], Fang Wang[3], Danielle Oprean[2], Yi Wang[1], and Kangwon Seo[1]

[1] Department of Industrial and Systems Engineering, University of Missouri, Columbia, MO 65211, USA
{wmmktn,kijung,smdqv,sara.mostowfi,yiwang,seoka}@missouri.edu
[2] School of Information Science Learning Technologies, University of Missouri, Columbia, MO 65211, USA
{yggcc,opreand}@missouri.edu
[3] Department of Engineering and Information Technology, University of Missouri, Columbia 65211, USA
wangfan@missouri.edu

Abstract. There is an increasing demand for the development of new metrics that can effectively measure physical demand experienced by users in augmented reality (AR) environments. In this study, we evaluated one of the recent metrics, called "slouching score," in AR-based biomechanics lecture. This study aims to uncover the correlation between the AR interaction and the physical demand of users in a difference setup compared to the earlier study. The slouching score, which evaluates posture changes that may indicate fatigue during AR interactions, is measured using Xsens motion capture equipment. These calculated scores are compared with responses to physical demand assessments surveyed using NASA-TLX questionnaires. One of the key differences between the current study and earlier ones is that participants had to physically move to access the next AR module in earlier studies. In contrast, this time, participants simply needed to click a virtual arrow button to view the next AR modules, eliminating the need for physical movement. Our preliminary findings show correlations between the slouching score from some modules and the NASA-TLX physical demand ratings.

Keywords: Augmented Reality · Motion Capture · Slouching Score

1 Introduction

Augmented Reality (AR) lies on the cutting edge of personal computing technology. Its immersive capabilities allow for a more integrated Human-computer interface than typical personal computers. The potential applications of AR are vast and span across numerous industries, highlighting its versatility and innovative power. However, what has yet to be comprehensively investigated is AR's impact on users during extended sessions,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 301–311, 2025.
https://doi.org/10.1007/978-3-031-93502-2_20

and how the rise in task load affects users' overall experience [1]. Dewi, Khairunnafi [2] discovered that academic stress and learning motivation significantly affect student's academic performance. The amount of workload given to students may not necessarily influence their academic performance, but it can significantly affect their stress levels. Therefore, it is crucial to check how students respond to the increased stress associated with this workload.

Measuring workload can be quite complex. It is multifaceted and can fluctuate depending on various factors. Furthermore, individuals perceive and articulate their workload in diverse ways [3]. The impact of workload pressures varies among individuals; for some, an increased workload may promote better performance, while for others, it could severely diminish their overall experience. In a scoping study performed in Lee, Lin [4], the commercial cleaning industry has emerged as a significant concern regarding ergonomic practices. Workers in this field often engage in repetitive tasks that require extensive use of various muscle groups, leading to an increased risk of injuries. Additionally, the demanding nature of their jobs often leaves little room for adequate rest periods, further worsening the physical strain. As a result, commercial cleaners experience a considerably higher incidence of work-related injuries than their office worker counterparts. This study concludes that firms should select proper methods to evaluate worker over-exposure.

Previous studies have highlighted the necessity for more quantitative methods to effectively assess the physical workload experienced by workers. In light of this, our research had developed a specific metric known as the slouching score [5], which quantifies the physical workload encountered during AR-based learning experiences. In earlier studies related to the slouching score metric, the experiments were designed to prompt users to perform physical movements when exposed to new AR content. However, the earlier findings may differ if users are not required to engage in physical movement while participating in AR-based learning activities. Hence, in this study, we investigated the relationship between slouching score and NASA-TLX physical demand when the user does not have to perform physical movements. We hypothesize that increased physical workloads can be predicted using the slouching scores.

2 Methodology

2.1 Experimental Design

The experiment was conducted on the University of Missouri–Columbia campus during the fall of 2024. A group of 21 junior students from the Industrial and Systems Engineering program participated in this study. Participants were engaged in AR learning activity that aimed to deepen their understanding of the principles of biomechanics, specifically in the context of the airline baggage handling task. The experiment was run over the course of three sessions on consecutive days for each participant. The material spanned the course of three lectures, to be given sequentially each session.

The first lecture includes simple, declarative information. Participants were asked to complete memorization activities to introduce them to the material. The second lecture

delved into more complex calculations, challenging participants to use both their declarative knowledge—what they know about the subject—and their procedural knowledge—how to apply that knowledge in practical scenarios—to tackle biomechanics problems. Lastly, the third lecture differs from lectures 1 and 2, as it asked the participant to apply what they learned in the previous lectures and solve real-world biomechanics case scenarios. In our study, we conducted a thorough analysis of the body movement data collected from participants during lecture 3. Our primary focus was to explore the relationship between the slouching score and the perceived physical workload experienced by the participants.

A Microsoft HoloLens2 headset was used for displaying lecture materials [6, 7]. The AR lectures were crafted with Unity to create an engaging learning atmosphere. The layout resembles a semi-circular blackboard, providing a dynamic and interactive environment. It was divided into five distinct panels, each dedicated to various aspects of the lecture (see Figs. 1, 3, and 4). Using "air-clicking" gestures, participants moved through eight distinct AR modules in a self-paced manner. As we mentioned before, the participants had been exposed to relevant biomechanics materials during the first two lectures. During the lecture 3, students solved practical biomechanical problems and engaged in relevant exercises within the AR environment. A virtual instructor provided more details to clarify the problems and their solutions. In contrast to earlier studies [5, 8] related to the slouching score, the participants in this study had engaged with the AR content without the necessity of physically relocating to a different position. This marks a significant departure from previous studies, where movement to a designated spot was required to begin interacting with the AR elements.

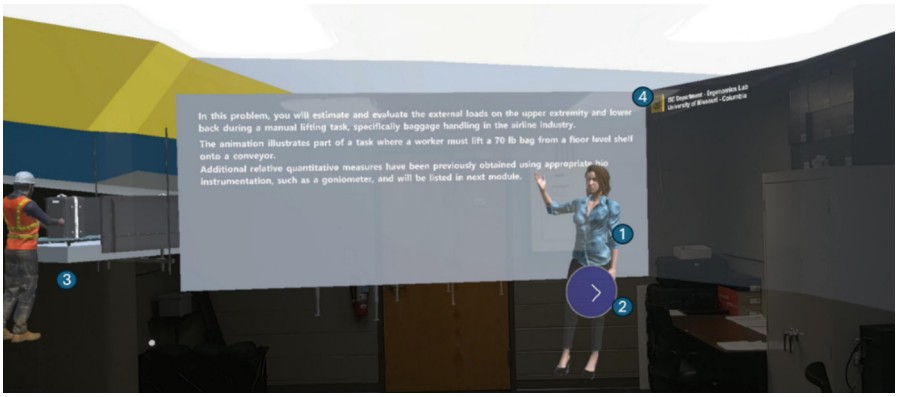


Fig. 1. AR Learning Environment: 1—Virtual Instructor; 2—"Air-Clicking" button; 3—Visualization of the task being analyzed; 4—Virtual blackboards.

During the experiment, participants used Xsens motion capture sensors to track their posture shifts while engaging in AR learning (see Fig. 2). The sensors were used to accurately monitor the postural responses of participants, enabling the precise calculation of their center of mass (COM) during AR learning.

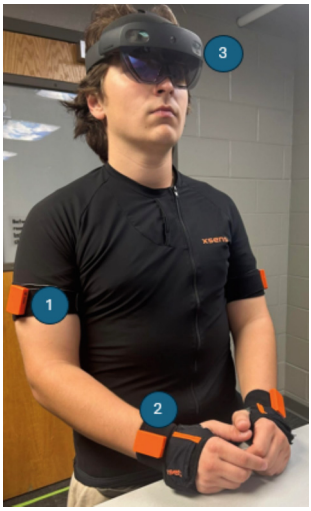


Fig. 2. Hardware utilized during the experiment: 1—Arm Mo-cap sensor; 2—Wrist Mo-cap sensor; 3—Hololens2 headset.

After completing the experiment, participants filled out a series of surveys aimed at gathering a range of insights on their experiences with the AR lectures and the system's usability. One of the surveys included the NASA-TLX [3], which was used to assess the participants' perceived workload during the lectures. This evaluation encompassed six dimensions of subjective workload: Mental, Physical, Temporal, Effort, Performance, and Frustration. We specifically used the physical demand data from the NASA-TLX to compare it with the motion capture data.

2.2 Module Contents and Design

Our study utilized a biomechanical AR lecture composed of seven modules. The scenario presented in this AR lecture focuses on luggage lifting operators subjected to physical stress due to incorrect lifting techniques while managing travelers' baggage. This AR training aims to highlight the ergonomic challenges these workers face and provide them with strategies to mitigate the risks associated with improper lifting. Below is additional information related to the AR content in each module:

- *Module 1:* This module is highly information dense, providing an overall review of the lecture material. It highlights the factors contributing to the rise in worker injuries while handling luggage. By engaging with this content, participants can effectively bridge gaps in their understanding of the lecture.
- *Module 2:* It provides detailed, step-by-step guidance and values during the first problem-solving exercise. It provides detailed descriptions of figures and measurements, giving a thorough overview of the expectations for participants during the exercise. Many details of this module are repeated from the provided figures. Participants will be required to engage with and answer the first question following their viewing of this module.

- *Module 3:* It describes the answer to the first question posed in the previous module, providing numerous figures and relating this issue to a real-world example. Slow-paced, descriptive, and heavily instructional details.
- *Module 4:* This part of the lecture explains the second question, outlining the requirements necessary to tackle it effectively. The module offers minimal detail. We gave ample time for everyone to work through the problem and presents numerous figures to assist in their efforts. (See Fig. 3).
- *Module 5:* Participants need frequent use of the formulas, particularly when verifying the solution for the second question. On the left panel, the free body diagram has been carefully labeled with the relevant forces and moments, and the corresponding values have been calculated as illustrated in Fig. 4.
- *Module 6:* This module addresses question #3. Like earlier sections, it describes the problem along with helpful figures and measurements.
- *Module 7:* This module features the most comprehensive reviews in this AR lecture. It offers detailed descriptions of each step taken in the question #3. Notably, this module is longer than the others.

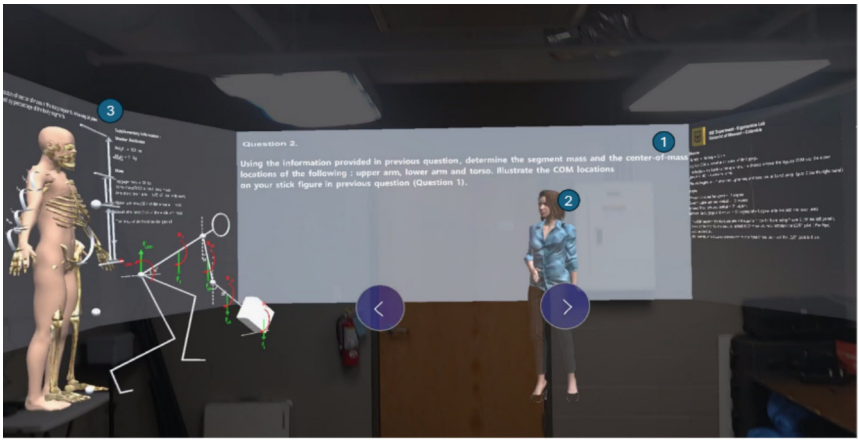


Fig. 3. Module 4 as seen by participants: 1—Description of the question two; 2—Virtual Instructor; 3—Figures for calculations.

3 Data Collection and Slouching Score

The slouching score, which assesses changes in posture that may indicate fatigue during AR interactions, is evaluated using Xsens motion capture technology. Pulipati, Kim [9] developed the slouching score metric. We calculated the slouching score for each AR module based on the equations and procedures from the paper by Pulipati, Kim et al., and compared these scores to the NASA-TLX physical demand scores. Through this analysis, we aim to find the factors that appear as significant predictors of the perceived physical demand experienced by participants during the experiment.

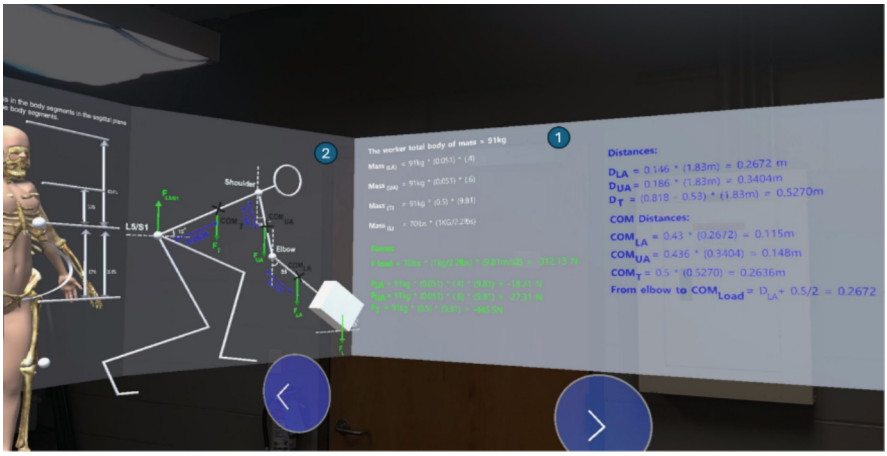


Fig. 4. Module 5 as seen by participants: 1—Formulas and calculations for question two's solution; 2—Labeled figure for clarity and reference.

The slouching score is derived from Xsens motion capture data. Before the start of the experiment, each participant was equipped with a full set of motion capture sensors and calibrated. This calibration process ensured the sensors aligned and functioning optimally for each participant. Once calibrated, the participants' body movements were recorded during the experiment. The slouching score is determined by measuring how much a participant's body posture has deviated from their baseline posture. This baseline is established during the initial 3–5 s of each module, where participants are guided to maintain a neutral position, looking straight ahead with their heads in a natural, upright alignment. By comparing the participant's subsequent posture to this established baseline, we can quantify the degree of change and assess their posture more accurately. According to the study done by Kim, Pulipati [10], the body movement data derived from the center of mass (COM) was identified as the most reliable data compared to other body parts. This stable COM data serves as the baseline for calculating the slouching score. Throughout the analysis, various calculated COM positions will be compared against this baseline to determine the degree of deviation from the baseline. A perfect slouching score, 100, indicates a participant's COM did not deviate from the baseline; a zero represents the most extreme COM deviation physically possible e.g. lying flat on the ground. COM deviation is expressed across all three coordinate planes (x, y, z) (See Fig. 5). Pulipati, Kim [9] study indicates that the most demonstrative of the three was the x-plane, representing a lateral shift in COM.

4 Results

Table 1 illustrates the slouching scores obtained across AR modules, alongside the NASA-TLX PD values. The summary statistics of the slouching score results, represented in Fig. 6, indicate that the mean (M) and standard deviation (SD) of the slouching score for AR modules are module 1—M: 93.1, SD: 7.61; module 2—M: 91.8, SD: 5.61;

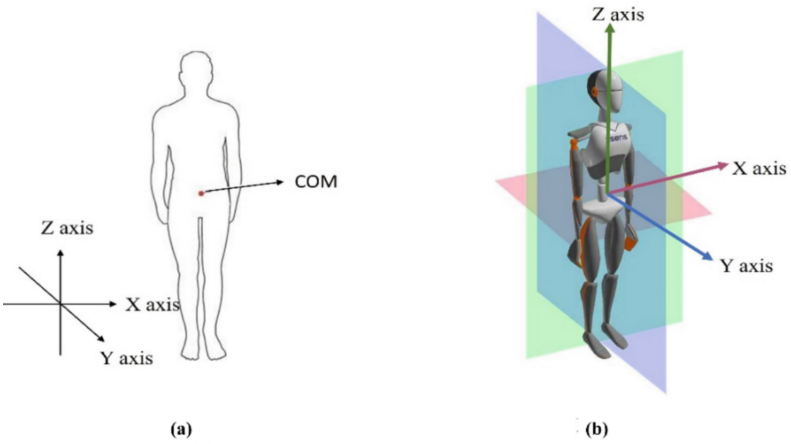


Fig. 5. Anatomical planes of the human body [5].

module 3—M: 91.2, SD: 7.60; module 4—M: 89.4, SD: 7.51; module 5—M: 88.4, SD:11.8; module 6—M: 89.3, SD: 9.44; module 7—M: 89.1, SD: 10.5. The mean (M) and standard deviation (SD) of NASA-TLX PD is M: 41.6, SD: 24.7.

We conducted a thorough correlation analysis between the NASA TLX PD values and the slouching scores recorded for each AR module. The findings revealed that modules 1, 2, 3, and 7 exhibited a correlation greater than 0.5 with the NASA-TLX PD values (refer to Table 2). This substantial correlation indicates a significant relationship between the slouching scores of these specific modules and their respective NASA-TLX PD values, suggesting that as the workload perception increases, the tendency for slouching among participants also rises in these modules.

Upon constructing the linear regression model (see Eq. 1) by using the correlation results from Table 2, we identified the slouching scores (SS) associated with AR modules 1, 2, 3, and 7 as significant factors influencing the outcome. These four modules emerged as strong predictors of perceived physical demand (PD). The model shows the modules 1 and 2 have negative relationship between SS and PD and the modules 3 and 7 shows positive relationship between SS and PD. The model is a statistically significant with p-value = 0.0118 and $R^2 = 0.53$ (See Fig. 7).

$$Est.PD = 340.9 - 4.159(ss_{\text{mod1}}) - 2.201(\text{SS}_{\text{mod2}}) + 1.850(ss_{\text{mod3}}) + 1.863(ss_{\text{mod7}}) \quad (1)$$

5 Discussion

Our findings revealed a notable difference from the study done by Pulipati, Kim [8], which observed a consistent trend among participants—a gradual decline in slouching scores calculated from the center of mass (COM) position along the X-axis across all modules. In contrast, our current study did not find this anticipated downward trajectory of slouching scores (see Fig. 6). This absence of decline suggests no increase in postural

Table 1. Average slouching scores across modules and NASA-TLX PD values.

Slouching Score in each AR module							NASA-TLX PD
1	2	3	4	5	6	7	
88	90	91	87	81	73	72	50
96	97	97	97	96	95	97	50
96	96	96	96	96	97	93	30
81	84	83	83	84	83	83	90
94	94	94	94	96	94	92	30
96	90	90	87	92	78	89	10
95	93	96	93	94	95	93	13
95	91	89	78	65	83	92	30
94	92	92	93	90	92	89	10
95	86	87	89	88	93	83	50
98	96	93	93	98	97	93	20
96	96	96	95	97	94	97	70
65	74	63	67	58	62	52	90
98	96	96	94	97	97	97	20
91	89	88	91	86	93	85	70
99	98	99	97	98	98	97	50
95	91	89	78	65	83	92	30
96	95	97	91	95	94	94	30
99	98	96	95	98	99	98	20
92	89	92	92	92	91	90	70
97	94	91	89	91	86	95	40

deviation among participants. In other words, the level of fatigue and physical demand perceived by the participants did not increase as they were exposed to the AR environment longer. This difference could be attributed to the nature of participants' body movements during the AR sessions. In the previous study [8], participants were required to physically relocate their bodies to proceed to the next AR module. Conversely, in the current experiment, participants remained stationary, simply pressing virtual arrow buttons to navigate between various AR modules. This difference in physical engagement played a vital role in the outcomes observed in our study.

The regression analysis revealed that the slouching scores for modules 1 and 2 exhibited a negative correlation with the estimated PD. This suggests that the estimated PD increased as the slouching scores for these modules decreased. In contrast, the slouching scores for modules 3 and 7 demonstrated a positive correlation with PD, indicating that lower slouching scores in these modules were associated with a lower estimated PD.

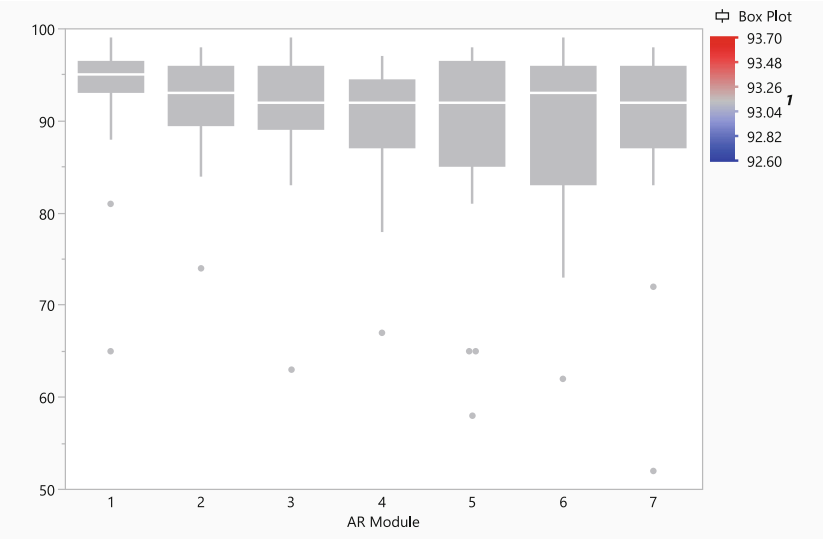


Fig. 6. Box Plot illustrating average slouching scores.

Table 2. Correlations between Slouching Scores and NASA-TLX PD (P-value <0.05*).

AR Module Number	Correlation (P-value)
1	−0.6758 (0.0008*)
2	−0.5902 (0.0049*)
3	−0.5548 (0.0090*)
4	−0.3582 (0.1109)
5	−0.3349 (0.1378)
6	−0.3855 (0.0844)
7	−0.5334 (0.0128*)

This relationship discrepancy could likely be attributed to the differing virtual content participants engaged within AR environments, which may have influenced their posture and movement patterns differently across the various modules.

We need further investigation of the outcomes observed in the regression model. Specifically, we need to explore the reasons behind the negative relationship in AR

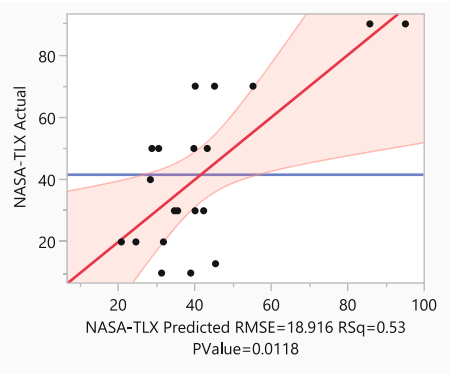


Fig. 7. A Scatter plot comparing actual NASA TLX PD values against predicted PD values.

modules 1 and 2, while modules 3 and 7 exhibit a positive relationship, and the other modules show no discernible relationships. One potential factor influencing these results could be the nature of the learning content presented in each module, the design of the virtual objects, and how the virtual instructor is positioned to convey the material. For instance, AR modules 1 and 2 are characterized by their focus on manual material handling task descriptions and background information related to biomechanics problems. So, participants who indicated experiencing lower levels of physical demand showed not much changes in their posture when compared to their baseline posture. This observed phenomenon may be linked to the fact that individuals with low physical demands were able to concentrate more effectively while engaging with the information presented by the virtual instructor and AR modules.

The learning materials contained in AR modules 3 and 7 focus predominantly on providing participants with the answers to the questions they need to tackle. For example, module 3 emphasizes drawing correct free body diagram of a worker who is doing manual material handling task while illustrating various forces and moments acting upon that diagram. The module 7 shows the mathematical procedures necessary for calculating the forces and moments affecting key joints, specifically the elbow, shoulder, and L5/S1. It means the participants who reported a low level of physical demand showed a greater postural change than those who reported high physical demand. This phenomenon may be attributed to the fact that individuals in the low-demand group were more actively engaged in closely following the instructions provided by the virtual instructor. Furthermore, they tended to compare their answers and move more thoroughly with the presented solutions, which likely contributed to alterations in posture.

6 Conclusion

The purpose of this study is to establish a quantitative measure of exhaustion and use this to predict the physical workload experienced by users in AR through the slouching score. The investigation into the relationship between slouching scores and the NASA-TLX PD revealed several significant connections. This study provides a unique inspection

of the impact of prolonged AR sessions on users. More specifically, this is the second study that shows the existence of a relationship between lateral deviations of COM and incurred physical demand. These findings are incredibly encouraging to use AR in their practices.

For the limitation of the current study, it is important to note that all participants belong to the same age group. Expanding the participant pool to include a variety of age groups would provide a more comprehensive understanding of how slouching scores are related to physical demands across different demographics. For future research, it would be beneficial to explore additional domains and a wider range of lifting tasks to deepen our understanding of the slouching score metric. This metric has the potential to serve as an effective indicator of physical workload, and examining various contexts could provide insights into how we can enhance its accuracy in reflecting a worker's fatigue level.

Acknowledgments. This study was funded by the National Science Foundation (IIS-2202108).

References

1. Guo, W., Kim, J.H.: How augmented reality influences student workload in engineering education. In: HCI International 2020–Late Breaking Papers: Cognition, Learning and Games: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22. Springer (2020)
2. Dewi, D.S., et al.: The effect of mental workload, stress, and learning motivation on student learning achievement during online courses. In: Proceedings of the International Conference on Industrial Engineering and Operations Management (2021)
3. Casner, S.M., Gore, B.F.: Measuring and evaluating workload: a primer. NASA Tech. Memo. **2010**, 216395 (2010)
4. Lee, W., et al.: Methods for measuring physical workload among commercial cleaners: a scoping review. Int. J. Ind. Ergon. **90**, 103319 (2022)
5. Pulipati, V., et al.: Utilizing motion capture to quantify physical workload in augmented reality learning environments. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, pp. 1181–1187. SAGE, Los Angeles, CA (2024)
6. Vidal-Balea, A., et al.: Analysis, design and practical validation of an augmented reality teaching system based on microsoft hololens 2 and edge computing. Eng. Proc. **2**(1), 52 (2020)
7. Yu, C.-Y., et al.: Developing an augmented reality-based interactive learning system with real-time location and motion tracking. In: International Conference on Human-Computer Interaction. Springer, Berlin (2023)
8. Pulipati, V., et al.: Measuring fatigue dynamics of augmented reality in the digital learning era using motion capture data. In: International Conference on Human-Computer Interaction. Springer, Berlin (2024)
9. Pulipati, V., et al.: Measuring physical demand in augmented reality learning environments. Ergonomics, 1–13 (2024)
10. Kim, J.H., et al.: Understanding the influence of fatigue on full arm gestures in augmented reality environments. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, pp. 1194–1199. SAGE, Los Angeles, CA (2024)



Assessment of Ergonomic Posture in Augmented Reality Environments Using Slouching Score

Siddarth Mohanty[1], Jung Hyup Kim[1(✉)], Sara Mostowfi[1], Yuanyuan Gu[2], Fang Wang[3], Danielle Oprean[2], and Kangwon Seo[1]

[1] Department of Industrial and Systems Engineering, University of Missouri, Columbia, MO 65211, USA
{smdqv,kijung,sara.mostowfi,seoka}@missouri.edu
[2] School of Information Science Learning Technologies, University of Missouri, Columbia, MO 65211, USA
{yggcc,opreand}@missouri.edu
[3] Department of Engineering and Information Technology, University of Missouri, Columbia, MO 65211, USA
wangfan@missouri.edu

Abstract. This study explores the application of slouching scores to assess ergonomic posture in augmented reality (AR) environments. Employing Microsoft HoloLens 2 with Xsens motion capture technology, participants engaged in interactive biomechanics tasks, including a practical luggage-lifting exercise. Real-time feedback guided users towards safe posture, emphasizing spinal alignment and reducing physical strain. Slouching scores functioned as quantitative measures of posture quality, establishing a connection between unsafe postures and the requisite postural adjustments. The results illustrate how AR-integrated systems can enhance posture awareness, improve user ergonomics, and promote active learning in both educational and professional settings.

Keywords: augmented reality · slouching scores · ergonomics · motion capture · biomechanics · posture assessment

1 Introduction

The rapid integration of augmented reality (AR) technologies into human factors and ergonomics research has opened new pathways for assessing and mitigating physical strain during task execution. By merging AR systems with motion capture technologies, researchers can deliver real-time feedback on user posture and biomechanics, thus advancing our understanding of ergonomics in dynamic environments. For example, devices such as Microsoft HoloLens enable immersive task simulations and offer immediate visual feedback, which can significantly improve posture correction and task performance. Studies have consistently highlighted the potential of these systems to enhance users' ability to visualize complex concepts and interact with virtual scenarios in real-time, thereby bridging theoretical knowledge with practical skill development [1–3]. This capability is particularly beneficial in manufacturing, logistics, and healthcare, where risk reduction, safety, and hands-on training are critical (Fig. 1).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 312–322, 2025.
https://doi.org/10.1007/978-3-031-93502-2_21

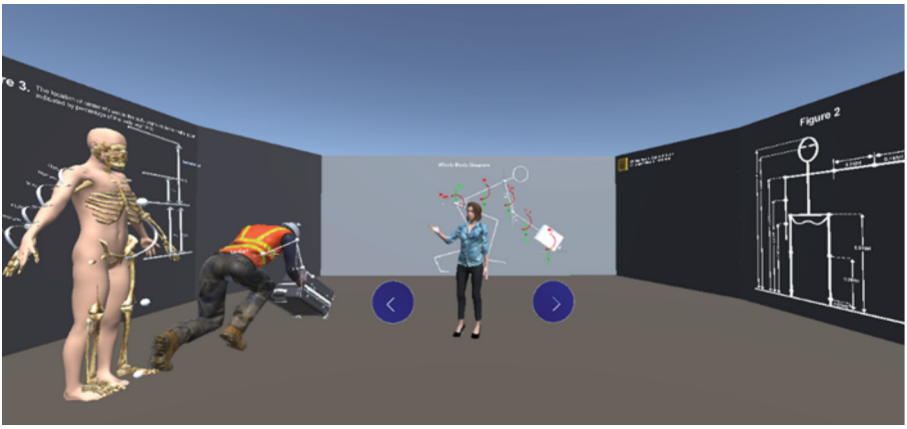


Fig. 1. AR scene in Microsoft HoloLens 2.

