SIGN SENSE: AN AI-BASED SYSTEM TO HELP DEAF, DUMB, AND BLIND PEOPLE TO COMMUNICATE

Bridging Communication Gaps Using Artificial Intelligence and Computer Vision

1Meka Seshu Babu, 1Allamsetty Chaitanya Kumar, 1Machineti Pavan, 2B. Ramya

1Students, Department of Computer Science and Engineering

2Assistant Professor, Department of Computer Science and Engineering

SRK Institute of Technology

Vijayawada, India

Abstract : This study has been undertaken to develop an intelligent assistive communication system named *Sign Sense* that helps Deaf, Dumb, Blind, and normal individuals to communicate effectively. The system is based on computer vision and deep learning techniques to recognize sign language gestures in real time. To achieve this, hand gestures are captured using a web camera and processed using image preprocessing and feature extraction methods. A Convolutional Neural Network (CNN) model is employed to classify the gestures and convert them into meaningful text and audio outputs. In addition to gesture recognition, speech input is also processed using speech recognition techniques to enable bidirectional communication. The system generates multiple forms of output such as text for normal users and audio feedback for blind users. The experimental analysis is carried out under real-time conditions, and the system demonstrates high accuracy and efficient response time. The analytical framework of the proposed system includes data acquisition, preprocessing, feature extraction, classification, and output generation. The results indicate that the system effectively reduces communication barriers and improves accessibility for differently-abled individuals.

I. INTRODUCTION

INTRODUCTION

Sign language communication is widely used by individuals with hearing and speech impairments as a primary means of expressing thoughts and interacting with others. However, the lack of awareness and understanding of sign language among the general population creates a significant communication gap. With the rapid advancement of artificial intelligence and computer vision, sign language recognition systems have gained popularity as assistive technologies. A sign language recognition system can be defined as a system that captures hand gestures, processes visual data, and converts them into meaningful text or speech output. In simple terms, it acts as a bridge between sign language users and non-sign language users, enabling effective and real-time communication.

The Sign Sense system focuses on improving communication accessibility by integrating gesture recognition, speech processing, and multimodal output generation. The system uses a web camera to capture hand gestures and applies image processing and deep learning techniques to recognize and interpret them. In addition, speech input from users is converted into text and corresponding gesture outputs to support bidirectional communication. This approach helps reduce communication barriers and provides an inclusive environment for Deaf, Dumb, Blind, and normal individuals. The system also minimizes the risk of social isolation and enhances interaction in areas such as education, healthcare, and public services, thereby contributing to a more accessible and connected society.

NEED OF THE STUDY.

The increasing number of individuals with hearing, speech, and visual impairments has created a significant challenge in effective communication within society. The lack of a common communication medium between sign language users and normal individuals has led to difficulties in daily interactions, education, healthcare services, and employment opportunities. The seriousness of this communication gap has become more evident in recent years as the need for inclusive technologies continues to grow. Studies indicate that a large portion of the population is not familiar with sign language, which creates barriers for deaf and speech-impaired individuals to express their needs and access essential services. Among all differently-abled individuals, Deaf/Dumb and Blind users face continuous challenges in communication due to the absence of real-time interpretation systems.

The need for an intelligent assistive communication system is essential to reduce these barriers and improve accessibility. The interaction between users requires effective handling of gestures, speech, and text through proper acquisition, processing, and output generation.

3.1Population and Sample

The Sign Sense system focuses on sign language gestures and speech inputs collected from users for recognition and communication purposes. The dataset used in this study consists of hand gesture images and real-time video inputs representing commonly used sign language gestures. These gestures are selected based on their frequency of use and relevance in daily communication. The dataset represents different categories of signs, including alphabets, numbers, and basic communication

gestures, which together form the population of the study. The collected data reflects variations in hand shapes, orientations, and lighting conditions, ensuring that the system performs effectively in real-world environments.
The study comprises a selected sample of gesture data and speech inputs used for training and testing the system. A subset of gesture classes is chosen based on clarity, usability, and recognition feasibility, and multiple samples for each gesture are collected from different users to improve model accuracy. The system also considers speech inputs from normal and blind users to enable bidirectional communication. Real-time data captured through a web camera and microphone is treated as the primary sample for evaluation. This approach ensures that the model is trained on diverse and practical data, thereby improving the reliability and performance of the Sign Sense system.

