Enabling Intelligent Immersive Learning using Deep Learning-based Learner Confidence Estimation

Mohammadreza Akbari Lor*, Shu-Ching Chen†, Mei-Ling Shyu*, Yudong Tao‡, Shahin Vassigh§
*School of Science and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA
†Data Science and Analytics Innovation Center (dSAIC), University of Missouri-Kansas City, Kansas City, MO 64110, USA
‡Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA
§College of Communication, Architecture and the Arts, Florida International University, Miami, FL 33199, USA
Email: {ma7fy, s.chen, shyum}@umkc.edu; yxt128@miami.edu; svassigh@fiu.edu

***Abstract*—In today's world, augmented reality and virtual reality (AR/VR) technologies have become more accessible to the public than ever. This brings the possibility of immersive learning to the forefront of education for future generations. However, there is still much to discover and improve in using these technologies to analyze and understand learning. This paper explores the utilization of data captured through AR/VR headsets during an immersive training program for industrial robotics. This includes data on time spent, eye gaze, and hand movement during a range of activities to track a learner's understanding of the content and intelligently estimate learner confidence within these environments using deep learning. Leveraging a dataset that comprises responses and confidence levels from 10 individuals across 35 questions, we aim to improve the uses and applicability of confidence estimation. We explore the possibility of training a model using learners' data to dynamically fine-tune lessons and activities for each individual, thereby improving performance. We demonstrate that a pre-trained compact LSTM classification model can be fine-tuned with relatively small data, for enhanced performance on an individual basis for better personalized learning.**

***Index Terms*—AR/VR; immersive learning; deep learning; personalized learning**

I. Introduction

As the Augmented Reality and Virtual Reality (AR/VR) technology becomes more popular and accessible, it is natural that it is also used as a training and learning tool for many different purposes [1]. Studies show that learning in a VR environment enhances learning, and immersion and interactivity of the experience play a major factor in the effectiveness of VR learning environments [2].

These technologies significantly increase learner engagement by transforming abstract concepts into interactive 3D models and simulations. For example, AR and VR enable students in scientific fields to visualize complex biological and chemical processes or physical mechanics more intuitively and interactively. They also provide a safe, controlled environment where students can conduct experiments or practice skills that would be too dangerous or costly in the real world, such as medical surgeries, operating industrial robots, or chemical experiments. Additionally, AR/VR can offer customized learning experiences tailored to the pace and style of individual learners, enhancing the accessibility and personalization of education. This capability ensures that learners from different backgrounds and learning styles can access high-quality educational resources.

Some examples of the use of these technologies for learning include using VR for immersive English as a foreign language training [3]. Other examples include the AIRLIT application, an innovative AR and VR-based learning tool, developed to enhance the practical learning of Airfield Lighting Systems among Airport Engineering Technology cadets [4], and using mixed reality in CPR training [5], among many others.

Automated monitoring in AR/VR learning environments can be very crucial for enhancing educational outcomes. Technology allows for real-time tracking and analysis of learner interactions, providing immediate feedback that helps students understand their performance and adjust their learning strategies accordingly. In the same way that educators benefit from detailed insights into each student's engagement and proficiency, which enables them to offer targeted support where needed, automated monitoring and adapting a learning system can achieve the same results. Automated systems can also facilitate the adaptation of learning content based on ongoing performance assessments, ensuring that each student faces the right level of challenges (i.e., personalized learning). This adaptability is key to maintaining an efficient and effective personalized learning environment that scales well, even with large numbers of students.

Regardless of the peculiarities of different fields and areas that may employ this technology, monitoring and estimating a learner's confidence in his/her performance or answers can be a powerful tool in designing smart and adaptive learning environments. While the issue of assessing a person's confidence when performing an activity or answering questions in an AR/VR environment is quite broad, focusing on specific aspects of it can enable us to make progress toward solving this fundamental requirement of adaptive learning.