One of the most promising uses of AR in ergonomics is the simulation of real-world conditions, such as repetitive lifting tasks, to evaluate physical strain and reduce musculoskeletal risks [4]. These simulations offer an opportunity to collect quantitative data on users' posture and bodily movements, making it possible to pinpoint elements of a task that could lead to injury if not addressed. A key aspect of this quantitative approach involves the use of slouching scores, which are derived from motion capture data and measure the degree of spinal alignment or deviation from optimal postures [5]. Such metrics provide a systematic and objective way to assess physical strain, enabling researchers and practitioners to identify unsafe movements before they result in chronic musculoskeletal disorders. Many of these disorders stem from strain on the lower back, where poor posture can place uneven stress on intervertebral discs, particularly in the L5/S1 region, increasing the likelihood of low back pain (LBP). By actively promoting spinal neutrality, bending at the hips and knees, and keeping loads close to the body, individuals can mitigate these risks and maintain better long-term health outcomes.

Real-time feedback systems in AR environments play a crucial role in reinforcing proper posture. In our study, motion sensors integrated within AR setups allow for the continuous capture of slouching scores, alerting users when deviations from neutral alignment occur. This immediate feedback loop empowers individuals to make dynamic posture adjustments rather than relying on post-hoc assessments that may come too late to prevent cumulative strain. Studies have shown that implementing such feedback mechanisms leads to a significant reduction in the risk of injuries associated with unsafe postures, thereby reinforcing ergonomic principles in both training and everyday work routines [6, 7]. Industries characterized by repetitive or high-risk tasks, such as construction or manufacturing, can particularly benefit from this technology, as it promotes consistent adherence to best practices.

A major catalyst behind these improvements is the granular data collected by modern motion capture systems, such as Xsens, which track detailed movement patterns, joint angles, and overall body movements [8]. By combining this precision with AR platforms, researchers can develop more accurate posture assessments and implement

highly effective interventions. For instance, analysing joint angles and spinal alignment in real time reveals how factors like task complexity, duration, and load weight influence posture. This knowledge feeds directly back into AR training simulations, making them more adaptive and responsive to the user's current performance. Several studies confirm that coupling motion capture data with AR-based training significantly enhances the accuracy of ergonomic assessments, ensuring that posture corrections are based on real, quantifiable deviations from optimal movement patterns [9, 10]. While the technological capabilities of AR systems are rapidly advancing, their success in ergonomics also hinges on user acceptance and comfort. Head-mounted displays (HMDs), which are commonly used to deliver AR content, must be lightweight, adjustable, and ergonomic in design to minimize discomfort, eye strain, and fatigue during prolonged use [11]. If these issues are not addressed, users may become resistant to the technology, negating the potential benefits of real-time ergonomic guidance. Additionally, intuitive interfaces that simplify navigation and interaction are essential. Overly complex systems can deter users, while gamified or visually engaging interfaces tend to enhance motivation, encourage consistent use, and make posture correction more enjoyable. These user-friendly features are crucial for ensuring the long-term adoption of AR-based ergonomics solutions across diverse work environments.

By providing immediate and context-specific feedback, AR supports iterative learning and continuous improvement of posture during task execution. This approach can be especially valuable in training contexts where users need to develop "muscle memory" for safe lifting techniques. Over time, consistent exposure to visual cues and corrective suggestions helps embed proper ergonomic practices into routine behaviour, reducing the likelihood of injury. Researchers have found that as learners become more adept at recognizing and correcting unsafe postures, they also exhibit better overall task performance and lower rates of fatigue [12]. This convergence of performance enhancement and injury prevention presents a powerful argument for integrating AR into workplace safety programs.

2 Methodology

2.1 Experimental Design

Participants participated 2 days of biomechanics lectures in an augmented reality (AR) environment to gain a foundational understanding of posture and movement dynamics [13, 14]. After completing the training, each participant was equipped with MovellaTM motion capture sensors, which were calibrated to their individual body dimensions for precise data collection. Later, performed the luggage lifting task as shown in Fig. 2.

In this study, the focus was primarily on the sensor placed at the L5/S1 vertebral junction, as it represents the center of mass (COM) of the human body, offering critical insights into posture and spinal alignment. The L5/S1 vertebral junction is a critical anatomical landmark in the human spine, often serving as a focal point for studies involving posture and spinal alignment.

This region, where the lumbar spine meets the sacrum, plays a pivotal role in maintaining the body's COM and overall balance [15]. Research has shown that the lumbosacral joint (L5-S1) is integral to the spine's structure, facilitating movements such



Fig. 2. Participant lifting luggage.

as twisting, bending, and standing upright. The accuracy of body motion data collection heavily relies on a thorough calibration process for motion sensors. This calibration is essential in ensuring that the sensors can effectively capture the movement, allowing for precise readings. The calibration process will be explained in subsequent sections of the paper.

In this study, 25 undergraduate students majoring in Industrial Engineering participated in assessing ergonomic posture and slouching scores within augmented reality environments. The experiment began with an AR-based lecture designed for the ergonomics class at the University of Missouri-Columbia. After completing the biomechanics lectures in the AR environment, participants took part in a hands-on exercise focused on manual material handling tasks. They interacted with engaging modules and worked through related problems to deepen their understanding. Following this, the practical aspect of the experiment involved a luggage-lifting activity. A virtual instructor, accessible through Microsoft HoloLens 2, demonstrated the correct posture for lifting and placing luggage in a designated area shown in Fig. 3. Participants mimicked the posture and performed the task while awaiting evaluation, whether the posture they mimicked was safe or unsafe. A posture recognition system was developed using motion capture sensors in the lab to analyze participants' postures and classify them as either "safe" or "unsafe." If an unsafe posture was detected, participants repeated the activity until achieving a safe posture. Xsens motion capture sensors recorded detailed body movement data, and slouching scores were calculated as quantitative indicators of posture quality, emphasizing ergonomic improvements and fostering active learning throughout the process.

2.2 Data Collection

Data collection for this study involved using MVN motion capture sensors to record participants' movements during the luggage-lifting activity. The motion capture data was exported in Excel format, providing detailed numerical records of body movement,



(a) (b)

Fig. 3. (a) A virtual instructor is demonstrating how to pick up luggage. (b) A virtual instructor is guiding participants to place their luggage in the designated area.

including posture alignment and slouching scores. Additionally, video recordings of the activity were captured to complement the motion capture data. These videos were crucial for correlating specific movements, such as lower back bending, with the numerical data by matching time frames. This dual approach ensured the accuracy and reliability of the posture analysis. The combined use of motion capture data and video allowed for precise evaluation of participant posture, enabling the identification of unsafe postures and facilitating iterative corrections.

2.3 Data Preprocessing

After data collection, the MVN motion capture file provided a COM dataset with positional deviations from the origin in X, Y, and Z coordinates. For this study, the focus was on the X coordinate (COM pos X). It has been used to calculate the slouching score in the previous study [16, 17]. Each luggage-lifting task consisted of a complete action cycle—from standing idle, lifting the luggage, and returning to the initial standing position before moving forward to place the luggage in the designated area. This entire process took approximately 5 s for each participant, corresponding to 300 frames (60 fps). Using the recorded MVN video and experiment data logs, the data was segmented into 5-s intervals for each participant to isolate the lifting action. Post-processing involved categorizing the data into" safe" and" unsafe" postures. This classification was based on observations of the participants' movements and corresponding slouching scores.

3 Data Analysis

After gathering all the data as discussed previously, the slouching score for each participant was calculated using Eq. 1 [12]. A slouching score is a quantitative measure used to assess an individual's posture, particularly focusing on spinal alignment during activities such as lifting. It is calculated by comparing the position of the center of mass (COM) during a specific movement phase to a predefined global baseline, typically established just before the movement begins. The time period of the global baseline was defined as

the 2 s right before the participant bent to lift the luggage. The remaining 3 s represented the lifting phase, where changes in the slouching score were observed. This enabled us to analyze the fluctuations in the slouching score, distinguishing between safe postural habits and those deemed unsafe.

$$Slouching\ score = 100 \times \left\{ 1 - \left(\frac{abs\ COM_x - Global\ Baseline}{\text{Max } Deviation} \right) \right\} \quad (1)$$

For each safe and unsafe posture of the 25 participants, slouching scores were calculated. These scores were compiled and analyzed using pooled t-tests to statistically evaluate differences between safe and unsafe postures. This analysis provided critical insights into the relationship between posture quality and slouching scores, helping to quantify the ergonomic strain on participants during the luggage-lifting activity.

4 Results

The slouching score comparison, calculated using Eq. (1), provided a quantitative measure of the difference in posture between safe and unsafe positions among the participants. The score ranged from 0 to 100, with 100 indicating no difference in posture compared to the baseline, while a score of 0 represented the maximum difference from the baseline posture. These scores were calculated for each safe and unsafe posture across 25 participants, segmented into distinct 5-s intervals based on data preprocessing. The analysis aimed to determine significant differences between safe and unsafe postures shown in Fig. 4.

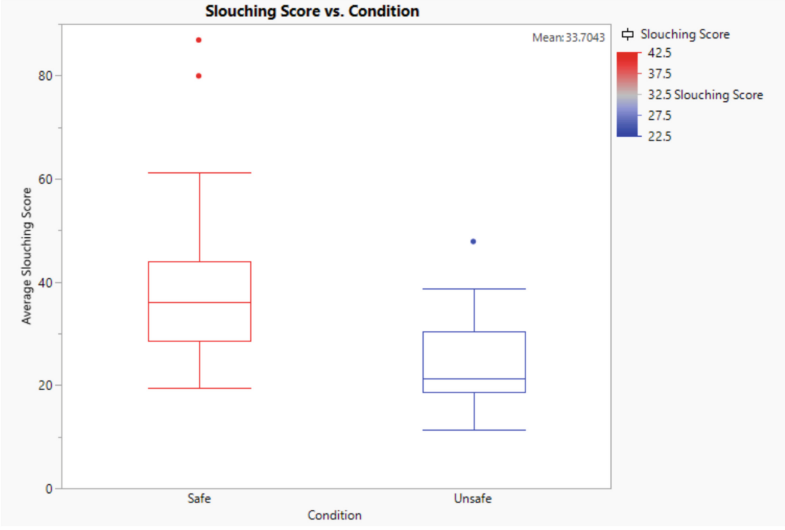


Fig. 4. Comparison of slouching score between safe and unsafe posture.

The mean slouching score for safe postures was 41.05 (Std. Error = 3.41), while the mean score for unsafe postures was considerably lower at 24.39 (Std. Error = 3.83). The standard errors and 95 percent confidence intervals (CI) further highlighted this difference, with safe postures demonstrating a CI range of 34.11 to 47.99 and unsafe postures ranging from 16.59 to 32.20. Refer to Table 1 for descriptive statistics.

Table 1. Descriptive Analysis of Slouching Scores.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Average Slouching Score Safe	19	41.0520	3.4059	34.114	47.990
Average Slouching Score Unsafe	15	24.3971	3.8332	16.589	32.205

A pooled t-test was performed to confirm the statistical significance of the difference of average slouching scores between safe and unsafe postures. The test revealed a mean difference of -16.65 ($t = -3.24798$, $p < 0.05$) refer Table 2, confirming that unsafe postures had significantly lower slouching scores compared to safe postures.

Table 2. Pooled t-Test Results.

Metric	Value
Difference	−16.655
t Ratio	−3.24798
Std Error Difference	5.128
Degrees of Freedom (DF)	32
Upper Confidence Limit Difference	−6.210
Lower Confidence Limit Difference	−27.100
p-value (Prob > |t|)	0.0027
Confidence Level	95%

The confidence interval for the mean difference ranged from -27.10 to -6.21, indicating a consistent and reliable difference across participants. This reinforces the hypothesis that slouching scores effectively differentiate between safe and unsafe postures. To gain a deeper understanding of the connection between slouching scores and various posture conditions, it is essential to conduct a more detailed analysis. Our study only examined two posture conditions—safe and unsafe—and their impact on slouching scores. Since 'posture condition' is a categorical variable with two levels (safe and unsafe), these conditions were compared directly to assess their influence on slouching scores.

The analysis revealed that slouching scores were consistently higher for safe postures compared to unsafe postures, with an average difference of 16.65. This difference

underscores the positive impact of adopting a safe posture, as higher slouching scores indicate better spinal alignment and reduced strain on the L5/S1 region. Approximately 22 percent of the variance in slouching scores could be attributed to differences between the posture conditions, as indicated by the statistical analysis. These findings highlight the effectiveness of slouching scores as a metric for evaluating posture. The statistically significant differences observed in t-tests validate the utility of slouching scores in distinguishing between safe and unsafe postures.

5 Discussion

This study aimed to evaluate ergonomic posture in AR environments using slouching scores as a quantitative metric. The findings, based on the results of statistical analyses, highlight several key insights regarding the effectiveness of slouching scores in assessing posture quality and informing ergonomic interventions.

5.1 Effectiveness of Slouching Scores

The slouching score has proven to be a reliable metric for assessing posture quality during heavy lifting tasks. Significant differences in slouching scores were observed between safe and unsafe postures, indicating its effectiveness in distinguishing between these conditions. Safe postures consistently yielded higher slouching scores, Indicating better spinal alignment and reduced strain on the L5/S1 region. Conversely, unsafe postures were associated with significantly lower slouching scores, reflecting increased spinal strain and deviations from ergonomic alignment [18].

This confirms the utility of slouching scores as a metric for identifying and quantifying ergonomic risks during lifting tasks. The distinction between safe and unsafe postures, as quantified by slouching scores, underscores the role of posture recognition systems in real-time assessment. The use of MVN motion capture sensors allowed precise measurements of the COM deviations in the X coordinate, providing an accurate representation of forward or backward bending tendencies. These measurements enabled a robust analysis of posture dynamics, highlighting the impact of unsafe postures on spinal health.

Figure 5 illustrates two postures: a safe posture (left) and an unsafe posture (right) during a luggage-lifting task, emphasizing the role of the COM and its impact on spinal strain, particularly at the L5/S1 vertebral junction. In the safe posture, the COM, represented by the green box, remains aligned with the base of support, ensuring minimal strain on the lower back [12]. The knees are bent, and the hips act as a pivot point, with the luggage kept close to the body, reducing the bending moment and evenly distributing force across the musculoskeletal system. Conversely, in the unsafe posture, the COM (red box) shifts forward, extending beyond the base of support due to excessive bending and straighter knees, which significantly increases the torque on the lower back. This misalignment places higher compressive and shear forces on the L5/S1 region, elevating the risk of injuries such as herniated discs and chronic LBP [11]. The unsafe posture forces the lower spine to bear the majority of the load, over activating spinal extensors and compromising stability.

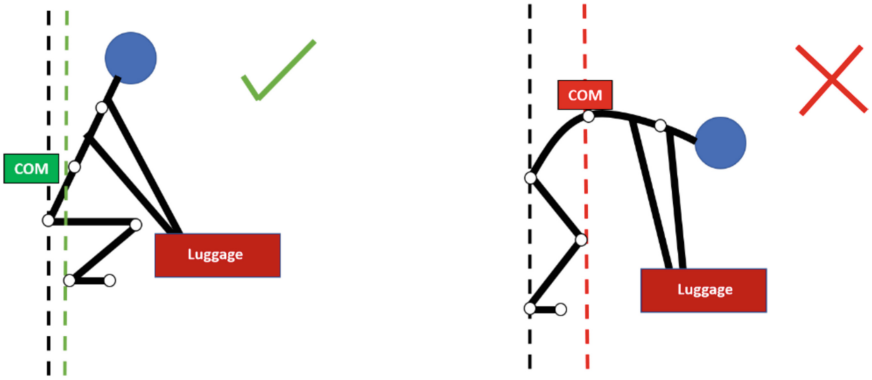


Fig. 5. Safe vs unsafe lifting posture with COM (center of mass).

These findings underscore the significance of training programs and AR feedback systems that offer real-time guidance on maintaining a neutral spine and keeping the center of mass within the base of support. This approach encourages ergonomic lifting techniques and helps to reduce the risk of workplace injuries.

5.2 Implications for Ergonomics in AR Environments

The integration of AR-based training and posture recognition systems offers a promising approach to improving workplace ergonomics. The use of Microsoft HoloLens 2 allowed participants to receive real-time feedback on their posture, guiding them to correct unsafe movements and adopt safer practices. This iterative feedback loop, coupled with the slouching score as a performance metric, facilitated active learning and reinforced ergonomic principles. One of the practical implications of this study is the potential for deploying AR-integrated posture recognition systems in industrial environments, where repetitive lifting tasks are common.

The ability to monitor and correct postures dynamically can reduce the risk of musculoskeletal disorders and improve overall task efficiency. Furthermore, the focus on the L5/S1 region, a critical area for spinal stability, ensures that interventions are targeted at reducing the most significant sources of strain.

5.3 Limitations and Future Directions

While the study provided valuable insights, several limitations should be addressed. First, the sample size of 25 participants, though sufficient for initial analysis, may limit the generalizability of the findings. Future studies should consider larger and more diverse participant groups to validate the results across different populations and workplace scenarios.

Additionally, the slouching score focused solely on the X coordinate of the COM. While this dimension is critical, other dimensions, such as the Y and Z coordinates, may provide complementary information about body movements. Incorporating these additional dimensions could enhance the comprehensiveness of the analysis.

Future research should also explore the long-term impact of AR-based ergonomic training on behavioral change. While this study demonstrated the immediate effectiveness of slouching scores in distinguishing between postures, longitudinal studies could examine whether participants retain these safer postures over time and apply them in real-world scenarios.

6 Conclusion

Overall, this study demonstrates the utility of slouching scores in assessing posture quality and supporting ergonomic interventions in AR environments. The findings highlight the potential of AR-based training and posture recognition systems to improve workplace safety and reduce musculoskeletal risks. By addressing the limitations and exploring new dimensions of analysis, future research can further enhance the applicability of these technologies, making them indispensable tools for modern ergonomics.

Acknowledgments. This study was funded by the National Science Foundation (IIS-2202108).

References

1. Dhaas, A.: Augmented reality in education: a review of learning outcomes and pedagogical implications. Am. J. Comput. Eng. **7**(3), 1–18 (2024)
2. Naveen, L., et al.: The influence of mobile augmented reality on consumer behavior: insights into affective, cognitive, and behavioral responses. Comput. Hum. Behav. **165**, 108558 (2025)
3. Dubreucq, E., et al.: Impact of virtual, augmented or mixed reality in basic life support training: a scoping review. Clin. Simul. Nurs. **99**, 101672 (2025)
4. Rolving, N., et al.: Effect of involving physiotherapists in the management of low back pain at emergency departments: a systematic review. Physiotherapy. **127**, 101454 (2024)
5. Du, S.-H., et al.: Spinal posture assessment and low back pain. EFORT Open Rev. **8**(9), 708–718 (2023)
6. Kim, W., et al.: A directional vibrotactile feedback interface for ergonomic postural adjustment. IEEE Trans. Haptics. **15**(1), 200–211 (2021)
7. Merino, G., et al.: Ergonomic evaluation of the musculoskeletal risks in a banana harvesting activity through qualitative and quantitative measures, with emphasis on motion capture (Xsens) and EMG. Int. J. Ind. Ergon. **69**, 80–89 (2019)
8. Wang, Y., Reiterer H.: The impact of augmented-reality head-mounted displays on users' movement behavior: an exploratory study. arXiv preprint arXiv:1905.10315 (2019)
9. Ghasemi, Y., et al.: Effects of head-locked augmented reality on user's performance and perceived workload. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 65, pp. 1094–1098. SAGE, Los Angeles, CA (2021)
10. Schmidt, M., et al.: Learner and User Experience Research: An Introduction for the Field of Learning Design & Technology. EdTech Books (2020)
11. Lee, H.R., et al.: Exploring the impact of mild-to-moderate foraminal stenosis at L5− S1 on clinical outcomes following L4–5 posterior lumbar interbody fusion. Spine J. **24**(5), 820–830 (2024)
12. Pulipati, V., et al.: Measuring fatigue dynamics of augmented reality in the digital learning era using motion capture data. In: International Conference on Human-Computer Interaction. Springer, Berlin (2024)

13. Yu, C.-Y., et al.: Developing an augmented reality-based interactive learning system with real-time location and motion tracking. In: International Conference on Human-Computer Interaction. Springer, Berlin (2023)
14. Kim, J.H., et al.: The effect of virtual instructor and metacognition on workload in a location-based augmented reality learning environment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 67, pp. 1550–1555. SAGE, Los Angeles, CA (2023)
15. Marras, W.S., et al.: Spine loading in patients with low back pain during asymmetric lifting exertions. Spine J. **4**(1), 64–75 (2004)
16. Pulipati, V., et al.: Measuring physical demand in augmented reality learning environments. Ergonomics, 1–13 (2024)
17. Pulipati, V., et al.: Utilizing motion capture to quantify physical workload in augmented reality learning environments. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, pp. 1181–1187. SAGE, Los Angeles, CA (2024)
18. Zhou, X., et al.: Construction activity analysis of workers based on human posture estimation information. Engineering. **33**, 225–236 (2024)

Biomechanical Evaluation of a Hybrid Upper-Body Exoskeleton During Lifting and Overhead Tasks in Occupational Settings: A Preliminary Study

Alessandro Piol[1(✉)], Emilia Scalona[2], Martina Mosso[1], Fatehia Bushara Garma Bushara[1], Maria Lucia Cavallo[3], Giacomo Valli[3], Noemi Pintori[3], Tiwana Varrecchia[4], Giorgia Chini[4], Alberto Ranavolo[4], Luca Falciati[3], Debora Brignani[3], Francesco Negro[3], and Nicola Francesco Lopomo[5]

[1] Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Brescia, Italy
alessandro.piol@unibs.it
[2] Centro Protesi INAIL, Vigorso di Budrio, Italy
[3] Dipartimento di Scienze Cliniche e Sperimentali, Università degli Studi di Brescia, Brescia, Italy
[4] Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, INAIL, Rome, Italy
[5] Dipartimento di Design, Politecnico di Milano, Milan, Italy

Abstract. Recently, the concept of "Industry 5.0" has been emerging, highlighting the importance of prioritizing the well-being of workers within the overall industrial environments. This shift in focus moves from traditional safety concerns to a more holistic approach that emphasizes operator's health and wellbeing. In this context, exoskeletons have gained significant attention as a potential solution to reduce the risks of developing work-related musculoskeletal disorders (WMSDs), particularly in tasks involving demanding and repetitive movements. This study explores the effectiveness of a hybrid upper-body exoskeleton in mitigating biomechanical strain during both dynamic and static tasks. An experimental protocol was developed, involving two main occupational task categories: a load-lifting task and a static overhead task, each one performed with and without the exoskeleton. The obtained results suggested that while the exoskeleton provided benefits in overall reducing muscular activation and joint range of motion, it may also increase biomechanical strain and fatigue for specific muscles and joint in certain conditions, particularly during asymmetrical tasks and while lifting high loads. These findings emphasize the importance of ensuring that the exoskeleton could dynamically adapt to individual user and task requirements for optimal performance. Additionally, the study calls for future research that incorporates cognitive and ergonomic considerations to better understand the long-term effects of exoskeleton use on the overall worker performance including motor coordination and attention.

Keywords: Hybrid Exoskeleton · Biomechanics · Muscular Activation · Kinematics

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 323–341, 2025.
https://doi.org/10.1007/978-3-031-93502-2_22

1 Introduction

In the field of occupational safety and health, a critical aspect is the implementation of effective strategies able to reduce the risk of injuries and the development of work-related musculoskeletal disorders (WMSDs). With the advent of Industry 4.0, and even more so in the emerging Industry 5.0, where automation and digitalization deeply integrate with industrial workflows, particular preventive measures have increasingly focused on exoskeletons as one of the prominent technological solutions [1–4]. In fact, whether active (i.e., including powered actuators) or passive (i.e., including mechanical springs, etc.), exoskeletons have been extensively studied as a mean to mitigate the biomechanical strain and, thus, reduce the incidence of WMSDs in occupational settings [5]. Their application has gained particular interest in sectors where workers are frequently subjected to biomechanical overload, such as those requiring repetitive manual handling of loads [6].

Although exoskeletons are designed to alleviate the strain on the musculoskeletal system, current scientific evidence regarding their overall biomechanical benefits remains inconclusive, mainly due to the fact that more rigorous and standardized research is needed to fully understand and optimize their biomechanical benefits [7]. In specific occupational scenarios, especially those involving asymmetric load handling, these devices may unintentionally increase the risks associated with WMSDs. This could be partially attributed to their limited ability to dynamically adapt to the user's needs, specific task requirements, and environmental constraints. Consequently, rather than reducing strain, exoskeletons may contribute to increased biomechanical load [8], heightened muscle activation [9], or elevated metabolic demand [10].

Expanding the research in this field requires a more comprehensive evaluation of exoskeleton effectiveness in real-world conditions. In particular, understanding how different exoskeleton designs influence movement strategies, muscular fatigue development, and long-term musculoskeletal adaptations remains crucial areas of investigation [11]. Additionally, the variability in users anthropometric characteristics and motor control strategies highlights the need for personalized solutions that can accommodate individual differences. Specific research should also take into account ergonomic and cognitive factors to assess the impact of exoskeleton use on concentration both in the short and long term [12]. Prolonged use of assistive devices could potentially lead to compensatory behaviors or altered motor coordination patterns. Addressing these challenges is essential for optimizing the design and implementation of exoskeletons in industrial applications, ensuring that their benefits outweigh potential drawbacks.

Considering these aspects, this study aims to investigate in a preliminary setting the effectiveness of a hybrid upper-body exoskeleton in terms of biomechanical support; to achieve this goal, a dedicated experimental protocol was designed, structured into two distinct sections to differentiate between possible use cases in specific occupational scenarios, both involving dynamic movements and static postural maintenance.

2 Materials and Methods

2.1 Participants

The preliminary analysis involved 3 healthy volunteers. None of them reported previous history of neurological disorders or recent orthopedic injuries. All the subjects were right-handed, according to the Edinburgh Handedness Inventory [13]. All the subjects were instructed about the experimental sessions and gave explicit consent to participate to the preliminary pilot study.

2.2 Experimental Setup

To assess the impact of an upper body hybrid exoskeleton on biomechanical domain, from a kinetic and kinematic point of view, two different motor tasks were implemented. Based on the specific features of the exoskeleton under analysis, two distinct motor tasks were specifically selected for implementation:

- Lifting Task (*Lft*): This task involved lifting and relocating a box from an initial position to a final position at a higher elevation, incorporating an in-place body rotation.
- Force Maintenance Task (*Mnt*): This task required the subject to sustain an overhead force in a static position.

In both conditions, the subject initially performed the tasks without any assistance and subsequently repeated them while wearing a hybrid upper-body exoskeleton. The identified exoskeleton was the LBE 30 model from Wearable Robotics; it can be defined as a "hybrid" solution, since the force are generated by passive components through the modulation performed by using an electric motor, that can be driven by exploiting a mobile application. Forces are provided at the wrist level, and their delivery is triggered by simultaneously pressing two flexible switches that are placed on the thumb tips.

2.3 Lifting Task Setup

The lifting task was designed and implemented to simulate a standard manual load-handling activity. It consists of three distinct phases, as illustrated in Fig. 1:

- Approach and grasp: Starting from a resting posture, with the body facing forward, the subject rotates towards the box, approaches it, and grasps it from its initial position (A).
- Lifting and placing: The subject lifts the box and places it at the final position (B).
- Returning to the initial posture: After placing the box, the subject returns to the original resting position.

The task is performed at a horizontal distance from point B equal to 80% of the subject's upper limb length, measured from the acromion to the center of the palm, as shown in Fig. 1. The task is executed under two load conditions:

- Low-load condition: using a 5 kg box (*Lft5*).

- High-load condition: using a 15 kg box (*Lft15*).

For each load condition, the movement is performed five times without the exoskeleton and five times while wearing the hybrid exoskeleton.

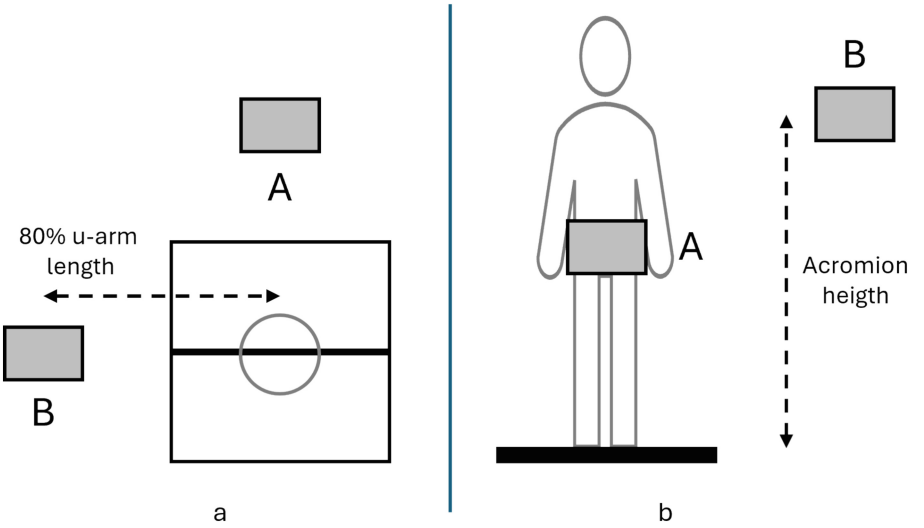


Fig. 1. Schematic representation of the lifting task setup, showing the top view (**a**) and the side view (**b**).

After the execution of each trial, an external operator relocates the box from position B to position A.

2.4 Overhead Force Appliance Setup

The second and final type of motor task was designed to simulate the static maintenance of an upper limb posture. Specifically, the task involves maintaining a posture with an upward load (i.e., overhead), simulating the use of a rivet gun or a drill. Like the previous task, this task can be divided into three phases:

- Achievement of the Force Application Posture: Starting from a resting position, the individual moves into the posture required for applying force.
- Constant Force Application: A consistent force is applied onto a load cell positioned at the end of a robotic arm. The force is applied with the aim of a personalized 3D printed handle which simulates the rivet gun. The height of the robotic arm is adjusted to be compatible with that of the participant, ensuring the task is performed at an appropriate ergonomic level.
- End of Force Application and Return to Rest Position: The force application phase ends, and the individual returns to the resting posture.

