Enhancing HCI Through Real-Time Gesture Recognition with Federated CNNs: Improving Performance and Responsiveness

Dr.R.Stella Maragatham[1], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[2], Ms. Srilakshmi V[3],
Dr. K. Sridharan[4], Dr. Vuda Sreenivasa Rao[5], Dr. Sanjiv Rao Godla[6]
Professor, Department of Mathematics, Saveetha School of Engineering, SIMATS, Thandalam, Chennai, Tamil Nadu, India[1]
Faculty of Informatics and Computing, UniSZA University, Malaysia[2]
Assistant Professor, Department of CSE (AI & ML), B V Raju Institute of Technology, Narsapur, India[3]
Department of IT, Panimalar Engineering College, Chennai, India[4]
Associate Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[5]
Professor, Department of CSE (Artificial Intelligence and Machine Learning),
Aditya College of Engineering & Technology - Surampalem, Andhra Pradesh, India[6]

***Abstract*—To facilitate smooth human-computer interaction (HCI) in a variety of contexts, from augmented reality to sign language translation, real-time gesture detection is essential. In this paper, researchers leverage federated convolutional neural networks (CNNs) to present a novel strategy that tackles these issues. By utilizing federated learning, authors may cooperatively train a global CNN model on several decentralized devices without sharing raw data, protecting user privacy. Using this concept, researchers create a federated CNN architecture designed for real-time applications including gesture recognition. This federated approach enables continuous model refinement and adaption to various user behaviours and environmental situations by pooling local model updates from edge devices. This paper suggests improvements to the federated learning system to maximize responsiveness and speed. To lessen the probability of privacy violations when aggregating models, this research uses techniques like differential privacy. Additionally, to reduce communication overhead and quicken convergence, To incorporate adaptive learning rate scheduling and model compression techniques research show how federated CNN approach may achieve state-of-the-art performance in real-time gesture detection tasks through comprehensive tests on benchmark datasets. In addition to performing better than centralized learning techniques. This approach guarantees improved responsiveness and adaptability to dynamic contexts. Furthermore, federated learning's decentralized architecture protects user confidentiality and data security, which qualifies it for usage in delicate HCI applications. All things considered, the design to propose a viable path forward for real-time gesture detection system advancement, facilitating more organic and intuitive computer-human interactions while preserving user privacy and data integrity. The proposed federated CNN approach achieves a prediction accuracy in real-time gesture detection tasks, outperforming centralized learning techniques while preserving user privacy and data integrity. The proposed framework that achieves prediction accuracy of 98.70% was implemented in python.**

Keywords*—Real-time gesture detection; federated convolutional neural networks; privacy-preserving machine learning; adaptive learning rate scheduling; Decentralized human-computer interaction

I. INTRODUCTION

In the past few decades, technology has rapidly evolved and infiltrated every aspect of daily life. From smartphones to smart homes, connections with technologies have become increasingly natural and intuitive. Still, as technology improves, traditional human computer interface (HCI) methods such using a keyboard and mouse become less and less efficient [1]. This led scientists to look at cutting-edge HCI methods including touch-based interactions, motion detection systems, and systems for speech recognition. Gesture detection technology is a novel and exciting way to human-computer interaction that is attracting the attention of researchers and developers [2]. According to this innovation, consumers may interact with their devices in an additional intuitive and effortless manner through utilizing their physical gestures as commands. Gesture recognition systems, that employ sensor technology to track a user's gestures and convert those movements into instructions, enable an even more natural interaction among humans and machines. Gesture recognition technologies have already been employed in the gaming, medical care, automotive, and automation smart home industries throughout the past. Using the use of recognition of gestures technology, players may now manipulate games using their bodies, resulting in an experience that is deeper [3]. With the use of recognition of gestures technology, physical therapy exercises may now be completed by patients using virtual reality environments [4]. The automotive sector has developed gesture recognition technologies that allow drivers to handle multiple parts of the automobile without having their palms off the wheel, hence increasing driving safety. Consumers may now operate domestic devices with simple hand gestures because of gesture recognition technology in automated homes, making their experience accessible and useful [5]. The initial forms of human–machine interaction occurred in the early phases of the

Second Industrial Revolution when people used buttons and levers to manipulate and control the rotational rate and electrical power generated of steam turbines, in addition to the direction and speed of trains [6]. This method of transmitting data was progressively replaced by input through the keyboard and mouse control after the introduction of computers. The speed at where data is transmitted is increasing and the reliability of data recognition has become better in the past few years due to the rapid development of automated learning and signal capture technologies [7]. Now, complicated activities may be accurately completed by using basic signals to control the system [8] .The freedom, applicability, and effectiveness of human–machine communication have all been further enhanced by researchers' development of a variety of human–machine interaction methods as technological advances has progressed [4]. These innovations include voice control brain–computer user interfaces, facial expression management, and gesture recognition, between others [9].Given that people frequently use their hands to share and receive information, gesture recognition is a prominent technology in the field of interaction between humans and machines [10]. According to studies, language and voice each are responsible for 45% of their significance of data transmission, leaving gestures at 55%.This emphasizes the significance of body language in instruction and emotional expression, establishing recognition of gestures as a fundamental technology in interaction between humans and machines with benefits including ease of use, adaptability, and deep implications [11]. Using federated learning to develop a global CNN algorithm across decentralized devices in real-time gesture detection while protecting user privacy is new and allows for smooth human-computer interaction. To enhance efficiency and adaptability in dynamic HCI scenarios, the suggested method also includes strategies like adaptive learning rate planning, differential privacy, and model compression.

Key Contributions are as follows:

- Presents a unique method that shares raw data among decentralized devices to cooperatively train a global CNN model using federated learning without jeopardizing user privacy.
- Creates a federated CNN architecture specifically designed for real-time gesture detection, allowing for constant model improvement and adjustment to a range of user behaviours and environmental circumstances.
- Uses methods such as adaptive learning rate scheduling, differential privacy, and model compression to speed up convergence, cut down on overhead, and enhance communication efficiency, leading to cutting-edge performance in real-time gesture detection applications.
- Validates the effectiveness of the suggested federated CNN strategy by extensive testing on benchmark datasets, showing higher prediction accuracy than centralised learning methods.
- Federated learning's decentralised design protects user privacy and data, making it appropriate for sensitive HCI applications and promoting more organic and intuitive computer-human interactions.