In this paper, we propose a framework that utilizes deep learning-based learner confidence estimation to enable personalized and intelligent immersive learning by leveraging the dataset in [6]. Our proposed framework highlights the versatility and efficiency of small, pre-trained models in immersive learning environments and sets the stage for future innovations in adaptive and personalized learning environments where model agility and rapid tuning are paramount.

979-8-3503-5118-7/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00023

The rest of this paper is organized as follows. Section II reviews the related work in immersive learning (II-A), machine learning approaches for temporal and spatial data (II-B), and confidence estimation research (II-C). Section III presents our proposed framework and explores data pre-processing (III-A), data augmentation (III-B), and model design details (III-C) to enable personalized and intelligent immersive learning. Section IV shows the experiment setup and results. Section V will be a discussion regarding the limitations of our experiments and the future directions of this research. Finally, section VI provides the conclusion of this paper.

II. Related Work

A. Immersive Learning

There are many notable studies and examples of using AR/VR for learning. As VR headsets become more advanced and accessible to institutions and a larger portion of the population own personal VR headsets, the possibility of integrating AR/VR technologies with education increases. Utilizing immersion in a VR setting for education in many fields can increase learning and lower the costs associated with training [1].

In the context of healthcare and medical education, a special issue [7] has highlighted several key advancements and empirical studies that corroborate the efficacy and increasing adoption of AR/VR. These diverse studies demonstrate the potential of these technologies in enhancing learning outcomes and student engagement in medical education.

Jayasundera et al. studied the effectiveness of VR in transitioning students to patient care [8]. Similarly, studies have explored the impact of VR on industrial design and manufacturing processes, particularly in enhancing worker training and factory planning. Notably, research has been conducted to assess the user experience (UX) with semi-immersive, haptic-centered virtual assembly systems, with results highlighting the necessity of optimizing the VR system usage to mitigate potential discomforts and enhance efficiency and satisfaction among new users [9].

B. Machine Learning with Temporal and Spatial Data

Classification using spatial and temporal data is an evolving field of research [10]–[12]. For eye gaze and pointer data, typically known as scanpaths, many frameworks have been proposed in various contexts. Scanpath modeling using variational hidden Markov models (HMMs) and discriminant analysis (DA) was developed, achieving notable classification accuracy in identifying observer tasks and stimuli characteristics from gaze patterns [13]. In [14], an SP-ASDNet model that combines convolutional neural networks (CNNs) and long short-term memory (LSTM) networks was designed to classify observers as typically developed (TD) or having Autism Spectrum Disorder (ASD) based on their gaze scanpaths. The use of LSTMs for temporal classification highlights the efficiency of the framework for eye gaze data.

Capinha et al. [15] explored the application of deep learning models to classify ecological temporal data as an alternative to traditional approaches which often rely on user-defined feature transformation. Their research highlights the advantages of deep learning techniques, such as their ability to classify directly from time series data. The research demonstrated the application of various deep learning architectures, including CNNs and LSTMs, in three ecological case studies: species identification from insect wingbeat spectrograms, species distribution modeling from climate time series, and phenological prediction from meteorological data. These models provided a standardized approach that enhanced predictive performance and automated much of the computational workflow.

An example of spatial and temporal data integration is trajectory classification which is critical for various location-based services. A Spatio-Temporal GRU, which effectively models both spatial and temporal dimensions, was proposed [16]. This is done by a novel segmented convolutional weight mechanism and a temporal gate that significantly enhanced trajectory classification performance over conventional deep learning approaches.

Pouyanfar et al. presented an innovative deep learning framework aimed at addressing the challenges of video classification, particularly for imbalanced datasets as traditional methods often overlooked the integration of spatial and temporal information simultaneously, leading to suboptimal classification performance [17]. Their research introduced a model that combined spatial and temporal data effectively using a spatio-temporal synthetic oversampling method alongside a pre-trained CNN for spatial feature extraction and a residual bidirectional LSTM for capturing temporal dynamics. The framework has been tested on imbalanced video datasets, demonstrating notable performance by effectively managing data skewness and leveraging both types of information to enhance classification accuracy.