The movement is performed from three different foot positions, as illustrated in Fig. 2: left (*MntL*), center (*MntC*) and right (*MntR*) relative to the vertical projection of the target force application point.

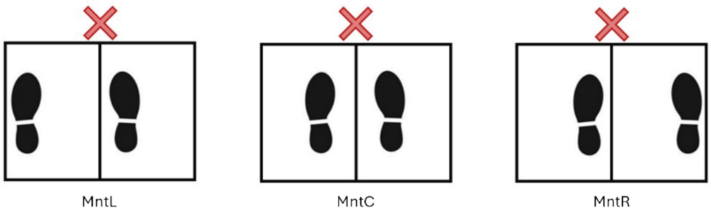


Fig. 2. Foot positions relative to the vertical projection of the application point, identified in the figure by the red cross.

For proper maintenance of the upper limb posture, the participant is provided with visual feedback regarding the force applied to the load cell. The force should not exceed 5 kg and must be kept as perpendicular as possible to the load cell, within a tolerance on the other directions of 10% (see Fig. 3).

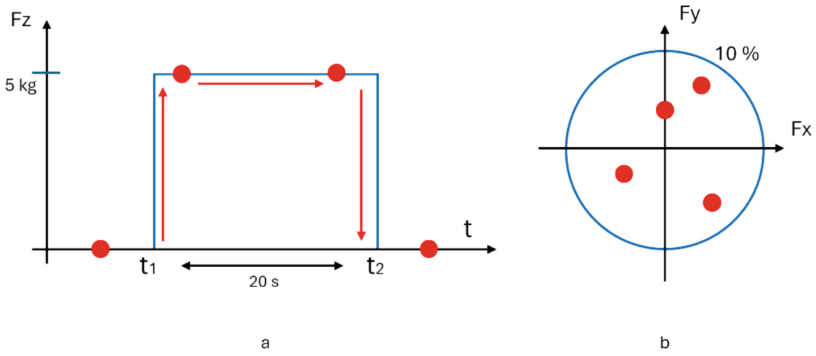


Fig. 3. An example of the ideal force feedback that the participant is required to follow perpendicularly to the load cell (**a**) and the tolerance for the other components (**b**). The data received from the sensor is displayed through a red disc, which "moves" over time.

For each foot position, five posture holds are performed, each lasting 20 s, with a 30-s rest interval in between.

2.5 Instrumentation

Motion Tracking. Whole-body kinematics was acquired by using a full-body motion tracking system based on inertial measurement units—IMUs (MVN Link and Biomech Awinda, XSens). The motion trackers were positioned within the suit according to the protocol required by the manufacturer; in particular, 17 sensors were positioned on the

subject to capture the movement of the following 23 body segments, which included the head, neck, eighth and tenth thoracic vertebra, third and fifth lumbar vertebra, right and left shoulder, right and left arm, right and left forearm, right and left hand, pelvis, right and left thigh, right and left shank, right and left foot, and right and left forefoot. Before starting the recording session, a calibration procedure was performed to align the motion trackers to the anatomical segments of the subject. The movement was acquired by using a sampling frequency of 60 Hz. Joint kinematics in terms of joint angles was exported for custom analysis.

Electromyography. A total of 16 upper body muscles were selected for surface electromyography (EMG) acquisition. The acquisition system used was the Cometa Mini-Wave (Cometa Systems), which employs bipolar sensors with a sampling frequency of 2000 Hz.

To ensure a more precise analysis of the motor task movement, the 16 sensors were strategically positioned on the muscles: right and left anterior deltoid (DA), right and left upper trapezius (UT), right and left posterior deltoid (DP), right and left medial triceps (MT), right and left pectoralis major (PM), right and left biceps brachii (BB), right and left flexor carpi radialis (CR), right and left longissimus dorsi (LD).

Exoskeleton. As previously reported, the exoskeleton utilized in this research was the model LBE 30 (Wearable Robotics srl). This device is an upper limb hybrid exoskeleton designed to assist the user during lifting tasks, particularly focusing even on high loads (up to 25 kg). Activation of the force, applied at the wrists level, can be initiated by the user through two pressure sensors that act as flexible switches located on the thumb tips. The applied force can be wirelessly adjusted by exploiting a mobile app and ranges from 0 to 25 kg. The exoskeleton was designed to distribute the force evenly between the two upper limbs. For the lifting task, a total force of 5 kg and 15 kg was selected respectively for *Lft5* and *Lft15* tasks. For the *Mnt* tasks, a total force of 5 kg was selected.

2.6 Experimental Protocol

Each subject undergoes two distinct experimental sessions, performed in two different days.

- The first session consists of the lifting tasks (*Lft*), performed with 5 kg (*Lft5*) and 15 kg (*Lft15*) loads.
- The second session focuses on the overhead force maintenance task (*Mnt*), executed in three different feet positions.

Lifting Session. The lifting session was structured as follows:

- Lifting without the exoskeleton
  - ***Lft5***: 5 trials with approximately 30 s of rest between each trial
  - 2-min rest period
  - ***Lft15***: 5 trials with approximately 30 s of rest between each trial

- Preparation for exoskeleton use
  - Exoskeleton donning, instructions, and system verification.
- Lifting with WR exoskeleton
  - ***Lft5***: 5 trials with approximately 30 s of rest between each trial
  - 2-min rest period
  - ***Lft15***: 5 trials with approximately 30 s of rest between each trial

For the exoskeleton-assisted task, subjects were instructed to activate the assistive force just before lifting the box using the switches; conversely, the force was deactivated immediately after releasing the box, before returning to the resting position. This ensures that the lifting movement was fully influenced by the exoskeleton. Figure 4 illustrates the task configurations with and without the exoskeleton.

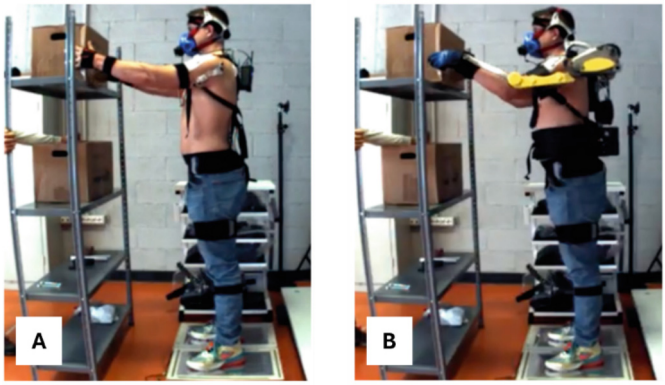


Fig. 4. Lifting task configuration without (A) and with (B) exoskeleton.

Overhead Force Maintenance Session. The overhead force maintenance session is structured as follows:

- Force maintenance without exoskeleton
  - ***MntL***: 5 trials with 30 s of rest between each trial
  - 2-min rest period
  - ***MntC***: 5 trials with 30 s of rest between each trial
  - 2-min rest period
  - ***MntR***: 5 trials with 30 s of rest between each trial
- Preparation for exoskeleton use
  - ***Exoskeleton*** wearing, instructions, and system verification.

- Force maintenance with the WR exoskeleton

  - ***MntL***: 5 trials with 30 s of rest between each trial
  - 2-min rest period
  - ***MntC***: 5 trials with 30 s of rest between each trial
  - 2-min rest period
  - ***MntR***: 5 trials with 30 s of rest between each trial

For the exoskeleton-assisted task, subjects were instructed to activate the assistive force after positioning the handle on the force cell, following the visual feedback. The exoskeleton was deactivated immediately after the 20-s holding phase, as indicated by the feedback system. Figure 5 illustrates the task configurations with and without the exoskeleton.

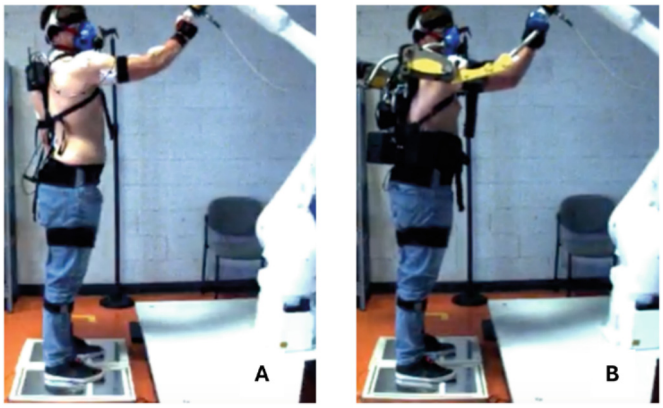


Fig. 5. Maintenance task configuration without (A) and with (B) exoskeleton.

2.7 Data Analysis

Both the lifting and maintenance tasks were divided into the different execution phases, and the reported data of this preliminary study refer to the main phase of each specific task, i.e., the lifting and the maintenance ones. In particular, Fig. 6 illustrates the phase segmentation and highlights the phase selected for analysis, specifically Phase #2 for each task.

All the joint angles were extracted from exported files via custom processing and analysis pipeline (Matlab2023a, MathWorks Inc.). The angles of the shoulder joint, elbow joint, wrist joint, and pelvis-T8 were chosen for this preliminary analysis based on the characteristics of the identified tasks. The range of motion (ROM) of each analyzed joint was quantified for both conditions, i.e., with and without the exoskeleton. Subsequently, the differences between the two conditions were analyzed.

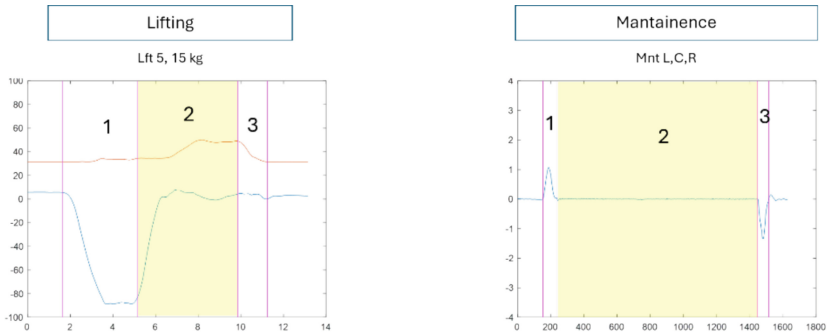


Fig. 6. Phase division based on the anatomical angles recorded during the lifting (left) and maintenance (right) trials.

Furthermore, the acquired raw EMG data were post-processed using a custom-made routine. Specifically, the raw EMG signal was filtered (band-pass between 5 Hz and 450 Hz, and notch at 50 Hz) to suppress undesired noise and interferences, and then full wave rectified. Finally, after extracting its envelope (low-pass moving average filter), the signal was normalized relative to the data acquired without the exoskeleton (i.e. = 0%). This normalization allowed the exoskeleton-free condition to serve as a reference, facilitating the quantification of differences. Muscle activation was quantified during the main phase of movement for each task by computing the integrals of the area under the processed EMG signals (iEMG) for each exoskeleton. For each participant, the integral was first calculated for each trial, and then the average value was obtained.

To obtain an overall view of the effect of the exoskeleton on muscle activation for each muscle, the variations in iEMG (*ΔiEMG)* relative to the trials without the exoskeleton were calculated for each participant. These variations were then averaged using Eq. (1):

$$\Delta iEMG_{exo_k} = mean_{subj}\left(\frac{iEMG_{exo_k} - iEMG_{NoExo}}{iEMG_{NoExo}} x100\right) \quad (1)$$

Possible outliers were removed from the dataset before analyzing the overall results.

3 Results and Discussion

3.1 Lifting Task

Low-Load Lifting. Focusing on shoulder muscles, Fig. 7 shows an example of EMG signals for the right DA (upper panel) and MT (lower panel) muscles during the low-load lifting phase, for the two conditions with and without the exoskeleton.

It can be observed that there is little difference in the overall amplitude of the signals, except for a minimum increase in activation for the MT signals when using the WR exoskeleton. However, a noticeable shift in muscle activation timing is evident, particularly for the anterior deltoid (DA), where activation occurs earlier in the WR condition.

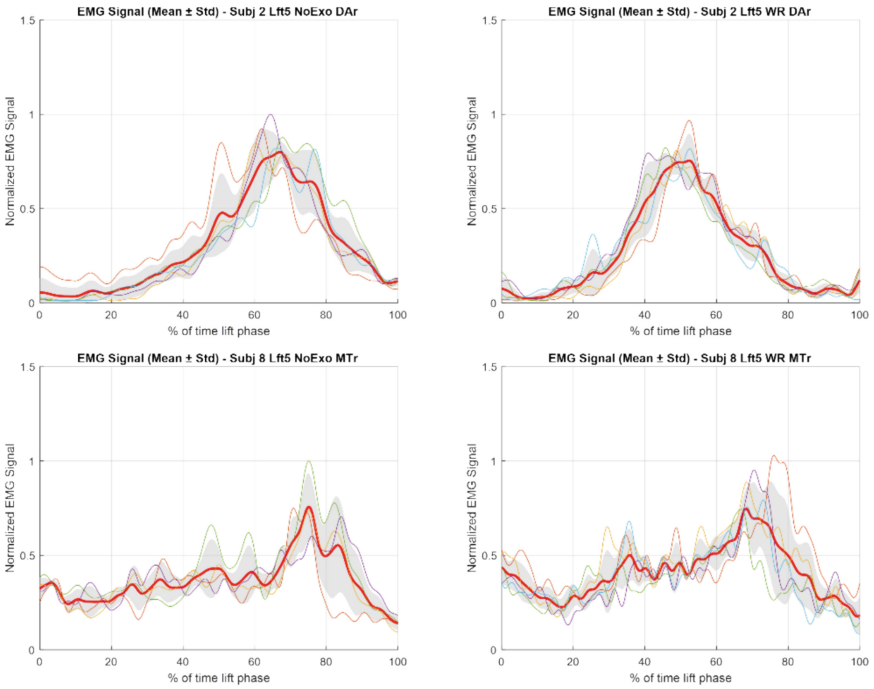


Fig. 7. Example of processed EMG signals for the right DA muscle (upper panel) and the right MT muscle (lower panel) in the *Lft5* task, comparing the conditions without (left panel) and with (right panel) the exoskeleton.

Figure 8 shows, for each participant, a graph representing the previously mentioned values corresponding to iEMG parameters; for each participant, the bars for the trials with and without the exoskeleton are displayed side by side.

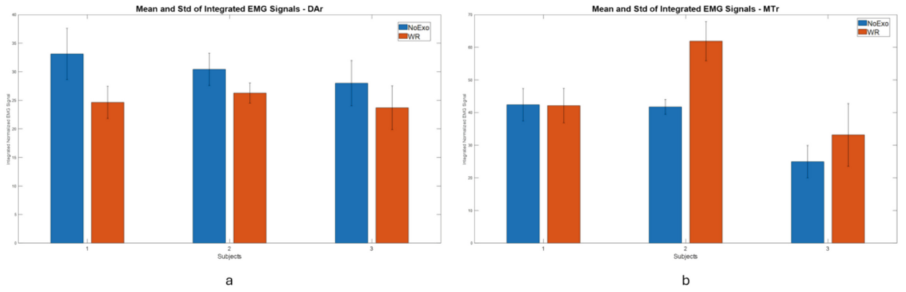


Fig. 8. Bar graph representing the values of iEMG, for the DAr (left) and MTr muscles (right).

By analyzing the graphs, it can be observed that while the exoskeleton generally led to a reduced overall activation of the right DA muscle during the task (for nearly all subjects), its use tended to increase muscle activation in the MT muscle.

Figure 9 shows the calculated $\Delta iEMG$, where the red line at y = 0 represents the threshold relative to the NoExo data.

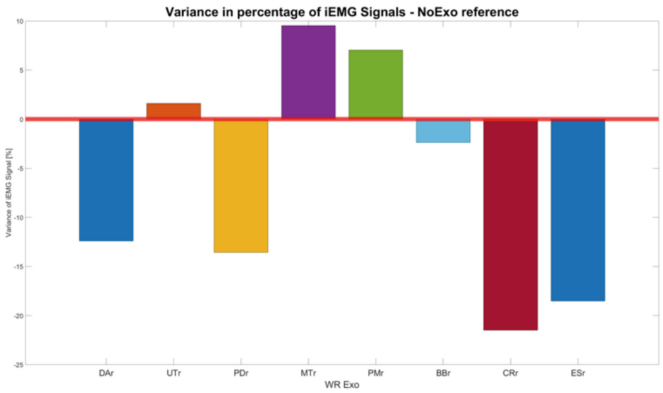


Fig. 9. ΔiEMG with respect to the data without exoskeleton for right limb.

Expanding the analysis to a wider group of muscles, it can be observed that, although below 10%, the right PM muscle also showed an increase in muscle activation. On the other hand, there is a clear decrease in activation for the CR and ES muscles, which, along with the previously mentioned DA, exhibit a reduction in activity.

Figure 10 presents an example of the flexion/extension angle of the right shoulder and elbow during the lifting phase for the two conditions, i.e., with and without the exoskeleton in the *Lft5* task.

Regarding the range of motion (ROM), a significant reduction was observed in the elbow joint; on the other side, the shoulder only exhibited a decreased range of movement, particularly when considering the target reaching height.

In this regard, Fig. 11 shows, for each participant, a graph representing the previously mentioned ROMs calculations. For each participant, the bars for the trials with and without the exoskeleton are displayed side by side.

Figure 12 shows the calculated ΔROM for the upper arm joints, where the red line at y = 0 represents the threshold relative to the NoExo data. The angles that are taken into consideration are the flexion/extension of shoulders and elbows and the pronation/supination of the wrists, which are critical in the definition of the biomechanical risks from the ergonomical point of view.

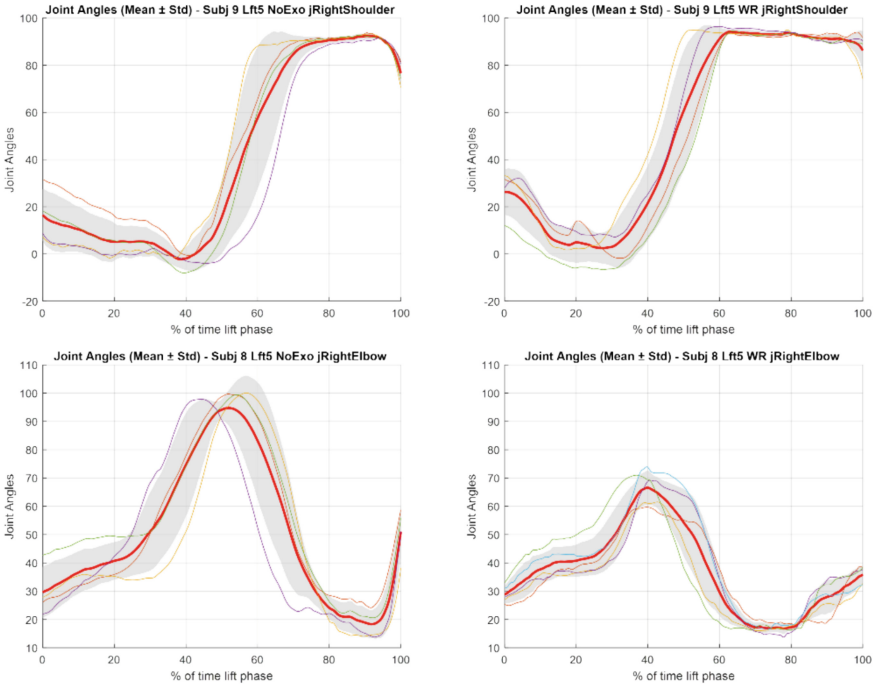


Fig. 10. Right shoulder flexion-extension angle for the *Lft5* task, without (left) and with (right) the exoskeleton.

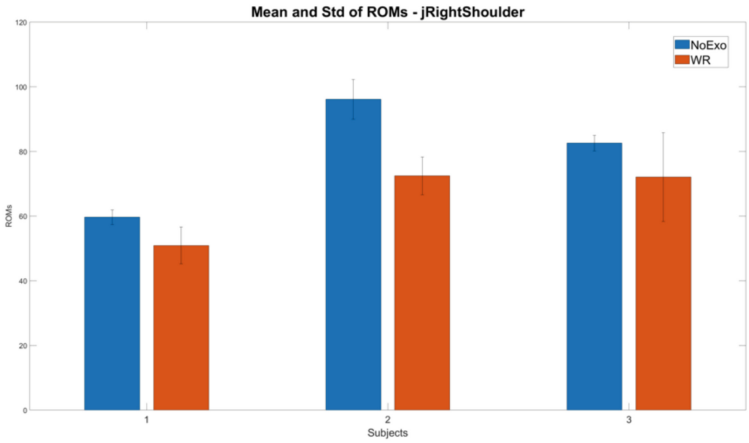


Fig. 11. Bar graph representing right shoulder flexion-extension ROMs for all the subjects.

A symmetry in ROM reductions can be observed between the right and left joints, except for the wrist, where a substantial difference is evident (where there was an increase in the dominant limb). This discrepancy may be attributed to the difficulty in clearly

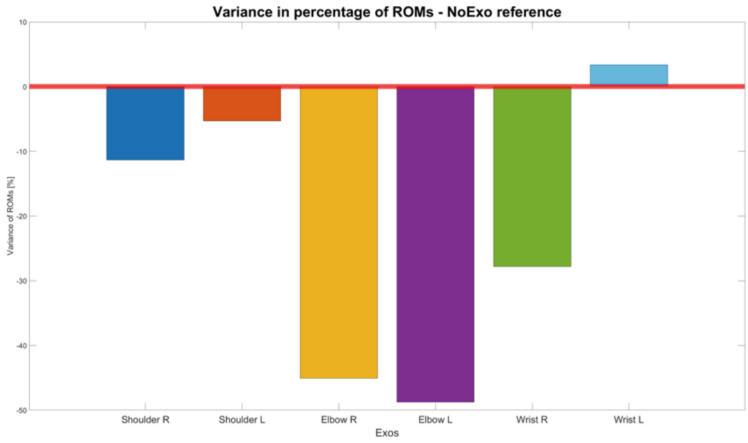


Fig. 12. ΔROM with respect to the data without exoskeleton.

supporting the box during the lifting phase, due to the type of wrist supports embedded in the exoskeleton; these supports seemed somehow to alter the natural grip due to awkward posture of the hand.

3.2 High-Load Lifting

Focusing on high loads (i.e., 15 kg), Fig. 13 shows an example of EMG signals for the right DA (upper panel) and MT muscles (lower panel) during the lifting phase, for the two conditions with and without the exoskeleton.

The differences observed in the EMG signals during the low-load task appeared more pronounced in the high-load condition. Specifically, a greater activation amplitude was observed for the DA muscle in the trial without the exoskeleton. Conversely, the MT muscle showed higher activation during the trial with the exoskeleton. In addition, increased signal instability was observed, which may indicate reduced control over the device due to the higher forces involved in the tasks (Fig. 14).

The conditions observed in the low-load case can be similarly underlined also in the high-load trials, and even increased for the DA. On the other side, we have to underline and increase in the MTr muscles for 2 out of 3 subjects. It is worth noting that all the subjects presented some difficulties in completing the trial due to physical discomfort caused by the exoskeleton fit, even if no significant issues were reported during the low-load trial.

Additionally, as seen in Fig. 15, with an increase in load, the differences in the total activation of the other muscles involved in the task became more pronounced in this high-load condition. As underlined before, particular attention had to be given to the activation of the MTr muscles.

For the same task, Fig. 16 presents an example of the flexion/extension angle of the right elbow during the lifting phase for the two conditions, i.e., with and without the exoskeleton in the *Lft15* task (Fig. 17).

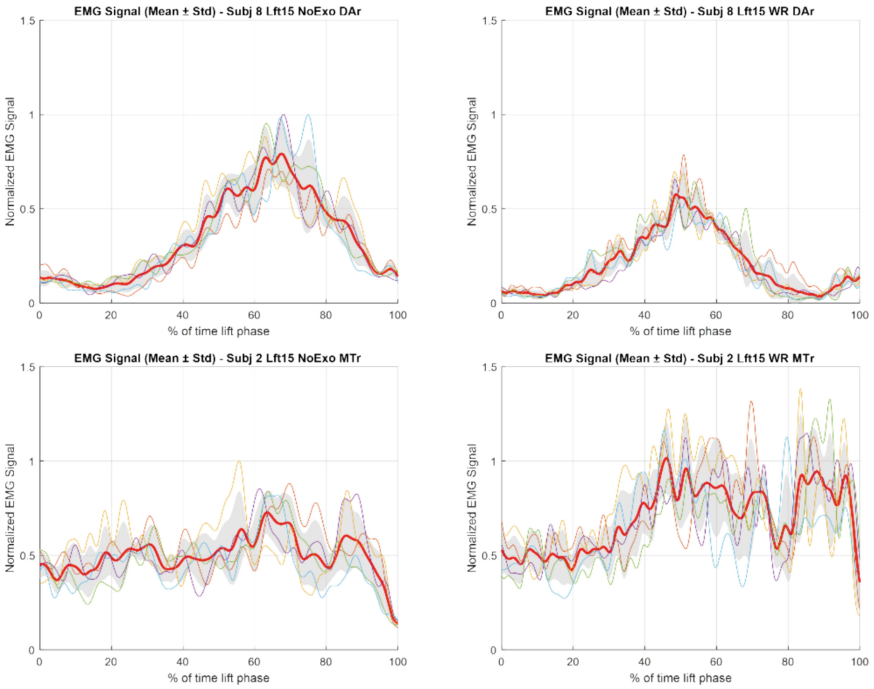


Fig. 13. EMG signals for the right DA muscle (upper panel) and the right MT muscle (lower panel) in the *Lft15* task, comparing the conditions without (left panel) and with (right panel) the exoskeleton.

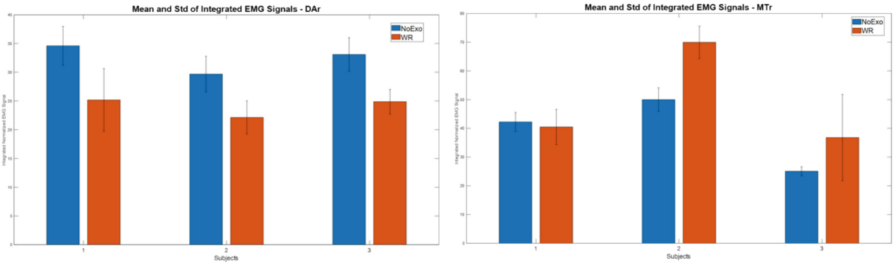


Fig. 14. Bar graph representing the values of iEMG parameter, for the DAr (left) and MTr muscles (right).

It is important to highlight that the elbow ROM significantly decreased as the load increased. This can be mainly due to the fact that, by applying more force at the wrist level, the subject tended to keep their arms more extended and offers less resistance to the exoskeleton. In fact, while the exoskeleton provides muscle-level benefits, it was observed a reduction in the overall range of motion of the elbow joint (Fig. 18).

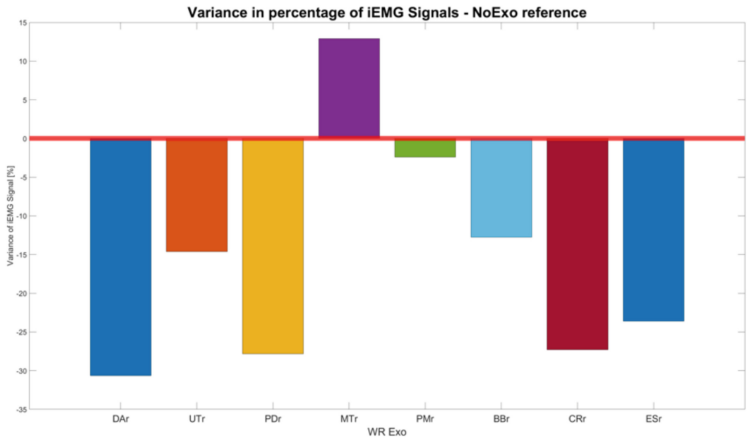


Fig. 15. ΔiEMG with respect to the data without exoskeleton for right limb.

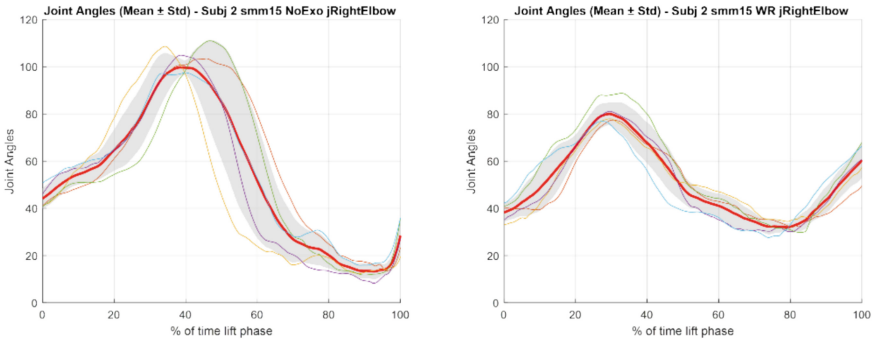


Fig. 16. Right elbow flexion-extension angle for the *Lft15* task, without (left) and with (right) the exoskeleton.

Indeed, at the shoulder joint, the flexion/extension ROM did not decrease, while at the elbow, there was a reduction of more than 40% during the movement. As observed previously, there was also a decrease in wrist ROM, both on the right and left sides. This could indicate a minimal learning curve in optimizing the grip on the box, despite the exoskeleton support.

3.3 Applied Force Maintenance Task

Focusing on the maintenance task, Fig. 19 shows an example of EMG signals for the right DA muscle under the two conditions, with and without the exoskeleton, in the MntL task.

