2023 IEEE International Conference on Smart Computing (SMARTCOMP) | 979-8-3503-2281-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SMARTCOMP58114.2023.00083

Teaching Humanoid Robots to Assist Humans for Collaborative Tasks

Julia Rodano
Department of Computer Science
Montclair State University
Montclair, USA
rodanoj1@montclair.edu

Omar Obidat
Department of Computer Science
Montclair State University
Montclair, USA
obidato1@montclair.edu

Jesse Parron
Department of Computer Science
Montclair State University
Montclair, USA
parronj1@montclair.edu

Rui Li
Department of Computer Science
Montclair State University
Montclair, USA
liru@montclair.edu

Michelle Zhu
Department of Computer Science
Montclair State University
Montclair, USA
zhumi@montclair.edu

Weitian Wang*
Department of Computer Science
Montclair State University
Montclair, USA
wangw@montclair.edu

***Abstract*—As technology has advanced, society has witnessed and participated in the creation of robots that can walk, talk, and recognize speech. To facilitate communication and collaboration between humans and humanoid robots, we develop a teaching-learning framework for human beings to teach humanoid robots to complete object identification and operation tasks. The robots learn from their human partners based on the transfer learning approach and can assist humans using their learned knowledge. Experimental results and evaluations suggest the success and efficiency of the developed approach in smart service contexts for human-robot partnerships. The future work of this study is also discussed.**

***Keywords*—Humanoid robots, smart service systems, human-robot collaboration, transfer learning, vision system.**

I. Introduction

Over the past several decades, robots have become increasingly popular in different fields [1]. Humanoid robots are becoming exceedingly more popular due to their human-like nature and ability to walk, talk, and even function as humans do. Essentially, they are borderline humans, but they cannot think and feel like humans. They may be governed by some algorithm that lets them think and process commands. One of the earliest applications of humanoid robot applications came in the 1940s. This robot had 6 degrees of freedom and was used to move highly radioactive objects [2]. This robot would lead to many robotics advancements, including the creation of modern humanoid robots. The application of robots ranges over many different fields. They can greatly benefit companies and governments to save time, reduce production costs, and make production efficient, which means more profit for companies. A typical workplace where people may be able to find humanoid robots is in factories and industrial plants. One example of a robot fit for such a setting is the ARMAR-6, a humanoid bot with over 27 degrees of freedom that is making way for humanoid robots in industrial spaces [3]. This robot was made for handling tools and helping humans in factories. It can recognize when people need help, carry heavy objects while simultaneously following a path, and can detect if something is near its arm to keep humans safe. Robots such as these can provide humans with the aid, they need to maintain a safe and efficient workspace.

Humanoid robots can also be used in healthcare, assisting doctors and nurses. These robots can handle communication with patients and move patients if needed. Robots are even proving helpful in education, where humanoid robots can act as teaching assistants and make learning environments more interactive [4, 5]. Human-robot collaboration can apply to many fields, even one's own home. Robots, with their assistive technology, can help those with disabilities or who are older live their lives independently. They will not need human aid, nor will they need to live in a nursing home. While these robots are beneficial here on Earth, they may also prove useful beyond Earth and into space. Humans and robots can work alongside each other in space, completing tedious tasks [6].

There are several different approaches to humanoid robot programming. One of them is through learning from human beings during human-robot interaction. In this case, a robot will watch a human perform a specific action, such as passing boxes or picking something up, and then try to replicate the action. Every time the robot watches or interacts with the human, it will slowly get "smarter" and better at practicing moving the item. An earlier example of this is the Association for the Advancement of Artificial Intelligence's (AAAI) Robot Competition held in 1999. In this competition, robots acted as if they were attendees at a conference and had to find the correct time and place of the panel they wished to attend [7]. Despite their groundbreaking technology, there are limitations on humanoid robots learning from human beings in collaborative tasks, especially in smart service contexts. This is due to several different factors. One of the main reasons is that humans all function differently from one another [8]. The habits of one person may be totally different from another. If one were to create a dataset for a humanoid robot to learn from, it would require hundreds, if not thousands, of samples of behavior from people who range in a multitude of different factors such as height, disability, and even the culture they come from. Achieving the dataset would be one task in itself, but conducting the research could prove more challenging.

