Grenze International Journal of Engineering and Technology, June Issue

HandSpeak: A Sign Language Learning Platform (ISL)

Swati Atul Joshi[1], Pranali D. Yelavikar[2], Sanchita N. Mote[3], Shreya S. Kodgire[4] and Sankalp A. Gaikwad[5]
[1-5]Department of Artificial Intelligence and Data Science Engineering, Vishwakarma Institute of technology, Pune, India
Email: swati.joshi@vit.edu, pranali.yelavikar25@vit.edu, sanchita.mote25vit.edu, shreya.kodgire25@vit.edu, Sankalp.1252010019@vit.edu

Abstract— **Indian Sign Language (ISL) is crucial in providing an effective communication mode for those who have hearing and speech difficulties. In most cases, it becomes hard to learn ISL due to non-availability of interactive means of practice. As a result, we have designed HandSpeak: A Sign Language Learning Platform, which involves the implementation of a vision-based sign language module along with a gesture recognition tool to help users learn effectively. This system makes use of techniques such as computer vision, MediaPipe for hand landmarks, and a Convolutional Neural Network for sign gesture classification. Users can practice different sign gestures using their cameras and receive real-time feedback in the form of text and speech output. Moreover, it supports quiz-based learning as a way of improving the process. In the experimentation stage, our MediaPipe-CNN method performs well compared with existing methods such as video-based learning and CNN-based learning.**

Keywords— **Hearing-impaired communication, Sign Language Recognition, Computer Vision, Assistive Technology, Gesture Recognition, Human-Computer Interaction.**

I. INTRODUCTION

Sign language serves as one of the most important modes of communication among people with hearing and speech disabilities. With sign language, individuals are able to convey their thoughts, feelings, ideas, and information through hand gestures and facial expressions. Sign language is not just a collection of hand signs but an actual language that possesses its unique grammar, structure, and spatial components [1]. Hence, learning sign language involves proper knowledge of hand shape, finger position, movement, facial expression, and body positioning. Indian Sign Language (ISL) serves as one of the most popular forms of sign language used by Indians who suffer from hearing and speech disability. Unfortunately, there exists a scarcity of ISL learning tools in India. Books, images, classroom teaching, and videos are some traditional modes of learning sign language, but they fail to validate the correctness of gestures made by the learner in real-time. Computer vision and deep learning have recently contributed to the creation of gesture recognition algorithms. Models based on CNNs work well in extracting spatial information from gestures in images and recognizing static hand signs [2]. Graph-based, LSTM-based, and landmark-based approaches are among the latest developments to recognize continuous dynamic gestures [3]. All these examples demonstrate that deep learning contributes to automatic gesture recognition effectively and efficiently.
It was also found that additional research is needed to develop standardized datasets and machine learning algorithms to facilitate Indian Sign Language recognition. iSign developed an extensive database for recognition tasks using Indian Sign Language-English language translation [4]. Other examples of research related to ISL use MediaPipe, CNNs, LSTMs, and graphs [5],[8].

Grenze ID: 01.GIJET.12.2.290_26

The use of MediaPipe proves especially convenient because the system extracts data on finger joints, palm location, and hand orientation from videos [10], [11].
Whereas most of the available systems tend to concentrate either on sign language recognition or translation, there are very few learning platforms that incorporate both learning materials and real-time gesture validation. While existing learning systems offer either video tutorials or quizzes, none can validate a learner's hand gestures. Hence, this paper seeks to fill this gap by proposing HandSpeak: a sign language learning platform (ISL). The system incorporates learning materials, gesture recognition from live camera feeds, quiz-based learning, and immediate text/audio feedback. HandSpeak utilizes the MediaPipe framework for detecting hand landmarks and employs a Convolutional Neural Network (CNN) algorithm for gesture classification.

II. Motivation

Learning Sign Language is still challenging due to the lack of trained teachers' availability and there are not many learning platforms and resources. Research noticed that the majority of hearing people are not familiar with sign language, despite the fact is that the hearing- impaired community uses it extensively. This creates a problem in communication at area like public places, educational institutions and workplaces.
The lack of real-time validation and instant feedback one of the major drawbacks of current Indian Sign Language (ISL) learning platform. It might be difficult for learners to identify whether their hand movement corresponds with correct sign. This drawback leads to less motivation and uneven learning outcomes.
Furthermore, a lot of current sign language recognition system ignore the Indian Sign Language (ISL), they mostly focus on American Sign Language (ASL). To overcome this, there is need to develop such a kind of learning platform which provides the best way of learning Indian Sign Language (ISL) to everyone.
The main goal of HandSpeak is to create a low-cost, vision-based system that combines recognition and learning into the one platform effectively. HandSpeak focuses on improving the accessibility, effectiveness and engagement of sign language instruction for both hearing-impaired people and the general people by providing the opportunity to learn, practice and get immediate feedback.