3.2 Data and Sources of Data

For this study, primary data has been collected. The dataset used for the Sign Sense system is a custom dataset created by capturing hand gesture images and real-time video using a web camera. The gestures are performed by different individuals and recorded under various lighting conditions and backgrounds to ensure diversity and improve system performance. In addition, speech data is collected using a microphone to support speech-to-text conversion and enable bidirectional communication. The dataset includes multiple samples for each gesture category to enhance the accuracy of the recognition model.
The real-time data is collected during the development and testing phase of the system. Gesture images and video frames are continuously captured and processed for training, validation, and testing of the model. The collected data is preprocessed, labeled, and organized into different gesture classes for effective learning.

3.3 Theoretical framework

Variables of the study consist of dependent and independent variables. The study uses a predefined approach for the selection of variables related to the Sign Sense system. The dependent variable in this study is the *gesture recognition output*, which represents the final predicted result of the system in the form of text or audio. This output is generated based on the input provided by the user through gestures or speech. The accuracy and effectiveness of the system are measured through this dependent variable. The independent variables include hand gestures, speech input, image features, and environmental conditions such as lighting and background.

RESEARCH METHODOLOGY

The methodology section outlines the plan and method of how the study is conducted. This includes the dataset of the study, sample of the study, data and sources of data, study variables, and analytical framework of the Sign Sense system. The study is based on a custom dataset collected using a web camera and microphone, which captures hand gestures and speech inputs from users. The collected data is preprocessed, labeled, and used for training and testing the system.

3.1Population and Sample

The Sign Sense system considers sign language gestures and speech inputs as the primary elements of the study. The dataset consists of various hand gesture categories selected based on commonly used signs in daily communication. These gesture categories represent different patterns of hand movements, shapes, and orientations, which together form the population of the study. The dataset reflects variations in users, environmental conditions, and gesture execution, making it suitable for real-time application. Therefore, it can be regarded as the universe of the study.
The study comprises gesture samples collected from different individuals using a web camera. A limited number of gesture classes are selected based on usability and recognition feasibility, and multiple samples are collected for each class. The system also includes speech inputs from users to support bidirectional communication. The collected samples are used for training and testing the model. The data collected during the development phase is treated as the base dataset for the Sign Sense system.

3.2 Data and Sources of Data

For this study, primary data has been collected. The dataset is created using a web camera by capturing hand gesture images and real-time video of sign language performed by different users. In addition, speech data is collected using a microphone to support speech-to-text conversion. The collected data includes multiple samples for each gesture category under different lighting conditions and backgrounds to improve the robustness of the system.
The real-time data is collected during the system development and testing phase. Gesture images and video frames are continuously recorded and processed for training and evaluation of the model. The collected data is preprocessed, labeled, and organized into different gesture classes.

3.3 Theoretical framework

Variables of the study consist of dependent and independent variables. The study uses a pre-specified method for the selection of variables related to the Sign Sense system. The study uses gesture recognition output as the dependent variable. From the input provided by the user in the form of hand gestures or speech, the system generates output in the form of text or audio. The accuracy and effectiveness of the system depend on how well the input is processed and recognized. The rate at which the system correctly identifies gestures and produces meaningful output is considered as the performance of the system.
Hand gestures, speech input, and image features are considered as independent variables for the system. The captured hand gestures using a web camera are processed through image preprocessing and feature extraction techniques. These features include hand landmarks, shape, position, and movement of the hand. It is assumed that better feature extraction leads to higher recognition accuracy. Variations in gesture execution, lighting conditions, and background may affect the performance of the system. Therefore, these factors play a significant role in determining the efficiency of gesture recognition.
Speech input is another important independent variable used in the system. Speech recognition techniques are applied to convert voice input into text. It is assumed that clarity of speech and noise levels in the environment influence the accuracy of speech-to-

text conversion. The system establishes a relationship between speech input and gesture output to enable bidirectional communication. The effectiveness of this conversion process directly impacts the usability of the system for blind and normal users. Deep learning models, particularly Convolutional Neural Networks (CNNs), are used as the core analytical component of the system. The model learns patterns from the dataset and establishes a relationship between input features and output classes. It is assumed that higher quality training data and optimized model parameters improve the system performance. The response time and recognition accuracy are used as key indicators to evaluate the relationship between input variables and system output. Thus, the theoretical framework highlights the interaction between gesture input, speech input, processing techniques, and output generation in achieving an efficient and reliable communication system.