C. Confidence Estimation

In the context of a learner's progress and confidence estimation, there have been various approaches. Yudelson et al. demonstrated that individualized Bayesian Knowledge Tracing (BKT) [18] models can be utilized to achieve better effects than general models for student learning and progress prediction [19].

Item Response Theory (IRT) [20] is another foundational statistical framework used in psychometrics to model the probability of correct responses to test items based on individual abilities and item characteristics. In confidence estimation, IRT is particularly relevant as it can help estimate a student's confidence in their answers by considering both the difficulty of the questions and the student's ability level. This method provides an understanding of how different factors influence a student's certainty in their responses, making it invaluable for adaptive testing scenarios [21].

In [6], a method for confidence estimation in immersive environments was presented. The authors studied the most deterministic and decisive features to be used for confidence estimation. On the other hand, in this paper, our proposed framework focuses more on personalized learning by utilizing confidence estimation.

III. Proposed Framework

The purpose of our research is to gauge whether a model trained on individuals for confidence estimation can be fine-tuned with the minimal amount of data for an unseen individual with even better performance. In this paper, we propose a framework that enables personalized and intelligent immersive learning using deep learning-based learner confidence estimation.

The proposed framework consists of three steps, namely data pre-processing, data augmentation, and model design (as depicted in Figure 1). As can be seen from Figure 1, after “Subject-based split,” the original dataset is split into two subsets, one for pre-training, and another one for fine-tuning and the model personalization. The arrow from “Subject-based split” to “Question-based split for each remaining subject” shows the independence between the subjects used for pre-training and fine-tuning. In “model design,” we conducted empirical studies and selected the best model design that obtains the best performance with the dataset in [6]. Using the selected model design to pre-train on the pre-training subset (half of the data in this study) will result in a performance worse than before due to the lower number of samples the model uses to train on and learn the confidence patterns. The rest of the subject’s data will be used to fine-tune the pre-trained models on a participant-to-participant basis to gauge the performance of each individual. A performance increase for all or a majority of participants demonstrates the effectiveness of our proposed framework for the confidence estimation of unseen individuals.

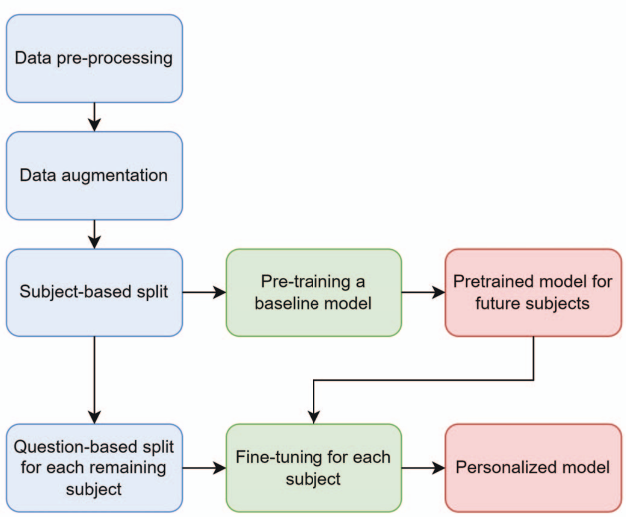

Fig. 1. Experiment Design

A. Data Pre-processing

The dataset used involves the case of participants answering multiple choice questions in two different scenarios, one with no previous training regarding the question topics which was dubbed Pre-test, and another one after some training dubbed Post-test, for the sake of comparison of confidence between the two different stages of learning. As this is an experiment in an immersive environment, eye gaze data is available in the time series format along with the time spent on each question. Ten (10) participants had their session data recorded and each answered 35 questions before and after watching some training material regarding the question topics. After each session, the participants gave a self-report of their confidence regarding each question. This resulted in 700 sequences before pre-processing.