III. Problem Domain

Indian Sign Language (ISL) is a significant communication tool used by hearing and speech-impaired individuals. Nonetheless, learning ISL poses a significant challenge to beginners since there is a lack of sufficient interactive learning tools, teachers, and practice assistance. Most of the traditional techniques of learning involve the use of books, videos, and in-classroom training methods. In this regard, learners are only provided with static information through these mediums. These techniques cannot check if the learner performs the gesture correctly.
The problem domain of this study can be identified under assistive technologies, human-computer interaction, computer vision, and deep-learning techniques based on gesture recognition. The system to be developed will enable learners to interactively learn ISL by incorporating cameras for gesture validation and checking if the user performs the gesture appropriately.

IV. Problem Definition

Most existing learning websites for sign language offer video, images, or text guides only. These help learners learn some basic gestures but lack the functionality to validate learners' hand gestures in real-time. In effect, learners keep on practicing wrong gestures without realizing they have been doing it all along. The effectiveness of their learning process thus remains in question.
It should be noted here that the majority of existing systems use American Sign Language or other types of gestures as opposed to Indian Sign Language. Hence, there is a requirement for a cost-effective gesture recognition system using cameras.

V. Problem Statement

Indian Sign Language Learning is a hard task for novices since the majority of current learning tools are mostly passive and cannot provide instantaneous feedback. Additionally, professional lessons are quite expensive and difficult to come by. The proposed research project will focus on providing an inexpensive and interactive learning system that can utilize computer vision technology to enable learners to practice sign language.

VI. INNOVATION IN APPROACH / RESEARCH CONTRIBUTIONS

The innovative aspect in HandSpeak is the inclusion of structured modules for Indian Sign Language with hand gesture recognition with instantaneous feedback. In contrast to conventional learning platforms which provide video lessons, images, and written information about gestures, the HandSpeak platform gives users the ability to practice making gestures by using a camera with feedback on whether the user's gesture is correct or not.
This solution consists of the MediaPipe hand landmark detection approach, CNN-based gesture classification, learning via quizzes, and feedback through text and audio. MediaPipe is helpful in providing real-time pipelines with landmark detection, while CNN-based models are useful for spatial feature learning of gestures [5],[6]. Thus, the proposed system provides an interactive, beginner-friendly solution to learning ISL.
Major Contributions:

- An Indian Sign Language Learning Platform is designed using a camera to minimize reliance on expensive hardware devices like sensors.
- Gesture recognition is synchronized with learning material in an organized manner.
- The inclusion of quizzes makes learning interesting.
- Feedback in text form and audio form is offered for enhanced learning experience.
- The system matches the gestures made by users against a trained gesture class.

VII. RELATED WORK

Sign language recognition has become one of the most vital areas of research in assistive technologies, computer vision, and human-computer interaction. Sign languages are full-fledged natural languages with grammatical structure and visual-spatial features; thus, sign language recognition requires identifying hand shape, hand movement, orientation, facial expressions, and body signals correctly [1]. Sensor-based devices including data gloves, accelerometers, and motion sensors have been used for the recognition process previously. While these approaches offered accuracy in tracking, they involved expensive external hardware.
Vision-based and deep learning-based techniques have emerged in recent research due to affordability and ease of use in camera-based techniques. Convolutional neural network models have been used extensively in static gesture recognition, whereas long short-term memory, gated recurrent unit, and graph-based models have been beneficial for dynamic gesture recognition [2], [3]. However, several challenges remain to be addressed in these techniques, namely inadequate lighting, complicated background environment, hand occlusion, and different hand orientation.
The datasets and research materials for Indian Sign Language are relatively less than other languages. One of the largest ISL-English datasets and the requirement for more research materials for ISL processing were identified using the iSign benchmark [4]. Some recent ISL recognition studies have applied MediaPipe, CNN, LSTM, and graph neural network models to improve real-time recognition accuracy [5],[9]. MediaPipe can be used due to its real-time feature of extracting landmarks and building pipelines in video applications [10], [11].