3.4Statistical tools and econometric models

This section elaborates the statistical and analytical techniques which are used to carry the study from data processing towards meaningful inferences. The methodology of the Sign Sense system involves evaluating the performance of gesture recognition and speech processing using standard evaluation metrics. The system processes input data such as hand gestures and speech, and produces output in the form of text and audio. The effectiveness of the system is measured through accuracy, precision, recall, and response time. The details of the methodology are given as follows.

3.4.1 Descriptive Statistics

Descriptive statistics have been used to analyze the performance of the system by measuring parameters such as accuracy, response time, and recognition rate. These measures help in understanding the overall behavior and efficiency of the system under different conditions. The variation in recognition accuracy across different gesture classes and environmental conditions is observed to evaluate system stability. When the performance metrics show consistent values, it indicates that the system is reliable and less sensitive to variations. However, fluctuations in accuracy and response time indicate sensitivity towards factors such as lighting conditions, background noise, and gesture variations.

3.4.2 Model for Gesture Recognition and System Evaluation

After the descriptive analysis, the methodology follows the next step in order to evaluate the performance of the Sign Sense system. The primary objective of the system is to determine how accurately the input gestures and speech are recognized and converted into meaningful outputs. The model focuses on identifying the relationship between input features such as hand gestures, speech signals, and extracted image features with the final predicted output. The deep learning model, particularly the Convolutional Neural Network (CNN), is used to learn patterns from the dataset and classify gestures into predefined categories. The effectiveness of the model is determined by how well it generalizes to new and unseen inputs.

The study follows a two-stage evaluation process. In the first stage, the model is trained using labeled gesture data, where input features such as hand landmarks and image patterns are mapped to specific gesture classes. In the second stage, the trained model is tested using real-time inputs to evaluate its prediction accuracy and performance. This approach helps in minimizing errors in prediction and improves the reliability of the system. The use of multiple samples and diverse data reduces the problem of overfitting and ensures better estimation of the model performance.

To further evaluate the system, performance metrics such as accuracy, precision, recall, and response time are used. The response time measures how quickly the system processes input and generates output. If the response time is low and accuracy is high, the system is considered efficient. Variations in performance may occur due to environmental factors such as lighting conditions, background noise, and gesture variations. Therefore, consistency in results indicates the robustness of the model.

Additionally, error analysis is performed to identify misclassified gestures and improve the system performance. The dataset is carefully selected to avoid missing values and ensure proper labeling of gesture classes. The samples are grouped and organized to improve training efficiency and model stability. This structured approach enhances the overall effectiveness and accuracy of the Sign Sense system in real-time communication scenarios.

3.4.2.1 Model for Gesture Recognition

n the first stage, a classification model is used to estimate the relationship between input features and gesture classes. The Convolutional Neural Network (CNN) model is applied to extract features from input images and classify them into predefined gesture categories.

$$y = f(x; \theta)$$

Where x represents the input features (such as hand landmarks, image pixels, and gesture patterns), θ represents the model parameters, and y represents the predicted gesture class.

The input data consisting of gesture images is processed through multiple layers of the CNN model, including convolution, pooling, and fully connected layers. These layers extract important spatial features from the images. The model is trained using labeled data, where each gesture corresponds to a specific class. The prediction accuracy depends on how well the model learns the mapping between input features and output classes.

In the second stage, the predicted gesture outputs are converted into meaningful communication forms such as text and audio. The system evaluates the performance of the model using metrics such as accuracy and loss function.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Where the number of correctly predicted gestures is divided by the total number of predictions made by the system. Higher accuracy indicates better system performance.

The model is tested using real-time input data, and the predicted outputs are compared with actual gesture classes. The error term represents the difference between predicted and actual outputs, which is minimized during the training process. This two-stage approach ensures that the system effectively recognizes gestures and generates accurate outputs for communication.

3.4.2.2 Model for Multimodal Recognition System

In the first stage, multiple input features such as hand gestures, speech signals, and environmental factors are used to compute the feature coefficients through the learning model. The system considers different factors that influence recognition performance, such as gesture features, speech clarity, lighting conditions, and background variations.

$$y = \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3 + \beta_4 f_4 + \epsilon$$

Where y represents the predicted output, f_1 represents gesture features, f_2 represents speech input features, f_3 represents image/environmental conditions, f_4 represents preprocessing and feature extraction effects, β_1 *to* β_4 represent the sensitivity of the system to these factors, and ε is the error term.