Instead of using the eye gaze feature in the dataset directly, the positions of the important items in the canvas displayed to the participants were used to create the label features, encoding at each time stamp, which part of the screen the participant’s gaze is at. A demonstration of this can be seen in Figures 2 and 3, where each region of the canvas is used to create a binary feature that encodes whether at any given time stamp in the dataset the participant’s gaze was in that region. These features and the time spent on each question (as another feature) have shown to be able to train the best-performing pre-trained and fine-tuned models.

The eye gaze with an example illustrated in Figure 3 was taken based on the pre-processing step, where the general coordinates were converted based on the canvas where the questions were displayed, and the fixation confidence feature was used to filter the low-confidence eye gaze data. This step, while done to make sure the sequences did not contain faulty data, has resulted in many of the sequences being completely removed. Of the 588 samples remaining, there was a high variability in sequence lengths. An analysis of the time elapsed field disproves the notion of this being solely the result of the data cleaning since some participants had also spent varying periods on different questions.

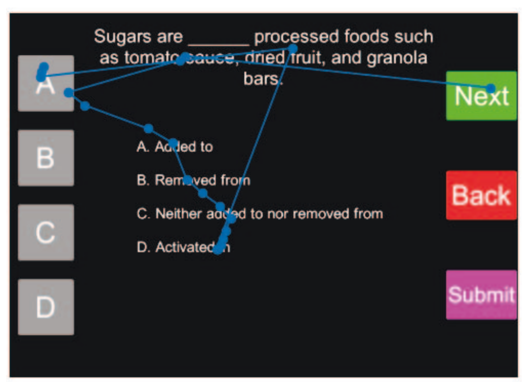

Fig. 2. Example of Question Canvas

B. Data Augmentation

Due to the large variability between sequence lengths and the low number of sequences for classification, data was augmented with all sequences broken into pieces with zero

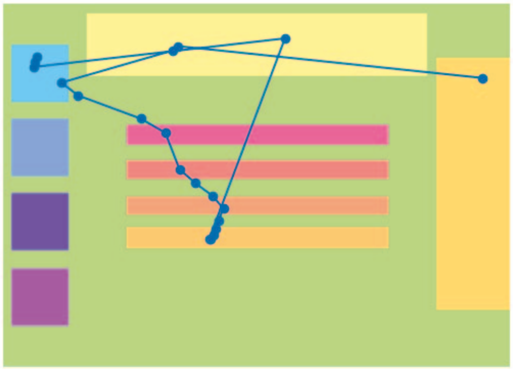

Fig. 3. Example of Scanpath of Eye Gaze on Masked Canvas

padding for the pieces with a shorter length for model training. The sequence length of 25 was chosen from the empirical studies and based on the hypothesis that a short sequence will not include enough data for the model to learn the necessary pattern for confidence estimation, and a long sequence length would result in fewer samples, limiting the model's ability to learn enough to generalize to unseen data.

For the pre-training of a baseline model to be used as a benchmark, the data of five participants were chosen for training. For the model fine-tuning, the data of five more subjects were utilized, where the data of each subject were used to tune a pre-trained model to see how well it could be adapted for a new subject. A 60%-40% split of questions was used for train and testing new models for each subject, respectively. This amounted to 42 questions for training and 28 questions for testing, where each set had an even representation of Pre-test and Post-test samples. However, this split became uneven in different ways for different subjects after the pre-processing and augmentation described above. As a result, subject 3 was excluded from the fine-tuning stage due to its very low sample numbers and not being useful for the purpose of the experiment.

C. Model Design

The proposed model, due to the sequential structure of the data, is an LSTM [22] with 2 layers, 100 hidden units and tanh activation function where the sequences of features were used as input. A dropout layer with a 0.5 rate was used for the robustness of training after the LSTM layers. Two fully connected layers with widths of 20 and 10 were used for the classification of the LSTM's output feature vector with ReLU activation functions and 0.5 dropout. The last layer classifies between confident and non-confident categories.