TABLE I. COMPARISON WITH OTHER SYSTEMS

Current System / Methodology	Technologies Used	Disadvantages	Enhancements to HandSpeak
Sensor-driven sign language technologies	Hand gloves, accelerometers, motion sensors	Hardware is expensive	Uses camera-based input
Conventional teaching methodologies	Textbooks, videos, imagery	Lack of real-time feedback	Instant verification of gestures
Mobile apps for sign languages	Mobile application, videos, quizzes	The user needs to validate themselves	Self-validation using cameras
Vision recognition technology techniques	Camera + Machine Learning/Deep Learning	Focuses more on recognition	Connects recognition with learning
CNN for gesture recognition	Deep learning	Does not necessarily involve teaching modules	Incorporates structured learning and test modules
Proposed HandSpeak system	MediaPipe, OpenCV, CNN, audio/text feedback	Present implementation is restricted to basic gestures	Offers ISL learning with live feedback

From the available literature, most of the existing systems concentrate either on gesture recognition, translation, or dataset creation. There is very little research done on combining structured Indian Sign Language learning, real-time gesture verification, quiz games, and instant audio/text feedback mechanisms. Hence, the proposed HandSpeak application intends to fill this gap by offering an interactive ISL learning environment with real-time gesture recognition and instant feedback support.

Deep learning techniques used to recognize and translate sign languages have been reviewed extensively in recent times [12],[15]. There are also studies involving hybrid models as well as approaches for real-time translation of static and dynamic sign languages [16],[20].

VIII. Proposed Solution and Methodology

The proposed HandSpeak: Sign Language Learning Platform aims to overcome the shortcomings of earlier methods through the fusion of vision-based gesture recognition with structured sign language learning modules. As opposed to sensor-based methods, HandSpeak does not rely on any expensive external devices like data gloves or motion sensors; only visual camera inputs are required, making the system cost-effective, accessible, and applicable to regular users. Moreover, the proposed system provides real-time feedback enabling users to make adjustments to their gestures.

The vision-based deep learning method employed in the system involves interactive sign language learning and gesture recognition. The entire process comprises several stages, namely data gathering, data preprocessing, model building, gesture recognition, integration of learning modules, provision of feedback, and finally, system implementation using open-source technologies.

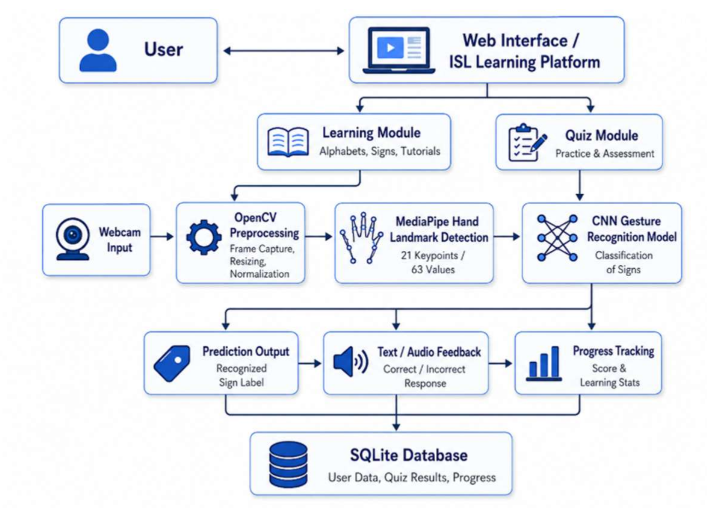

Fig. 1. System architecture diagram of HandSpeak Sign Language Platform

For building the machine learning model, a custom dataset and publicly available Kaggle dataset have been used. There are approximately 1300 images for each alphabet gesture. The images are available in PNG format and JPG format and are labelled with respect to their gesture class. The dataset consists of hand gestures that are taken under various lighting conditions and under various hand positions with respect to background images. The variety helps the model learn different gestures.

The images are pre-processed before feeding them into the training algorithm. All the images are resized to certain dimensions to have uniformity. Data preprocessing steps like normalizing and noise removal help in improving the quality of dataset. The simplest data augmentation methods like adjusting brightness, rotating, and horizontal flipping have been used.