In the second stage, the system performs evaluation using the combined effect of these factors on the final output. A regression-like relationship is established between the predicted outputs and the influencing feature coefficients to analyze the system performance.

$$\hat{y} = \gamma_0 + \gamma_1\beta_1 + \gamma_2\beta_2 + \gamma_3\beta_3 + \gamma_4\beta_4 + \epsilon$$

Where $\hat{y}$ represents the average predicted output performance, γ_0 is the intercept, γ_1 *to* γ_4 represent the contribution of each factor, β_1 *to* β_4 are the estimated coefficients from the first stage, and ε is the error term.

The model evaluates how different input factors influence the overall system performance. It is assumed that gesture features and speech input have a direct impact on recognition accuracy, while environmental conditions may affect the system either positively or negatively. The combination of these factors determines the efficiency and reliability of the Sign Sense system. This approach helps in analyzing the contribution of multiple variables in improving the accuracy and robustness of the system.

3.4.3 Comparison of the Models

The next step of the study is to compare the different models used in the Sign Sense system to evaluate which model is more effective and accurate in recognizing gestures and processing speech inputs. The study compares the performance of individual models such as gesture recognition (CNN-based model) and multimodal recognition (combined gesture and speech model). The comparison is based on evaluation metrics such as accuracy, response time, and reliability. This helps in determining which model is more suitable for real-time communication. The study follows analytical comparison methods to evaluate the effectiveness of these models.

.

3.4.3.1 Model Comparison Equation

The gesture-based model can be considered as a specific case of the multimodal system, as the multimodal model includes both gesture and speech inputs. These models are non-identical because they use different input features and processing techniques. To compare the performance of these models, a combined evaluation approach is used. The comparison equation is as follows:

$$y = \alpha y_{multi} + (1 - \alpha) y_{gesture} + e$$

Where y represents the overall system performance, $y_{(multi)}$ represents the performance of the multimodal model, $y_{(gesture)}$ represents the performance of the gesture-only model, and α measures the effectiveness of the models. If α is close to 1, the multimodal model is considered more accurate and efficient compared to the gesture-only model. If α is closer to 0, the gesture-based model performs better.

The comparison helps in identifying the most suitable model for real-time communication. The results indicate that combining multiple input sources generally improves system performance, making the multimodal model more reliable and efficient for practical applications.

3.4.3.2 Posterior Odds Ratio

A standard assumption in system performance evaluation is that the output errors of the models follow a consistent and independent distribution. In the Sign Sense system, it is assumed that the residual errors obtained from gesture recognition and multimodal recognition models are independently and identically distributed. Based on this assumption, it is possible to compare the performance of the two models using the posterior odds ratio. This method provides a more formal and theoretically sound comparison compared to simple model evaluation techniques.

The comparison is performed using the posterior odds ratio in favor of one model over another. The formula for posterior odds ratio is given as follows:

$$R = \left(\frac{ESS_1}{ESS_2}\right)^{\frac{N}{2}} \cdot N^{\frac{K_1 - K_2}{2}}$$

Where ESS_1 represents the error sum of squares of the multimodal model, ESS_2 represents the error sum of squares of the gesture-based model, N is the number of observations, K_1 is the number of independent variables in the multimodal model, and K_2 is the number of independent variables in the gesture-based model.

.

This comparison helps in determining the most efficient model for the Sign Sense system by evaluating the contribution of multiple input variables and their impact on overall system performance.

IV. RESULTS AND DISCUSSION

4.1 Results of System Implementation

The results of the Sign Sense system are presented through real-time gesture recognition using a web-based interface. The system captures hand gestures using a camera and processes them to generate predictions along with confidence levels. The interface includes camera feed, control buttons, and recognition history, which helps in monitoring system performance.