Figure 20 shows, for each participant, a graph representing the previously mentioned iEMG integrals for the DAr muscle. For each participant, the bars for the trials with and without the exoskeleton are displayed side by side.

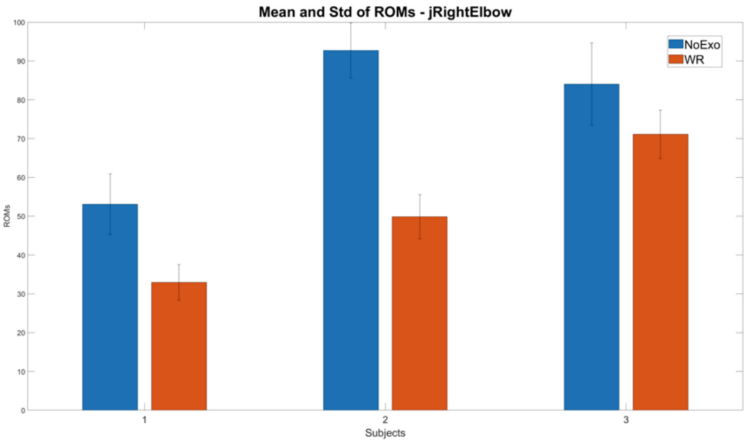


Fig. 17. Bar graph representing right elbow flexion-extension ROMs for all the subjects.

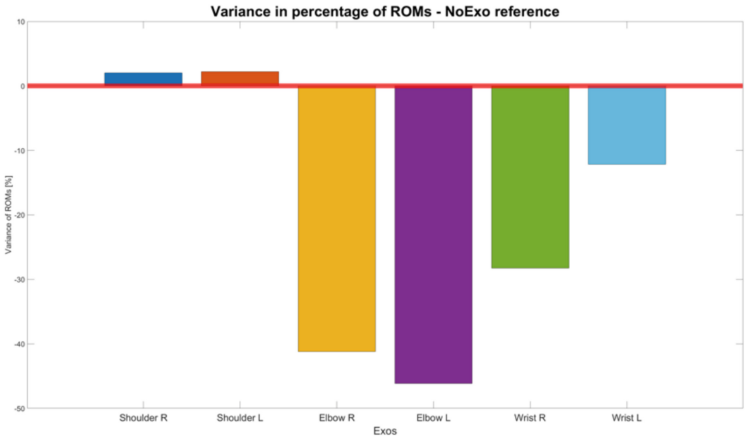


Fig. 18. ΔROM with respect to the data without exoskeleton.

From the previous two figures, a clear reduction in muscle activity at the DA (of the actively used arm) was observed for each subject. The increased stability and likely further learning in the use of the device led to better control over it, as the task was much more static with respect to the load-lifting ones.

Figure 21 presents a comparison of ΔiEMG parameters across the different positions (*MntL*, *MntC*, and *MntR)*. The left ES muscle has been included in the graph, as it plays a crucial role in maintaining the posture across the various feet positions.

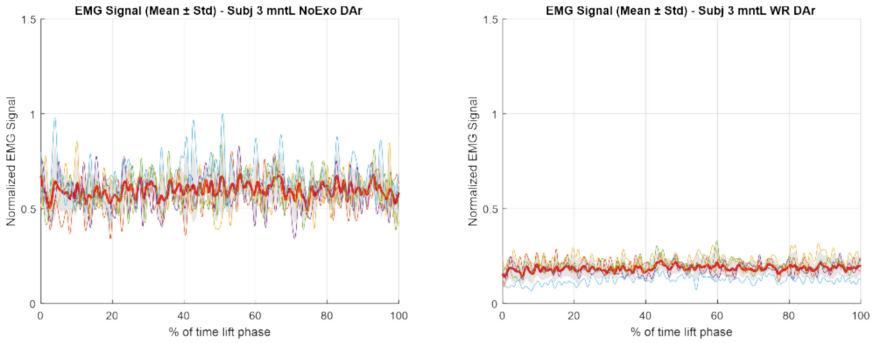


Fig. 19. EMG signals for the right DA muscle in the MntL task, comparing the conditions without (left panel) and with (right panel) the exoskeleton.

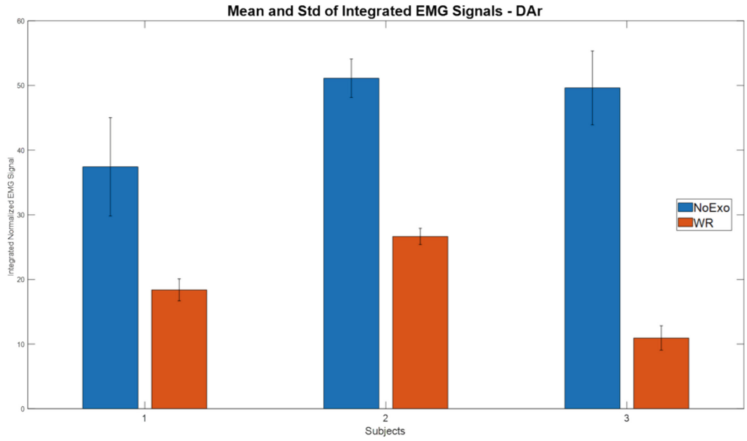


Fig. 20. Bar graph representing the iEMG integrals, for the DAr muscles.

The most significant decrease in muscle activity was observed at the biceps level, with over a 60% reduction across all three positions of the task. Other substantial reductions were noted in the UT and, as mentioned earlier, in the DA, with values around 50%. On the other hand, increases in electromyographic activity could also be observed in some trials for the Carpal Radialis (CR) and, especially, in both the Erector Spinae (ES). In the case of the CR, the EMG activity may be more unstable as the subject needs to ensure staying within the tolerance limits imposed by the force feedback by controlling the force provided by the exoskeleton. For the ES, the increase in muscular activity, which approaches 40% in the *MntC* task, may be probably due to the reduced overall mobility when wearing the exoskeleton, leading to a compensatory postural adjustment that places additional load on that specific muscular group.

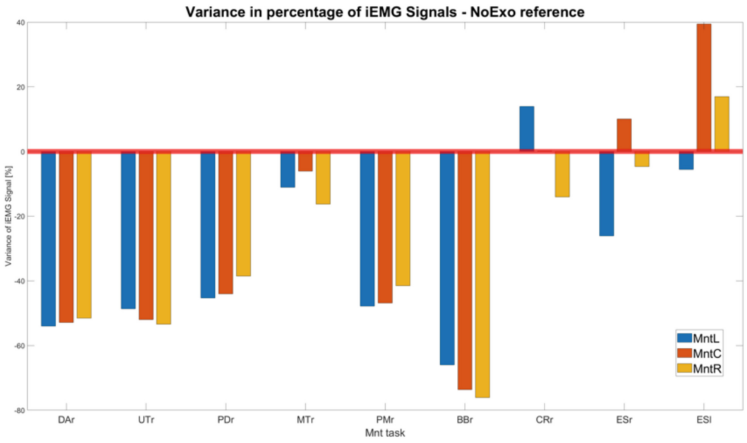


Fig. 21. ΔiEMG with respect to the data without exoskeleton for the Mnt task.

4 Conclusions

In conclusion, the present study highlighted the possibility of introducing a specific setup able to underline several key findings regarding the impact of exoskeleton use on muscle activation and range of motion (ROM) during lifting and maintenance tasks. One notable limitation concerns the overall wearability of the exoskeleton and its specific fit with respect to the high-load trial; indeed, this emphasized the importance of proper ergonomic adaptation for any kind of similar devices in the field.

Regarding muscle activation, those of the anterior deltoid (DA) were constantly and significantly reduced using the exoskeleton, in both lifting and overhead tasks; another muscle supported by the exoskeleton was in particular was the biceps (BB), which showed significant reductions in activity during the maintenance tasks, with decreases reaching over 60% for all three positions. Conversely, other muscles like the carpal radialis (CR) and erector spinae (ES) exhibited increased activity, likely due to the compensatory postural adjustments required when wearing the exoskeleton, particularly in response to reduced mobility in the shoulder and elbow joints.

It is also worth noting that the analysis of the kinematic data lacked a detailed evaluation of the joint movements in the trunk and hip regions, which could provide further insights into the overall biomechanical effects of exoskeleton use. These data points represent a gap for future studies aimed at offering a more comprehensive understanding of the full body interaction with exoskeletons during static and dynamic tasks. Moving forward, in addition to including complementary kinematic and kinetic data (e.g., force platforms) for a complete analysis, there is a need to include also information concerning metabolic consumption, which could allow to generate a comprehensive picture of the biomechanical effects of using this exoskeleton for the analyzed tasks.

Acknowledgments. This study was funded by INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro), grant number BRIC 2022, ID-58.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Scalona, E., et al.: (2023). Human Ergonomic Assessment Within "Industry 5.0" Workplace: Do Standard Observational Methods Correlate with Kinematic-Based Index in Reaching Tasks? https://doi.org/10.1007/978-3-031-35741-1_17
2. Xun, X., et al.: Industry 4.0 and industry 5.0—inception, conception and perception. J. Manuf. Syst. **61**, 530–535 (2021). https://doi.org/10.1016/j.jmsy.2021.10.006
3. Leng, J., et al.: Industry 5.0: prospect and retrospect. J. Manuf. Syst. **65**, 279–295 (2022). https://doi.org/10.1016/j.jmsy.2022.09.017
4. Maddikunta, P.K.R., et al.: Industry 5.0: a survey on enabling technologies and potential applications. J. Ind. Inf. Integr. **26**, 100257 (2022). https://doi.org/10.1016/j.jii.2021.100257
5. Bortolini, M., et al.: Automatic assessment of the ergonomic risk for manual manufacturing and assembly activities through optical motion capture technology. Procedia CIRP. **72**, 81–86 (2018). https://doi.org/10.1016/j.procir.2018.03.198
6. Landau, K., et al.: Musculoskeletal disorders in assembly jobs in the automotive industry with special reference to age management aspects. Int. J. Ind. Ergon. **38**(7–8), 561–576 (2008). https://doi.org/10.1016/j.ergon.2008.01.006
7. Gao, G., et al.: Effect of exoskeleton devices on work-related musculoskeletal disorders (WMSDs) among healthcare workers: a scoping review. Ergonomics, 1–13 (2024). https://doi.org/10.1080/00140139.2024.2413150
8. Weston, E.B., et al.: Biomechanical evaluation of exoskeleton use on loading of the lumbar spine. Appl. Ergon. **68**, 101–108 (2018). https://doi.org/10.1016/j.apergo.2017.11.006
9. Theurel, J., et al.: Physiological consequences of using an upper limb exoskeleton during manual handling tasks. Appl. Ergon. **67**, 211–217 (2018). https://doi.org/10.1016/j.apergo.2017.10.008
10. Gregorczyk, K., et al.: Effects of a lower-body exoskeleton device on metabolic cost and gait biomechanics during load carriage. Ergonomics. **53**, 1263–1275 (2010). https://doi.org/10.1080/00140139.2010.512982
11. Bär, M., et al.: The influence of using exoskeletons during occupational tasks on acute physical stress and strain compared to no exoskeleton—a systematic review and meta-analysis. Appl. Ergon. **94**, 103385 (2021). https://doi.org/10.1016/j.apergo.2021.103385
12. Piol, A., et al.: The effects of the use of exoskeletons for manual handling on cognitive abilities: a mixed reality approach. In: Duffy, V.G. (ed.) HCI International 2024—Late Breaking Papers. HCII 2024. Lecture Notes in Computer Science, vol. 15376. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-76809-5_21
13. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. **9**(1), 97–113 (1971). https://doi.org/10.1016/0028-3932(71)90067-4

Impact of Exoskeleton Support on Discharge Statistics of Motor Units During Quasi-Isometric Tasks

Giacomo Valli[1(✉)], Martina Mosso[2], Maria Lucia Cavallo[1], Alessandro Piol[2], Emilia Scalona[2], Tiwana Varrecchia[3], Giorgia Chini[3], Alberto Ranavolo[3], Luca Falciati[1], Debora Brignani[1], Nicola Francesco Lopomo[2], and Francesco Negro[1]

[1] Department of Clinical and Experimental Sciences, Università of Brescia, Brescia, Italy
giacomo.valli@unibs.it
[2] Department of Information Engineering, Università of Brescia, Brescia, Italy
[3] Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, INAIL, Monte Porzio Catone, 00078 Rome, Italy

Abstract. The industry 5.0 paradigm emphasizes a human-centric approach to industrial processes, integrating technology to optimize safety, ergonomics, and productivity. Among these technologies, exoskeletons have emerged as promising tools for alleviating physical strain during labor-intensive tasks. While previous research has focused on the impact of exoskeletons on muscle activation, biomechanics, and metabolism, their effects on the neuromuscular system remain unexplored. This pilot study aimed at testing a protocol to investigate the influence of a hybrid exoskeleton on neuromuscular control during quasi-isometric tasks at varying mechanical loads and support levels. High-Density Surface Electromyography (HDsEMG) and single motor unit (MU) analysis were used to assess changes in MU discharge rate and discharge rate variability in the *biceps brachii* and *anterior deltoid* muscles. A single participant performed tasks involving holding 4 kg and 8 kg loads, both with and without exoskeleton assistance, where support levels were calibrated at 75% (for the 4 kg load) and at 50% and 75% (for the 8 kg load). Results showed that during the 4 kg task, MU discharge characteristics were stable in the biceps, while the deltoid exhibited a decrease in discharge rate and an increase in variability. For the 8 kg task with 50% and 75% support, biceps MU discharge remained consistent, but the deltoid showed a further decrease in discharge rate and increased variability, especially at 75% support. These preliminary results show that HDsEMG can be used to analyze the influence of an exoskeleton on neuromuscular control and indicate that exoskeletons can alter neural control strategies.

Keywords: Industry 5.0 · Exoskeletons · Neuromuscular Control · High-Density Surface Electromyography · Motor Unit Discharge Rate

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 342–352, 2025.
https://doi.org/10.1007/978-3-031-93502-2_23

1 Introduction

Work-related musculoskeletal disorders (WMSDs) represent a significant health issue in the European Union, impacting the professional and personal lives of millions of workers. These disorders not only burden workers but also pose significant costs to businesses and society. According to the 2015 European Working Conditions Survey, approximately three out of five workers reported musculoskeletal disorders. Among these, lower back pain and muscle discomfort in the upper limbs are the most common, followed by pain in the lower limbs.

Despite increasing automation, many occupational tasks still require the execution of manual material handling (MMH) activities with an active worker participation, exposing individuals to biomechanical overload risks. Physical risk factors such as repetitive movements, prolonged awkward postures, heavy lifting, vibrations, and exposure to low temperatures are strongly associated with the onset of WMSDs.

WMSDs significantly impact workers' overall health, causing pain, functional limitations, and disability. These disorders often lead to increased absenteeism, reduced productivity, and a decline in quality of life. Moreover, they frequently coexist with other health issues such as anxiety, sleep disturbances, and fatigue, creating a vicious cycle that further deteriorates workers' health and well-being. Furthermore, the economic impact of WMSDs is substantial, both for businesses and society. For example, in Germany, WMSDs were estimated to result in a production loss of €17.2 billion and a gross value-added loss of €30.4 billion in 2016. These data underscore the importance of investing in the prevention and management of WMSDs to ensure workers' health and safety [5].

The industry 5.0 paradigm emphasizes a human-centric approach to industrial processes, integrating advanced enabling technologies to optimize occupational health and safety, ergonomics, and risk management [8]. Among these technologies, exoskeletons have gained attention as potential solutions to alleviate physical strain during labor-intensive tasks due to their potential for unloading muscles and joints, reducing physical effort, and enhancing workers' well-being and productivity.

Occupational exoskeletons are wearable devices designed to support and assist body movements and are proposed as a solution to prevent WMSDs. These devices aim to reduce musculoskeletal load and improve posture during strenuous tasks. Following traditional ergonomic interventions, exoskeletons are gaining attention as a potential innovation in various occupational scenarios, including automotive, aerospace, logistics, construction, and agriculture. Exoskeletons can be broadly classified as "active" or "passive." Active exoskeletons use actuators powered by external energy sources such as batteries and are controlled by processing units that adjust assistance based on sensors input. Passive exoskeletons, in contrast, do not use actuators but rely on elastic components like springs to support posture or movement. Additionally, a "hybrid" category includes these exoskeletons that combine elastic energy storage with smaller actuators, which modify spring characteristics to optimize task-specific support.

In order to allow for proper integration of these enabling technologies with a human-centric approach, it is necessary to have a complete understanding of their multifaceted impacts on workers. Previous research has explored how the use of different types of exoskeletons could affect muscle activation levels, biomechanics, and metabolism.

However, the specific effects of exoskeletons on the neuromuscular system remain unknown.

In humans, the neuromuscular function can be investigated, non-invasively, using High-Density Surface Electromyography (HDsEMG) recordings and single motor unit (MU) detection techniques. HDsEMG is the state-of-the-art technique to record muscle electrical activity over a large muscle area using a grid of closely spaced electrodes [12, 13]. HDsEMG allows for the decomposition of signals into contributions from individual MUs and for the analyses of the MUs discharge behavior [7]. When looking at single MU discharges, the neural drive to the muscle can be quantified by the MU discharge rate, which can be considered as a marker of the intensity of the muscle activation signal. Conversely, the discharge variability can be used as a marker of the consistency/perturbation of the motor commands [10].

A recent pre-norm indicates HDsEMG as a possible approach to estimate the risk of biomechanical overload [2].

The aim of this preliminary investigation was to develop and test a methodology for analyzing changes in neuromuscular control induced by using an upper-body hybrid exoskeleton during load support, with a focus on single MU behavior. Such insights could inform exoskeleton development and help to establish guidelines that ensure the safe and effective adoption of exoskeleton technologies.

2 Materials and Methods

2.1 Participant

This preliminary, descriptive study involved the enrollment of a single, 30 years old, healthy male volunteer. The participant reported no prior history of neuromuscular disorders or any contraindications to performing the task. The volunteer was informed on the details of the experimental session and provided explicit consent to participate in this pilot study.

In a preliminary visit, the participant was familiarized with the use of the exoskeleton and with the motor task. The participant also completed the Global Physical Activity Questionnaire (GPAQ) which was used to assess the physical activity level and sedentary behavior [1].

2.2 Exoskeleton

The exoskeleton used in this study was the Wearable Robotics LBE 30 (Wearable Robotics srl, Ghezzano, San Giuliano Terme—PI, Italy), a hybrid upper-limb device designed to assist with shoulder flexion movements. The device applies force at the wrist when triggered by the user through pressure sensors on the thumbs. The force, which can be wirelessly adjusted via a smartphone, ranges from 2 to 25 kg with a 1 kg step. The system evenly distributes the force between both upper limbs; for example, when set to 10 kg, it applies 5 kg of support to each limb.

2.3 Experimental Setup

The data collection took place during a single experimental session, which was divided into two parts: the first part without the assistance of the exoskeleton and the second part with exoskeleton support. A break was provided between the two parts to allow for complete recovery. The motor task consisted in quasi-isometric contractions where the participant was asked to suspend 4 and 8 kg loads at waist height with the elbows slightly bent in a comfortable position. The loads consisted in 2 paper boxes without handles filled with properly fixed weights up to the target load. The loads were initially positioned on a shelf set at a standard height of 95 cm. The isometric task required participants to lift and then hold the two previously described loads at waist height for 20 seconds. The participant first performed the task without assistance, completing two 20-second isometric contractions with a 4 kg load, separated by a 60-second recovery period. This was followed by two additional 20-second contractions with an 8 kg load, each separated by a 90-second recovery period. Afterward, the participant wore the exoskeleton and repeated the same contractions, maintaining the same timing and protocol. For the 4 kg load, the exoskeleton assistance was set to 75% (i.e., 3 kg total support, 1.5 kg support per limb), and for the 8 kg load, the assistance was set to both 50 and 75% (i.e., 4 and 6 Kg total support, 2 and 3 kg support per limb, respectively).

Tasks with different loads and support levels were performed in a random order.

During the tasks, HDsEMG and inertial sensors data were collected. However, for the purpose of this investigation, only the HDsEMG data was analyzed.

2.4 High-Density Surface Electromyography

While performing the isometric contractions, the electromyographic activity of the *biceps brachii* and of the *anterior deltoid* muscles was recorded. The recordings were performed using matrices of 64 equally spaced electrodes (GR08MM1305; OT Bioelettronica, Torino, Italy), one for each muscle, filled with conductive cream (Ac cream; OT Bioelettronica) for a total of 4 grids.

The grids were placed in the muscle belly oriented with the anatomical direction of muscle fascicles. The muscle borders were initially identified via palpation and subsequently confirmed through ultrasound imaging [4]. Once the position and orientation of the muscle fibers were visualized via ultrasound, palpation alone proved sufficient to locate the correct area for applying the electrode arrays.

Before placing the matrix, the skin was shaved, cleaned with 70% ethanol, and then with abrasive-conductive paste (Spes medica, Salerno, Italy) [9]. Individual reference electrodes were placed on the nearest prominent bone (e.g., two on the olecranon and 2 on the acromion).

The evaluate the signal quality, 10 seconds of contraction was acquired at rest and the noise level was quantified. The experimental setup (i.e., cleaning of the skin and position of the reference electrodes) was adjusted until a noise level below 20 μV was reached for all the muscles [14].

HDsEMG signals were sampled at 2048 Hz with the EMG-Quattrocento (OT Bioelettronica). The signal was recorded in monopolar configuration, amplified ($\times$ 150), and band-pass filtered (10–500 Hz) at source.

The recorded signals were stored offline for further processing (Fig. 1).

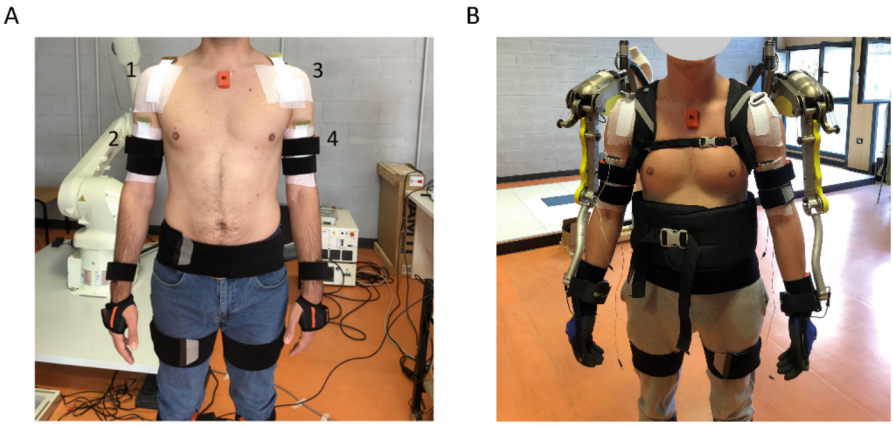


Fig. 1. Representation of the HDsEMG grid positioning. The position of the four HDsEMG grids is depicted in panel **A**. Two grids were positioned on the right and left *biceps brachii* muscles (numbers 2 and 4) while other two on the *anterior deltoid* muscle (numbers 1 and 3). The position of the grids with respect to the exoskeleton is depicted in panel **B**.

2.5 Signal Processing

For the scope of this preliminary investigation, only the signals collected from the dominant (right) limb will be processed and analyzed.

The acquired HDsEMG signal was preprocessed in MATLAB (R2021b, Natick, Massachusetts: The MathWorks Inc.). The HDsEMG signals recorded during the task performed without exoskeleton support was first band-pass filtered between 20 and 500 Hz with a second-order Butterworth filter and the powerline interference was removed with a 50 Hz notch filter. Subsequently, non-functioning channels or channels presenting an anomalous level of noise were removed. The resulting signals were decomposed into discharge times of the individual MUs using a validated convolutive blind source separation technique [7].

The results of the decomposition were visually inspected and manually edited by removing erroneous firing times and by adding undetected firings. To ensure an accurate firing detection, the cleaned MUs with a silhouette score below 0.88 were excluded [10]. During this phase, we aimed at obtaining the best separation between the pulses and the noise, in order to extract an optimal separation vector, which will be later used for the tracking of the MUs in the task performed with the exoskeleton [3].

After completing the MU cleaning, the obtained separation vectors have been transferred to the corresponding contraction performed with exoskeleton support, therefore allowing us to detect the same MUs in both contractions (Fig. 2).

An additional confirmation that the same MU was really detected was also obtained via the comparison of the MU action potentials obtained from the same MU detected in both contractions. For this scope, the MU action potentials have been generated by spike-triggered averaging (which means averaging segments of the EMG signal around each firing event) of the single differential spatial filtering applied on the direction of muscle fibers. The estimation of similarity between the MUAPs is then achieved via

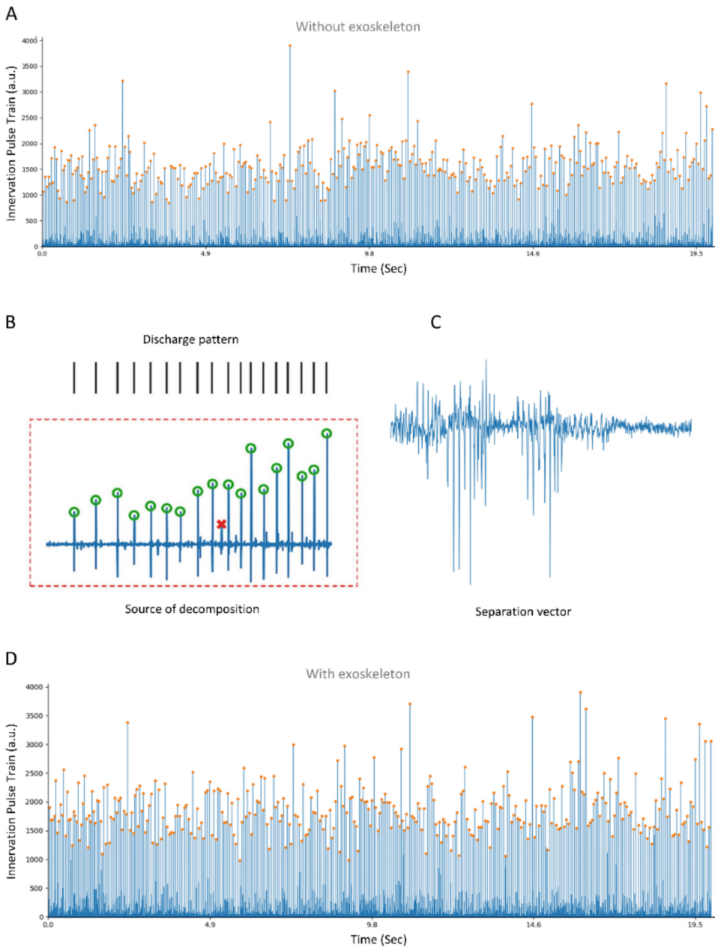


Fig. 2. HDsEMG signal processing. The HDsEMG signals acquired during the non-assisted task were decomposed into individual motor unit (MU) discharge times using a validated convolutive blind source separation technique (**A**). Decomposition results were visually inspected and refined by correcting firing errors and detecting missed firings (**B**). MUs with a silhouette score below 0.88 were excluded to ensure high accuracy. Optimal separation vectors (**C**) derived from the cleaned signals were then applied to contractions performed with exoskeleton support, enabling the tracking of the same MUs across both conditions (**D**). a.u.: Arbitrary units.

two-dimensional cross-correlation analysis. A minimum cross-correlation coefficient of 0.8 was set as a threshold to establish the accurate MU detection [6] (Fig. 3).

2.6 Analysis of Motor Unit Properties

To enhance the accuracy of the analysis, MU discharge statistics were computed only for MUs that were detectable in both the contractions performed with and without exoskeleton support. While this approach reduced the total number of observable MUs, since not

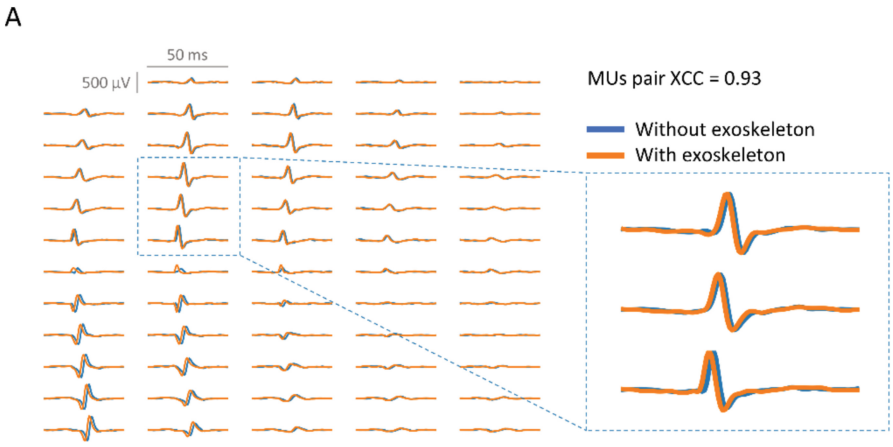


Fig. 3. Motor units tracking validation. Comparing the shape and spatial distribution of motor unit action potentials gives an additional confirmation of the identification of the same motor unit in different recording. In this example are represented the action potentials obtained from the same motor units displayed in Fig. 2. MU: Motor Unit, XCC: Cross Correlation Coefficient.

all MUs could be reliably identified in both conditions, it increased the validity of the findings by ensuring that the same MU population was compared across the intervention. This method provided a more reliable basis for evaluating changes in MU behavior between the two conditions.

MU discharge rate was calculated as the mean value of the instantaneous discharge rates across all the contraction [11]. Discharge rate variability was calculated as the coefficient of variation of the interspike interval across all the contraction.

The analysis of MU discharge statistics was performed in Python (V3.11.6, Python Software Foundation, USA) with the *openhdemg* V0.1.1 library [10] (Fig. 4).

3 Results

3.1 Participant

The participant performed moderate physical activity 2 times per week and resulted moderately active from the GPAQ test.