A. AI Model and Gesture Recognition System

HandSpeak adopts a gesture recognition system based on computer vision implemented using MediaPipe, OpenCV, and deep learning with convolutional neural networks (CNN). In this case, OpenCV will be applied to obtain live video frames while MediaPipe will be used for detecting and identifying landmarks from these hand frames. The system does not need any hardware equipment to identify key hand points such as joint areas, the location of the palm, and hand orientation.

The landmarks are then fed into the CNN for classifying gestures. CNN is a good fit for the problem at hand because it can effectively detect hand shapes from inputted images by learning spatial characteristics of the data. The CNN model will be developed using TensorFlow and Keras with an Adam optimizer and categorical cross-entropy loss function.

B. Mathematical Formulation of CNN for Gesture Recognition

The architecture of CNN comprises several stages, including convolution, pooling, and classification layers. The convolutional layer plays an essential role in extracting significant features from the input image.

$$X^{(l+1)} = f\left(X^{(l)} * W^{(l)} + b^{(l)}\right) \quad (1)$$

Where:

- $X^{(l)}$= input feature map at layer l
- $W^{(l)}$= convolution kernel weights
- $b^{(l)}$= bias term
- $f(\cdot)$= activation function (ReLU used in this model)
- $*$= convolution operation

Pooling layer reduces spatial dimensions and helps in avoiding overfitting.

$$X_{pooled} = \max\left(X_{window}^{(l+1)}\right) \quad (2)$$

Fully connected layer transforms the learned features to probabilities of each class.

$$\hat{y} = softmax\left(W_{fc}x + b_{fc}\right) \quad (3)$$

Loss function used during training is categorical cross-entropy.

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y_i}) \quad (4)$$

Where,

- C = number of gesture classes
- y_i = true label in one-hot encoded format
- $\hat{y}_i$ = probability of class i.

C. Learning Module Design

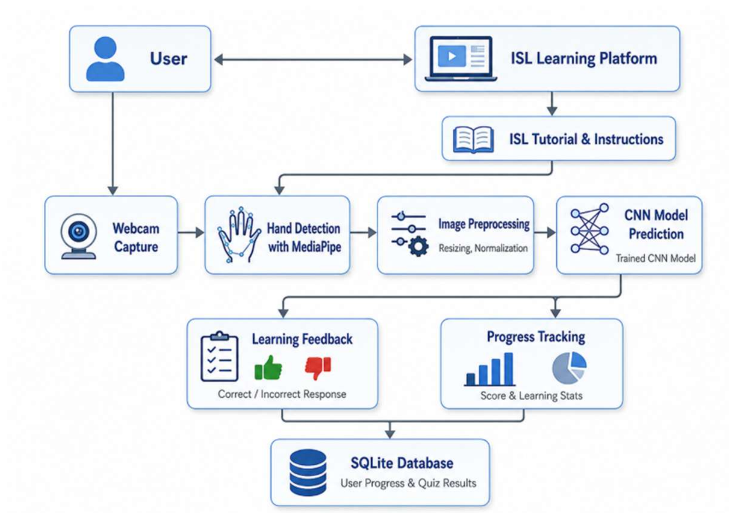

Fig. 2. Learning workflow of HandSpeak sign language Platform

The learning module aims at ensuring a controlled learning environment where the Indian Sign Language is taught. Visual images and instructions are provided on how to form different signs in terms of hand orientation, finger positioning, and movement. The learner can know how to perform the gestures correctly before trying to sign something.

Another important feature offered by HandSpeak is the quiz module. In this case, learners get an opportunity to take quizzes and find out about the performance immediately after. If the learner makes some movements in front of the camera, the system will capture such gesture and compare it with the trained gesture model. Based on the classification done, the learner gets notified whether he/she did something correctly or not.

D. Real-Time Feedback and Interactive Functions

Real-time feedback is considered one of the crucial aspects of the HandSpeak platform. Once the gesture is successfully recognized, its equivalent sign is translated into text and speech. If the gesture was performed correctly, then the user will receive a response like "Correct." On the other hand, if the gesture was not performed correctly, then a message such as "Wrong" is displayed on the screen.

The text-based feedback allows the user to verify the recognition accuracy visually, whereas speech feedback can establish connections between gestures and words. The text-to-speech engine can be easily built using libraries like pyttsx3 and Google Text-to-Speech.