Participants may be needed to run some of these experiments, but some people may have trust concerns with the robots during the interaction process [9]. For example, trying to test robots in a healthcare setting could be difficult

2693-8340/23/$31.00 ©2023 IEEE
DOI 10.1109/SMARTCOMP58114.2023.00083

since patients may not trust the robots. Robots do not have empathy, emotion, or the same decision-making processes that humans do. This may result in a difficult challenge for researchers who need human participants to validate their proposed approaches. In social situations, humans will always act more naturally. Humans are comforting in nature, especially with touch. A simple pat on the back or a hug can make people feel better. Robots, however, cannot produce the same "empathy" that humans can produce. Their touch is not comforting and may even be awkward for some people [10]. Robots may have the upper hand in automation tasks, but they are unable to surpass humans when it comes to comfort. When it comes to medical fields, robots may be at a disadvantage because they are unable to produce the emotional response that humans create. Overall, these different issues combined make human-robot interaction research, especially on humanoid robots, be confronted with uncertain challenges.

Motivated by the above issues, in this study, to facilitate communication and collaboration between humans and humanoid robots, we develop a teaching-learning framework for human beings to intuitively teach humanoid robots to complete collaborative tasks in smart service contexts. NAO robots will work alongside humans in collaborative teaching tasks. The robots learn from their human partners based on the transfer learning approach and can assist humans using their learned knowledge to identify, pick up, and deliver objects that the human requested by speech instructions. Experimental results and evaluations demonstrate the success and efficiency of the developed approach in smart service environments for human-robot partnerships.

II. Approaches

A. Overview of the Proposed Approach

In this study, the focus will be on expanding human-robot interaction. As shown in Fig 1, the human will teach a NAO robot to identify different kinds of objects. Then the robot will employ its obtained knowledge to pick up an item and hand it off to its human partner. In the learning process, the developed approach begins running and waits for an object name to be received through a verbal command for the human. All listening is done through a speech recognition system [11]. This process will loop through different kinds of objects (in this work chips, plates, cups, and toys are used). Through a computer vision system developed by a camera, the robot will observe and identify the quadrant that the item is in. In the learning process, the transfer learning method is used to train the robot to recognize the class of each object. The module that controls object recognition will use a TCP socket to send the quadrant number to the NAO's motion control system. Once the NAO receives the human requests after learning, it will then begin the retrieval process using its learned strategies. Depending on the item's quadrant, it will move to a predetermined position, either between quadrants 1 and 2 or between quadrants 3 and 4. The NAO will then pick up the requested object. Depending on the object, the NAO will perform a specific hand and arm movement. If the NAO needs to pick up a cup, plate, or chips, it will use the general grasping function. If the item is a toy, the NAO will use a grasping function specific to the toy, since it's a bit of a different shape compared to the other three objects, and then return it to the requested user.

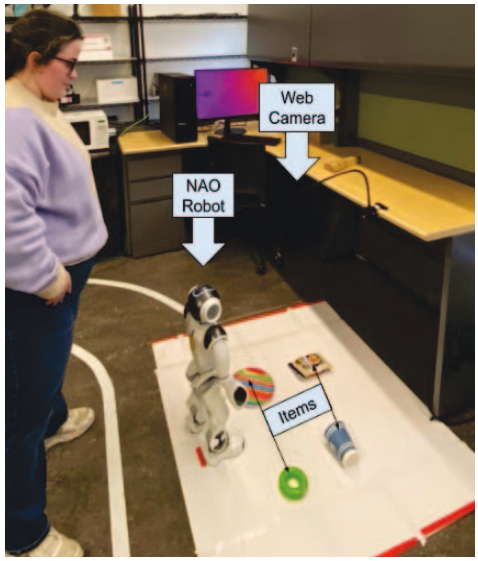

Fig. 1. The experimental platform.

B. Teaching Robots through Transfer Learning

This project requires a specific model that could accurately and successfully predict given objects in the human-robot collaboration process. Transfer learning is used by the human to teach the robot to identify different categories of objects. Transfer learning is one way to transfer data from one domain to another to improve learning capabilities [12]. It is an example of machine learning that can be used with robots to make human-robot collaboration run smoothly. Transfer learning involves creating models that are specific to a certain data set. Typically, transfer learning relies on the deep convolutional neural network (CNN) that provides a pre-trained model. The CNN's pre-trained model compares the data set it is fed to its own data set and begins building a new model based on older learning models [13]. The pre-trained model is used to create the new model and its weights can be adjusted based on the data set. The new model will contain the necessary data based on its pre-trained model and the data set provided. The data is then transferred to become a new model that can be used in other tasks. The newly built model for this study contains all the data related to a cup, plate, toy, and chips, so each object can be identified in real-time through an external camera.