After a first familiarization session of approximately 10 minutes of practice with moving with the exoskeleton and autonomously enabling/disabling the actuators, the participant was fully confident in using it.

The participant successfully completed all the experimental procedures without reporting pain or tiredness of the upper limb muscles. However, discomfort was reported at the level of the back, neck, shoulders and wrist due to the weight of the exoskeleton and due to the exoskeleton anchor point at the wrist. Five minutes of rest were sufficient to achieve the participants self-perceived full recovery between the tasks without and with exoskeleton support.

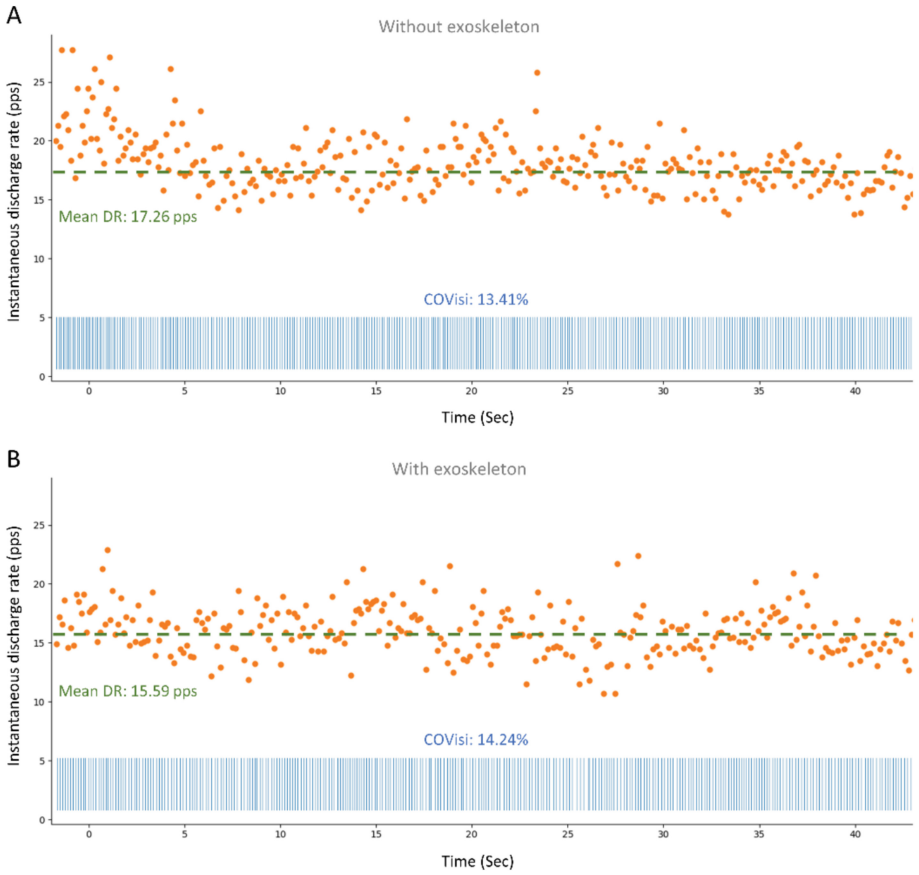


Fig. 4. Motor unit discharge rate and discharge rate variability estimation. Representation of the instantaneous motor unit discharge rate (orange dots) and of the mean instantaneous discharge rate (green horizontal dashed line) together with discharge variability estimated as the coefficient of variation of the interspike interval (discharge times are represented by blue vertical lines). Panel A and B represent the same motor unit without or with exoskeleton support, respectively. DR: Discharge Rate, COVisi: Coefficient of Variation of the Insterspike Interval, pps: Pulses Per Second.

3.2 Motor Unit Decomposition and Tracking

For the *biceps brachii* muscle, the decomposition and tracking process successfully identified and tracked MUs across both experimental conditions (without and with exoskeleton support). Specifically, during the contraction performed with a 4 kg load and 75% exoskeleton support, 8 MUs were accurately identified in both conditions. During the contraction performed with the 8 kg load and 50% support, 6 MUs where identified and tracked. Similarly, also during the contraction performed with the 8 kg load and 75% support, 6 MUs where identified and tracked.

For the *anterior deltoid* muscle, 5 MUs were accurately identified and tracked during the contraction performed with a 4 kg load and 75% exoskeleton support. During the

contraction performed with the 8 Kg load and 50% support, 9 MUs where identified and tracked. Similarly, also during the contraction performed with the 8 kg load and 75% support, 9 MUs where identified and tracked.

3.3 Motor Unit Analysis

Biceps Brachii. During the 4 kg task, the MU discharge rate for the *biceps brachii* was 12.49 ± 1.88 pulses per second (pps) without the exoskeleton and 12.42 ± 1.64 pps with the exoskeleton. Discharge rate variability increased from $17.36 \pm 4.20\%$ without the exoskeleton to $22.01 \pm 3.59\%$ with the exoskeleton.

During the 8 kg task without the exoskeleton, the MU discharge rate was 15.58 ± 2.17 pps while discharge rate variability was $24.69 \pm 13.89\%$. With the exoskeleton providing 50% support, the discharge rate of the *biceps brachii* increased slightly to 16.09 ± 1.36 pps. Discharge rate variability remained similar and was $25.04 \pm 5.23\%$. With the exoskeleton providing 75% support, the discharge rate for the *biceps brachii* decreased to 12.38 ± 1.86 pps, while the discharge rate variability remained at $24.03 \pm 4.56\%$.

Anterior Deltoid. During the 4 kg task, the MU discharge rate for the *anterior deltoid* was 10.80 ± 0.76 pulses per second (pps) without the exoskeleton and 9.68 ± 0.28 pps with the exoskeleton. Discharge rate variability was $19.74 \pm 6.79\%$ without the exoskeleton and $14.16 \pm 1.32\%$ with the exoskeleton.

During the 8 kg task without the exoskeleton, the MU discharge rate was 14.03 ± 1.15 pps while discharge rate variability was $17.90 \pm 2.07\%$. With the exoskeleton providing 50% support, the discharge rate of the *anterior deltoid* decreased to 12.68 ± 0.82 pps. Discharge rate variability increased to $21.56 \pm 3.01\%$. With the exoskeleton providing 75% support, the discharge rate for the *anterior deltoid* decreased to 11.96 ± 3.07 pps, while the discharge rate variability increased to $27.57 \pm 3.41\%$.

4 Discussion

In this preliminary investigation, we successfully assessed non-invasively the effects of a hybrid upper-limb exoskeleton on neuromuscular control, with a focus on MU discharge behavior during quasi-isometric contractions performed with different loads and levels of support.

4.1 Motor Unit Discharge Statistics

The MUs discharge, which reflects the neural drive to the muscle, differed across conditions and muscles.

In the *biceps brachii* muscle, the discharge rate slightly increased with 50% exoskeleton support during the 8 kg task but decreased with 75% support. This trend suggests a possible reduction in neural drive detectable only with higher support levels. Conversely, discharge rate variability showed minimal change across conditions, indicating consistent motor command stability during tasks involving the *biceps brachii*.

In the *anterior deltoid* muscle, the MU discharge rate consistently decreased with exoskeleton support across all tasks, although progressively greater differences were noticeable at higher load and support levels. This decrease, together with increased discharge rate variability during higher support levels, suggests that the use of the exoskeleton could have an effect on the neural strategies used to control muscle force production.

The differences observed between the two muscles may be attributed to the mechanical design and function of the exoskeleton. Specifically, as the exoskeleton is engineered to support shoulder flexion movements, its assistance likely targets the shoulder musculature more directly, exerting a greater influence on the *anterior deltoid* compared to the *biceps brachii* muscle. This differential effect highlights the role of task-specific mechanical interactions in modulating neural adaptations during exoskeleton-supported activities.

4.2 Limitations

Since this is a preliminary study based on a single participant, any observation and assumption must necessarily be validated on a larger sample to establish statistical robustness and generalizability. Additionally, inter-individual variability in MU behavior and adaptation to exoskeleton support remains unexplored.

It must also be noticed that given the peculiar mechanics and design of each type of exoskeleton, different exoskeleton models may impose varying mechanical constraints and levels of assistance, potentially leading to different neuromuscular responses. Future studies should investigate whether the observed effects hold across a broader range of exoskeleton types, tasks, load levels, and muscles.

5 Conclusion

Based on the results of this study, the use of HDsEMG proved to be effective in the identification of MU discharge patterns in the *biceps brachii* and *anterior deltoid* muscles during quasi-isometric tasks. Additionally, MU tracking via reapplication of the separation vectors proved to be effective in recognizing the same MU across the tasks performed without or with exoskeleton support.

Preliminary findings suggest that a hybrid upper-limb exoskeleton influences motor unit behavior, with distinct effects observed between muscles and support levels. Furthermore, our findings highlight the potential of HDsEMG to assess exoskeleton induced neuromuscular changes, informing their design and integration in occupational settings.

Acknowledgments. This study was funded by INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro), grant number BRIC 2022, ID-58.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armstrong, T., Bull, F.: Development of the World Health Organization global physical activity questionnaire (GPAQ). J. Public Health. **14**, 66–70 (2006)
2. CWA17938: CWA 17938:2023, Guideline for Introducing and Implementing Real-Time Instrumental-Based Tools for Biomechanical Risk Assessment (2023). https://www.cencenelec.eu/get-involved/research-and-innovation/cen-and-cenelecactivities/cwa-download-area/
3. Frančič, A., Holobar, A.: On the reuse of motor unit filters in high density surface electromyograms recorded at different contraction levels. IEEE Access. **9**, 115227–115236 (2021)
4. Hug, F., Del Vecchio, A., Avrillon, S., Farina, D., Tucker, K.: Muscles from the same muscle group do not necessarily share common drive: evidence from the human triceps surae. J. Appl. Physiol. **130**, 342–354 (2021)
5. Kok, J., Vroonhof, P., Snijders, J., Roullis, G., Clarke, M., Peereboom, K., Dorst, P., Isusi, I.: Work-Related Musculoskeletal Disorders—Prevalence, Costs and Demographics in the EU. European Agency for Safety and Health at Work (2019)
6. Martinez-Valdes, E., Negro, F., Laine, C.M., Falla, D., Mayer, F., Farina, D.: Tracking motor units longitudinally across experimental sessions with high-density surface electromyography. J. Physiol. **595**, 1479–1496 (2017)
7. Negro, F., Muceli, S., Castronovo, A.M., Holobar, A., Farina, D.: Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. J. Neural Eng. **13**, 026027 (2016). https://doi.org/10.1088/1741-2560/13/2/026027
8. Ordieres-Meré, J., Gutierrez, M., Villalba-Díez, J.: Toward the industry 5.0 paradigm: increasing value creation through the robust integration of humans and machines. Comput. Ind. **150**, 103947 (2023)
9. Piervirgili, G., Petracca, F., Merletti, R.: A new method to assess skin treatments for lowering the impedance and noise of individual gelled Ag–AgCl electrodes. Physiol. Meas. **35**, 2101–2118 (2014)
10. Valli, G., Ritsche, P., Casolo, A., Negro, F., De Vito, G.: Tutorial: analysis of central and peripheral motor unit properties from decomposed high-density surface EMG signals with openhdemg. J. Electromyogr. Kinesiol. **74**, 102850 (2024a)
11. Valli, G., Sarto, F., Casolo, A., Del Vecchio, A., Franchi, M.V., Narici, M.V., De Vito, G.: Lower limb suspension induces threshold-specific alterations of motor units properties that are reversed by active recovery. J. Sport Health Sci. **13**, 264–276 (2024b)
12. Varrecchia, T., Ranavolo, A., Chini, G., De Nunzio, A.M., Draicchio, F., Martinez-Valdes, E., Falla, D., Conforto, S.: High-density surface electromyography allows to identify risk conditions and people with and without low back pain during fatiguing frequency-dependent lifting activities. J. Electromyogr. Kinesiol. **73**, 102839 (2023)
13. Varrecchia, T., Ranavolo, A., Conforto, S., De Nunzio, A.M., Arvanitidis, M., Draicchio, F., Falla, D.: Bipolar versus high-density surface electromyography for evaluating risk in fatiguing frequency-dependent lifting activities. Appl. Ergon. **95**, 103456 (2021)
14. Del Vecchio, A., Holobar, A., Falla, D., Felici, F., Enoka, R.M., Farina, D.: Tutorial: analysis of motor unit discharge characteristics from high-density surface EMG signals. J. Electromyogr. Kinesiol. **53**, 102426 (2020). https://doi.org/10.1016/J.JELEKIN.2020.102426

Decision Trees and IMU Sensors for Risk Prediction in Manual Material Handling

Hardik Vora[1], Fatemeh Davoudi Kakhki[1,2(✉)], and Armin Moghadam[3]

[1] Machine Learning and Safety Analytics Laboratory, School of Engineering, Santa Clara University, Santa Clara, CA 95053, USA
{hvora,fatemeh.davoudi}@scu.edu
[2] Department of General Engineering, Department of Electrical and Computer Engineering, Santa Clara University, Santa Clara, CA 95053, USA
[3] Department of Technology, College of Engineering, San Jose State University, San Jose, CA 95053, USA
armin.moghadam@sjsu.edu

Abstract. Manual material handling (MMH) tasks, involving activities like lifting, carrying, and holding various loads, contribute significantly to lower back fatigue, pain, injuries, and musculoskeletal disorders in occupational settings. This study employs machine learning algorithms to assess ergonomic risks in repetitive lifting tasks, aiming to enhance injury prevention, refine training protocols, and inform workload design for improved worker health. Using data from eight Inertial Measurement Unit (IMU) sensors, we developed predictive models to classify lifting tasks based on their biomechanical strain. Ten participants performed repetitive lifts with two box weights, with acceleration and gyroscope data collected. Feature engineering included total acceleration, total angular velocity, and statistical measures computed over a one-second rolling window. Three decision tree-based models achieved classification accuracies of 79.35% (gradient boosting), 91.49% (random forest), and 95.45% (XGBoost). The superior performance of XGBoost highlights the potential of IMU-based machine learning for real-time ergonomic risk assessment. These findings support the development of wearable safety systems for instant feedback, enabling proactive interventions that promote safer work environments and reduce musculoskeletal disorders.

Keywords: Manual material handling · machine learning · ergonomic risk assessment · Inertial measurement unit (IMU) · wearable sensors · Occupational safety

1 Introduction

Work-related Musculoskeletal disorders (WMSDs) caused by manual material handling, such as repetitive or improper lifting, are a significant occupational concern, leading to injuries and reduced productivity [1]. WMSDs are a primary cause of absenteeism and chronic health issues among workers in physically demanding roles [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 353–364, 2025.
https://doi.org/10.1007/978-3-031-93502-2_24

Traditional ergonomic assessments are often subjective, relying on manual observation and retrospective feedback, which can delay necessary interventions [3]. Accurate prediction and classification of ergonomic risks are crucial for improving workplace safety and ergonomics [4]. The application of wearable sensor technology for occupational ergonomics has gained significant traction in recent years.

Various studies have explored the use of wearable devices, particularly Inertial Measurement Units (IMUs), to assess biomechanical risks associated with repetitive lifting, postural deviations, and WMSDs. These studies apply machine learning techniques to improve risk prediction, enhance real-time interventions, and refine workplace safety strategies. Baklouti et al. [5] proposed an IMU-based system for evaluating WMSD risks, tested in a cable manufacturing facility. The system demonstrated high agreement with standard ergonomic assessment methods. By enabling precise joint angle measurements and detecting risky postures, the study highlighted the potential of IMUs in personalized ergonomic interventions. Donisi et al. [6] integrated wearable IMUs with the Revised NIOSH Lifting Equation (RNLE) to evaluate biomechanical risks in manual lifting tasks. Using machine learning classifiers, their model achieved an accuracy exceeding 90%, demonstrating the feasibility of automated lifting risk classification. Zhao et al. [7] focused on the construction industry, where workers often perform repetitive and awkward postures. Their wearable IMU system was designed to detect risky postures, analyze ergonomic risks, and provide feedback via a user-friendly interface. The study emphasized the importance of real-time intervention for injury prevention.

A broader review of sensor technologies for occupational health and safety further supports the growing interest in wearable devices. Fanti et al. [8] provided a systematic review of modern sensing technologies, highlighting the emergence of real-time monitoring systems that integrate multiple sensors, including air quality, noise, and biomechanical exposure monitoring for designing Internet of Things (IoT)-enabled wearable sensors with applications in workplace safety.Sabino et al. [9] explored the role of wearable sensors in healthcare ergonomics, particularly among healthcare professionals who frequently engage in patient handling and other physically demanding tasks. Their study confirmed that wearable IMU and electromyography (EMG) sensors significantly improve ergonomic risk assessment in clinical settings. Davoudi et al. [10, 11] contribute to the field of wearable sensor-based ergonomic risk assessment by integrating EMG sensors with deep learning models to classify biomechanical risks in repetitive lifting tasks. Their approach expands on prior work utilizing IMUs for lifting risk assessment [6, 7] by incorporating high-resolution muscle activation data to enhance risk prediction accuracy. Their findings reinforce the effectiveness of machine learning-driven sensor systems in providing real-time ergonomic risk assessment, aligning with previous studies that highlight the potential of sensor-driven analytics for workplace safety and proactive injury prevention [5, 8].

1.1 Objective of the Study

While previous studies have explored the use of wearable IMUs and machine learning algorithms for ergonomic risk assessment in manual material handling tasks, significant gaps remain. There is a scarcity of research focusing on real-time predictive modeling that incorporates dynamic biomechanical features such as total acceleration and angular

velocity over rolling time windows. For instance, Mudiyanselage et al. [12] utilized surface electromyography (sEMG) sensors combined with machine learning to detect harmful body movements during material handling, achieving high accuracy in risk level prediction. However, their approach primarily focused on muscle activity data rather than dynamic motion features. Similarly, Matijevich et al. [13] developed a wearable solution for monitoring low back loading using IMUs and pressure insoles, emphasizing lumbar load estimation but not real-time predictive modeling of lifting strain.

While prior research has primarily utilized wearable IMUs for postural assessments and generalized risk evaluation, our approach addresses the gap in real-time predictive modeling. This study advances occupational ergonomic risk assessment by developing machine learning models that leverage specific biomechanical features extracted from IMU sensor data to classify and differentiate levels of lifting strain. By incorporating total acceleration, total angular velocity, and statistical measures over a rolling time window, we capture dynamic biomechanical patterns during lifting tasks with greater precision. This enables automated risk differentiation, supporting real-time interventions and personalized ergonomic recommendations. Our method offers a scalable, sensor-driven solution for proactive injury prevention, reducing reliance on subjective assessments and informing adaptive workload design in physically demanding occupations such as manual material handling and repetitive lifting tasks.

2 Methodology

The primary objective of this study was to distinguish between low-risk and high-risk tasks using machine learning algorithms. This proactive methodology aims to prevent injuries, enhance ergonomics, and inform targeted interventions to correct improper lifting techniques and refine training protocols. Employers can also apply these insights to design workloads that minimize fatigue and promote worker health.

2.1 Experimental Set-Up and Data Collection

The Santa Clara University Institutional Review Board (IRB) approved this study (Approval No: 23-12-2086). Prior to participation, all individuals provided informed consent and were informed of their right to withdraw at any time; however, all participants completed the assigned tasks.

This study utilized data from eight Inertial Measurement Unit (IMU) sensors to develop predictive models for assessing risk levels associated with lifting tasks. The dataset was collected from ten participants, each performing repetitive lifting tasks with a 10 lb (light) and an 18 lb (heavy) box to represent varying levels of biomechanical strain.

Each lifting task was classified as either a low-strain or high-strain task based on biomechanical load, with risk levels determined using the Revised NIOSH Lifting Equation (RNLE), as comprehensively discussed in [11]. The experimental setup, including sensor placement and task execution, is illustrated in Fig. 1.

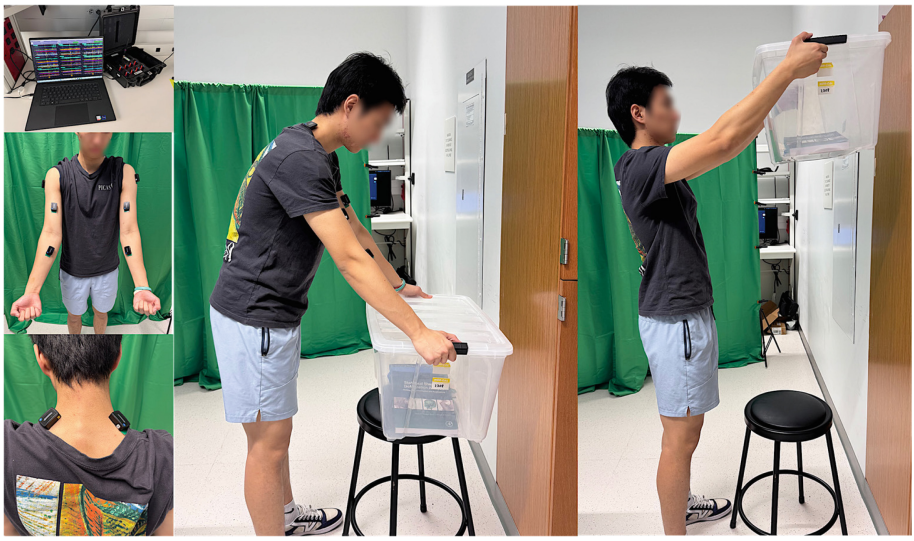


Fig. 1. Experimental setup for assessing biomechanical strain during lifting tasks.

Each participant executed 20 lifting cycles, repeated three times per condition. IMU sensors recorded acceleration and gyroscope data at a sampling frequency of 148 Hz across the X, Y, and Z axes. Participants were equipped with IMU sensors placed on key muscle groups, including the biceps, deltoids, and levator scapulae, as shown in the images below. The lifting tasks involved overhead lifting, placing objects on elevated surfaces, and transferring loads onto waist-level platforms to simulate real-world occupational lifting scenarios.

The dataset comprises 51 columns, including timestamps, participant IDs, acceleration and gyroscope readings for all three axes across all sensors, and a binary target label indicating biomechanical strain level. Each row represents a single recorded moment, capturing instantaneous acceleration and angular velocity, facilitating a detailed temporal analysis of lifting biomechanics.

2.2 Data Preparation for Modeling

Feature engineering was performed to enhance model performance by capturing key movement dynamics. Total acceleration and total angular velocity were computed to represent combined motion across all three axes. Additionally, four statistical measures (mean, standard deviation, maximum, and mean absolute deviation) were calculated over a one-second rolling window to capture temporal variations in both acceleration and gyroscope data.

These engineered features were used as inputs for three decision tree-based machine learning models to classify task-related risk levels. The extracted features, their mathematical formulas, and their significance in distinguishing lifting strain levels are summarized in Table 1. These features quantify movement intensity, variability, peak exertion, and deviations, contributing to ergonomic risk differentiation. Using a rolling window

of 148 data points (1 s) ensures that time-dependent variations in lifting motion are captured, improving the model's ability to distinguish between lower and higher biomechanical strain levels. These engineered features enhance the predictive accuracy of machine learning models for real-time ergonomic risk assessment.

Table 1. Extracted feature and contribution to biomechanical risk prediction.

Feature	Formula	Contribution to risk prediction
Total acceleration magnitude	$ACC_{mag} = \sqrt{(ACC_X)^2 + (ACC_Y)^2 + (ACC_Z)^2}$	Overall movement intensity during lifting
Total angular velocity magnitude	$Gyro_{mag} = \sqrt{(Gyro_X)^2 + (Gyro_Y)^2 + (Gyro_Z)^2}$	Total rotation motion of the body
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	Average movement magnitude over time
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	Variability in motion patterns
Maximum	$x_{max} = \max(x_1, x_2, \ldots, x_N$	Identifies peak movement intensity
Mean absolute deviation	$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \mu\rvert$	Average variability of data points from the mean, and fluctuations around the average movement

2.3 Machine Learning Classifiers for Strain Level Prediction

In this study, we employ four classification models to assess binary risk classification including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. Each model was trained and evaluated to gain robust comparison of their capabilities in binary risk classification of the manual material handling task in this study.

Logistic Regression is a statistical method used for binary classification which estimates the probability of an instance belonging to a particular class using the sigmoid function. The probability of the target class is calculated, where β_0 represents the intercept, and β_i are coefficients learned during the model training, as in [13]:

$$P\left(Y = 1 \mid X = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}\right)$$

Random Forest [14] is an ensemble learning technique that constructs multiple decision trees and aggregates their outputs to improve accuracy and reduce overfitting. The final

prediction is obtained by averaging the predictions of individual trees, where T is the number of tress and $f_t(X)$ represents the prediction score for each tree.

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X)$$

Gradient Boosting follows a sequential learning process where decision trees are built iteratively, correcting errors from previous trees [15]. The model updates its predictions using the gradient of the loss function:

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X)$$

where $F_m(X)$ is the updated model, $h_m(X)$ is the newly added weak learner and γ_m is the step size optimized at each iteration.

XGBoost extends Gradient Boosting with optimized computation and regularization. Its objective function consists of both the loss function and a regularization term to prevent overfitting:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(T_k)$$

where L $(y_i, \hat{y}_i)$ shows the loss between the predicted and actual values, and Ω (T_k) is the regularization term to control model complexity [16].

3 Results

This section presents the comparative performance of different machine learning models in classifying biomechanical risk levels based on inertial sensor data. Model accuracy, precision, recall, and F1-score were evaluated to assess classification effectiveness. Additionally, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) analysis were conducted to examine each model's discriminatory power. Finally, the implications of these findings for real-time ergonomic risk assessment and workplace safety are discussed, highlighting the potential integration of machine learning-driven assessments into wearable monitoring systems.

3.1 Model Comparison

The overall accuracy of the four models is deicted in Fig. 2. It shows that XGBoost has superior performance in classifying biomechanical risk levels. The classification accuracies of the tree-based machine learning models were 79.35% for gradient boosting, 91.49% for random forest, and 95.45% for XGBoost decision trees. However, accuracy alone does not provide a complete assessment of model effectiveness, particularly in scenarios where the cost of misclassification is high.

To gain deeper insights, we examined precision, recall, and F1-score, which together offer a more comprehensive evaluation of the models classification reliability.

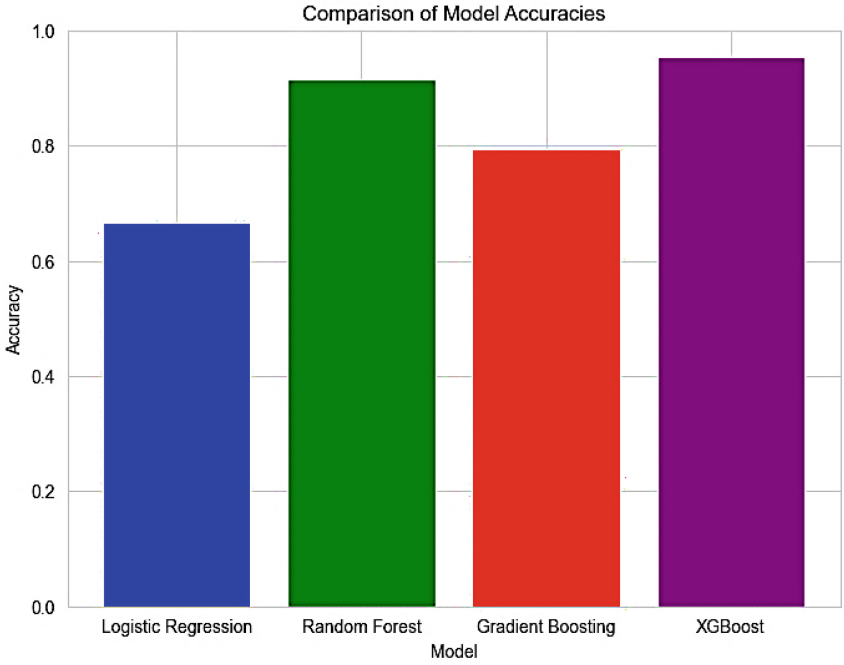


Fig. 2. Comparison of model accuracies.

Precision quantifies the proportion of movements classified as high-strain that were actually high-strain. A high precision score indicates that the model minimizes false positives, reducing the likelihood of incorrectly flagging low-strain movements as high-strain. XGBoost (96%) and Random Forest (92%) demonstrated the highest precision, signifying their ability to make high-confidence classifications with minimal false alarms. In contrast, Gradient Boosting (78%) and Logistic Regression (65%) exhibited lower precision, indicating a greater tendency to erroneously classify low-strain movements as high-strain, potentially leading to unnecessary ergonomic interventions.

Recall (Sensitivity) measures how well the model correctly identifies high-strain movements, ensuring that truly risky lifting tasks are flagged for intervention. A model with high recall effectively minimizes false negatives, reducing the chances of overlooking hazardous movements. XGBoost (96%) and Random Forest (93%) excelled in recall, demonstrating their capability to capture a wide range of high-strain movements, making them highly effective for preventing workplace injuries. On the other hand, Gradient Boosting (84%) and Logistic Regression (68%) showed lower recall, indicating that they failed to detect a significant number of high-strain movements, which limits their applicability for real-time ergonomic assessments. The F1-score, which is the harmonic mean of precision and recall, provides a balanced metric that accounts for both false positives and false negatives. XGBoost (0.96) and Random Forest (0.92) achieved the highest F1-scores, confirming their consistency in differentiating movement risks

with both high precision and high recall. Gradient Boosting (0.79) and Logistic Regression (0.67) performed considerably worse, underscoring the limitations of traditional classification methods in analyzing sensor-based movement data.

Overall, the high F1-scores of XGBoost and Random Forest validate their suitability for real-time ergonomic risk assessment, ensuring that both low- and high-strain movements are classified with high accuracy while minimizing misclassification errors. These models demonstrate strong potential for deployment in wearable safety monitoring systems, offering reliable biomechanical risk detection that can aid in injury prevention and ergonomic optimization.