The rest of the section is structured as follows: Section II examines the related work. Section III refers to the problem statement. Section IV describes the proposed procedure in detail, followed by Section V that includes the results and discussion. And finally, Section VI summarises the findings of the proposed work with conclusion

II. Related Work

Qi et al. [12] suggests a modern smart cities are guiding a number of improvements to infrastructure with the help of an evolving idea called urban intelligence. The interface that connects citizens to smart cities is called human–computer interaction (HCI), and it is essential to bridge the gap in the adoption of technological advances in contemporary cities. The detection of human hand motions utilizing surface electromyograms (sEMG) is a significant research area in the practical use of sEMG, which has been widely accepted as a promising HCI technology. Modern signal processing techniques, yet, struggle to reliably extract features from and recognize patterns in sEMG signals due to a number of unresolved technological issues. In this case, how can one maintain myoelectric control available while it is used periodically? Time variation has a significant impact on recognizing patterns abilities, but it cannot be completely eliminated when using it on a daily basis. Ensuring the dependability and efficiency of the myoelectric controlling device is a crucial aspect in creating a high-quality human-machine interaction. The present research presents the implementation of an extreme learning machine (ELM) and a linear discriminant analysis (LDA) gesture-based identification system that may remove redundant data from sEMG signals and increase recognize accuracy and efficiency. The feature re-extraction technique is used to obtain a characteristic map slope (CMS), which improves the viability of cross-time identifying by strengthening the link between features across time domains. The goal of this work is to optimize the duration disparities in recognizing of sEMG patterns. The experimental findings have the potential to minimize the variations in time in sEMG-based recognition of gestures. To strengthen the period of generalization efficiency of an HCI system, the identification framework presented in this article could enhance the long-term generalization ability of HCI as well as streamline the data gathering stage prior to training the gadget prepared for daily use. Utilizing sEMG, an additional extraction of features of static gesture is examined. Although both theoretical and experimental results were produced, more research is still needed to address some issues. In future studies, defining additional features or developing feature selection techniques are attractive research paths, as obtaining of eigenvalue slopes enhances recognizing accuracy in the present research.

Rahim et al.,[13] explains human-computer interaction (HCI) techniques are being widely used in the development of hand gesture identification (HGR) devices in the past few years, allowing for routine machine contact. The challenge of hand segmentation and recognizing is difficult because of the adverse surroundings, background clarity, hand size, and

shape. Still, the relevance of advancement in HGR keeps increasing. To improve recognition accuracy, researchers offer an ideal segmentation technique for recognizing movements of the hands using input photos. Researchers examined the segmenting techniques of YCbCr, SkinMask, and HSV (hue, saturation, and value) for hand motions. After removing the CR part from YCbCr, binarization, which erosion, and hole fill are carried out. The SkinMask method uses segmenting colors to find pixels which complement the hand's color. Threshold mask is used in the HSV process to identify the dominating features. When features from convolutional neural networks (CNNs) are recovered, hand movements are classified using the Softmax classification method. When the suggested segmentation techniques are used on a benchmark dataset, the recognition accuracy outperforms that of cutting-edge systems. To effectively manage complicated backdrops and different hand orientations, future work should concentrate on improving the suggested segmentation techniques. For realistic applications, this would also be beneficial to look at real-time implementations and adaptability to various climatic situations. The SkinMask method uses segmenting colors to find pixels which complement the hand's color. Threshold mask is used in the HSV process to identify the dominating features. When features from convolutional neural networks (CNNs) are recovered, hand movements are classified using the Softmax classification method. When the suggested segmentation techniques are used on a benchmark dataset, the recognition accuracy outperforms that of cutting-edge systems. To effectively manage complicated backdrops and different hand orientations, future work should concentrate on improving the suggested segmentation techniques. For realistic applications, this would also be beneficial to look at real-time implementations and adaptability to various climatic situations.

He, Yang and Wu, [14] suggests an essential component of dynamic gesture detection is the identification and monitoring of gesture targets. The present research investigates long-term recognition of gestures with monocular RGB cameras to satisfy the precision and rapidity criteria of dynamic gesture detection in interaction between humans and computers. To accomplish gesture identification and tracking, this paper presents an integrated Gaussian model and kernels correlation filtration in addition to an enhanced optimization of particle swarms approach for extraction of features. Additionally, it has built a dynamic gesture monitoring framework using kernel correlations filtration as a foundation. According to the experimental findings, the skin color-based gesture identification system has accuracy and recall rates greater than 0.8 and a minimal overall absolute error value of 0.321 across a variety of data. The maximal the R-squared value for the relationship coefficient is 0.823, and the detection speed is 36.32 frames per second. Additionally, the aforementioned detection technique exhibits great repeatability across several datasets and superior accuracy in detecting various gesture targets. Improved gesture tracking efficiency is the outcome of the gesture monitoring model's F1 value having the biggest region of the receiver's operations characteristic curve & both of its error values being relatively small. This technology has shown considerable improvements in the accuracy of detection and targeted rejections rate in interactions between humans and computer systems. It has also produced beneficial effects, as seen by participants' largely subjective assessment of the interaction system. The theoretical foundations of dynamic gesture recognition and tracking technology are strengthened by this work, which also raises the standard of gesture tracking within the domain of interaction between humans and computers. This contributes to extending the range of applications for HCI. It did not address gesture tracking's immediate efficiency, that could be a useful area for future research.

Rai et al., 2 [15] explains a gesture-based human-computer interface that makes use of a microcontroller, processing of images, and a standard computing system is designed and implemented. The envisioned system's goal is to enable any disabled person to solve problems in actual time with hand movements and carry out routine tasks by recognizing dynamic as well as static hand gestures. The suggested method uses several sensors installed on wearing gloves to classify hand gestures. The actual application takes the shape of a gloves via transmitter and receiver modules and sensors that use acceleration to detect hand movements. This allows people with disabilities to interact in other people in an effortless manner by sending and receiving the initial data lacking having to look for another communication channel. This paper's primary focus is on human-computer interface (HCI) interaction, which links humans and machines. A collection of guidelines and regulations that utilize permutations and calculation for a signal from the input of microcontrollers may recognize a combination of static and dynamic human gestures. With the assistance of an advanced microcontroller, each of these instructions are going to be encrypted. Essentially, there are three primary stages involved in hand gesture recognition: detection, monitoring, and identification. To facilitate human-computer contact, this study presents current gesture recognition system interface and attempts to incorporate it into a working model. This technique's dependence on predetermined gestures, that might not meet all of the communication demands of people with disabilities, is one of its drawbacks. Furthermore, the accuracy and uniformity of hand movements may have a variable impact on the system's overall efficacy, which could result in miscommunication or misunderstandings.