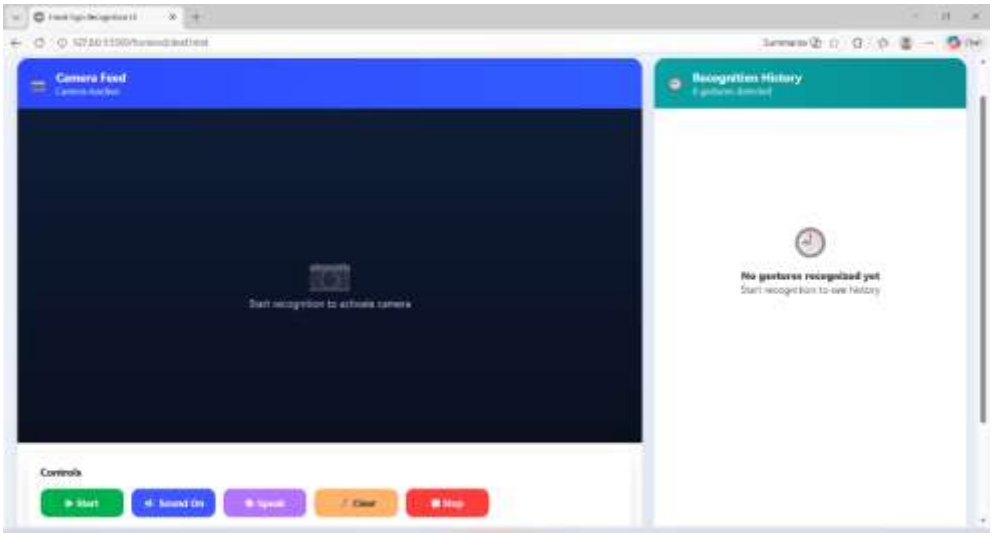


Fig4.1:Initial State

Fig. 4.1 shows the initial state of the system where the camera is inactive and no gestures are detected. This indicates that the system remains in standby mode until the user initiates the recognition process.

.