E. Tools and Technology Stack

HandSpeak is constructed using free and open-source software to provide scalability and implementation simplicity. Python is utilized as the primary programming language due to its extensive support for computer vision and machine learning. The CNN architecture and model are created and trained by TensorFlow and Keras. OpenCV is utilized for capturing live video feeds and performing image processing operations. MediaPipe is employed for hand detection and extracting landmarks. Flask serves as the bridge between the machine learning model and the website application. HTML, CSS, and JavaScript are used for the front-end development process.

In terms of the speech feedback mechanism, pyttsx3 or Google Text-to-Speech can be used to produce the audio output corresponding to the detected sign.

F. Overall Workflow

HandSpeak's overall workflow starts from the camera, which takes in the live video frame. After processing the frame by using OpenCV, MediaPipe detects the landmarks of the hand. From there, the landmarks and gesture features will undergo CNN classification. After the process, the result will be shown in text form and then in audio output. If the gesture is wrong, the learner can do it again.

Overall workflow summary:

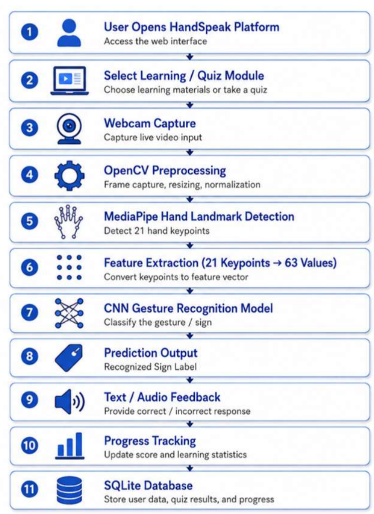

Fig. 3. Overall workflow of the proposed HandSpeak System.

HandSpeak is able to link sign language learning to gesture recognition by incorporating instruction, recognition, and feedback into one package. HandSpeak serves as an excellent interactive platform for Indian Sign Language learners. It allows the learner to learn through practice in real-time while eliminating the need for expensive hardware equipment.

IX. PROBLEM FORMULATION AND SYSTEM DESIGN

The HandSpeak system is considered a gesture recognition and learning task that operates in real-time. The inputs of the system are the video streams that are obtained from a camera device. This video stream is analysed per frame to detect the hand region and extract hand landmarks that are essential for recognizing the performed gesture. These landmarks include fingertips, palms, and joints.

There are four key components of the system design. First, the input video stream of a hand gesture is acquired using the camera. Second, the input frame is processed using the MediaPipe and OpenCV libraries to detect hand landmarks. Third, the CNN model that has been trained identifies the performed gesture and assigns it to one of the sign classes. Lastly, the identified sign is displayed on the screen as text and audio is provided as feedback to the learner.
Such a design enhances the process of learning because apart from observing the gestures, the learner also executes them while receiving prompt feedback. Consequently, the technology serves as an interactive learning system as well as a gesture recognition system.

X. Data Model

The data model of the HandSpeak system describes how the data flows from the initial data input stage to the stage of generating the final feedback. The input data comes in the form of either a picture of the hand gesture or live video frames collected through the camera. The live frames are analysed by OpenCV and MediaPipe to identify the hand area and then extract the key landmarks of the hand. The landmark or image features obtained are sent to the CNN classifier model for classification. Lastly, the identified sign is shown on-screen as text and also provided as an audio response.

TABLE II. DATA MODEL FOR HANDSPEAK SYSTEM

Input Data	Processing Procedure	Feature Extraction	Output
Camera input frame	OpenCV pre-processing	Location of hand	Pre-processed frame
Hand input image	Mediapipe	Digital location of finger joints and palm points	Location vector
Location/feature vector	CNN model	Spatial hand gestures	Predicted sign language symbol
Predicted sign	Conversion system	Sign language meaning	Converted text and audio
Answer from quiz	Validation system	Correctness of answer	Scoring result