In transfer learning, the task (T) is made up of the label space Y and the prediction function F, which can be expressed as $T = \{Y, F\}$ whereas the domain (D) is made up of a feature space X and the marginal probability distribution $P(X)$ express as $D = \{x, P(X)\}$. Reusing data from the source domain to enhance the performance of the target domain is how knowledge is transferred from the source domain and task to the target domain and task. Transfer learning speeds up the learning process and makes the model more accurate [14]. To increase learning efficiency and performance, transfer learning essentially takes the knowledge acquired in one domain and applies it to another that is related.

C. ResNet50

To train the model, ResNet50 is used in this work. It is a deep convolutional neural network that helps train datasets for image classification. Convolutional neural networks work well for image classification because they can pick out details about items, such as edges or curves [15]. This helps with image recognition, and using a model produced with ResNet50 in combination with image processing tools such as OpenCV allows for real-time object detection. ResNet stands for residual network and is a way to create models through datasets of images. When creating a model based on images for this study, there is no need for advanced machine learning processes. ResNet can solve complex tasks and increase the model's accuracy with little time and effort [16].

D. Data Collection

The data set for this study contains five different objects, a cup, plate, toy, chips, and “blank”. Blank serves as the control of the group and is shown when the camera does not see any of the four main items. The data collection process is elaborated in Fig. 2. The data set was created by our developed program that would automatically store photos into a file called “dataset”. The program would first check if the data set folder existed. If it did not, a folder would be created but if the folder did exist the program would continue. From there a connection between the computer running the program and the camera would be established through OpenCV and the region of interest (ROI) would be built as 180 by 180 pixels. Images would then be created and stored into the dataset folder. Each item contains about 1700 images and includes different rotations of the object as well as different background colors.

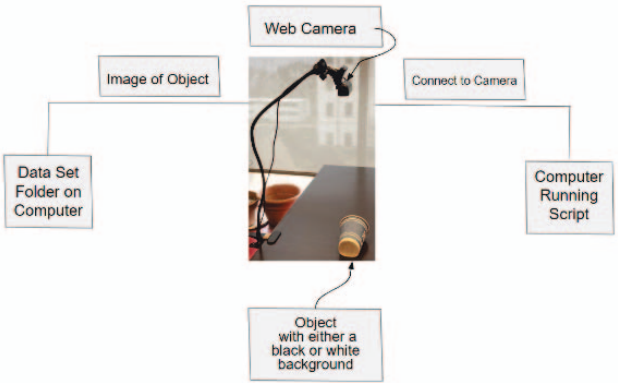

Fig. 2. Data collection for robot learning.

E. Robot Learning

The objective of this study is to have the robot work in conjunction with its human companion via a computer vision system to recognize and find their requested object. Transfer learning, along with ResNet50, helps create a new model for the NAO robot to learn from. A top camera aids in the computer vision system and object recognition since there is a limit to accessing the NAO's camera directly to run OpenCV. The NAO will receive information through a program that controls the motion of the NAO. This program will only execute once the computer vision system sends the quadrant number and item name from the external camera. The robot can locate and retrieve items for its human companion in combination with its learned strategies based on transfer learning. In the robot learning process, the ResNet50 produces *(l - 1)* outputs, which are then used in the next layer denoted by $(x_l - 1)$. Layers and activation functions are then added, which can vary [17]. This study used ReLU 512, 256, 128, 64, and Softmax 5 as the activation functions. Adding the activation functions results in an output of $F(x_l\text{-}1)$. From there, the output becomes x_l, which can be represented as $x_l = F(x_l\text{-}1) + x_l\text{-}1$.