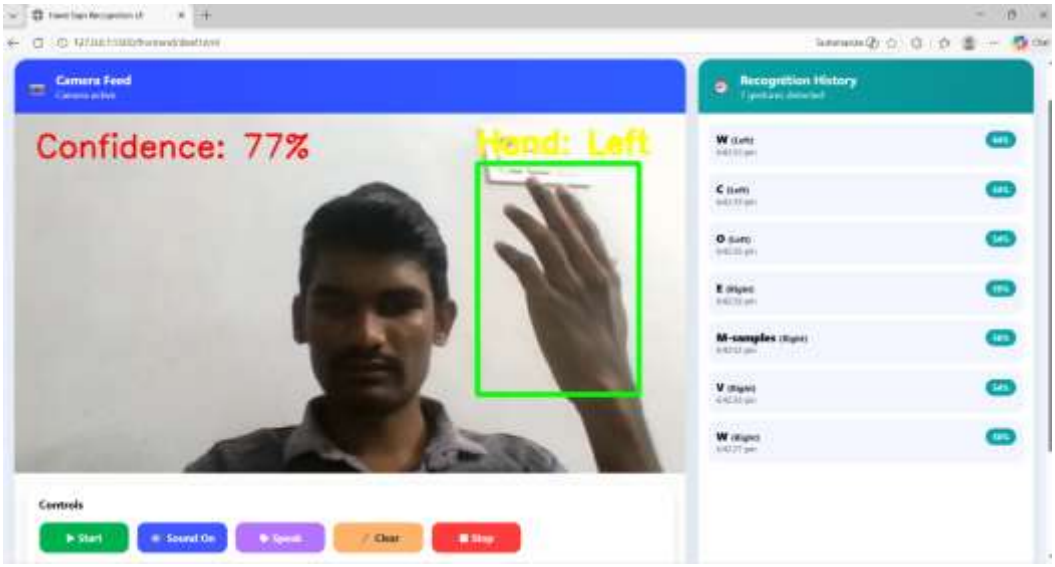


Fig 4.2

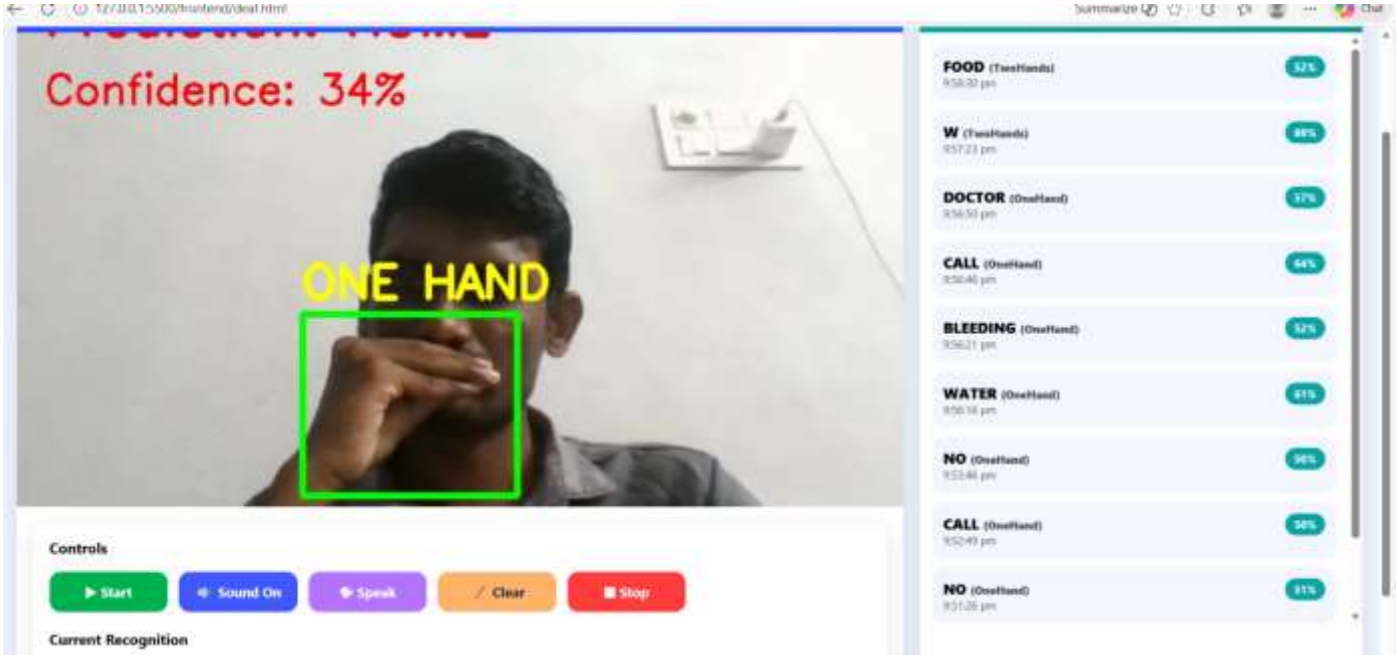
Fig. 4.2 shows the successful recognition of a hand gesture with a confidence level of approximately **77%**. The system correctly identifies the hand (left/right) and records the detected gestures in the recognition history panel. This demonstrates that the system performs accurately under normal conditions

Fig. 4.3 shows a gesture detected with a confidence level of around **47%**. This indicates moderate accuracy, which may be affected by environmental conditions such as lighting, hand positioning, or background variations. However, the system is still able to recognize the gesture.

Fig 4.3

Fig. 4.4shows another gesture recognition result where the system detects a gesture labeled as **"ONE HAND"** with a confidence level of approximately **34%**. The system successfully identifies the gesture category (one-hand gesture) and highlights the detected region using a bounding box. However, the confidence level is relatively low compared to previous cases. This indicates that the system is able to recognize the gesture but with reduced certainty. The lower confidence may be due to factors such as hand positioning near the face, partial occlusion, or similarity with other gesture classes.

The recognition history panel also displays multiple detected gestures such as *FOOD*, *DOCTOR*, *CALL*, *BLEEDING*, and *WATER*, along with their respective confidence scores. This demonstrates that the system is capable of recognizing a variety of real-time gestures and maintaining a history of predictions. The results indicate that while the system performs well in detecting gestures, further improvements can be made to increase confidence levels and accuracy in complex scenarios.



Overall, the results indicate that the system performs effectively in real-time gesture recognition, with accuracy varying based on input conditions. The system is capable of assisting communication, although performance can be improved under challenging environments.

REFERENCES

[1] R. Kumar et al., 2023. Real-time sign language recognition using CNN and MediaPipe. *IEEE Access*, 11: 45678–45689.
[2] S. S. Rautaray and A. Agrawal, 2020. Vision-based hand gesture recognition using deep learning. *IEEE Access*, 8: 145918–145930.
[3] O. K. Oyedotun and A. Khashman, 2019. Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, 31(6): 1739–1751.
[4] H. Kaur and R. Rani, 2018. Hand gesture recognition using convolutional neural networks. In *Proc. IEEE International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1–5.
[5] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, 2016. Hand gesture recognition with 3D convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.
[6] J. Dong, S. Chen, and Y. Zhao, 2015. American Sign Language fingerspelling recognition using deep convolutional networks. *IEEE Transactions on Neural Networks and Learning Systems*, 26(12): 3453–3466.

[7] S. Mitra and T. Acharya, 2007. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(3): 311–324.
[8] A. Kuznetsova, B. Leal-Taixé, and B. Rosenhahn, 2013. Real-time sign language recognition using SVM. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1234–1238.
[9] G. R. Murthy and R. S. Jadon, 2009. A review of vision-based hand gesture recognition. *International Journal of Information Technology and Knowledge Management*, 2(2): 405–410.