3.2 Evaluating Model Discrimination and AUC Analysis

The ROC curve (Fig. 3) provides a graphical representation of the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different classification thresholds for each model. The AUC score serves as a quantitative measure of a model ability to distinguish between different biomechanical strain levels. A model with an AUC closer to 1.0 demonstrates superior discriminatory power, whereas a model with an AUC closer to 0.5 performs no better than random guessing.

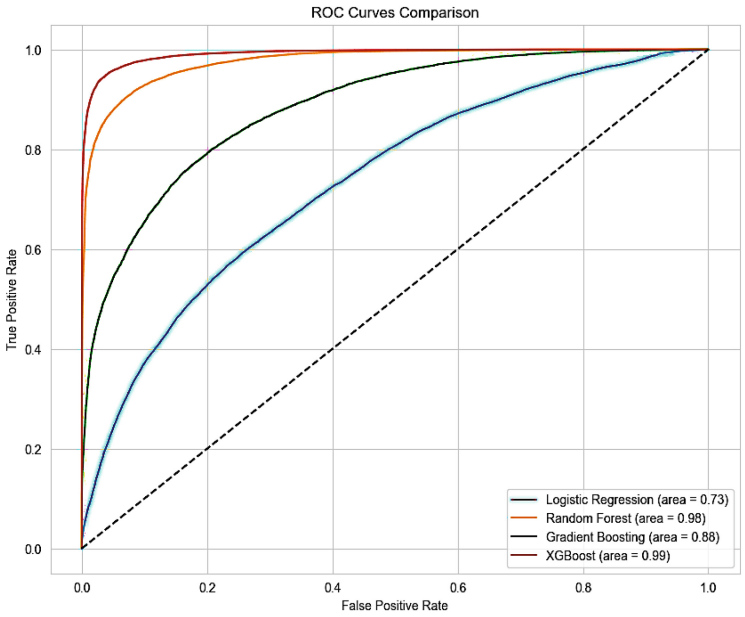


Fig. 3. Comparison of model ROC-AUCs.

Among the models tested, XGBoost achieved an AUC of 0.99, reflecting near-perfect classification performance. This high AUC confirms XGBoost's ability to clearly separate high-strain and low-strain lifting movements, minimizing both false positives and

false negatives. Random Forest followed closely with an AUC of 0.98, indicating high reliability in classifying lifting strain, though slightly less optimized than XGBoost. Gradient Boosting, with an AUC of 0.88, outperformed traditional methods but fell short of the ensemble-based models, suggesting it is a viable yet suboptimal choice for real-time ergonomic risk classification. In contrast, Logistic Regression had the lowest AUC of 0.73, revealing its limited ability to capture complex relationships in IMU-based movement data. A higher AUC score signifies a more effective risk classification model, as it reflects the model ability to correctly distinguish between safe and potentially hazardous lifting movements. The near-optimal AUC values of XGBoost and Random Forest reinforce their suitability for real-time workplace monitoring and ergonomic risk assessment applications. These models offer greater precision in movement classification, ensuring that risky lifting behaviors are accurately detected, thereby enhancing injury prevention strategies in physically demanding occupational environments.

3.3 Implications for Ergonomic Risk Assessment and Workplace Safety

The superior accuracy and classification performance of XGBoost (95.5%) and Random Forest (91.5%) underscore their potential for deployment in real-world ergonomic assessments. Integrating these models into smart wearable devices can provide real-time feedback on lifting techniques, alerting workers and supervisors to potentially hazardous movements before injuries occur. The high precision and recall scores of XGBoost ensure minimal false alarms, making it particularly suitable for preventive safety interventions. Through implementing machine learning-driven sensor analysis, organizations can transition from manual, subjective ergonomic evaluations to automated, data-driven assessments. This shift enables several key advancements such as real-time risk monitoring, data-driven training programs, and injury prevention frameworks.

As for the real-time risk monitoring, deploying these models in industrial, healthcare, and logistics settings allows for the detection and mitigation of improper lifting techniques before injuries occur. For instance, wearable devices equipped with inertial measurement units (IMUs) can monitor workers' movements and provide immediate feedback, reducing the risk of musculoskeletal disorders (MSDs) [12, 17].

Furthermore, developing personalized ergonomic training protocols based on actual sensor data helps workers adopt safer movement patterns. Machine learning models can analyze complex movement data to identify risky behaviors and inform targeted interventions [18]. Regarding the injury prevention frameworks, safety practitioners and supervisors and occupational health professionals can use these models to identify high-risk individuals and implement targeted interventions to reduce the incidence of MSDs. Automated assessments provide objective data, enhancing the effectiveness of injury prevention strategies [19].

4 Conclusion

This study demonstrates the effectiveness of machine learning models in differentiating biomechanical strain levels using IMU sensor data, advancing real-time ergonomic risk assessment in occupational settings. The high classification accuracy of XGBoost (95.5%) and Random Forest (91.5%) highlights their potential for deployment in smart wearable monitoring systems to prevent musculoskeletal disorders (MSDs) and improve workplace safety. By leveraging automated, sensor-driven assessments, organizations can shift from traditional manual ergonomic evaluations to real-time, data-informed interventions, optimizing both worker well-being and operational efficiency.

Despite its strong performance, this study has several limitations. First, the dataset was collected from only 10 participants, which may not fully capture inter-individual variations in lifting techniques and biomechanical responses. A larger and more diverse sample population is necessary to enhance the generalizability of the model. Second, the study focused exclusively on IMU sensor data, omitting other potential biomechanical indicators such as EMG signals, force measurements, and heart rate variability, which could provide a more comprehensive risk assessment. Third, while XGBoost and Random Forest performed well, real-time deployment in industrial settings requires further validation to assess computational efficiency and adaptability across different work environments. Additionally, the current approach does not account for long-term fatigue accumulation, which is a critical factor in ergonomic risk assessment.

Future research should address these limitations by expanding the participant pool to include workers from diverse occupations and demographics, ensuring robust generalizability of the models. Integrating multi-modal sensor data, such as EMG, pressure sensors, and computer vision-based motion tracking, could further enhance model performance by capturing muscle activity, joint angles, and load distribution. Moreover, deep learning techniques, including recurrent neural networks and transformer-based models, could be explored to improve temporal pattern recognition in lifting movements. Another important avenue for research is the development of real-time feedback mechanisms, where machine learning models not only classify risk but also provide instant corrective suggestions through wearable devices or augmented reality interfaces. Finally, future studies should examine the longitudinal impact of these models on reducing workplace injuries, ensuring that data-driven ergonomic interventions translate into tangible health benefits for workers.

By addressing these challenges, future research can further refine the role of machine learning in ergonomic risk assessment, ultimately contributing to safer workplaces, improved worker productivity, and the development of intelligent occupational safety systems.

Acknowledgments. We thank the students who volunteered and participated in this research. We also thank other research assistants who contributed to the data collection process and conducting experiments: Arnold Nieto and Paul Li.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lazzaroni, M., et al.: Improving the efficacy of an active Back-support exoskeleton for manual material handling using the accelerometer signal. IEEE Robot. Autom. Lett. **7**(3), 7716–7721 (2022). https://doi.org/10.1109/LRA.2022.3183757
2. Daruis, D.D.I., Rosly, A.L., Abd Aziz, I., Hishamuddin, N.S., Md Deros, B.: Ergonomic risk assessment of manual material handling at an automotive manufacturing company. Dent. Press. **5**(1), 317–324 (2017). https://doi.org/10.17261/pressacademia.2017.606
3. Davoudi Kakhki, F., Moghadam, A., Nieto, A., Vora, H.: Evaluating the efficacy of a passive exoskeleton for enhancing ergonomics in manufacturing. Appl. Sci. **14**(13), 5810 (2024). https://doi.org/10.3390/app14135810
4. Stoidner, M., Brundl, P., Nguyen, H.G., Baechler, A., Franke, J.: Workplace analysis and ergonomics in engineer-to-order production sites: a study on the workplace Design of Control Cabinet Manufacturing Enterprises. In: 2023 IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2023, pp. 838–842. Institute of Electrical and Electronics Engineers Inc. (2023). https://doi.org/10.1109/IEEM58616.2023.10406509
5. Baklouti, S., Chaker, A., Rezgui, T., Sahbani, A., Bennour, S., Laribi, M.A.: A novel IMU-based system for work-related musculoskeletal disorders risk assessment. Sensors. **24**(3419) (2024). https://doi.org/10.3390/s24113419
6. Donisi, L., Cesarelli, G., Coccia, A., Panigazzi, M., Capodaglio, E.M., D'Addio, G.: Work-related risk assessment according to the revised NIOSH lifting equation: a preliminary study using a wearable inertial sensor and machine learning. Sensors. **21**(2593) (2021). https://doi.org/10.3390/s21082593
7. Zhao, J., Obonyo, E., Bilén, S.G.: Wearable inertial measurement unit sensing system for musculoskeletal disorders prevention in construction. Sensors. **21**(1324) (2021). https://doi.org/10.3390/s21041324
8. Fanti, G., Spinazzè, A., Borghi, F., Rovelli, S., Campagnolo, D., Keller, M., Borghi, A., Cattaneo, A., Cauda, E., Cavallo, D.M.: Evolution and applications of recent sensing technology for occupational risk assessment: a rapid review of the literature. Sensors. **22**(4841) (2022). https://doi.org/10.3390/s22134841
9. Sabino, I., Fernandes, M.C., Cepeda, C., Quaresma, C., Gamboa, H., Nunes, I.L., Gabriel, A.T.: Application of wearable technology for the ergonomic risk assessment of healthcare professionals: a systematic literature review. Int. J. Ind. Ergon. **100**, 103570 (2024). https://doi.org/10.1016/j.ergon.2024.103570
10. Davoudi Kakhki, F., Vora, H., Moghadam, A., Kyrarini, M.: Musculoskeletal risk modeling using sensors and machine learning. In: IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (2024). https://doi.org/10.1109/IEEM62345.2024.10857250
11. Davoudi Kakhki, F., Vora, H., Moghadam, A.: Biomechanical risk classification in repetitive lifting using multi-sensor electromyography data, revised National Institute for Occupational Safety and Health lifting equation, and deep learning. Biosensors. **15**(84) (2025). https://doi.org/10.3390/bios15020084
12. Mudiyanselage, S.E., Nguyen, P.H.D., Rajabi, M.S., Akhavian, R.: Automated workers' ergonomic risk assessment in manual material handling using sEMG wearable sensors and machine learning. Electronics. **10**(20), 2558 (2021)
13. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. John Wiley & Sons (2013)
14. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
15. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)

16. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining, pp. 785–794 (2016)
17. Matijevich, E.S., Volgyesi, P., Zelik, K.E.: A promising wearable solution for the practical and accurate monitoring of low back loading in manual material handling. Sensors. **21**(2), 340 (2021)
18. Stefana, E., Marciano, F., Rossi, D., Cocca, P., Tomasoni, G.: Wearable devices for ergonomics: a systematic literature review. Sensors. **21**(3), 777 (2021)
19. Kuber, P.M., Kulkarni, A.R., Rashedi, E.: Machine learning-based fatigue level prediction for exoskeleton-assisted trunk flexion tasks using wearable sensors. Appl. Sci. **14**(11), 4563 (2024)

Simulation and Analysis of Cervical Biomechanics and Injury Risk in Fighter Pilots Under Different Seatback Angles

Jiatao Wang[1,2], Chuancai Sun[1,2], Weisheng Jiang[1,2], Qianxiang Zhou[1,2(✉)], and Zhongqi Liu[1,2]

[1] School of Biological Science and Medical Engineering, Beihang University, Beijing, China
zqxg@buaa.edu.cn

[2] Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing, China

Abstract. This study investigates the biomechanical effects of different seat inclinations on the cervical spine of fighter jet pilots under high-load conditions. It assesses the stress, strain, and injury risk to the cervical spine during two flight scenarios: rapid pull-up and stable circling maneuvers. Based on the 50th percentile pilot anthropometric data from Chinese military standards, a detailed finite element model of the cervical spine, including vertebrae, intervertebral discs, and ligaments, was constructed. The model's validity was confirmed by comparing the experimental data of vertebral range of motion and axial impact. By applying acceleration profiles corresponding to rapid pull-up and stable circling maneuver conditions, the biomechanical responses of pilots were simulated at seat inclinations of 17° and 22°. The study analyzes the effects of different seat angles on cervical spine stress, strain, and displacement. It employs the Neck Injury Criterion (NIC) to assess the risk of spinal cord injury in the neck. Under both conditions, a seat inclination of 22° significantly increased the stress and displacement in the vertebrae and intervertebral discs, though it did not reach the threshold for direct injury. The C5-C6 intervertebral disc experienced the highest stress during rapid pull-ups, indicating that the lower cervical spine bears greater loads under high-load conditions. The C3–C4 and C4–C5 intervertebral discs showed a notable increase in stress during stable circling maneuvers, reflecting differences in loading patterns under varying flight conditions. The NIC values for both seat angles did not reach the threshold for spinal cord injury; however, higher seat inclinations resulted in elevated NIC values, suggesting potential risks associated with prolonged exposure to high-load environments. Greater seat back inclinations increase the biomechanical load on pilots' cervical spines, elevating the risk of neck-related ailments. This study provides scientific evidence for the design of fighter jet seats and pilot protection strategies. It is recommended that seat inclination effects on the cervical spine be comprehensively considered in the design process to ensure pilots' long-term health and safety.

Keywords: pilot · seatback angle · biomechanics · neck injury · finite element analysis

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 365–377, 2025.
https://doi.org/10.1007/978-3-031-93502-2_25

1 Introduction

Advancements in the speed and maneuverability of modern fighter aircraft have subjected pilots to immense physiological challenges during high-intensity maneuvers, particularly under high-acceleration environments. Elevated loads not only threaten pilots' cardiopulmonary and circulatory systems but also significantly impact regions subjected to substantial stress, notably the cervical spine. Specifically, structural tissues such as cervical muscles, ligaments, and intervertebral discs are susceptible to injuries [1]. Surveys reveal that over 51% of fighter pilots experience varying degrees of neck pain and discomfort throughout their careers [2], with the incidence of neck pain among Chinese pilots reaching as high as 33.7% [3]. These injuries are predominantly associated with high-acceleration tasks and commonly present as muscle strains, intervertebral disc herniation, and ligament tears [1, 4].

Research indicates that cervical injuries are not solely attributable to high loads but are also considerably influenced by multiple factors, including seat-back angles, helmet weight, and flight maneuvers [5–7]. These external factors can amplify the cervical load induced by high acceleration, thereby elevating the risk of injury. For instance, varying seat-back angles alter the pilot's head position, leading to changes in the activation of cervical muscles and affecting the mechanical load distribution across the cervical spine [8]. Additionally, the weight of the flight helmet imposes extra pressure on the neck, especially during high-maneuver flights where rapid posture changes further exacerbate cervical stress [5, 9]. Neck pain not only diminishes pilots' operational capabilities but can also severely impact mission efficiency, potentially necessitating rehabilitation or surgical interventions, interrupting flight training, and jeopardizing the continuity of combat missions [10].

Modern fighter aircraft seats are increasingly designed with larger seat-back angles to enhance pilots' visibility and operational comfort. Currently, seat-back angles of 17° and 22° are relatively common, with some models, such as the F-16, featuring ejection seats with angles as steep as 30° [3]. However, while increasing the seat-back angle can somewhat improve visibility and comfort, it raises concerns about increased cervical load and injury risk. Zhang et al. [11] found that a 20° seat-back angle, compared to a vertical seat-back, causes pilots' heads and necks to extend backward, significantly stressing the cervical flexor muscles. Similarly, using finite element model, Liu Chunjie et al. [12] validated that larger seat-back angles can lead to ligamentous strains during emergency ejections.

Nevertheless, existing studies predominantly focus on specific seat-back angles or flight postures, lacking systematic numerical simulations of complex flight conditions (such as rapid ascents and turns) under varying seat-back angles. This limitation hinders the optimization of seat design to a certain extent. Hence, this study constructs a detailed finite element model of the cervical spine to simulate the biomechanical responses of pilots under high-acceleration rapid ascent and stable turning conditions. By integrating injury assessment criteria, the study predicts the risk of cervical injuries. The findings aim to provide scientific evidence for optimizing fighter aircraft seat designs, reducing the risk of cervical injuries among pilots, and enhancing flight safety.

2 Methods

2.1 Finite Element Model Establishment

Geometric Model Establishment. This study recruited a single healthy male volunteer whose anthropometric measurements align with the 50th percentile of Chinese military-standard fighter pilots, specifically with a height of 170.4 cm (±1 cm) and a weight of 56.5 kg (±2 kg), and who was rigorously screened to exclude any cervical spine pathologies, deformities, degenerative changes, or previous neck injuries. Data was acquired using a 64-slice spiral CT scanner with parameters set to 120 V voltage, 30 mA current, 1.3 mm slice thickness, and a 0.625 mm slice interval. The scanning range extended from the skull base to the first thoracic vertebra (T1), and the resulting axial images served as the experimental data.

The DICOM images obtained from the CT scans were imported into Mimics 17.0 software, where bone gray-scale information was extracted through threshold segmentation to generate a three-dimensional skeletal model spanning C1 to T1. This model was subsequently imported into Geomagic Studio software, utilizing the "Polygon" and "Precision Surface" modules to convert the point cloud data into a geometric surface model. Two adjacent vertebrae were selected to construct complete intervertebral discs, with disc boundaries appropriately trimmed at the base of the upper vertebra. Shell operations were then employed to generate the disc models. Following this, the "Scaling" function delineated the annulus fibrosus and nucleus pulposus regions, achieving precise modeling of the intervertebral discs (Fig. 1).

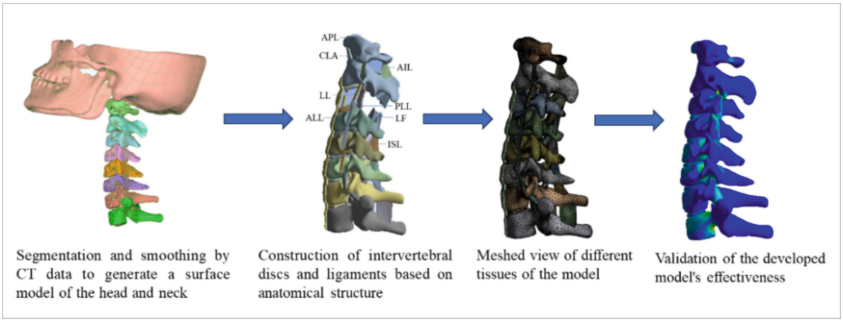


Fig. 1. Cervical spine finite element model main flow chart.

Ligaments are critical soft tissues within the cervical region, primarily responsible for bearing tensile or traction forces. Due to the limited soft tissue resolution of CT scans, this study simulated cervical ligaments using solid elements based on anatomical knowledge. The ligaments modeled include the cruciate ligament of the atlas, alar ligaments, anterior longitudinal ligament, posterior longitudinal ligament, ligamentum flavum, interspinous ligaments, intertransverse ligaments, and the apical ligament of dens.

Finite Element Model Establishment. Based on the structural characteristics of cervical spine tissues, the cortical bone of the vertebrae was modeled using Shell elements with

a thickness of 1.5 mm. All other tissues were represented using 10-node second-order tetrahedral elements. Element sizes ranged between 0.2 mm and 2 mm to enhance computational efficiency without compromising accuracy, resulting in a mesh comprising approximately 508,500 elements and 781,300 nodes. Subsequently, mesh partitioning parameters were optimized, and lower-quality meshes were refined to ensure adherence to required quality standards.

Material Attribute In the finite element model of the cervical spine, material properties for each tissue were selected based on extensive experimental data from domestic and international studies and relevant literature on existing human finite element models. Material attributes for various model components were determined by integrating experimental data available in references [13–15]. Specific parameters are detailed in Table 1.

Table 1. Materials properties of cervical tissue.

Biological tissue	Density (kg/m^3)	Poisson's ratio	Young modulus (MPa)	Cross sectional area (mm$^{2)}$)
Cortical bone	2000	0.3	1200	–
Cancellous bone	1000	0.45	0.5	–
Cancellous bone	1000	0.49	5	–
Annulus fibrosus	1200	0.45	500	–
Apical ligament of dens	800	0.4	20	5
Cruciate ligament of the atlas	800	0.3	40	30
Alar ligament	800	0.3	40	20
Anterior longitudinal ligament	800	0.3	40	6.1
Posterior longitudinal ligament	800	0.3	40	5.4
Interspinous ligament	800	0.3	20	5.1
Intertransverse ligament	800	0.3	20	20
Ligamentum flavum	800	0.3	20	50.1

Assembly of Model. Intervertebral discs and ligaments primarily connect the cervical vertebrae, with intervertebral discs linking the cartilage endplates between adjacent vertebral bodies to provide significant stability. In the model, contact interactions between the annulus fibrosus of the intervertebral discs and the vertebrae and between the ligaments and the vertebrae were defined as bonded connections. To accurately simulate the tensile behaviour of the ligaments, relative sliding between vertebrae was permitted. Based on these contact interactions, a pinball radius of 1 cm was applied to regulate sliding friction within this contact range.

2.2 Verification of the Finite Element Model

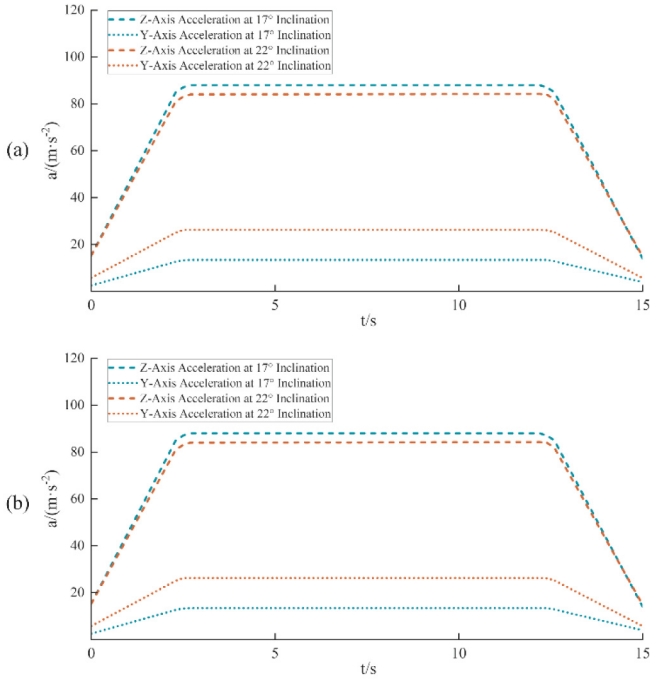
After constructing the finite element model of the pilot's cervical spine, it was essential to validate its effectiveness prior to conducting simulation analyses. Current experimental methods for validating head and neck finite element models primarily include cranio-caudal (head-to-foot) and thoracic-dorsal (chest-to-back) validations. This study focuses primarily on axial load directions; therefore, Panjabi et al.'s [16] research on the mechanical characteristics of the human cervical spine and Nightingale et al.'s [17] dynamic response study of cadaveric head and neck axial impact tests were selected to validate the model's effectiveness from both static and dynamic perspectives.

Static Verification. Panjabi et al. [16] conducted load applications on fresh cadaveric cervical spines to obtain load-displacement curves in flexion-extension, rotation, and lateral bending directions. To validate the model's static performance, constraints were applied to the lower surface of the T1 thoracic vertebra, restricting all translational and rotational degrees of freedom. A reference point was established at the center of rotation, and pure moments of 1.0 N m were applied in six directions to simulate cervical spine movements.

Dynamic Verification. Nightingale et al. [17] performed cervical axial impact experiments at various angles, with the 0° impact test results closely aligning with the experimental conditions of this study. Based on this experimental data, an initial velocity of 3.2 m/s was set, and a 16 kg mass point representing torso weight was coupled to the T1 vertebra in the model. The connection between the head-neck region and the rigid platen was established using a frictionless contact method to simulate the dynamic impact process accurately.

2.3 Flight Load Curve

This study selected acceleration curves from two typical operational scenarios to simulate flight conditions under high-load environments: rapid ascent and stable circling maneuvers. These acceleration profiles were used as boundary conditions and applied to the head-neck model for dynamic analysis. These scenarios reflect the high-stress environments that fighter pilots frequently encounter during mission execution (Fig. 2).



(a) Rapid pull-up acceleration (b) Stable circling maneuvers acceleration

Fig. 2. Acceleration load curve.

3 Results and Discussion

3.1 Verification Result

Result of Static Verification. As examples of flexion and extension processes, the simulation calculated the range of motion (ROM) between adjacent vertebrae and conducted a comparative analysis with experimental data, as shown in Fig. 3. The figure indicates that the ROM of the C1-C2 cervical vertebrae is significantly greater than that of the other cervical vertebrae. The numerical differences in ROM between the remaining vertebrae are minimal and consistent with the experimental results. The primary reason for the substantial ROM differences is that the C2–C7 vertebrae are connected by intervertebral discs, which provide stability. In contrast, no intervertebral discs exist between the C1 and C2 vertebrae. The connection between C1 and C2 is established through the inferior articular facets of the lateral mass of the atlas (C1) and the superior articular facets of the axis (C2). Additionally, the anterior articular facets of the odontoid process of C2 connect to the posterior portion of the anterior arch of C1, resulting in fewer constraints and a more extensive range of motion.

The finite element model constructed in this study exhibits a higher degree of refinement than in vitro experiments. The ROM values obtained from the simulations are slightly smaller but remain within the error margins of the in vitro experimental data.

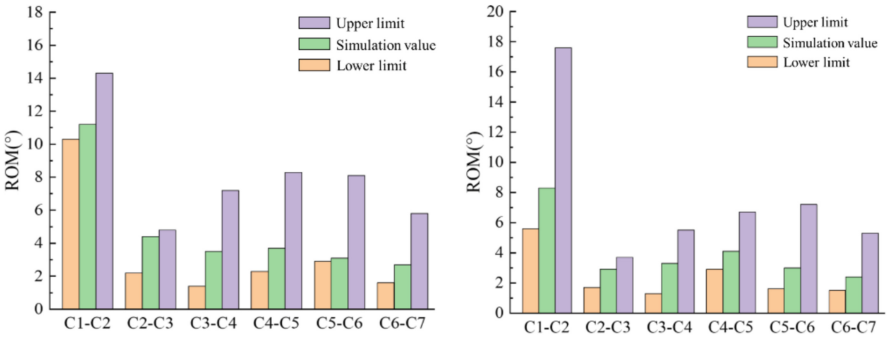


Fig. 3. The comparison of ROM of flexion.

Overall, the trends observed in the simulation data are broadly consistent with the experimental data, thereby validating the model's accuracy under static conditions.

Result of Dynamic Verification. According to records in reference [17], after the head and neck are subjected to impact forces, the upper cervical segments C2-C3 and C3-C4 experience extension deformations, while the lower cervical vertebrae exhibit more significant flexion deformations. By comparing the simulation results with the movement trends observed in human experiments, it was found that the simulated cervical flexion deformation patterns closely align with the experimental outcomes. However, the magnitude of deformation in the simulations is slightly smaller than that observed in cadaveric tests.

Comparing the first 10 ms of the simulation results with the experimental curves from the literature (as shown in Fig. 4), it was observed that the simulated peak values of head collision forces are slightly higher than those of the experimental curves, and the response times exhibit a slight delay. Although the simulation curves do not entirely fall within the upper and lower bounds of the experimental data, their variation trends are highly consistent with the experimental results. This indicates that the model effectively reflects the biomechanical responses of the head and neck under dynamic impacts.

Through comparative validation, the cervical model's simulation results maintain a consistent trend with classic experimental data. This demonstrates that the constructed finite element model of the pilot's cervical spine possesses good biological fidelity and can accurately represent the biomechanical responses of the head and neck during maneuvering flights. Therefore, this model can be utilized for dynamic simulation analyses of sustained load conditions during maneuvering flights.

3.2 Simulation Result

In the pilot's cervical finite element model simulation, different seatback angles appreciably influenced the vertebrae and intervertebral discs' stress, strain, and displacement under two operational conditions: rapid ascent and stable circling maneuvers. The following sections provide a detailed analysis of the main results.

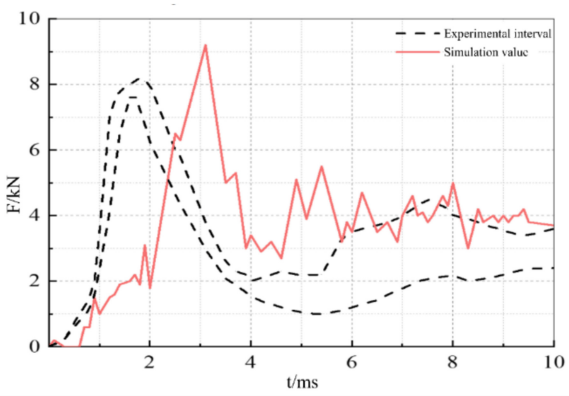


Fig. 4. The comparison of head impact force data.

Vertebral Result. Under rapid pull-up conditions, the maximum stress on the vertebrae at a seat inclination angle of 17° is 40.85 MPa, which increases to 48.86 MPa at a 22° inclination, representing an approximate increase of 19.6%. Although the stress substantially increases, it remains well below the threshold for direct vertebral injury (221 MPa) [18]. However, in a long-term high-load environment, this stress increase may lead to fatigue accumulation in the vertebral microstructure, thereby enhancing the potential risk of fractures. Regarding displacement, vertebral displacement is 12.04 mm at a 17° inclination and increases to 14.51 mm at a 22° inclination, reflecting an approximate increase of 20.5%. The increase in vertebral displacement indicates that, at larger inclination angles, the relative movement between vertebrae is amplified, which may cause excessive stretching of the intervertebral discs and surrounding ligaments, leading to abnormal changes in the uncovertebral joints and further compromising the overall stability of the cervical spine.