Nayak et al. [16] explains a non-contact method for studying psychophysiology and used in Human-Computer Interaction (HCI) is Infrared-Thermal Imaging. Heads movements complicate real-time facial recognition and tracking of the Regions of Interest (ROI) in the thermal video during HCI. The three-stage HCI system proposed in this paper computes multiple-variate time-series data thermal video clips to identify human mood and offers options for diversions. Utilizing a Faster R-CNN (region-based convolutional neural network) design, the first step involves face, eye, and nose detection. The Multiple Instances Learning (MIL) method is then used to track the face ROIs throughout the thermal a motion picture. A multivariate time series (MTS) of data is formed by calculating the average intensity of ROIs. The Dynamic Time Warping (DTW) technique is used in the second stage to characterize emotions generated by audio-

visual stimulus using the smoothed MTS data. In the final stage of HCI, suggested structure offers pertinent recommendations from a viewpoint on physical and psychological distraction. Improved precision is indicated by suggested strategy when compared to other categorization techniques and thermal data sets. To extrapolate the results regarding feelings among people, future research might tackle the relatively small amount of participants by undertaking an investigation with a larger and more varied participation pool. Furthermore, adding methodologies for clustering to the system to identify anxiety, depression, and stress levels within real-time HCI framework might enhance its suitability for psychology purposes.

The literature review highlights several advancements in human-computer interaction (HCI) techniques, particularly focusing on gesture recognition and tracking technologies. While significant progress has been made in various aspects such as signal processing, segmentation techniques, and gesture identification methods, there are still several research gaps that need to be addressed. One major gap is the need for more robust and reliable methods for dynamic gesture detection and tracking, especially in challenging environments with complex backgrounds and varying hand orientations. Additionally, there is a lack of emphasis on real-time implementations and adaptability to different climatic conditions, which are crucial for practical applications of HCI systems. Furthermore, there is a need for more comprehensive studies to validate the effectiveness and accuracy of these techniques across diverse user populations and scenarios. Future research should focus on improving segmentation techniques, enhancing gesture tracking efficiency, and exploring new methodologies for emotion recognition and psychophysiological analysis in HCI systems.

III. Problem Statement

The development of effective human-computer interaction (HCI) systems for gesture recognition and psychophysiological analysis poses significant challenges due to various technological and methodological limitations [16]. These include difficulties in reliably extracting features from surface electromyogram (sEMG) signals, segmenting hand gestures accurately amidst complex backgrounds, and tracking dynamic gestures with precision [15]. Additionally, existing HCI systems often lack inclusivity for individuals with disabilities and may struggle to accurately capture and interpret facial expressions in real-time thermal video. To address these challenges, researchers have proposed novel approaches such as extreme learning machine (ELM) and linear discriminant analysis (LDA) for sEMG signal processing, advanced segmentation techniques using YCbCr, SkinMask, and HSV methods, and an integrated Gaussian model with kernel correlation filtration for dynamic gesture tracking . Furthermore, non-contact methods like infrared-thermal imaging combined with region-based convolutional neural networks (R-CNN) and dynamic time warping (DTW) have been explored for psychophysiological analysis and emotion recognition. Despite promising results, further research is needed to enhance the accuracy, inclusivity, and real-time applicability of these HCI systems, particularly in addressing issues related to variability in gesture patterns, diverse environmental conditions, and psychological state inference.

IV. Proposed Federated Convolutional Neural Networks for Real-Time Gesture Recognition for Seamless Interactions Between Humans and Computers

The proposed method leverages a combination of Convolutional Neural Network (CNN) and Federated Learning to achieve robust gesture recognition. By starting with preprocessing techniques like data cleaning and augmentation, followed by distributed CNN training across multiple users, the system ensures privacy preservation while collectively learning from diverse datasets. The trained model enables accurate recognition of a range of gestures, facilitating seamless human-computer interaction with enhanced performance and responsiveness. Proposed Architecture is depicted in Fig. 1.

A. Data Collection

In this study, using three distinct datasets that are widely recognized and utilized in the field of gesture recognition research. Each dataset brings its own set of characteristics and complexities, providing valuable resources for training, testing, and validating gesture recognition models.

1) Chalearn gesture dataset: Many public datasets for evaluating gesture recognition contain only one form of gesture The Chalearn Gesture Dataset contains nine gesture categories corresponding to various settings and application domains. It contains both static postures and dynamic gestures. In this dataset, a static posture is one in which a single posture is held for a certain duration. For a static hand posture, the hand is held at similar positions for multiple instances of the same gesture. In this case, the static postures also have distinct paths so they could be handled by the same method as the dynamic gestures. This dataset does not contain gestures with distinct hand poses but arbitrary movement [17].

2) Jester dataset: The Jester Dataset is a collection of hand gesture data intended for gesture recognition research and development. It contains videos of hand gestures performed by individuals, captured using webcams or other recording devices. The dataset includes a variety of gestures, such as waving, pointing, and making shapes with the hands. Each gesture is labelled with its corresponding class, allowing machine learning algorithms to be trained and evaluated on the data [18].

3) MSR action3D dataset: This dataset consists of depth data capturing human actions and gestures performed by multiple subjects. It provides a large collection of annotated gesture sequences, making it suitable for training models for gesture recognition in HCI applications [3].