XI. Results and Discussion

In order to determine the efficacy of the proposed HandSpeak system, several gestures of the Indian Sign Language were analysed based on camera input data. The criteria for assessment included real-time efficiency, gesture recognition accuracy, precision, recall, and learning efficiency. The testing was carried out primarily for basic sign language alphabets and frequently used gestures that could be employed by beginner learners.
HandSpeak was shown to provide adequate performance of gesture recognition in terms of controlled light settings and plain backgrounds. The integration of MediaPipe-based hand landmark detection and CNN-based classification allowed for the extraction of critical features of hands, including finger orientation, palm structure, and joint locations. As a result, the system successfully identified gestures performed by the user with minimal latency and gave instant feedback.
It is crucial for feedback to be given in real-time because of the benefits that it provides towards the learning process. The difference between video-based learning techniques and handspeak is the fact that unlike video-based learning technique, HandSpeak will actively monitor the gesture made by the user and notify whether the gesture was correct or wrong.
After recognizing a gesture, it would be translated into text output as well as audio output for better understanding. Text output allows the learner to visually recognize the gesture whereas audio output would allow the user to connect with spoken languages.
There were also certain limitations noted during the testing phase. The presence of complex background, dim lighting, partial obstruction by hands, variations in hand size, and varied speeds of gestures influenced the performance accuracy to some extent. Limitations noted here are quite typical for vision-based gesture recognition techniques. Yet under controlled conditions, the proposed system worked well.

TABLE III. PRECISION AND RECALL CALCULATION

Letter	Precision (%)	Recall (%)
A	85.7	90.0
B	84.2	80.0
C	85.0	85.0
Average	84.97	85.0

The bar chart depicted in Fig. 4 below gives an indication of the precision and recall rates attained by the alphabet letter gestures, namely A, B, and C. From the results presented in the chart below, gesture A obtained precision accuracy of 85.7% and recall accuracy rate of 90.0%; gesture B obtained precision accuracy of 84.2% and recall accuracy of 80.0%, while gesture C had both recall and precision rates at 85.0%.

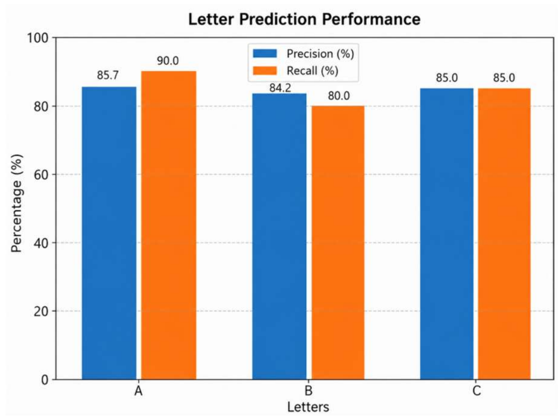

Fig. 4. Letter prediction performance of selected gestures

TABLE IV. COMPONENT TEST OUTCOMES

Component	Input	Expected Output	Result
Camera	Hand gestures	Video feed	Pass
MediaPipe	Hand image	Hand landmarks	Pass
ML model	Landmarks	Sign label	Pass
Flask API	Label	JSON/text output	Pass
Frontend	API data	Display change	Pass
Database	score	Stored data	Pass

TABLE V. CONFUSION MATRIX OF SELECTED GESTURES

Actual / Predicted	A	B	C
A	54	3	3
B	6	48	6
C	3	6	51

The confusion matrix indicates the performance of the gesture recognition model with respect to the selected alphabet gestures A, B, and C. The rows indicate the gesture classes of the test set, whereas the columns indicate the prediction gesture classes made by the model.

For instance, letter A is made up of 54 correctly classified samples and 6 misclassified samples as other data. Similarly, letters B and C also contain correct data, but some misclassification occurs because of the similarities in the shapes of hands and fingers positioning.

Precision and recall are used to evaluate the performance of the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+F}} \tag{6}$$

Where, TP = True Positive

FP = False Positive

FN = False Negative

For letter A:

$$\text{Precision} = \frac{54}{54+9} = 0.857 = 85.7\%$$

$$\text{Recall} = \frac{54}{54+6} = 0.900 = 90.0\%$$

Similarly, precision and recall values have been computed for letters B and C.

A. Consistency of Result Explanation

In the case of the MediaPipe-CNN model proposal, the results yielded a level of 94.23% as the overall accuracy rate for the gesture recognition dataset as a whole during controlled experimental tests. As for the class-level performance evaluation, the confusion matrix was generated for a selection of the alphabet gestures A, B, and C. The average precision and recall obtained from the generated confusion matrix were 84.97% and 85.0%, respectively.

Thus, the value 94.23% can be considered as the accuracy rate of the suggested model, whereas the precision and recall values indicate the class-wise performance of the suggested hand gesture classification model.