III. Experimental Setup

A. Experimental Platform

As shown in Fig. 1, the experimental platform involves several components including a white background, household objects, a NAO robot, a web camera, and a camera stand. A whiteboard is laid out on the floor, and each item is placed in a quadrant. The objects include a cup, a plate, a toy ring, and a small bag of chips. Each item has a piece of nylon string taped to both sides to form a handle. This acts as a way for the NAO to grab the object without having to elevate the white platform and figure out the exact heights at which the NAO could grab each object. This modification is made because, although the NAO robots can "sit" and "crouch", they do not have the physical capability to bend over, which limits their range of motion in terms of picking things up off the ground. To combat this issue, we attach a nylon string to each object. This way, the NAO can assume the crouching position to pick up objects and move them from their position to the final objective. The camera stand is attached to a table, and the camera looks down to give an aerial view of the four objects. In front of the objects is the NAO robot. Our developed approach is run on a workstation (through Ubuntu 16.04) configured with a 3.60 GHz Intel® Xeon® W-2223 Processor and an NVIDIA® RTX™ A4000 Graphic Card.

B. Task Description

At the beginning of the experiment, voice recognition will start through the camera's microphone. The human user will request an object by saying its name, such as "Toy", "Cup", "Plate", or "Chips". Once the system recognizes the word through the speech recognition library, it will begin the search process. The frame is divided into four smaller quadrants, and the system will continuously loop through them to locate the requested object. Once the object is identified, the loop will break. The quadrants are labeled 1 - 4 when viewed from an aerial perspective. The name and quadrant number of the current item will be recorded. For example, quadrant 3 contains a bag of chips, so "quad3" will be stored in the quadrant variable, and "chips" will be stored in the item variable. This data will be sent through a TCP socket to the program controlling the NAO. Upon receiving the name of the item and the quadrant, the NAO will move to the quadrant where the object is found. Depending on the quadrant number, the NAO will move a certain distance so it is either between quadrants 1 and 2 or 3 and 4. The NAO will then crouch, and its arm will move behind the object. It will swing the arm forward to catch the nylon string in its hand and then close its hand. It will return to its standing position and then move back to the starting point by walking backward. The NAO will then turn,

lift the object, rotate its wrist so that its hand is facing upwards, and open its hand for its human companion to take the object.

IV. Results and Analysis

A. Training and Cross-Validation Accuracy

To train the new model, a dataset of different objects was used, which were split into five categories: toy, cup, plate, chips, and blank. As presented in Fig. 3, both the training accuracy and the cross-validation accuracy reached up to 100%. The model can predict images on both black and white backgrounds, as well as predict the item without a background (i.e., the item is so close to the camera that the background is not visible). The figure resulting from the model training shows that the model could work well and accurately. Occasionally, the model may declare an object as something it is not, but such occasions are rare and infrequent.

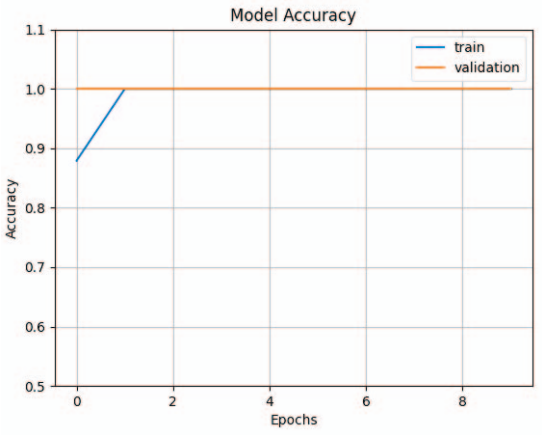

Fig. 3. Training and cross-validation accuracy.

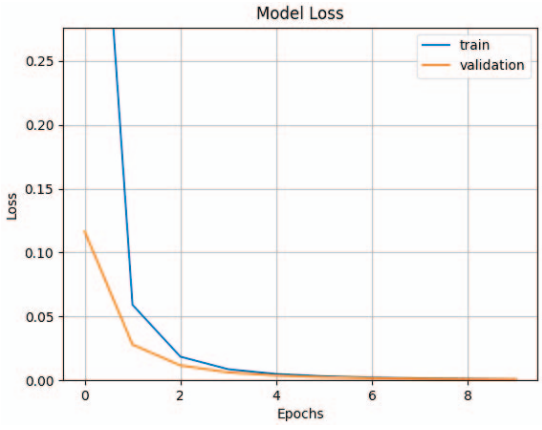

Fig. 4. Training and cross-validation loss.