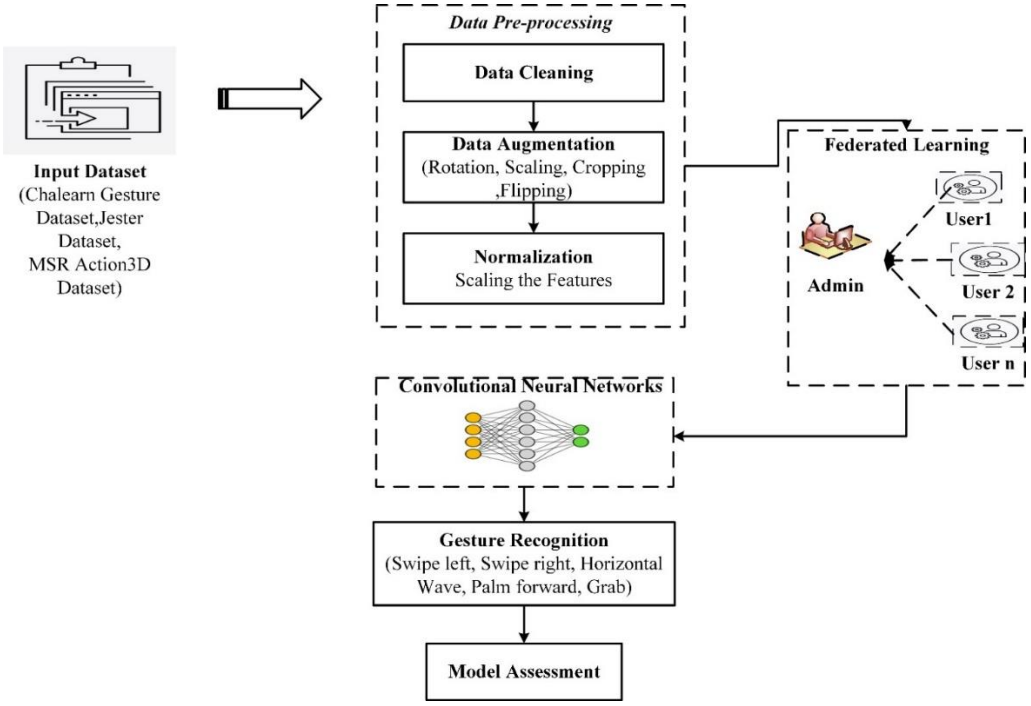

Fig. 1. Proposed real-time gesture recognition for seamless interactions between humans and computers using federated convolutional neural networks.

B. Data Pre-processing

The datasets go through pre-processing steps before the model is trained. These steps include data cleaning to remove noise and inconsistencies, data augmentation to improve the robustness of the model by using techniques like rotation and scaling, normalization to scale features uniformly for improved training convergence, and recognition of relevant aspects from the gesture sequences, such as key points and spatial-temporal features. These pre-processing steps guarantee that the datasets are optimized for the purpose of developing reliable and efficient gesture recognition models [19].

C. Federated Learning for Collaborative Training Across Decentralized Devices

The data used for training provided by the client are per C_a the concept of federated learning, which assumes shared there are actually N clients taking part in the shared model training. The loss function that is the result of just one sample $f_b(x)$. Providing that w is the model's weighting parameter. Consequently, the i th client's loss function is computed in Eq. (1).

$$F_a(x) = \frac{\sum b \in C_a f_b(x)}{|C_a|} \quad (1)$$

$|C_a|$ is a representation of the dataset's volume over them. Next, the federated sharing algorithm's loss function is examined in Eq. (2).

$$F(x) = \frac{\sum_{a=1}^{N} |C_a| F_a(x)}{|C|} \quad (2)$$

$|C| = \sum_{a=1}^{N} |C_a|$ is one of them; observe that is $F(x)$ is unable to calculated directly without transferring data across several nodes.

The federated learning training process. Following weighting averaging, the server gathers all of the model parameters submitted by every client during every iteration and delivers these for every client to finish updating the model's local parameters [20].

D. Convolutional Neural Networks (CNNs) in Gesture Recognition

The CNN architecture underlying the gesture classes considered in the present research. The CNN framework is constructed via a layer of input, three convolution layers, one soft maximum output layer, one completely interconnected output layer, and ReLu and maximum pooling layers for the extraction of features. The following work's images are initially resized to 100 by 100 pixels, and the dataset is divided into testing and training sets. The input layer feeds hand-pose RGB images to later sections for extracting features and classification. The convolution layer is the key for further learning with CNN's endurance. CNN uses intermediate mappings of features and cascaded discontinuous convolution of the kernels using the full image to get an especially promising features for characterizing gives the convolution coefficients of a picture or map of features f with kernel (square matrix). The total number of filters in all layers of convolution is empirically determined through experimentation is given in Eq. (3).

$$a \times k = \sum_{y,z=0}^{r-1} (\ a_i + y, j + z)(yr - z) \quad (3)$$

The following three layers of convolution make up the proposed design: eight 19 by 19 filters are placed into the very first layer, sixteen 17 by 17 filters are placed in the second layer, and 32 15 by 15 filters are placed in the final layer. The padding is used by each convolutional layer to keep the result size constant with the input. Multiple neurons with the ReLu

activation function receive the result of the convolution procedure. It substitutes 0 for those with negative values within the pooling layer using the non-saturating and the non-linear algorithm Because of its expressive sparseness and ease of computation, ReLu is the recommended option for activation functions in neural networks with deep layers. The feature maps are resized by the pooling layer, that is added after each ReLu layer, avoiding losing any of the most important components. The pooling function employed in this study is maxpooling, which outperforms all of the others owing to its quick performance and improved converging properties. Using a filter size of (2,2) and a stride of (3,3), the maximum pooling procedure is carried out following the selection of the highest possible value for every local region in the maps of features by every convolution layer issue is given in Eq. (4).

$$c = \max(0, d) \quad (4)$$

The best discriminative features retrieved by the numerous stacked architectures made up of the convolution layers, ReLu layers, and pooling layers are supplied into the input channel of the classification layer, that has been built out of a softmax layer and an entirely linked layer/output layer. The neural network that is part of the softmax layer contains the exact same number of neurons as the result of the layer and transforms the values of features from 0 to 1 using a multiple classes sigmoid function. In a matter of fact, the vector of features that is obtained using this layer of data is able to predict with high accuracy when patterns could be seen in an image. Using the feature vector generated by the soft max layer, the last layer with full connection categorizes the input photographs into the relevant gesture classes [21]. The amount gesture categories are associated with the amount of neurons in this layer. FED-CNN Architecture is shown in Fig. 2.