TABLE VI. COMPARISON PERFORMANCE OF MODELS IN TERMS OF ACCURACY

Model/Method	Performance (in %)	Comments
Conventional Learning (video-based)	Not applicable	No real-time gesture validation
CNN (without MediaPipe)	82.5	Not real-time, slower recognition process
MediaPipe-CNN (proposed Handspeak)	94.23	Real-time hand gesture recognition process

The results indicate that the proposed model performs better than conventional learning and CNN. Traditional learning does not have any real-time gesture validation capability, while CNN may perform poorly in case of noisy background and unreliable hand region detection. In the proposed method, MediaPipe detects key landmarks of hands before being classified using the CNN model.

The system was also evaluated with varied scenarios including controlled lighting, low lighting, and complex background. In the scenario of controlled lighting, high accuracy was achieved. Nonetheless, a minor degree of misclassification occurred in cases of complex background and low lighting. This demonstrates that the proposed system performs excellently in real-time learning. Further improvements can be achieved through increasing the number of samples in the dataset.

In summary, from the analysis of the performance of the system, it is evident that HandSpeak effectively integrates gesture recognition and sign language learning. Through the integration of real-time validation, multimodal feedback, and interactive learning, the proposed system could be applied as a powerful learning instrument among beginners, hearing people, and speech-hearing impaired learners intending to learn Indian Sign Language.

XII. JUSTIFICATION OF RESULTS

The results from the experiment indicate that the proposed HandSpeak can detect basic sign language gestures accurately. In using the MediaPipe model, hand landmark extraction is easier because it identifies crucial finger and palm key points. The extracted landmarks will enable the CNN model to classify the gestures accurately.

In examining the confusion matrix, one can deduce that there is misclassification of some of the gestures because some have similarities in hand shape. The similarity occurs since some alphabets have closely similar hand shapes. The model works efficiently for the chosen gestures, considering the value of recall and precision. However, the model detects better in conditions where the environment has adequate lighting and plain background.

When compared to the conventional method of video learning, HandSpeak offers an interactive mode of learning since the user's gesture is confirmed instantaneously. When compared to other methods based on sensors, it is cheaper since there is no need for any extra equipment to be attached to it. The findings confirm that HandSpeak is a good educational device for teaching Indian Sign Language.

XIII. COMPARISON

The suggested HandSpeak was analysed in comparison with the other existing methods for learning and recognizing sign languages. The traditional video-based learning methods fail to offer accuracy measurement since they just deliver the learning material. The sensor-based methods will help recognize gestures with the use of special hardware, namely gloves or motion sensors.

The CNN-based gesture recognition without MediaPipe will help classify gestures, but it may lack stability since hand landmark recognition will not work properly.

The suggested MediaPipe and CNN combination is faster since the former recognizes hand landmarks first before sending the data for the latter. It filters out unnecessary information from the background thus allowing the CNN to concentrate on hand features.

TABLE VII. COMPARISON OF HANDSPEAK WITH THE OTHER METHODS

Methods	Feedback	Hardware Needs	Accuracy /Performance	Limitations
Traditional video-based learning	No real-time feedback	None	not applicable	Learner cannot validate hand gesture
Sensor-based approach	Real-time is feasible	Necessitates use of gloves/sensors	Effective when used in controlled environment	Expensive and inconvenient to use
CNN without using MediaPipe	Limited feedback	None	82.5%	Low consistency in recognition
HandSpeak	Feedback in text and audio format	Camera	94.23%	Lighting and background affect recognition

The above comparison clearly reveals that HandSpeak offers an effective solution.

XIV. CONCLUSION

The HandSpeak system effectively implements a vision-based and interactive approach to learning and recognizing the Indian Sign Language (ISL). The designed approach increases the engagement and effectiveness of learning ISL by integrating the concept of structured learning modules with that of real-time gesture recognition. Learners can perform gestures through their webcam and get instant feedback on their performance. This feature increases the effectiveness of this approach in comparison with other approaches, including video-based and book-based learning methods.

MediaPipe, OpenCV, and a CNN deep learning algorithm have been used to detect and recognize the hand gestures in real-time. The use of expensive hardware like sensor gloves and wearable gadgets has not been required because of implementing the camera-based approach in the system. The identified gestures are then translated into text and audio outputs.

LIMITATIONS

While the suggested HandSpeak system offers real-time gesture recognition and interactive learning features, there are certain limitations to consider. The existing version primarily concentrates on basic alphabets and some hand signs. Moreover, the system’s effectiveness may be hampered by poor lighting, complicated backgrounds, and partial occlusion of the hand.