With the learning rate set at 0.001 and the Adam optimizer, each model was trained for 200 epochs. For each, the epoch with the highest validation accuracy and the lowest difference from training accuracy was selected as the optimal step, and weights were saved.

For the model design selection, considering the aforementioned data pre-processing and data augmentation steps, the best performance was achieved with the LSTM classification network shown in Figure 4. The selected model, when trained on the entire dataset with 8 subjects for training and 2 subjects for testing, achieves a performance of 85.6% that matches with the performance reported in [6], but it has an increase in the layer number and an increase in the hidden units, due to the change in the sequence length that was used in the proposed framework.

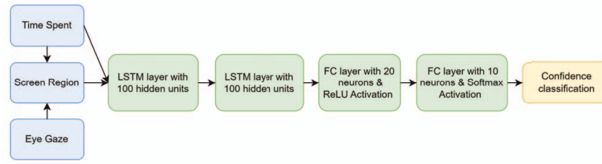

Fig. 4. Model Design

IV. Experiment Results

For our experiment setup, to make sure that our fine-tuning improvements are not specific to a particular data split and to avoid overfitting, a form of repeated cross-validation was performed. We randomly generate 10 different splits where the 10 subjects are divided into 2 sets of 5 such that the first set is used for pre-training and the second set for fine-tuning. An extra condition for the splits was that each subject must appear in each of the 2 sets exactly 5 times. This is to ensure that all subjects are used at different stages equally and that the improvement results are robust and not split-dependent.

Due to the lower number of subjects to learn from, the 10 pre-trained models achieved accuracies between 66.05% and 72.52%. In addition, because of the low number of samples for each subject during the fine-tuning experiments, the model's testing accuracy and training accuracy started with some distance that remained near-constant as the model was further trained before overfitting and divergence. For each model, we ensured the best-selected epoch did not have a train and validation accuracy distance greater than the starting distance. We selected the epoch where training and testing increased equally, symbolizing genuine learning of the subjects' confidence patterns rather than the memorization of the specific training samples. As mentioned in III-B, subject 3 is not considered for the fine-tuning step of the experiments.

For the fine-tuned models of each subject, the average of their results was used to demonstrate an aggregated performance increase and showcase the robustness of our results. Three performance metrics were used for each subject across the 5 splits when the subject was in the fine-tuing set. The 'Average Accuracy' is the mean of the validation accuracies per subject. For each subject, the 'Average Accuracy Increase' as shown in Equation (1) is the mean of the ratio between the accuracy of the fine-tuned model and the accuracy of its respective baseline, while the 'Average Absolute Accuracy Increase' in Equation (2) is the difference instead of the ratio. Finally, the 'Average Error Decrease' as seen in Equation (3) is the mean of the ratio between the fine-tuned model and its respective baseline.

$$\text{Average Increase} = \left(\frac{\sum_{i=1}^{n} \frac{\text{Acc}_{\text{Subject},i}}{\text{Acc}_{\text{Base},i}}}{n} - 1 \right) \times 100 \quad (1)$$

$$\text{Average Absolute Increase} = \frac{1}{n} \sum_{i=1}^{n} (\text{Acc}_{\text{Subject},i} - \text{Acc}_{\text{Base},i}) \quad (2)$$

$$\text{Average Error Decrease} = \frac{\sum_{i=1}^{n} (B_i - S_i)}{\sum_{i=1}^{n} B_i} \times 100 \quad (3)$$

where n is the number of observations, $\text{Acc}_{\text{Subject},i}$ is the accuracy of the fine-tuned model of a subject in the ith split, $\text{Acc}_{\text{Base},i}$ is the accuracy of its respective baseline model, $B_i = 100 - \text{Acc}_{\text{Base},i}$ and $S_i = 100 - \text{Acc}_{\text{Subject},i}$.