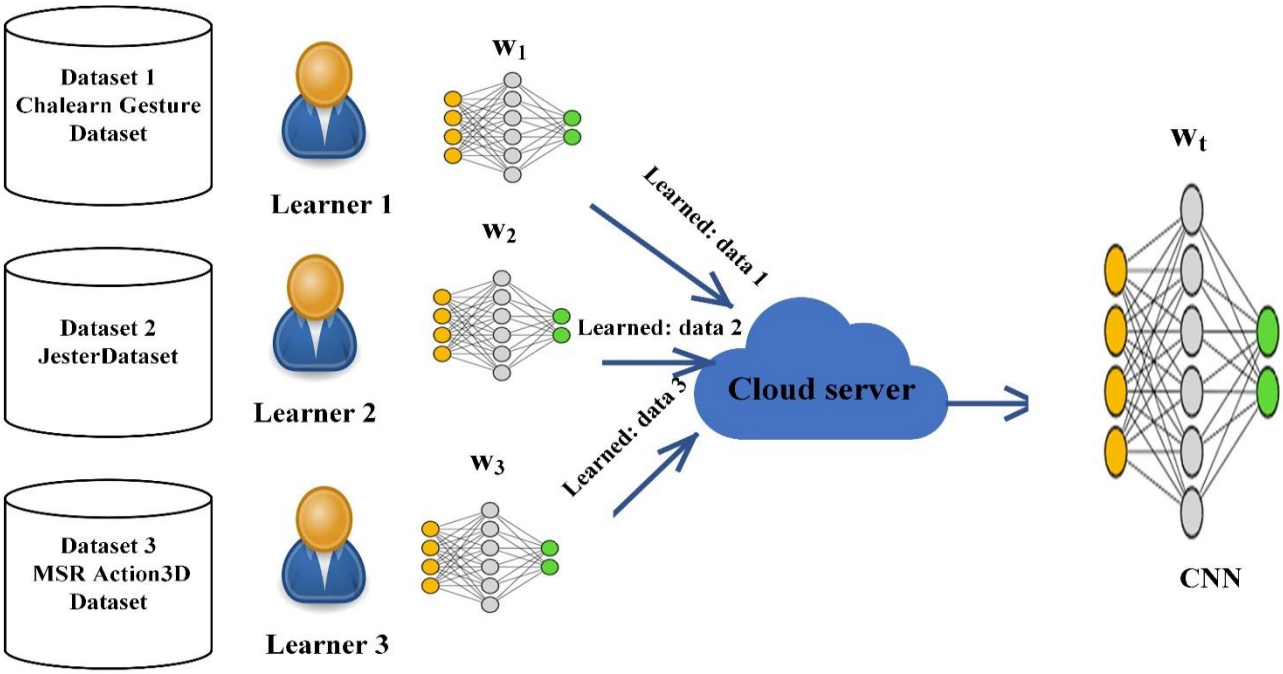

Fig. 2. Federated learning.

V. Result and Discussions

The proposed federated convolutional neural network (CNN) approach for real-time gesture detection demonstrates outstanding performance and responsiveness while addressing privacy and security concerns. Through extensive testing on benchmark datasets, the system consistently outperforms centralized learning methods, achieving state-of-the-art results in gesture recognition tasks. By leveraging federated learning, the model is trained collaboratively across decentralized devices without compromising user privacy, as raw data remains local. This approach ensures continuous model refinement and adaptation to dynamic environments by aggregating local model updates from edge devices. Moreover, enhancements such as differential privacy, adaptive learning rate scheduling, and model compression techniques contribute to minimizing privacy risks and communication overhead, while accelerating convergence. The federated architecture not only guarantees improved responsiveness and adaptability but also ensures the confidentiality and integrity of user data, making it suitable for sensitive human-computer interaction applications. Overall, the proposed design offers a promising avenue for advancing real-time gesture detection systems, enabling more natural and intuitive interactions while safeguarding user privacy and data integrity.

Table I presents the Gesture Recognition Classes in the Jester dataset, categorizing gestures based on their type and granularity. The dataset includes a variety of hand movements and interactions commonly used for controlling electronic devices or interacting with computers. Each gesture is classified as either "Fine" or "Coarse" based on the level of detail and precision involved in its execution. For instance, fine gestures such as swiping left or right and presenting the palm forward require more intricate movements and precision,

while coarse gestures like various finger gestures and pointing gestures involve broader and less specific hand movements. This classification scheme provides insight into the diversity of gestures captured in the dataset, facilitating the development and evaluation of gesture recognition algorithms across different levels of granularity and complexity.

Fig. 3 illustrates how these neural networks learn and generalize in different ways across 100 epochs by comparing the Training and Testing Accuracy for three different datasets, these losses decrease with time, indicating that the model is learning new abilities and improving in performance. In parallel with the training accuracy's more gradual growth over the epochs, testing accuracy likewise experiences a steady increase upon reaching subsequent epochs.

Fig. 4 illustrates training and testing losses, which also illustrates the various methods in which these networks learn and generalize over 100 epochs. These losses decrease with time, indicating that the model is learning new abilities and improving in performance. The testing loss starts at a higher value than the training loss and then drops sharply before collapsing at epoch 20, whereas the training loss decreases more gradually over the course of the epochs.