In stable circling maneuver conditions, vertebral stress is 25.63 MPa at a 17° inclination and dramatically increases to 55.39 MPa at a 22° inclination, representing a substantial increase of 116.1%. This escalation indicates that larger inclination angles markedly increase the stress load on the cervical spine structure in lateral and axial load environments, potentially due to the augmented shear forces experienced during stable circling maneuvers. Although these stress values do not reach the threshold for direct injury, the sustained increase in stress may have chronic effects on the vertebral microstructure, such as localized reductions in bone density and stress concentration in the cortical bone layer, potentially leading to decreased mechanical performance of the bone.

Vertebral displacement is 3.89 mm at a 17° inclination and increases to 5.33 mm at a 22° inclination, reflecting an approximate increase of 37.1%. This increase in displacement indicates that, at larger inclination angles, the relative displacement between vertebrae is significantly amplified, which may result in chronic wear of the uncovertebral joints, intervertebral discs, and ligaments, thereby reducing cervical spine stability and increasing the risk of long-term injuries (Fig. 5).

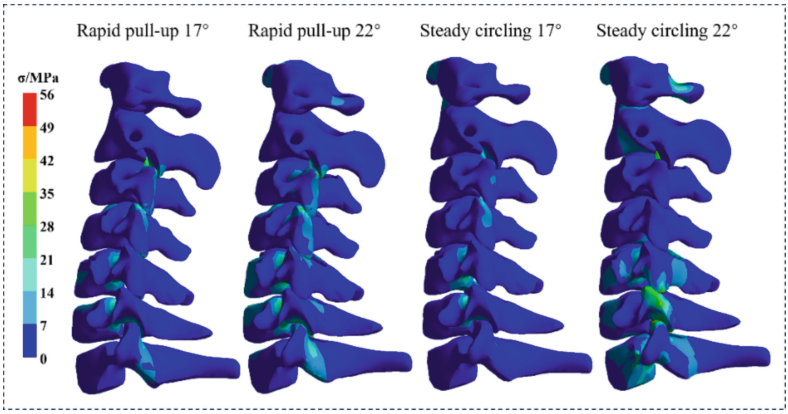


Fig. 5. The maximum stress nephogramin of cervical cortex of pilots at different seatback angles.

Intervertebral Disc Result. Since the nucleus pulposus is gelatinous, the stress on the intervertebral discs is primarily concentrated on the annulus fibrosus. To further understand the load conditions of the intervertebral discs under different seat inclination angles, a detailed analysis of the maximum stress in each annulus fibrosus was conducted, as shown in Fig. 6. Simulation results indicate that under rapid pull-up conditions, the maximum stress on the intervertebral discs ranges from 5.88 MPa (C4–C5) to 10.04 MPa (C5–C6) at a 17° seat inclination, increasing to 7.58 MPa (C4–C5) to 12.90 MPa (C5–C6) at a 22° inclination, representing an approximate increase of 27% to 29%. The C5–C6 intervertebral discs exhibit the highest stress levels and the most significant increases under both inclination angles, likely due to more excellent axial compression and shear forces experienced at larger angles, making them stress concentration regions, which may lead to microstructural injury in the annulus fibrosus of the intervertebral discs, resulting in the accumulation of microcracks and increasing the risk of nucleus pulposus herniation. This stress distribution pattern suggests that the lower cervical spine is particularly susceptible to tissue injury during rapid elevation, and long-term accumulation may lead to degenerative disc disease in pilots.

Under stable circling conditions, the maximum stress on the intervertebral discs increases significantly with the seat inclination angle; however, the magnitude of this stress increase varies markedly among different discs. At a 17° inclination, the stress ranges from 3.02 MPa (C3–C4) to 8.56 MPa (C5–C6), increasing to 5.89 MPa (C5–C6) to 9.97 MPa (C4–C5) at a 22° inclination. Notably, the C3–C4 and C4–C5 intervertebral discs exhibit stress increases of 182.5% and 96.3%, respectively, which are substantially higher than those of other discs. This disparity may be attributed to the location of these discs in the mid-upper cervical spine, where they are subjected to more significant multiaxial stresses due to rotational and lateral shear forces. Consequently, the annulus fibrosus experiences multiaxial tensile stresses, exacerbating stress concentration effects. Conversely, the stress on the C5–C6 intervertebral disc decreases by 31.2% at a 22° inclination as the upper discs bear more load, alleviating the pressure on the lower segments. These findings indicate that the loading patterns of the intervertebral discs

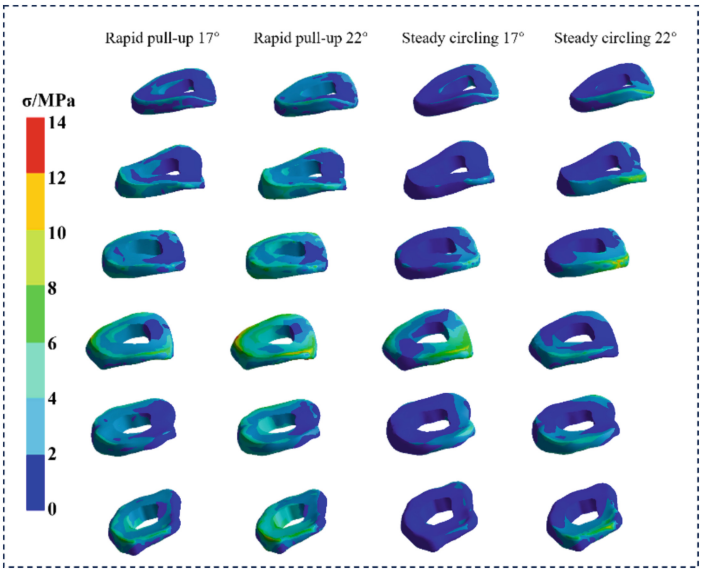


Fig. 6. The maximum stress nephogramin of Intervertebral disc of pilots different seatback angles.

are more complex during stable circling, and changes in inclination angles significantly affect the mechanical responses of the cervical spine.

In addition to stress, the strain and displacement of the intervertebral discs also increase markedly with the seat inclination angle. Under rapid pull-up conditions, strain increases from 0.15 to 0.19, an approximate increase of 26.7%, and displacement increases from 4.68 mm to 6.22 mm, an approximate increase of 32.9%. Under stable circling conditions, strain increases from 0.17 to 0.20, an approximate increase of 17.6%, and displacement increases from 1.87 mm to 2.96 mm, a substantial increase of 58.3%. The more considerable strain and displacement indicate that the intervertebral discs undergo more significant deformation at higher inclination angles, potentially causing the annulus fibrosus to experience stretching beyond physiological limits. Such excessive stretching can lead to micro-injury accumulation and compromise the integrity of the nucleus pulposus and annulus fibrosus. Furthermore, deformation of the intervertebral discs reduces their stability and increases the risk of injury, which may also affect the pilot's motor coordination and ability to resist external forces (Fig. 7).

3.3 Injury Risk Determination

The Neck Injury Criterion (NIC) is utilized to evaluate the impact of anterior-posterior neck movements on spinal cord injuries. Its calculation formula is as follows:

$$NIC = 0.2a_{rel} + v_{rel}{}^2 \tag{1}$$

where $_{\text{rel}}$ represents the anterior-posterior acceleration of the head's center of mass relative to the T1 thoracic vertebra, and v_{rel} denotes the anterior-posterior velocity of the head's center of mass relative to the T1 thoracic vertebra.

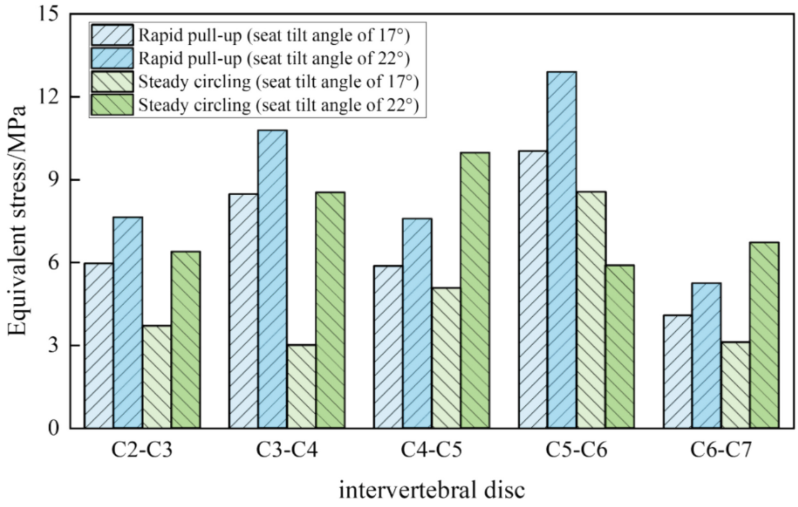


Fig. 7. Extreme stress values of intervertebral disc at different seat inclinations.

Simulation results indicate that under the rapid pull-up conditions, the maximum NIC values at seatback angles of 17° and 22° are 0.61 m^2/s^2 and 0.85 m^2/s^2, respectively. The NIC value at a 22° seatback angle is the highest, reflecting that larger seatback angles generate more significant anterior-posterior impacts on the cervical spinal cord under high-load environments. However, these NIC values are well below the spinal cord injury threshold of 15 m^2/s^2, suggesting that even at larger seatback angles, the risk of cervical spinal cord injury for pilots during rapid pull-up maneuvers remains low.

Under the stable circling conditions, the maximum NIC values at seatback angles of 17° and 22° are 0.34 m^2/s^2 and 0.57 m^2/s^2, respectively. Although the NIC value at a 22° seatback angle is relatively higher, the overall NIC values are lower than the rapid ascent condition. This may be due to the more gradual changes in acceleration and velocity during stable circling, resulting in more negligible anterior-posterior impacts on the neck.

Although NIC values under all operational conditions and seatback angles do not reach the spinal cord injury threshold of 15 m^2/s^2, indicating a low risk of spinal cord injury in single-flight missions, these incremental increases in NIC may have potential long-term effects on pilots' neck health due to prolonged and repeated high-load exposures. These slight increments may cause micro-injuries to neck tissues, accumulating over time and increasing the risk of spinal cord or nerve root injuries, affecting pilots' long-term health. Therefore, despite controllable short-term risks, attention must be paid to the cumulative effects of these loads on pilots' necks over their extended professional careers.

4 Conclusion

This study systematically analyzed the biomechanical responses of the pilot's neck under different seatback angles (17° and 22°) during rapid pull-up and stable circling conditions by constructing a detailed finite element model of the cervical spine. The simulation results indicated that seatback angles significantly influenced the vertebrae and intervertebral discs' stress, strain, and displacement. Nevertheless, these values did not reach the thresholds for direct injury.

Under the rapid pull-up conditions, a seatback angle of 22° resulted in a noticeable increase in vertebral stress and displacement, suggesting potential cumulative microstructural injury to the vertebrae. The intervertebral discs bore greater biomechanical loads at larger seatback angles, with the C5–C6 disc experiencing the highest stress. In stable circling maneuvers, the C3–C4 and C4–C5 intervertebral discs showed substantial stress increases, indicating that different operational conditions influence the loading patterns on the cervical spine differently. Although the NIC values under both conditions remained below the injury threshold, the observed increases highlight the need to consider the potential long-term effects of prolonged exposure.

In summary, this study validated the impact of varying seatback angles on the biomechanical performance of the cervical spine and provided scientific insights for the design of fighter aircraft seats. The findings suggest that while enhancing pilot comfort, it is crucial to carefully balance seatback angles to mitigate the long-term effects on neck health. Optimizing seat design can reduce the risk of cervical injuries for pilots in high-load environments, improving operational safety and mission efficiency.

Acknowledgments. This study was supported by the National Key Research and Development Program of China (NO. 2023YFF0615904) and Henan Key Laboratory of Imaging and Intelligent Processing (No. HKLIIP2023-A01).

References

1. Wang, L.Z., Fan, Y.B.: The biomechanics of injury and prevention. Adv. Mech. **50**(00), 124–168 (2020) (in Chinese)
2. Honkanen, T., Sovelius, R., Mäntysaari, M., et al.: + Gz exposure and spinal injury-induced flight duty limitations. Aerospace Med. Hum. Perform. **89**(6), 552–556 (2018)
3. Liu, S.Y., Cong, H., Wang, H.: Study on neck injuries in military pilots. J. Med. Biomech. **25**(04), 262–265 (2010) (in Chinese)
4. Keskimölö, T., Pernu, J., Karppinen, J., et al.: Degenerative cervical spine changes among early career fighter pilots: a 5-year follow-up. BMJ Mil. Health. **169**(4), 291–296 (2023)
5. Harrison, M.F., Forde, K.A., Albert, W.J., et al.: Posture and helmet load influences on neck muscle activation. Aerospace Med. Hum. Perform. **87**(1), 48–53 (2016)
6. Riches, A., Spratford, W., Witchalls, J., et al.: A systematic review and meta-analysis about the prevalence of neck pain in fast jet pilots. Aerospace Med. Hum. Perform. **90**(10), 882–890 (2019)
7. Sovelius, R., Mäntylä, M., Huhtala, H., et al.: Head movements and neck muscle activity during air combat maneuvering. Aerospace Med. Hum. Perform. **91**(1), 26–31 (2020)

8. Riches, A., Spratford, W., Witchalls, J., e t al.: A systematic review and meta-analysis about the prevalence of neck pain in fast jet pilots. Aerospace Med. Hum. Perform. **90**(10), 882–890 (2019)
9. Sovelius, R., Mntyl, M., Huhtala, H., et al.: Head movements and neck muscle activity during air combat maneuvering. Aerospace Med. Hum. Perform. **91**(1), 26–31 (2020)
10. Fan, W.Y.: A critical review on the biomechanical study of cervical interbody fusion cage. Med. Novel Technol. Devices. **3**, 47–84 (2021)
11. Zhang, H., Meng, L., Gong, Y., et al.: The influence of backrest angles on the passenger neck comfort during sleep in the economy class air seat without head support. Int. J. Ind. Ergon. **84**, 103074 (2021)
12. Liu, C.J.: The Study of Dynamic Response and Injuries of Pilot's Neck during Emergency Ejection under High G Loading. Tianjin University of Technology, Tianjin (2022) (in Chinese)
13. Yoganandan, N., Kumaresan, S., Pintar, F.A.: Biomechanics of the cervical spine part 2. Cervical spine soft tissue responses and biomechanical modeling. Clin. Biomech. **16**(1), 1–27 (2001)
14. Wang, J.J.: Research on Human Cervical Finite Element Modeling and Bionic Cervical Fusion Cage. Jilin University, Changchun (2014) (in Chinese)
15. Maylia, E., Nokes, L.D.M.: The use of ultrasonics in or thopaedics—a review. Technol. Health Care. **7**(1), 1–28 (1999)
16. Panjabi, M.M., Crisco, J.J., Vasavada, A., et al.: Mechanical properties of the human cervical spine as shown by three-dimensional load–displacement curves. Spine. **26**(24), 2692–2700 (2001)
17. Nightingale, R.W., McElhaney, J.H., Richardson, W.J., et al.: Dynamic responses of the head and cervical spine to axial impact loading. J. Biomech. **29**(3), 307–318 (1996)
18. Carter, D.R., Hayes, W.C.: Bone compressive strength: the influence of density and strain rate. Science. **194**(4270), 1174–1176 (1976)



Effects of Shared Control Modalities on Task Performance and User Experience in Robotic Teleoperation: A Path-Tracking Case Study

Dewei Zhu, Uhjin Hur, Seonghyeok Park, and Shuping Xiong(✉)

Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
{deweizhu,shupingx}@kaist.ac.kr

Abstract. This study investigated the effects of different shared control modalities—Direct Control (DC), State-Fusion Shared Control (SFSC), State-Guidance Shared Control (SGSC), and Semi-Autonomous Control (SAC)—on robotic teleoperation performance and user experience. Through a path-tracking experiment, we evaluated these modalities across two task difficulty levels (low/high). Results showed that SFSC mostly outperformed other modalities, reducing motion jitter and workload while increasing user trust and satisfaction, especially in high-difficulty tasks. SAC also demonstrated superior motion stability in high-difficulty tasks and reduced physical demand and user effort compared to DC, though it resulted in higher workload and lower satisfaction than SFSC. Similarly, SGSC improved tracking accuracy compared to DC but led to higher workload and lower satisfaction than SFSC. Correlation analysis revealed significant relationships between workload reduction and enhanced trust and satisfaction. These findings provide valuable insights for designing and selecting the appropriate shared control modalities based on desired outcomes and task difficulty, which can help optimize task performance and user experience in different scenarios in robotic teleoperation systems for path-tracking tasks. Future research should explore these modalities with larger participant groups and diverse teleoperation tasks.

Keywords: Teleoperation · Shared control · Human-robot collaboration · User experience · Workload

1 Introduction

The expansion of human activities into hazardous and inaccessible environments, such as space [1], undersea and nuclear facilities [2, 3], has created a growing need for teleoperation systems. Teleoperation, where robots are remotely controlled by human operators, has become indispensable for performing complex tasks in these challenging settings [4]. As technological advancements continue, the role of teleoperation is set to grow, with future progress relying on the ability of humans to handle increasingly complex tasks in these environments.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 378–390, 2025.
https://doi.org/10.1007/978-3-031-93502-2_26

Conventional teleoperation, which requires human operators to fully control robotic actions, faces significant limitations, including reduced situation awareness, coordination challenges, time delays, and high cognitive workload [5]. These factors make executing precise and complex tasks difficult. To address these issues, researchers have developed shared control strategies that combine human decision-making capabilities with the precision and durability of robots [6]. These approaches improve task efficiency and enable robots to handle more dynamic and demanding teleoperation tasks [7].

As shared control teleoperation continues to gain attention, various shared control modalities—such as Semi-Autonomous Control, State-Guidance Shared Control, and State-Fusion Shared Control—have been proposed to address diverse operational challenges [8]. In Semi-Autonomous Control, humans and robots share responsibility for controlling different variables of the robot. For instance, humans may handle the robot's general operation while the robot takes charge of tasks like collision avoidance, reducing the human operator's workload [9]. Similarly, in a grasping task, the human controls the robot's position, while the robot autonomously adjusts the gripper's orientation for more precise object grasping [10]. In State-Guidance Shared Control, the human fully controls the robot's actions but receives guidance from the robot's controller via haptic device, such as haptic cues, to refine their inputs [11]. A common example is the use of virtual fixtures, where virtual fixtures provide force guidance during teleoperation to enhance accuracy in delicate tasks [12]. Finally, in State-Fusion Shared Control both the human and the robot controller jointly control the robot's actions through an arbitration mechanism. For example, Ezeh et al. [13] proposed a probabilistic shared control approach to enhance the safety and efficiency of smart wheelchairs by fusing autonomous motion planning with user commands, particularly for users with mobility impairments. Despite this progress, a crucial unresolved issue is the systematic study of how different shared control modalities affect performance and usability in teleoperation systems. Such research is necessary to provide systematic guidance for choosing the suitable shared control modality for robotic teleoperation based on the requirements of teleoperation tasks.

This study investigates the impact of different shared control modalities on task performance and user experience in robotic teleoperation, using a path-tracking task as a case study. Additionally, it examines the interaction between task difficulty and shared control modalities. The findings offer valuable insights and practical guidance for selecting the most suitable shared control modality in teleoperation tasks, particularly those involving path-tracking scenarios.

2 Methodology

2.1 System Setup

Hardware. The robotic teleoperation system for path tracking tasks (Fig. 1) integrates a 6-DOF robotic manipulator (xArm 6, Ufactory), VR hand controller (Oculus Quest 2, Meta), RGB-D Camera (RealSense D435i, Intel), a 2D monitor, and a high-performance PC with Intel(R) Core(TM) i5–13,500 Processor(s), 16G RAM, and Nvidia GeForce RTX 3060Ti GPU.

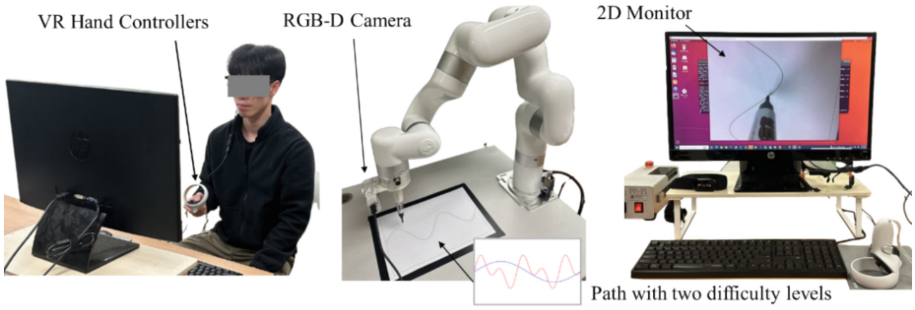


Fig. 1. Setup of the robotic teleoperation platform for path-tracking tasks.

System Architecture. The robotic teleoperation platform for path-tracking tasks (see Fig. 2) consists of three main components: the Unity module, the ROS (Robot Operating System) module, and a data transmission module based on the ROS-Unity Bridge.

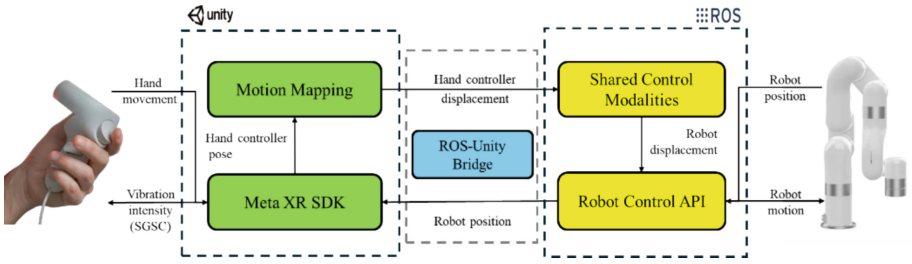


Fig. 2. Architecture of shared control teleoperation system for path-tracking tasks.

Unity Module. The Unity module functions as a runtime environment for VR devices. During teleoperation, it retrieves real-time posture data from the VR hand controller via the Meta XR SDK. This data is processed using the proposed Motion Mapping Strategy (see section: Motion Mapping), where it is transformed into displacement vectors of the VR hand controller relative to its initial position at the start of a tracking task. The displacement vectors are then remapped to the robot's coordinate system. Finally, the processed data is transmitted to the ROS module through the ROS-Unity Bridge.

ROS Module. The ROS module is responsible for robot control. Upon receiving the displacement vector of the VR hand controller from the Unity module, it processes the data through different shared control modalities, converts it into the final displacement vector for the robot, and invokes the robot control APIs (xArm API) to move the robot to the target position with maximum acceleration. This ensures synchronized motion between the VR hand controller and the robot.

Data Transmission. The ROS-Unity Bridge facilitates real-time, efficient data exchange, transmitting VR hand controller displacement data from Unity to ROS and the real-time robot position from ROS to Unity.

Motion Mapping. To enable intuitive teleoperated path tracking, this study employs a motion mapping strategy that aligns human-robot motion in both direction and scale (See Fig. 3). As the path tracking tasks in this study are confined to a two-dimensional plane, the VR hand controller is used to control the robot's motion within its XY plane. Due to the inherent differences in coordinate system definitions between the VR hand controller and the robot, we developed an incremental coordinate-based motion mapping strategy. At the initiation of each tracking task, the system records the operator's VR hand controller position as a reference point and continuously extracts positional increments relative to this origin. These incremental values are remapping to align with the robot's coordinate system before transmission to the ROS module. Simultaneously, the robot's reference origin is set to the starting point of the tracking path. Upon receiving the remapped increments from the Unity module, the ROS module operates the robot in servo mode at maximum acceleration, adding these increments to its reference origin. This process achieves real-time synchronized motion, enabling intuitive teleoperated path-tracking.

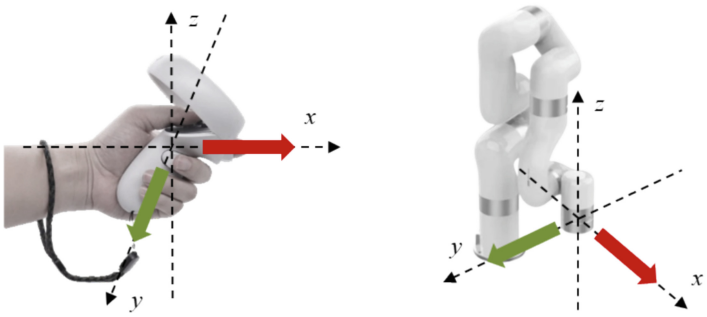


Fig. 3. Motion mapping between the VR hand controller and the robot arm.

Shared Control Modalities Following control modalities were developed for path-tracking tasks:

Semi-Autonomous Control (SAC). Restrict the robot's motion within a predefined safety boundary along the tracking path. Within this boundary, the robot remains fully controlled by the human operator. The safety boundary is defined by an upper and a lower bound, both of which follow the same shape as the tracking path and are offset by a predefined distance O_{safe}.

State-Guidance Shared Control (SGSC). Provides vibration feedback through a VR controller during the path tracking task to guide the operator to track the path accurately, with vibration intensity increasing exponentially as the robot deviates from the tracking path. We define a function to control the vibration intensity as follows:

$$f(x) = \frac{\tan^{-1}\left[q(x-1)\right] + \tan^{-1} q}{\tan^{-1} q}, x \in [0, 1] \subset \mathbb{R} \tag{1}$$

The x represents the normalized robot deviation δ_{norm} relative to the tracking path in this control modality, defined as $\delta_{\text{norm}} = \frac{\delta_t}{\delta_{\text{threshold}}}$, where δ_t is the real-time robot deviation,

and $\delta_{\text{threshold}}$ is the predefined deviation threshold. When $\delta_t \geq \delta_{\text{threshold}}$, the vibration intensity remains at its maximum value of 1.0. The influence of different q values on the steepness of the intensity curve as shown in Fig. 4.

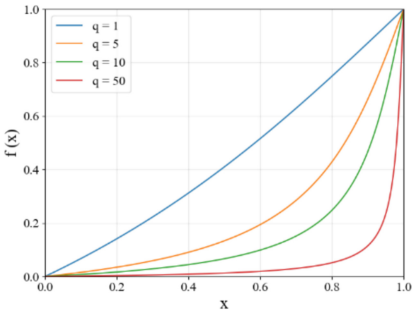


Fig. 4. Function curves of $f(x)$ for tuning vibration intensity (SGSC) and robot control authority (SFSC).

State-Fusion Shared Control (SFSC). Dynamically adjusts control authority between the robot and the operator during teleoperation, increasing robot's authority as the deviation from the tracking path grows larger. Similar with the SGSC modality, when the robot deviation relative to the tracking path reaches the predefined threshold deviation $\delta_{\text{threshold}}$, the robot control authority will reach the maximum value $\gamma_{\max}$. The contribution of human input $\boldsymbol{u}_{\mathbf{h}}$ and robot input $\boldsymbol{u}_{\mathbf{r}}$ to the final robot movement are calculated by the following equation:

$$\boldsymbol{u} = \gamma \boldsymbol{u}_{\mathbf{r}} + (1 - \gamma)\boldsymbol{u}_{\mathbf{h}}, \gamma \in [0, 1] \subset \mathbb{R} \tag{2}$$

The weight of control authority γ is calculated by multiplying the result of Eq. 1 by $\gamma_{\max}$. When the robot deviation reaches $\delta_{\text{threshold}}$, the robot control authority will remain at $\gamma_{\max}$.

Direct Control (DC). The human operator directly controls the robot arm's motion.

The parameter setting of each modality in this study is shown in Table 1.

Table 1. Parameter settings for shard control modalities (SAC, SGSC, SFSC).

Shared control modalities	Parameter settings
SAC	$O_{\text{safe}} = 7\text{mm}$
SGSC	$\delta_{\text{threshold}} = 7\text{mm}; q = 5$
SFSC	$\delta_{\text{threshold}} = 7\text{mm}; \gamma_{\max} = 0.6; q = 5$

Path Template The tracking paths in this study (see Fig. 5) are categorized into two difficulty levels: *low* and *high*. These paths are generated using the following multi-sine

function, defined by the parameters a_1, a_2, $a_3 \in \mathbb{N}$ and θ_1, θ_2, $\theta_3 \in \mathbb{R}$. The variable $x \in [0, 277]$ (in mm) represents the span of the tracking path along the robot's X-axis. Additionally, $T = 277$ is a fixed parameter that controls the number of periods in the multi-sine curve.

$$\psi(x) = a_1 \sin\left(\frac{x^*\theta_1}{T}\right) + a_2 \sin\left(\frac{x^*\theta_2}{T}\right) + a_3 \sin\left(\frac{x^*\theta_2}{T}\right) \tag{3}$$

The parameters for the *low* and *high* difficulty paths $\{(a_1, a_2, a_3), (\theta_1, \theta_2, \theta_3)\}$ are $\{(36, 0, 0), (2\pi, 0, 0)\}$and $\{(24,12,6), (5\pi, 10\pi, 10\pi)\}$, respectively.

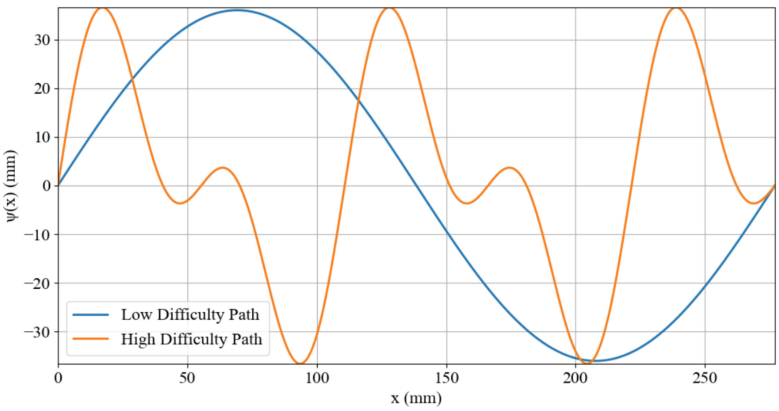


Fig. 5. Tracking paths with two difficulty levels.