B. Training and Cross-Validation Loss

The dataset used to create the model holds over 8500 images, around 1700 per item with different background colors and rotations to ensure the most accurate results when using the computer vision system. The pictures need to be more general; they cannot be specific to a certain condition. If they were, that certain condition would need to be recreated every single time which makes the work tedious. Through general pictures, the model can be accurate to less specific conditions. The conditions in which the model is created, or the experiment is run do not have to be specific. Training and cross-validation losses are significant to verify the robustness of the model and ensure that it can generalize on new input images. The smaller the loss is, the better the trained model will be. We utilized the cross-entropy function to evaluate the training and cross-validation losses. As presented in Fig. 4, both the training loss and the cross-validation loss attain a value of 0.

C. Confusion Matrix

To evaluate which object the model has an issue with interpreting, a confusion matrix can be used to help this. The rows of the confusion matrix represent the true values of the objects, while the columns represent the predicted object information. As shown in Fig. 5, the numbers of the correctly identified objects are denoted by the middle diagonal. The numbers outside of the diagonal stand for misrecognized objects. If the model's prediction for an item is different from its true value, a number would appear in place of zero on the confusion matrix. For example, if there is a number in place of a zero between the intersection of true 'chips' and predicted 'blank', the model would be confusing the number of 'blank' images with its true value, 'chips'. The confusion matrix in Fig. 5 indicates that the trained object recognition model for the NAO robot is robust.

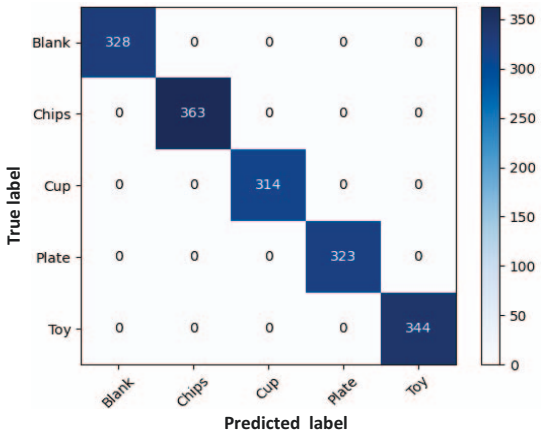

Fig. 5. The confusion matrix.

D. Real-World Human-Robot Collaboration

As presented in Fig. 6, the developed approach is validated in real-world human-robot collaboration. This process involves the instructions that the NAO robot receives for the object it is asked to pick up. Fig. 6 provides an example of the NAO robot working through the object recognition, picking up, and delivery processes. Fig. 6(a) shows the NAO robot is receiving its human partner's command "Cup". Then the robot will identify the requested object based on its learned knowledge through the camera. After finding the object, as presented in Fig. 6(b)-(d), the NAO moves forward to the corresponding quadrant where the item is in. The robot crouches once it has reached the specific position. It then positions its arm behind the nylon string and quickly moves it forward to catch and close its hand around the nylon to pick up the cup. From here, The NAO returns by walking backward to its original position

and delivers the requested object to its human partner by rotating its wrist and opening its hand (Fig 6.(e)-(h)).

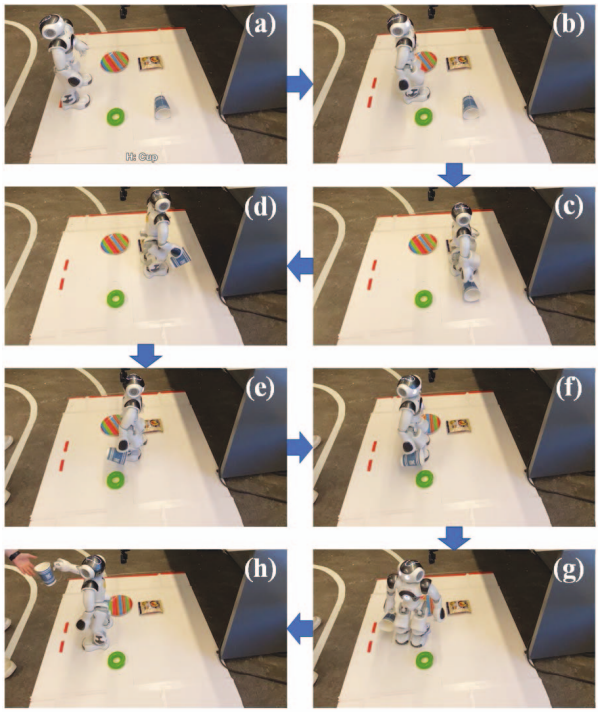

Fig. 6. Real-world verification in human-robot collaboration.