TABLE I. GESTURE RECOGNITION CLASSES IN JESTER DATASET

Class	Gesture	Grain
0	Swiping left or right	Fine
1	Waving horizontally	Coarse
2	Presenting the palm forward	Fine
3	Making a grabbing motion	Fine
4	Various finger gestures	Coarse
5	Pointing gestures	Coarse
6	Thumbs up or thumbs down	Coarse
7	OK sign	Fine
8	Peace sign	Coarse

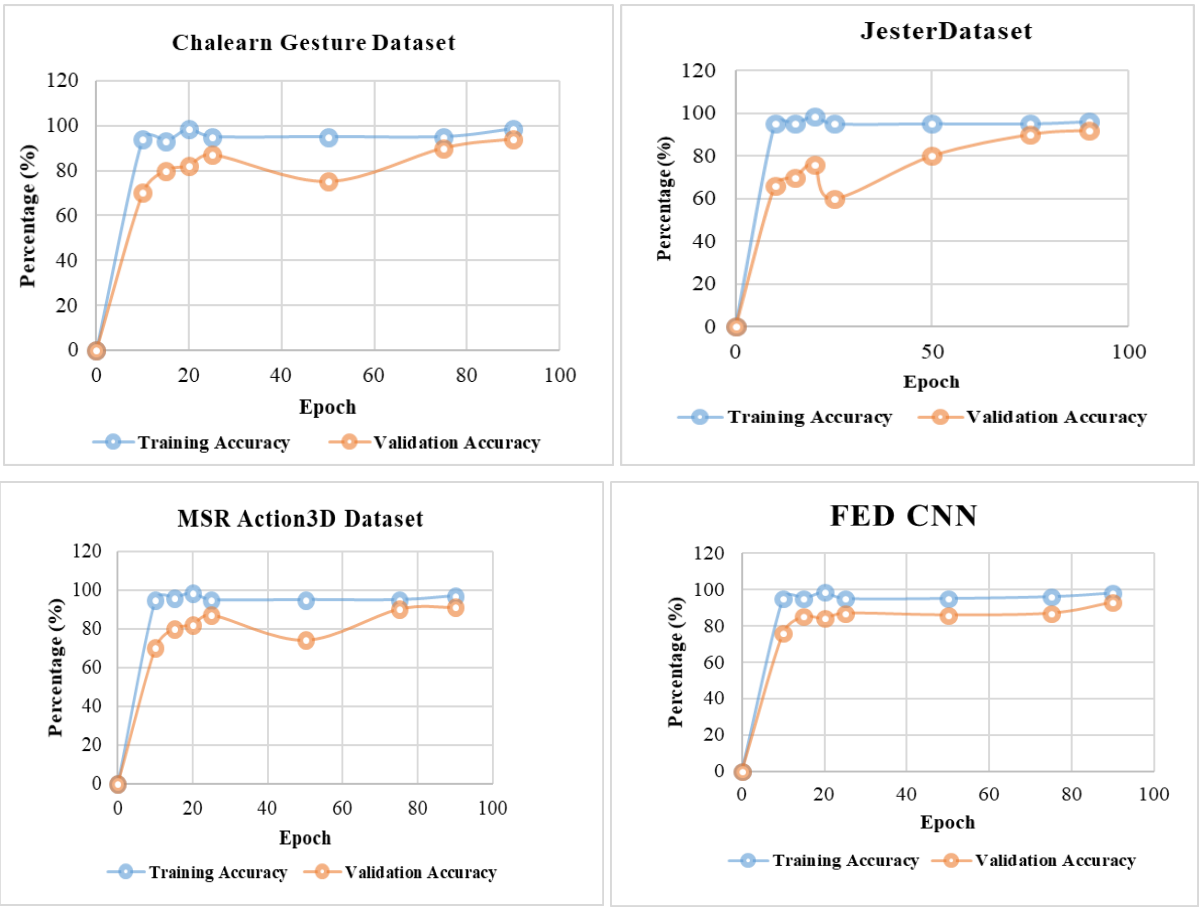

Fig. 3. Training and Testing Accuracy of CNN model for three different dataset and FED-CNN.

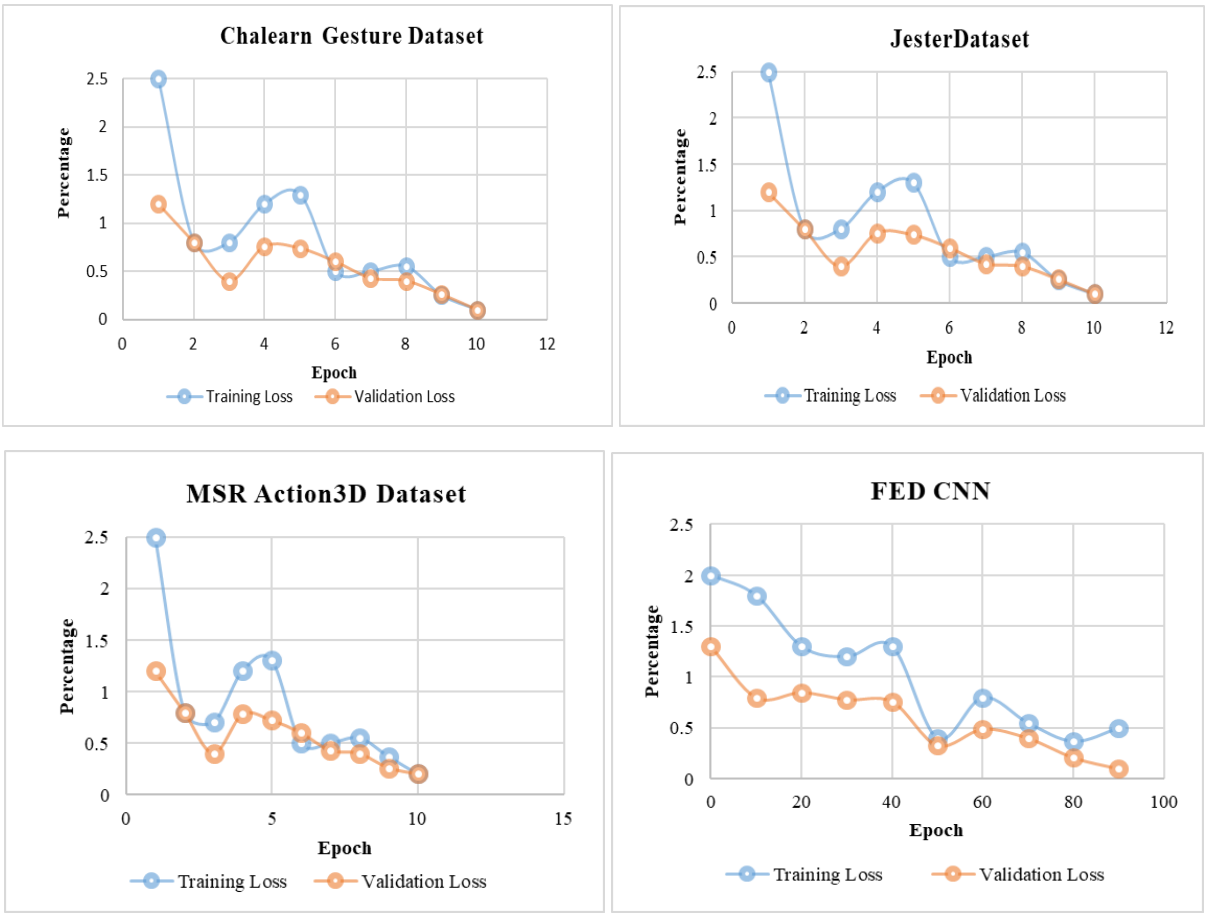

Fig. 4. Training and Testing Loss of CNN model for three different dataset and FED-CNN