As can be seen in Table I, we have an increase in performance ranging up to 16.65% absolute accuracy increase for 8 of the 9 considered subjects and a performance decrease for only one. Due to the small size of the network, the fine-tuned models took between 5 and 30 seconds to train for 200 epochs so that the best epoch's weights could be selected as the finalized model of that subject, showcasing the speed and therefore, the viability of implementing this solution to a real-world environment. This model design can learn in real time from a learner's feedback in an immersive environment, leading to a better estimation of learners' confidence.

TABLE I
COMPARISON OF PERFORMANCE ACROSS ALL SUBJECTS, EXCLUDING SUBJECT 3

Subject	Average Accuracy	Average Accuracy Increase	Average Absolute Accuracy Increase	Average Error Decrease
1	71.09%	4.45%	2.99%	9.66%
2	72.08%	6.57%	4.44%	14.33%
4	73.22%	4.54%	3.20%	10.31%
5	85.83%	24.16%	16.65%	53.69%
6	77.69%	11.13%	7.80%	25.15%
7	82.50%	19.03%	13.10%	42.24%
8	77.96%	12.01%	8.35%	26.92%
9	71.99%	4.75%	3.20%	10.33%
10	62.00%	-8.11%	-5.64%	-18.19%

V. DISCUSSION AND FUTURE DIRECTION

The dataset utilized in this study was relatively small, and the experimental procedures employed to gather data were basic. To advance the scope of research outlined in this paper and to contribute more substantially to the body of knowledge, several strategies could be implemented. One approach is to expand the dataset by involving more participants and maintaining a similar experimental setup. Future studies would benefit from collecting confidence data on a larger scale, encompassing a diverse group of individuals from various backgrounds and fields. This would enhance the dataset's variety and make it more representative of a broader demographic, thereby improving the generalizability and robustness of the findings.

While the use of multiple-choice questions was advantageous for this experiment due to their straightforward and definitive structure, it is important to acknowledge that learning processes are complex and varied. To further the research in confidence estimation and to provide deeper insights, more nuanced and sophisticated experiments should be designed to incorporate different ways of measuring knowledge and therefore the confidence of learner. These experiments should involve detailed recording and testing criteria that are capable of capturing the subtleties and complexities of learning processes. Different types of assessments, beyond multiple-choice questions, could be explored to simulate real-world learning conditions more accurately.

Furthermore, confidence can also be defined on a more granular scale as opposed to only a binary classification. Having a more refined view of these characteristics, whether discrete and encompassing more nuanced categories, or continuous and on a scale, can better reflect the complex nature of confidence in learners. This direction will involve considerations regarding the definition of different categories of confidence or the scale used if defined as continuous. The same applies to data gathering methods for this more refined and expansive view of confidence as a feature.

Implementing these enhancements will require careful consideration of the experimental design and the methodologies for data analysis. By addressing these elements, future research can make meaningful advances in understanding and measuring learner confidence and contribute significantly to personalized and intelligent immersive environment education and pedagogy.

VI. CONCLUSION

The integration of AR/VR into learning environments presents a promising path for adaptive learning systems. Our research has demonstrated the feasibility of deploying a pre-trained, compact LSTM model within such immersive settings to be easily fine-tuned to accurately estimate an individual learner's confidence. By focusing on the nuances of learner interactions and self-reported confidence levels in the dataset, this study has highlighted the capability of fine-tuned models to adapt to individual variability and improve confidence estimation significantly over baseline performances. The quick training times and the model's ability to learn from a limited number of samples emphasize the practicality of implementing such technology in live settings, where rapid adaptation to individual learner profiles is crucial. Not only do these results support the viability of using fine-tuned LSTM models for confidence estimation in VR/AR-based learning, but they also suggest broader applicability for similar methods in other adaptive systems where understanding and reacting to user confidence could enhance interaction quality and learning outcomes.

Ultimately, this research lays foundational work for future investigations into personalized learning environments. It encourages ongoing refinement of models and methods to harness the full potential of AR/VR technology in education, making learning more engaging, efficient, and tailored to individual needs.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation (NSF) Grant Numbers IIS-2202610 and CNS-2301552, and Microsoft Accelerating Foundation Models Research.