2.2 Experimental Design and Procedures

A within-subject design was employed with 10 volunteer participants (3 females, 7 males; mean age: 26.5 ± 2.5 years), all of whom were university students at KAIST, South Korea. The study examined the effects of shared control modality (4 levels: three shared control modalities and one direct control baseline) and task difficulty (2 levels: low and high) on teleoperated path tracking tasks.

The experiment consisted of training and formal phases.

Train Phase. During the training phase, participants were first introduced to the teleoperation path-tracking system setup (see Fig. 1) and briefed on the experimental objectives. The experimenter demonstrated how each control modality providing assistance during path tracking. For instance, in the SGSC, varying intensities of vibration feedback were delivered through the VR hand controller to indicate path deviation magnitude. The experimental procedure and questions in questionnaires were then explained. Training was conducted for each control modality and task difficulty, where a training trial involved teleoperating the robot to track the path from start to end point and back. Participants completed at least two trials for each combination of modality and path difficulty

($2 \times 4 = 8$ conditions). To accommodate individual differences in learning speed, additional training trials were provided as needed, until participants satisfied two criteria: (1) fluent path-tracking without prolonged pauses, and (2) maintaining acceptable deviation or demonstrating valid correction when significant deviations occurred.

Formal Phase. In the formal phase, participants performed path-tracking tasks at two levels of difficulty using four different control modalities, presented in a randomized sequence to minimize potential carryover effects. Each task required participants to track a given path as accurately as possible within 60 s.

During the tracking process, three key performance metrics were recorded: *Tracking Error*, which measures the deviation of the robot's motion from the target path, commonly used in teleoperated robotic surgery scenarios where precise path-tracking is crucial for tasks such as tissue cutting; *Motion Jitter*, which evaluates the stability of the robot's movement, relevant in robotic welding where uneven motion impacts weld quality; and *Tracking Speed*, which assesses the speed of perform the path tracking.

After each trial, participants complete questionnaires to evaluate user experience. These include the *NASA-TLX*, which measures workload, and three single-scale questions assessing their sense of *Control*, *Trust*, and *Satisfaction* for each shared control modality and task difficulties. Participants rate their responses using a seven-point Likert scale.

2.3 Statistical Analysis

To evaluate the effects of shared control modalities and task difficulty on task performance and user experience, a repeated-measures ANOVA was employed. The analysis included four shared control modalities (DC, SAC, SGSC, SFSC) and two levels of task difficulty (*low*, *high*) as within-subject factors. Dependent variables were *Tracking Error*, *Motion Jitter*, *Tracking Speed*, *NASA-TLX score*, the sense of *Control*, *Trust*, and *Satisfaction*. The Bonferroni correction was applied to adjust for multiple comparisons in the post-hoc tests.

Additionally, Pearson correlation analysis was conducted to investigate the relationships between the task performance and user experience metrics. The significance level for all statistical tests was set at $p < 0.05$.

3 Results

The ANOVA analysis results are summarized in Table 2.

3.1 Task Performance

For *Tracking Error*, significant effects were observed for control modality ($F_{3,63} = 2.86$, $p = 0.044$) and task difficulty ($F_{1,63} = 107.87$, $p < 0.001$), with no significant interaction between the two factors (see Table 2). As illustrated in Fig. 6, Post-hoc tests indicated that SGSC resulted in significantly lower *Tracking Error* compared to DC ($p = 0.038$), while high task difficulty led to higher *Tracking Error*. For *Motion Jitter*,

Table 2. Summary of ANOVA results for performance and user experience metrics.

Measure	Effect	(DFn, DFd)	*F*	*p*
Tracking Error	SCM	(3, 63)	2.855	0.044
	TD	(1, 63)	107.871	<0.001
	SCM × TD	(3, 63)	1.114	0.350
Motion Jitter	SCM	(3, 63)	7.921	<0.001
	TD	(1, 63)	92.808	<0.001
	SCM × TD	(3, 63)	4.518	0.006
Tracking Speed	SCM	(3, 63)	2.406	0.076
	TD	(1, 63)	2.940	0.091
	SCM × TD	(3, 63)	0.541	0.656
NASA-TLX	SCM	(3, 63)	7.416	<0.001
	TD	(1, 63)	0.418	0.520
	SCM × TD	(3, 63)	0.160	0.923
Sense of Control	SCM	(3, 63)	1.213	0.312
	TD	(1, 63)	0.461	0.500
	SCM × TD	(3, 63)	0.361	0.782
Trust	SCM	(3, 63)	6.274	<0.001
	TD	(1, 63)	0.120	0.730
	SCM × TD	(3, 63)	0.559	0.644
Satisfaction	SCM	(3, 63)	11.603	<0.001
	TD	(1, 63)	0.538	0.466
	SCM × TD	(3, 63)	0.090	0.966

Note: SCM: Shared Control Modality, TD: Task Difficulty. DFn and DFd denote degrees of freedom for group variability (numerator) and within-group variability (denominator)

significant effects were identified for control modality ($F_{3,63} = 7.92$, $p = 0.001$), task difficulty ($F_{1,63} = 92.81$, $p < 0.001$), and their interaction ($F_{3,63} = 4.52$, $p = 0.006$). SFSC demonstrated greater motion stability compared to DC ($p < 0.001$), and high task difficulty increased *Motion Jitter*. The interaction effects further showed that in high-difficulty tasks, SFSC resulted in greater motion stability than both DC ($p < 0.001$) and SGSC ($p = 0.040$), while SAC achieved higher stability than DC ($p = 0.007$). However, no significant differences in *Motion Jitter* were found between control modalities in low-difficulty tasks. *Tracking Speed* did not show any significant effects.

3.2 User Experience

Workload. Significant effects of control modality on the workload measure, represented by the weighted average NASA-TLX score ($F_{3,63} = 17.42$, $p < 0.001$), were

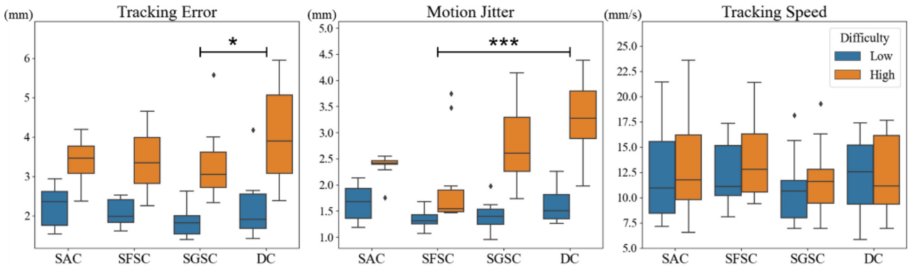


Fig. 6. Effects of control modality and task difficulty on task performance metrics.

found (see Table 2). Post-hoc tests revealed that SFSC resulted in significantly lower NASA-TLX score compared to all other modalities (DC: $p < 0.001$; SGSC: $p < 0.001$; SAC: $p = 0.018$), while SAC also achieved a significantly lower workload than DC ($p = 0.002$). As illustrated in Fig. 7, For mental demand, significant effects of control modality ($F_{3,63} = 5.39$, $p < 0.002$) were observed, with SFSC resulting in significantly lower demand compared to DC ($p = 0.036$) and SGSC ($p = 0.002$) in post-hoc tests. For physical demand, significant effects ($F_{3,63} = 6.49$, $p < 0.001$) indicated that SFSC ($p < 0.001$) and SAC ($p = 0.005$) required significantly lower physical demand than DC. Significant effects were also found for performance demand ($F_{3,63} = 6.48$, $p < 0.001$), with SFSC showing significantly lower performance demand than DC ($p < 0.001$) and SGSC ($p = 0.028$). For effort demand, SFSC demonstrated significantly lower effort compared to all other modalities ($F_{3,63} = 17.63$, $p < 0.001$; DC: $p < 0.001$; SGSC: $p < 0.001$; SAC: $p = 0.021$), while SAC also required significantly less effort than DC ($p = 0.004$). For frustration demand, significant effects ($F_{3,63} = 7.73$, $p < 0.001$) were found, with SFSC resulting in significantly lower frustration compared to DC and SGSC ($p < 0.001$). Temporal demand did not show any significant effects.

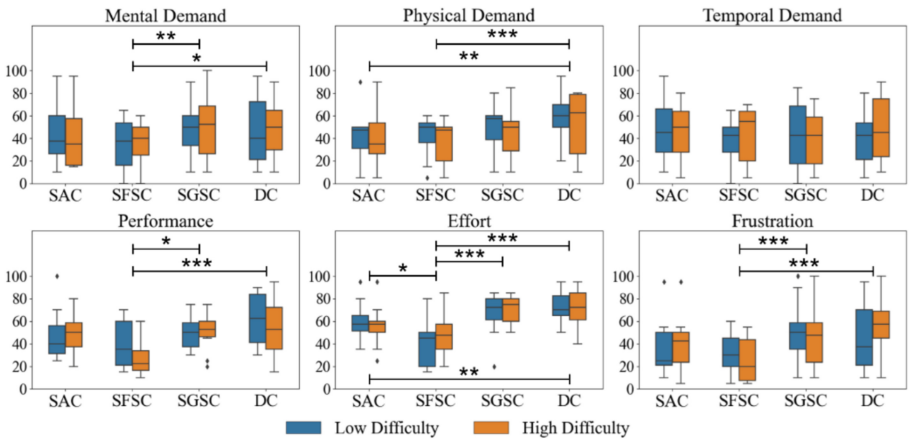


Fig. 7. Effects of control modalities and task difficulty on workload (NASA-TLX scores).

Other UX Metrics. For sense of *Control*, no significant effects were found (see Fig. 8). For *Trust*, significant effects ($F_{3,63} = 6.27$, $p = 0.001$) indicated that SFSC fostered higher *Trust* than DC ($p < 0.001$). *Satisfaction* was also significantly influenced by control modality ($F_{3,63} = 11.60$, $p < 0.001$), with SFSC achieving higher ratings than all other modalities (DC: $p < 0.001$; SGSC: $p < 0.001$; SAC: $p = 0.017$).

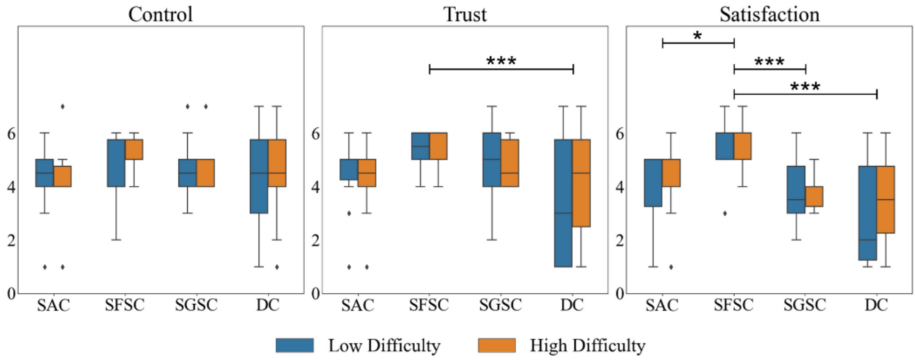


Fig. 8. Effects of control modalities and task difficulty on user experience (UX) metrics.

3.3 Correlation Analysis

The correlation analysis revealed several key relationships. *Tracking Error* exhibited a very strong positive correlation with *Motion Jitter* ($r = 0.825$, $p < 0.001$) and a weak positive correlation with *Tracking Speed* ($r = 0.261$, $p = 0.019$). *Tracking Speed* demonstrated a moderate negative correlation with *NASA-TLX Score* ($r = -0.403$, $p < 0.001$). *NASA-TLX Score* also showed a moderate negative correlation with *Trust* ($r = -0.556$, $p < 0.001$) and *Satisfaction* ($r = -0.526$, $p < 0.001$), while a strong positive correlation was found between *Trust* and *Satisfaction* ($r = 0.780$, $p = 0.001$), emphasizing the close relationship between these metrics.

4 Discussion

This study examined the effects of shared control modalities on task performance and user experience in robotic teleoperation through a path-tracking task. The results provide important insights into the efficacy of different shared control modalities, under different task difficulty conditions.

4.1 Performance Analysis of Shared Control Modalities

SFSC outperformed other modalities in mitigating *Motion Jitter*, particularly under high-difficulty conditions, demonstrating its effectiveness in dynamically adjusting robot intervention based on tracking errors. In contrast, SGSC achieved significantly higher

tracking accuracy than DC, suggesting that providing feedback to human operators to refine their input—rather than directly intervening in robot control—can enhance tracking accuracy. Moreover, both SAC and SFSC demonstrated notable improvements in task performance as difficulty increased. This implies that when tracking deviations become larger and more frequent under higher task difficulty levels, direct robot intervention in SAC and SFSC proves more effective than the indirect shared control modality of SGSC, which only provides feedback to influence human input. In general, SFSC excels in tracking stability, while SGSC offers advantages in tracking accuracy. SAC shows substantial improvements in both accuracy and stability as task difficulty rises. These findings highlight that each control modality has unique strengths depending on specific task performance aspects. Therefore, selecting an appropriate modality should be based on the specific requirements of the real-world application.

4.2 Relationship Between User Experience Metrics

The *NASA-TLX score* showed that SFSC significantly reduced workload across all dimensions compared to other modalities, except for temporal demand. These reports from the participants indicate that they experienced lower mental, physical, and frustration demands with SFSC and felt that they performed better with less effort during teleoperation. The reduced workload likely contributed to the higher levels of *Trust* and *Satisfaction* observed for SFSC in the ad-hoc survey. Notably, there was a moderate negative correlation between workload and *Trust*, suggesting that lower cognitive and physical demands enhance the perceived reliability of the system. Similarly, the strong positive correlation between *Trust* and *Satisfaction* highlights the interdependence of these user experience dimensions in robotic teleoperation.

4.3 Influences of Parameter Changes Within Shared Control Modalities

In this study, each shared control modality was tested using fixed parameters (see Table 1). However, the impact of parameter variations within these modalities remains unclear. To gain preliminary insights for future research, two participants performed path-following tasks under conditions where the internal parameters of the modalities were changed. Specifically, we examined the task performance and user experience of the SAC and SFSC modalities under different parameter settings. The SAC modality was tested with varying safe boundary values O_{safe}= {3 mm, 7 mm}, while the SFSC modality was evaluated with different threshold deviations $\delta_{\text{threshold}}$ = {3mm, 7mm} and maximum robot control authority $\gamma_{\max}$ = {0.3, 0.6}, under a low task difficulty condition.

In real-world teleoperation scenarios, the exact path to be followed by the robot is not always known in advance and must be identified and estimated using robotic vision techniques [14, 15]. This introduces inevitable estimation errors. The parameters O_{safe} and $\delta_{\text{threshold}}$ were designed to simulate such errors, where larger values indicate greater discrepancies between the estimated and actual paths. Experimental results from the two participants revealed significant improvements in task performance and reported user experience when the parameters O_{safe} and $\delta_{\text{threshold}}$ were reduced from 7 mm to 3 mm. This suggests that lower uncertainty in path estimation contributes positively

to the effectiveness of shared control modalities. Furthermore, in the SFSC, increasing γ_{max} from 0.3 to 0.6 resulted in moderate improvements in both task performance and user experience. However, these improvements were less pronounced compared to the impact of reducing $\delta_{threshold}$, indicating that precise estimation of the desired path is more critical to effectiveness than shared control authority alone. This underscores the importance of accurate environmental perception in robotic teleoperation, particularly through advanced computer vision technologies like object recognition and pose estimation.

4.4 Limitations and Future Work

While the study provides important insights, it is limited by the small sample size and the controlled experimental setup, which are inherent to preliminary investigations. Future research should involve larger participant groups, consider a wider range of intra-modality parameter settings, and explore different interface types, such as joysticks and haptic devices. Additionally, further investigations should include a broader spectrum of teleoperation tasks, such as grasp-and-place, peg-in-hole, and object rotation. Incorporating advanced physiological measures, such as EEG or eye-tracking data for real-time workload monitoring could offer a more comprehensive understanding of the effect of shared control modalities on user experience.

5 Conclusion

The study examined the effects of different shared control modalities on task performance and user experience in robotic teleoperation through a path-tracking task. The results indicate that SFSC consistently outperforms other modalities by reducing motion jitter, and workload, leading to higher user trust and satisfaction, especially under high task difficulty conditions. SGSC, while improving tracking accuracy, resulted in higher workload and lower satisfaction compared to SFSC. SAC demonstrated notable performance improvements with increasing task difficulty, highlighting its potential in challenging tasks. Correlation analysis revealed that reducing workload enhances perceived trust and satisfaction, emphasizing the close relationship between these users experience metrics. These findings suggest that selecting an appropriate shared control modality should consider the specific performance requirements and user experience expectations of the application. Despite the valuable insights provided, the study is limited by its small sample size and controlled environment. Future research should involve larger participant groups, diverse teleoperation tasks, and the integration of advanced physiological measures to gain a deeper understanding of the effect of shared control modalities.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (NRF-2022R1F1A1061045, NRF-2022M3J6A1063021). The authors extend their sincere gratitude to Tiejun Ma, Haejun Kim, and others for their participation in the preliminary study and valuable feedback, which contributed to enhancing the shared control teleoperation system.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Weber, B., Balachandran, R., Riecke, C., Stulp, F., Stelzer, M.: Teleoperating robots from the international space station: microgravity effects on performance with force feedback. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8144–8150. IEEE (2019)
2. Yan, J., Gao, T., Yang, X., Chen, C., Guan, X.: Teleoperation control of autonomous underwater vehicle toward human on the loop: needs, analyses, and solutions. IEEE Syst. Man Cybernet. Mag. **10**(3), 2–13 (2024)
3. Stedman, H., Kocer, B.B., Van Zalk, N., Kovac, M., Pawar, V.M.: Evaluating immersive teleoperation interfaces: coordinating robot radiation monitoring tasks in nuclear facilities. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 11972–11978. IEEE (2023)
4. Moniruzzaman, M.D., Rassau, A., Chai, D., Islam, S.M.S.: Teleoperation methods and enhancement techniques for mobile robots: a comprehensive survey. Robot. Auton. Syst. **150**, 103973 (2022)
5. Chen, J.Y.C., Haas, E.C., Barnes, M.J.: Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37**(6), 1231–1245 (2007)
6. Li, Y., Tee, K.P., Chan, W.L., Yan, R., Chua, Y., Limbu, D.K.: Continuous role adaptation for human–robot shared control. IEEE Trans. Robot. **31**(3), 672–681 (2015)
7. Selvaggio, M., Cacace, J., Pacchierotti, C., Ruggiero, F., Giordano, P.R.: A shared-control teleoperation architecture for nonprehensile object transportation. IEEE Trans. Robot. **38**(1), 569–583 (2021)
8. Li, G., Li, Q., Yang, C., Su, Y., Yuan, Z., Wu, X.: The classification and new trends of shared control strategies in telerobotic systems: a survey. IEEE Trans. Haptics. **16**(2), 118–133 (2023)
9. Gottardi, A., Tortora, S., Tosello, E., Menegatti, E.: Shared control in robot teleoperation with improved potential fields. IEEE Trans. Hum. Mach. Syst. **52**(3), 410–422 (2022)
10. Adjigble, M., Stolkin, R., Marturi, N.: Haptic-guided assisted telemanipulation approach for grasping desired objects from heaps. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4142–4149. IEEE (2023)
11. Cheng, M., Li, R., Ding, R., Xu, B.: Multitarget adaptive virtual fixture based on task learning for hydraulic manipulator. J. Field Robot. **41**(8), 2715–2731 (2024)
12. Mühlbauer, M., Hulin, T., Weber, B., Calinon, S., Stulp, F., Albu-Schäffer, A., Silvério, J.: A probabilistic approach to multi-modal adaptive virtual fixtures. IEEE Robot. Autom. Lett. **9**, 5298–5305 (2024)
13. Ezeh, C., Trautman, P., Devigne, L., Bureau, V., Babel, M., Carlson, T.: Probabilistic vs linear blending approaches to shared control for wheelchair driving. In: 2017 International Conference on Rehabilitation Robotics (ICORR), pp. 835–840. IEEE (2017)
14. Li, Y., Yuan, X., Liu, Y., Cohn, R., He, H., Karamouzas, I.: Scene editing as teleoperation: a case study in 6DOF kit assembly. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9071–9078. IEEE, Kyoto (2022)
15. Zhang, J., Liu, L., Xiang, P., Fang, Q., Nie, X., Ma, H., Lu, H.: AI co-pilot bronchoscope robot. Nat. Commun. **15**(1), 241 (2024)

Author Index

A
Abt, Vincent II-250
Amith, Muhammad "Tuan" I-72
Arredondo, Maria Teresa 53
Ashrafi, Navid III-147
Auguszt, Tillmann II-268

B
Balatsoukas, Panagiotis III-332
Baytar, Fatma II-3
Bergdahl, Nina I-3
Bolaji, Josephine II-233
Boldi, Arianna III-280
Brignani, Debora I-323, I-342
Brolin, Erik I-121
Brown, David J. III-200
Bruce, Cody III-14
Brukamp, Kirsten III-3
Bu, Yitong I-191
Buchholz, Maximilan II-268
Bushara, Fatehia Bushara Garma I-323

C
Cabrera, Maria Fernanda 53
Cai, Yuqing II-281, III-163
Cao, Jinyun I-145
Cao, Yudi III-89
Castellano, Elena I-53
Cavallo, Maria Lucia I-323, I-342
Chablat, Damien III-41
Chang, Kuei-Yuan I-132
Charron, Camilo III-41
Cheah, Kenny S. L. III-175
Cheminat, Olivier III-41
Chen, Ding I-145
Chen, Junjie II-210
Chen, Keliang I-22
Chen, Xingyu III-175
Chen, Yali II-18
Chen, Yu I-275
Chini, Giorgia I-323, I-342
Connor, Bonnie III-200
Copot, Cosmin III-224
Cui, Licong I-72

D
Davoudi Kakhki, Fatemeh I-353
De, Soureesh I-289
Demirel, H. Onan I-191
Deng, Lili III-212
Deng, Yan I-210
Di Serio, Arianna II-153
Díaz-Castro, Diego II-77
Donado, Ana María II-48
Drefahl, Paulo I-160
Du, Aiping III-257
Duffy, Vincent G. I-178

F
Falciati, Luca I-323, I-342
Fang, Hao III-175
Feng, Bingbing II-172
Fico, Giuseppe I-53

G
Ganaha, Wataru III-185
Gao, Jie III-267
Gao, Jinjin II-101
Gao, Yiyun III-135
Garg, Aakash III-14
Gensichen, Jochen II-192
Gómez del Moral Herranz, Rodrigo Martín I-53
Godbole, Mihir III-14
Graf, Philipp III-147
Gu, Yuanyuan I-301, I-312
Günter, Nadia II-250
Guo, Wenbin I-160, I-257
Guo, Yong Yan I-85, III-107
Guo, Zi-Yi I-22
Gütlich, Daria II-250

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

V. G. Duffy (Ed.): HCII 2025, LNCS 15791, pp. 391–394, 2025.
https://doi.org/10.1007/978-3-031-93502-2

H
Haddick, Sean III-200
Hamilton, Jane I-72
Hannay, Jo E. III-28
Hao, Xiaodong I-257
Hasanzadeh, Sogand I-178
Hernandez-Palma, Hugo II-35, II-48, II-77
Hioki, Shota III-185
Hou, Serena I-72
Hua, Yang III-135
Huang, Yitian I-22
Hui, Jizhuang I-257
Hur, Uhjin I-378

I
Iriondo Pascual, Aitor I-121

J
Janssen, Ulrike III-313
Jia, Wei Quan III-267
Jiang, Lijun II-62, III-212
Jiang, Nina I-178
Jiang, Weisheng I-365
Jiménez-Coronado, Angélica II-77
Jiménez-Delgado, Genett Isabel II-35, II-48, II-77
Jin, Lin Guo I-85
Joffe, Gabrielle B. I-191
Johansson, Björn I-289

K
Kastner, Patrick I-238
Kim, Ji-Eun I-227
Kim, Jung Hyup I-301, I-312
Kim, Woojoo II-281, III-163
Kittelmann, Luca II-192
Kostage, Kevin I-160
Kostopoulou, Konstantina I-53
Krcmar, Helmut II-192
Kroknes, Monica III-28
Kroll, Nicole III-14
Kwon, Cassandra II-233
Kyosev, Yordan III-224

L
Lao, Chi-Cheng I-22
Lee, Chang-Franw I-132
Lei, Tian I-210
Lenat, Antoine III-41
Lewis, James III-200
Li, Jiaxin I-227
Li, Zhaoxu I-210
Li, Zhelin III-212
Li, Zhengxuan II-120
Li, Zongze I-22
Liang, Jing I-22
Liang, Xiang II-62
Lin, Yi II-133
Lin, Yu Syuan I-41
Liu, Jun I-22
Liu, Weilin III-267
Liu, Zhongqi I-365
Livitckaia, Kristina I-53
Loitsch, Claudia II-268
Lopomo, Nicola Francesco I-323, I-342
Lorenz, Sebastian III-313
Luderschmidt, Johannes II-250
Lyu, Wei II-210

M
Ma, Changda I-238
Ma, Feiyu I-257
Ma, Tiancong I-97
Maguire, Martin III-332
Maher, Mona II-3
Man, Siu Shing III-212
Marquardt, Manuela III-147
Martin, Ludger II-250
Mastrantuono, Will I-301
Mazzola, Luca I-160
McQuerry, Meredith II-233
Merino-Barbancho, Beatriz I-53
Miehling, Jörg III-298
Moghadam, Armin I-353
Mohanty, Siddarth I-301, I-312
Moore, Nicole I-72
Mori, Giulio II-153
Morris, Andrew III-332
Mosso, Martina I-323, I-342
Mostowfi, Sara I-301, I-312

N
Negro, Francesco I-323, I-342
Neira-Rodado, Dionicio II-35, II-48, II-77
Neumann, Ana C. I-72
Nieto-Granados, Anderson II-35, II-48
Nishiyama, Masashi III-185
Nouri, Jalal I-3

O
Oprean, Danielle I-301, I-312
Ouyang, Ruolei II-18

P
Pantelidou, Konstantina I-53
Park, Seonghyeok I-378
Patel, Saira II-3
Paternò, Fabio II-153
Peppers, Sean I-160
Pérez Luque, Estela I-121
Peraza, Ingrid Estefany III-224
Pintori, Noemi I-323
Piol, Alessandro I-323, I-342
Prommegger, Barbara II-192

Q
Qiu, Ziwen II-233
Qu, Chengyi I-160, I-257

R
Ranavolo, Alberto I-323, I-342
Rapp, Amon III-280
Reindl-Spanner, Philipp II-192
Rhee, Daniel I-72
Rivera, Ivania II-3
Rujas, Miguel I-53

S
Savas, Lara S. I-72
Scalona, Emilia I-323, I-342
Scherb, David III-298
Schmidt, Ann-Malin III-224
Schofield, Simon III-200
Schorlemmer, Julia III-147
Schulz, Trenton W. III-28
Seo, Jinsil Hwaryoung III-14
Seo, Kangwon I-301, I-312
Shahtahmassebi, Golnaz III-200
Shen, Chao II-210
Shen, Yaodong III-367
Shi, Yichao I-238
Singleton, Jessica III-238
Sjöberg, Jeanette I-3
Söderlund, Henrik I-289
Spelly, Gwen III-298
Staab, Sergio II-250
Striegl, Julian II-268
Sun, Chuancai I-365
Sun, Mingyu II-18
Sun, Yifan II-210
Sun, Yijie III-61
Sun, Zhiting II-18

T
Tan, Litun I-22
Tan, Tianyang I-22
Tan, Zhengtang II-89
Tang, Lu I-72
Tang, Shunshun II-89
Tanilkan, Sinan S. III-28
Tao, Cui I-72
Thai, Lauren III-14
Tian, Xuetao I-22

V
Valli, Giacomo I-323, I-342
van der Ploeg Feriche, Jose Maria I-289
van Remmen, Judith III-298
Varrecchia, Tiwana I-323, I-342
Voigt-Antons, Jan-Niklas III-147
Vona, Francesco III-147
Vora, Hardik I-353

W
Wang, Chenfeng II-281
Wang, Chunlan I-238
Wang, Fang I-301, I-312
Wang, Hong I-257
Wang, Jiatao I-365
Wang, Kaige I-145
Wang, Li-Jen III-79
Wang, Shanshan II-101
Wang, Wei II-133
Wang, Yi I-301
Wang, Yuqing III-89
Wang, Zebin II-89
Wang, Zihan III-107
Wanta, Helge III-313
Wartzack, Sandro III-298

Wells-Beede, Elizabeth III-14
Weng, Chenying I-72
West, Andre III-238
Weston, Cindy III-14
Wong, Seng Yue III-175
Wu, Di II-89
Wu, Lei III-120, III-350
Wu, You Yang I-85

X
Xiao, Wenxin I-227
Xie, Yuqin III-257
Xie, Zijie III-332
Xiong, Shuping I-378
Xu, Qiongyu II-210
Xu, Shuo I-97
Xu, Yan-Dian I-22
Xu, Yingjiao II-101
Xu, Zhenhao II-210

Y
Yang, Rui I-210
Yang, Tianqi II-18
Yang, Yoon II-3
Yang, Yujie III-120, III-350
Yang, Zhiyuan II-281
Ye, Zi Hong I-85
Yoo, Sujin II-3
Yu, Xiaoqun II-281, III-163
Yuan, Zhuyun II-89

Z
Zambrano Roa, Paola II-35
Zeng, Ling I-22
Zhang, Mengxin II-281
Zhang, Zhipeng II-62
Zhang, Ziyi III-257
Zhao, Ruisi III-120, III-350
Zhao, Yulin II-172, III-61
Zhao, Yuyue III-89
Zheng, Li II-120, III-367
Zheng, Meng-Cong I-41, III-79
Zhou, Jing II-133
Zhou, Meiling I-275
Zhou, Qianxiang I-365
Zhou, Wei I-178
Zhou, Yihan III-163
Zhou, Zi Xin I-85
Zhu, Dewei I-378
Zhu, Yanzhi II-210