Misclassifications may occur due to similar shapes of hands among various alphabet gestures. Furthermore, the model relies on hand gestures captured using a camera, meaning that appropriate hand positioning is crucial for correct predictions. Future studies could enhance the database by incorporating multiple users, varied lighting conditions, dynamic gestures, facial expressions, and full-sentence signing.

The results show that the proposed MediaPipe-CNN approach achieves good recognition performance under controlled conditions. The system can be used by beginners, students, hearing individuals, and hearing and speech-impaired learners who are interested in learning Indian Sign Language. Overall, HandSpeak demonstrates how computer vision and artificial intelligence can support accessible and inclusive assistive learning systems.

FUTURE WORK

In HandSpeak: Sign Language Learning Platform following future improvements can be included:

- There can be a sentence-based gesture detection
- It can be converted into mobile application and deployed it.
- There can be additional module for expression recognition.
- To increase the accuracy of model there can be use of another deep learning models.

REFERENCES

[1] K. Emmorey, “Ten Things You Should Know About Sign Languages,” Current Directions in Psychological Science, vol. 32, no. 5, pp. 387–394, 2023.

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[3] S. Patra, A. Maitra, M. Tiwari, K. Kumaran, S. Prabhu, S. Punyeshwarananda, and S. Samanta, “Hierarchical Windowed Graph Attention Network and a Large Scale Dataset for Isolated Indian Sign Language Recognition,” arXiv preprint arXiv:2407.14224, 2024.

[4] A. Joshi, R. Mohanty, M. Kanakanti, A. Mangla, S. Choudhary, M. Barbate, and A. Modi, "iSign: A Benchmark for Indian Sign Language Processing," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 10827–10844, 2024.
[5] K. Goyal and V. G., "Indian Sign Language Recognition Using MediaPipe Holistic," arXiv preprint arXiv:2304.10256, 2023.
[6] G. Khartheesvar, M. Kumar, A. K. Yadav, and D. Yadav, "Automatic Indian Sign Language Recognition using MediaPipe Holistic and LSTM Network," Multimedia Tools and Applications, vol. 83, no. 20, pp. 58329–58348, 2024.
[7] S. Srivastava, S. Singh, Pooja, and S. Prakash, "Continuous Sign Language Recognition System Using Deep Learning with MediaPipe Holistic," Wireless Personal Communications, vol. 137, no. 3, pp. 1455–1468, 2024.
[8] A. S. Nayak, N. Subudhi, T. Rana, M. Sahu, and R. C. Balabantaray, "Indian Sign Language Recognition and Translation into Odia," in Proceedings of the Workshop on Sign Language Processing, Association for Computational Linguistics, pp. 10–17, 2025.
[9] R. Damdoo, A. S. Varde, and S. K. Ghosh, "An Integrative Survey on Indian Sign Language Recognition and Translation," IET Image Processing, 2025.
[10] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.
[11] Google AI Edge, "MediaPipe Solutions Guide," Google Developers, 2025.
[12] S. Tan et al., "A Review of Deep Learning-Based Approaches to Sign Language Recognition, Translation, and Generation," Advanced Robotics, 2024.
[13] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, and M. Sabokrou, "A Survey on Recent Advances in Sign Language Production," Expert Systems with Applications, vol. 243, p. 122846, 2024.
[14] I. M. Violet and R. Leena Sri, "A Comprehensive Survey on Recent Advances and Challenges in Sign Language Recognition Systems," Discover Artificial Intelligence, 2025.
[15] S. Ingoley et al., "Efficient Models Based on Deep Learning Technique for Sign Language Recognition," Procedia Computer Science, 2025.
[16] P. Shah et al., "Indian Sign Language Communication Using LSTM, MediaPipe Holistic and Large Language Models," in Atlantis Highlights in Intelligent Systems, 2025.
[17] S. Sharma et al., "Recognition of Indian Sign Language Using Deep Learning," Wireless Personal Communications, 2022.
[18] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, A. Akbar, and B. F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," Mathematics, vol. 11, no. 17, p. 3729, 2023.
[19] B. Alsharif, A. S. Altaher, A. Altaher, M. Ilyas, and E. Alalwany, "Deep Learning Technology to Recognize American Sign Language Alphabet," Sensors, vol. 23, no. 18, p. 7970, 2023.
[20] S. Singhal, J. Gupta, A. Sharma, A. Gupta, and N. Sharma, "Indian Sign Language Detection for Real-Time Translation using Machine Learning," arXiv preprint arXiv:2507.20414, 2025.