REFERENCES

[1] S. Hauze and J. Frazee, "Virtual immersive teaching and learning: How immersive technology is shaping the way students learn," in *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2019, pp. 1445–1450.

[2] K. Mahmoud, I. Harris, H. Yassin, T. J. Hurkxkens, O. K. Matar, N. Bhatia, and I. Kalkanis, "Does immersive vr increase learning gain when compared to a non-immersive vr learning experience?" in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 480–498.

[3] S.-M. Lee, X. Wang, I. Park, and R. Lestiono, "It feels so real! situated authentic language learning in immersive virtual reality," *Education and Information Technologies*, pp. 1–23, 2024.

[4] D. Amalia, V. Suryan, V. Septiani, Y. Komalasari, R. Rizko, A. Febriansyah, S. S. Ristumanda, M. Kristiawan, and O. Hendra, "Development of a game-based learning: Airfield lighting system simulator using virtual reality and augmented reality," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 2, pp. 978–994, 2024.

[5] H. Sungur, Z. M. van Berlo, and L. M. Lüwa, "Enhancing cardiopulmonary resuscitation training with mixed reality: Improving cardiopulmonary resuscitation performance and enjoyment," *Cyberpsychology, Behavior, and Social Networking*, 2024.

[6] Y. Tao, E. Coltey, T. Wang, M. Alonso, M.-L. Shyu, S.-C. Chen, H. Alhaffar, A. Elias, B. Bogosian, and S. Vassigh, "Confidence estimation using machine learning in immersive learning environments," in *2020 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2020, pp. 247–252.

[7] F. D. Buono, A. Marks, and D. Lee, "Virtual reality in medical education," 2024.

[8] M. Jayasundera, M. Myers, K. Pandian, and G. Gingell, "Virtual reality simulation: evaluating an experiential tool for the clinical application of pathophysiology," *Medical Science Educator*, vol. 32, no. 6, pp. 1575–1577, 2022.

[9] S. Macias-Velasquez, H. I. Medellin-Castillo, and A. Garcia-Barrientos, "New-user experience evaluation in a semi-immersive and haptic-enabled virtual reality system for assembly operations," *International Journal of Human-Computer Studies*, p. 103312, 2024.

[10] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715–734, 2001.

[11] S. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "Indexing and searching structure for multimedia database systems," in *Storage and Retrieval for Media Databases 2000*, vol. 3972, International Society for Optics and Photonics. SPIE, 1999, pp. 262 – 270.

[12] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[13] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior research methods*, vol. 50, no. 1, pp. 362–379, 2018.

[14] Y. Tao and M.-L. Shyu, "Sp-asdnet: Cnn-lstm based asd classification model using observer scanpaths," in *2019 IEEE International conference on multimedia & expo workshops (ICMEW)*. IEEE, 2019, pp. 641–646.

[15] C. Capinha, A. Ceia-Hasse, A. M. Kramer, and C. Meijer, "Deep learning for supervised classification of temporal data in ecology," *Ecological Informatics*, vol. 61, p. 101252, 2021.

[16] H. Liu, H. Wu, W. Sun, and I. Lee, "Spatio-temporal gru for trajectory classification," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1228–1233.

[17] S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Deep spatio-temporal representation learning for multi-class imbalanced data classification," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2018, pp. 386–393.

[18] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, pp. 253–278, 1994.

[19] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized bayesian knowledge tracing models," in *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer, 2013, pp. 171–180.

[20] W. M. Yen and A. R. Fitzpatrick, "Item response theory," *Educational measurement*, vol. 4, pp. 111–153, 2006.

[21] J. Johns, S. Mahadevan, and B. Woolf, "Estimating student proficiency using an item response theory model," in *International conference on intelligent tutoring systems*. Springer, 2006, pp. 473–480.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.