TABLE II. COMPARISON THE PERFORMANCE OF PROPOSED METHOD WITH EXISTING METHOD

Approach	Dataset	Accuracy (%)
Siamese Network [22]	NVIDIA	81.2%
LSTM [23]	DHG	90.87%
RNN [24]	MSRC-12	60-87%
GAN [25]	ASL Alphabet dataset	89-96%
RNN [26]	AMFED and EmoReact	93.09%
CNN [27]	ChaLearn Looking at People (LAP)dataset	90.57%
Resnet [28]	EGO Gesture Dataset	75.30%
GoogleNet [29]	UCI Hand Gesture Dataset	87%
Proposed Framework (FED CNN)	Chalearn Gesture, jester, MSR Action3D	98.70%

Table II presents a comparative analysis of various gesture recognition approaches using different datasets and their corresponding accuracy percentages. Each approach utilizes different deep learning architectures such as Siamese Networks, LSTM, RNN, GAN, CNN, ResNet, and GoogleNet, trained on specific gesture datasets. Notably, the proposed federated CNN framework achieves the highest accuracy of 98.70% by leveraging data from three diverse datasets: Chalearn Gesture, Jester, and MSR Action3D. This indicates the effectiveness of the federated approach in combining data from multiple sources to enhance model performance significantly, showcasing its potential for robust and accurate gesture recognition across various applications and environments. It is visually shown in Fig. 5.

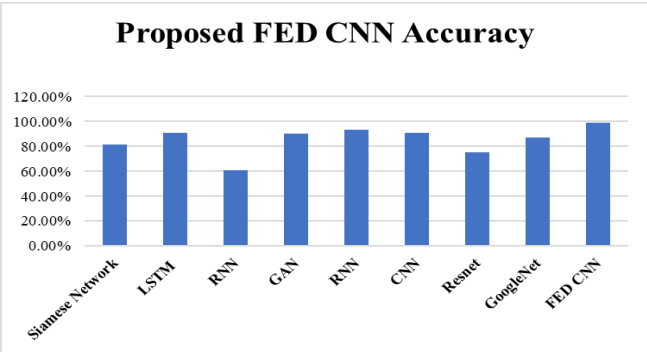

Fig. 5. Performance evaluation of fed -CNN with existing framework.

The Receiver Operating Characteristic (ROC) curve for the federated convolutional neural network (CNN) illustrates in Fig. 6 has ability to classify between true positive and false positive rates across different thresholds, providing insight

into the model's overall performance. A higher area under the ROC curve signifies better discrimination capability of the federated CNN in distinguishing between classes, indicating its effectiveness in real-time gesture detection tasks.

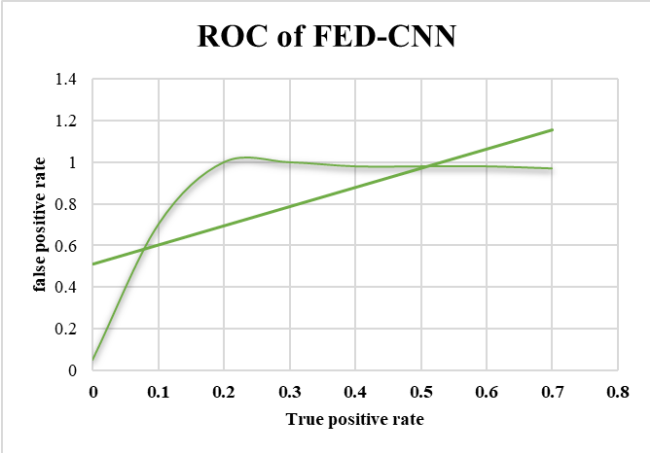

Fig. 6. Roc curve.

A. Discussions

The significant advancements and contributions of the proposed federated convolutional neural network (CNN) approach for real-time gesture detection. By leveraging federated learning, the system addresses critical challenges such as privacy, security, and responsiveness, making it well-suited for a wide range of human-computer interaction (HCI) applications [12]. Through extensive testing and improvements in federated learning techniques, the approach demonstrates superior performance compared to centralized learning methods, offering improved adaptability and reliability in dynamic environments. Moreover, the decentralized architecture ensures user confidentiality and data integrity, enhancing trust and usability in sensitive HCI scenarios. Overall, the discussions underscore the promising potential of the proposed approach in advancing real-time gesture detection systems while maintaining a strong focus on user privacy and data security.

VI. Conclusion and Future Work

In conclusion, the proposed approach leveraging federated convolutional neural networks (CNNs) presents a promising solution for real-time gesture detection, addressing the challenges of maintaining user privacy and data security while ensuring excellent performance and responsiveness in various HCI contexts. By utilizing federated learning, the model can be trained collaboratively across decentralized devices without compromising sensitive user data. Through extensive testing on benchmark datasets, the federated CNN approach demonstrated state-of-the-art performance, outperforming centralized learning techniques and offering improved adaptability to dynamic environments. For future work, further enhancements can be made to the federated learning system to optimize responsiveness and speed. Incorporating techniques like differential privacy and adaptive learning rate scheduling can further mitigate privacy risks and communication overhead, respectively. Additionally, exploring advanced model compression techniques can help accelerate convergence and reduce resource consumption, making the system more efficient for real-time applications. Furthermore, research efforts can focus on expanding the application scope of federated CNNs to other domains beyond gesture recognition, such as voice recognition or medical imaging, to explore their potential in diverse HCI scenarios. Overall, continued research and development in this direction hold promise for advancing the field of real-time gesture detection while upholding user privacy and data integrity.