V. Conclusions and Future Work

In this work, we have developed a teaching-learning framework for human beings to intuitively teach humanoid robots to complete collaborative tasks in smart service contexts. The proposed approach can effectively facilitate communication and collaboration between humans and humanoid robots. Transfer learning and ResNet50 are used for the NAO robot to create a newly learned model based on certain household objects. In real-world human-robot collaboration experiments, the NAO robot can pick up objects and return them to the human who requested the object. The results and evaluations suggest the success and efficiency of the developed approach in smart service environments for human-robot partnerships. In the future, this study could be expanded upon to add more advanced techniques where the NAO's built-in camera is used as the primary vision system and the distance between the NAO and objects can be evaluated. This would make the experiments run smoothly and be more accurate since the NAO could position itself right next to the object. The developed approach could also be improved upon. A new methodology could be created so the NAO can pick the item up off the floor. This would eliminate the nylon string that was used and decrease the chance of the NAO falling over. Spatial awareness could also be implemented. Currently, the NAO cannot calculate its surrounding environment. If this information were to be determined, the NAO could correctly position itself and be aware of other objects around it. That way it can pick the object up without stepping on objects and possibly falling.

Acknowledgment

This work is supported in part by the National Science Foundation under Grant CNS-2117308 and in part by the National Science Foundation under Grant CNS-2104742.

References

[1] E. Appleton and D. J. Williams, *Industrial robot applications*. Springer Science & Business Media, 2012.

[2] T. B. Sheridan, "Human–robot interaction: status and challenges," *Human factors,* vol. 58, no. 4, pp. 525-532, 2016.

[3] T. Asfour *et al.*, "Armar-6: A collaborative humanoid robot for industrial environments," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018: IEEE, pp. 447-454.

[4] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: A new kind of tool," *IEEE Intelligent Systems and Their Applications,* vol. 15, no. 4, pp. 25-31, 2000.

[5] C.-H. Ting, W.-H. Yeo, Y.-J. King, Y.-D. Chuah, J.-V. Lee, and W.-B. Khaw, "Humanoid robot: A review of the architecture, applications and future trend," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 7, no. 7, pp. 1364-1369, 2014.

[6] K. Hambuchen, J. Marquez, and T. Fong, "A review of NASA human-robot interaction in space," *Current Robotics Reports,* vol. 2, no. 3, pp. 265-272, 2021.

[7] M. A. Goodrich and A. C. Schultz, *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[8] C. Matuszek, H. Soh, M. Gombolay, N. Gopalan, R. Simmons, and S. Nikoladis, "Machine Learning in Human-Robot Collaboration: Bridging the Gap," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022: IEEE, pp. 1275-1277.

[9] C. Hannum, R. Li, and W. Wang, "A Trust-Assist Framework for Human-Robot Co-Carry Tasks," *Robotics,* vol. 12, no. 2, pp. 1-19, 2023.

[10] C. J. Willemse, A. Toet, and J. B. Van Erp, "Affective and behavioral responses to robot-initiated social touch: toward understanding the opportunities and limitations of physical contact in human–robot interaction," *Frontiers in ICT,* vol. 4, p. 12, 2017.

[11] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating Human–Robot Collaborative Tasks by Teaching-Learning-Collaboration From Human Demonstrations," *IEEE Transactions on Automation Science and Engineering,* vol. 16, no. 2, pp. 640-653, 2018.

[12] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data,* vol. 3, no. 1, pp. 1-40, 2016.

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv Neural Inf Process Syst,* vol. 27, 2014.

[14] B. Huang, X. Chen, Y. Sun, and W. He, "Multi-agent cooperative strategy learning method based on transfer Learning," in *2022 13th Asian Control Conference (ASCC)*, 2022: IEEE, pp. 1095-1100.

[15] S. Ayadi and Z. Lachiri, "Deep Neural Network for visual Emotion Recognition based on ResNet50 using Song-Speech characteristics," in *2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, 2022: IEEE, pp. 363-368.

[16] I. Z. Mukti and D. Biswas, "Transfer learning based plant diseases detection using ResNet50," in *2019 4th International conference on electrical information and communication technology (EICT)*, 2019: IEEE, pp. 1-6.

[17] L. Alzubaidi *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data,* vol. 8, pp. 1-74, 2021.