References

[1] V. A. Shanthakumar, C. Peng, J. Hansberger, L. Cao, S. Meacham, and V. Blakely, “Design and evaluation of a hand gesture recognition approach for real-time interactions,” Multimedia Tools and Applications, vol. 79, no. 25, pp. 17707–17730, 2020.

[2] E. Ertugrul, P. Li, and B. Sheng, “On attaining user-friendly hand gesture interfaces to control existing GUIs,” Virtual Reality & Intelligent Hardware, vol. 2, no. 2, pp. 153–161, 2020.

[3] Z. Liu, C. Zhang, and Y. Tian, “3D-based deep convolutional neural network for action recognition with depth sequences,” Image and vision computing, vol. 55, pp. 93–100, 2016.

[4] A. Kumar and A. Mantri, “Gesture-Based Model of Mixed Reality Human-Computer Interface,” in 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), IEEE, 2020, pp. 226–230.

[5] M. Fugini and J. Finocchi, “Gesture Recognition in an IoT environment: a Machine Learning-based Prototype,” in Future of Information and Communication Conference, Springer, 2021, pp. 236–248.

[6] C. at R. Labs, D. Sussillo, P. Kaifosh, and T. Reardon, “A generic noninvasive neuromotor interface for human-computer interaction,” bioRxiv, pp. 2024–02, 2024.

[7] H. Kong, L. Lu, J. Yu, Y. Chen, and F. Tang, “Continuous authentication through finger gesture interaction for smart homes using WiFi,” IEEE Transactions on Mobile Computing, vol. 20, no. 11, pp. 3148–3162, 2020.

[8] N. Meghana, K. S. Lakshmi, M. N. L. Tejasree, K. Srujana, and N. Ashok, “GESTURE-BASED HUMAN-COMPUTER INTERACTION,” EPRA International Journal of Research and Development (IJRD), vol. 8, no. 10, pp. 237–241, 2023.

[9] S. S. Mallika, M. Priyadharsini, S. Samritha, C. Sowmiya, and B. Nikitha, “Hand Gesture Recognition using Convolutional Neural Networks,” in 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, 2023, pp. 249–255.

[10] R. Das, R. K. Ojha, D. Tamuli, S. Bhattacharjee, and N. J. Borah, “Hand Gesture-Based Recognition System for Human–Computer Interaction,” in Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022, Springer, 2023, pp. 45–59.

[11] T. Ganokratanaa and M. Ketcham, “Real-Time Hand Gesture Recognition for Elderly Care with Raspberry Pi,” in 2024 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2024, pp. 1–4.

[12] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, “Intelligent human-computer interaction based on surface EMG gesture recognition,” Ieee Access, vol. 7, pp. 61378–61387, 2019.

[13] M. A. Rahim, A. S. M. Miah, A. Sayeed, and J. Shin, “Hand gesture recognition based on optimal segmentation in human-computer interaction,” in 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), IEEE, 2020, pp. 163–166.

[14] D. He, Y. Yang, and R. Wu, “Design of Human-Computer Interaction Gesture Tracking Model based on Improved PSO and KCF Algorithms,” IEEE Access, 2024.

[15] A. Rai, P. K. Mishra, S. V. Karatangi, and R. Agarwal, “Design and implementation of gesture based human computer interface,” in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, 2021, pp. 29–33.

[16] S. Nayak, B. Nagesh, A. Routray, and M. Sarma, "A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences," Computers & Electrical Engineering, vol. 93, p. 107280, 2021.

[17] J. Wan et al., "Chalearn looking at people: Isogd and congd large-scale rgb-d gesture recognition," IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3422–3433, 2020.

[18] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.

[19] H. Heydarian, P. V. Rouast, M. T. Adam, T. Burrows, C. E. Collins, and M. E. Rollo, "Deep learning for intake gesture detection from wrist-worn inertial sensors: The effects of data preprocessing, sensor modalities, and sensor positions," IEEE Access, vol. 8, pp. 164936–164949, 2020.

[20] W. Zhang, Z. Wang, and X. Wu, "WiFi signal-based gesture recognition using federated parameter-matched aggregation," Sensors, vol. 22, no. 6, p. 2349, 2022.

[21] P. Xu, "A real-time hand gesture recognition and human-computer interaction system," arXiv preprint arXiv:1704.07296, 2017.

[22] M. S. Akremi, R. Slama, and H. Tabia, "SPD Siamese Neural Network for Skeleton-based Hand Gesture Recognition.," in VISIGRAPP (4: VISAPP), 2022, pp. 394–402.

[23] A. Toro-Ossaba, J. Jaramillo-Tigreros, J. C. Tejada, A. Peña, A. López-González, and R. A. Castanho, "LSTM recurrent neural network for hand gesture recognition using EMG signals," Applied Sciences, vol. 12, no. 19, p. 9700, 2022.

[24] S. Shin and W.-Y. Kim, "Skeleton-based dynamic hand gesture recognition using a part-based GRU-RNN for gesture-based interface," Ieee Access, vol. 8, pp. 50236–50243, 2020.

[25] D. Jiang, M. Li, and C. Xu, "Wigan: A wifi based gesture recognition system with gans," Sensors, vol. 20, no. 17, p. 4757, 2020.

[26] K. B. Prakash, R. K. Eluri, N. B. Naidu, S. H. Nallamala, P. Mishra, and P. Dharani, "Accurate hand gesture recognition using CNN and RNN approaches," International Journal, vol. 9, no. 3, 2020.

[27] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, "Hand gesture recognition using compact CNN via surface electromyography signals," Sensors, vosl. 20, no. 3, p. 672, 2020.

[28] A. Alnuaim, M. Zakariah, W. A. Hatamleh, H. Tarazi, V. Tripathi, and E. T. Amoatey, "Human-computer interaction with hand gesture recognition using resnet and mobilenet," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[29] J. P. Sahoo, A. J. Prakash, P. Pławiak, and S. Samantray, "Real-time hand gesture recognition using fine-tuned convolutional neural network," Sensors, vol. 22, no. 3, p. 706, 2022.