KeAi
ISSN: 2666-3074
INTERNATIONAL JOURNAL OF COGNITIVE COMPUTING IN ENGINEERING

# INDIAN SIGN LANGUAGE INTERPRETATION USING CNN AND MEDIAPIPE WITH TEXT-TO-SPEECH INTEGRATION

**[1]Kaveri A. Chandan, [2]Dr.Vijay B. More**
[1]PG Student, Department of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India
[2]Professor, Department of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India
[1]kaverichandan0503@gmail.com, [2]vbmore2005@rediffmail.com

---

**Abstract:** Indian Sign Language (ISL) is of great importance in communication for deaf and speech-impaired people from all around the nation of India. However, the lack of universally available interpretation tools has resulted in an obstacle to communication between ISL users and the general population. This research presents a real-time ISL interpretation system that consists of CNN and MediaPipe hand tracking to process gestures and convert them to natural language text. Finally, Text-to-Speech (TTS) technology is integrated to allow for the output of the gestured information to hearing individuals for supportive interaction. The proposed model takes in live video input, extracts hand landmarks based on MediaPipe output, and passes these landmarks into a trained CNN to make gesture classification. Analytically, with high accuracy on a broad, machine-readable ISL dataset of commonly used signs, the system achieves high accuracy. After classification, it translates the output into readable text and converts it into speech to form a high-fidelity communication bridge between hearing and non-hearing people. The way it integrates gesture detection, deep learning-based recognition, and speech synthesis, is to enable real-time intuitive and efficient two-way communication. As a lightweight, scalable system suitable for deployment on mobile and wearable devices, it largely conforms to what is found in daily use as an educational, social, and professional tool. The system shows great promise in improving social inclusion, accessibility, and independence of hearing and speech-disabled individuals in India.

**Keywords:** Indian Sign Language (ISL), Convolutional Neural Network (CNN), MediaPipe, Text-to-Speech (TTS), Gesture Recognition, Real-Time Communication, Assistive Technology.

## I. INTRODUCTION

Language is a basic human tool for sharing experiences, speaking, and exchanging ideas and feelings. But for millions of people with hearing and speech impairments, verbal communication isn't an option. Sign language, particularly Indian Sign Language (ISL), serves as a critical medium for expression within this community. Despite its rich structure and expressiveness, ISL is still difficult and poorly understood by most hearing populations because it is not widely promoted and is not taught as a formal course. As a result, ISL users are socially excluded, lack employment opportunities, and experience educational disadvantages. Deep

learning and computer vision technologies have made such rapid progress to now enable us to form intelligent systems that can bridge this gap through real-time gesture recognition and natural language generation.

Automatic sign language recognition (SLR) is in high demand, as several years ago there was a new need for inclusive technologies that translate sign language to assist the hearing impaired in conveying their messages to the community around them. Current approaches to sign language recognition are based on hardware devices, e.g. data gloves and motion sensors. While these methods were moderately efficacious, they were very intrusive, expensive, and not practical for daily use. The development of vision-based techniques has resulted in noninvasive scalable and video-based alternatives, which can discern hand gestures or body movements using video inputting, machine learning, and computer vision.

State of the art in image classification and object detection tasks have been revolutionized by recent advances in CNNs, allowing models to learn complex visual patterns. CNN has been proven to extract important spatial features from hand gestures and can be used to classify with high accuracy for sign language recognition. Although accurate hand localization is still a challenge especially in real-world environments with varying lighting, objects occluding the hands, and various hand orientations, it can be achieved even in these settings. However, these challenges can be overcome with frameworks like MediaPipe developed by Google that provide robust and lightweight real-time hand landmark detection solutions based on RGB video input. 21 hand key points per frame can be tracked with MediaPipe, and the data offers a degree of richness for gesture analysis without the need for special hardware.

The contributions of this research are in the form of a real-time Indian Sign Language interpretation system created using a unified framework that combines MediaPipe hand tracking, CNN-based gesture classification, and Text-to-speech (TTS) synthesis. It will capture the video live, detect and track the hand movements with a media pipe algorithm, and finally will classify the ISL gesture using a trained CNN model. First, the recognized gesture is converted into readable text and then through a TTS engine to create spoken output. This complete pipeline makes two-way communication possible, so while nonsigners are talking neither is an interpreter required to understand ISL gestures and aid in promoting inclusive communication.

The portability and scalability are one of the distinguishing features of this system. This framework is lightweight and can be applied on devices such as mobile and wearable. Having high applicability to the real world such as schools, hospitals, workplaces, and public service centers, it makes the system. It provides immediate interpretation, empowering the users to participate in the conversation in real-time without latencies.

The paper provides an important contribution to the ongoing efforts of gesture-aware communication systems by leveraging deep learning with real-time hand tracking and natural language generation. CNNs and RNNs have been explored before as classifiers of sign language, but most are extremely computationally complex, or cannot operate in real-time. Access is restricted for the broader population due to several others that do not provide output in spoken language. Using TTS enables this research to go beyond simply improving communication, and creates additional opportunities for multi-modal accessibility tools for education, emergency response, and assistive learning.

For training and evaluation, the proposed system was trained and evaluated on a curated ISL dataset consisting of a variety of static and dynamic gestures, which include common words and phrases. The model perceives variations in gesture speed, orientation, and signer differences with high accuracy and robustness. Additionally, using MediaPipe to reduce detections significantly shrinks the occurrence of false detections since we leverage anatomical consistency in hand landmark positioning to ensure that the CNN model retains performance despite the changes in environments.

**Contributions to the Research Work**

The main contributions of this work are as follows:

1. A real-time ISL interpretation system was developed by integrating CNN-based gesture recognition with MediaPipe hand tracking and Text-to-Speech (TTS) conversion for seamless sign-to-speech communication.
2. MediaPipe's hand landmark extraction was utilized to improve gesture localization accuracy without the need for external sensors, enhancing system efficiency and scalability.
3. A custom-trained CNN model was implemented on a curated ISL dataset, achieving high recognition accuracy across varied gesture styles, lighting conditions, and signer differences.
4. The system was designed for portability and real-world application, enabling deployment on mobile and wearable devices to support inclusive communication in educational, professional, and public settings.

The research paper is broken into four main parts so you can see a logical progression. Literature Survey (Section 2): This section explores current approaches to sign language recognition: the inadequacies of traditional CNN RNN models, increasing usage of Transformer architectures, and the lack of real-time scalable solutions designed for sign recognition of Indian Sign Language (ISL). Section 3: The proposed Work describes a real-time ISL interpretation system based on MediaPipe for hand landmark detection, CNN for gesture classification, and Text to Speech (TTS) synthesis for the audio output, and demonstrates how it is portable and can be used in the real world. Results and Discussion in Section 4 describe the experimental setup, compare the system's performance with existing models, and visualize key metrics, namely accuracy, precision, recall, and F1-score, for the proposed system's effectiveness and efficiency. Section 5: Conclusion summarizes the contributions, impact, and future scope of the research by discussing the importance of intelligent, inclusive deep learning technologies to make the deaf and mute community more accessible to communication.

## II. LITERATURE SURVEY

Indian Sign Language (ISL) is a vital means of communication for the deaf and hard-of-hearing community in India. However, the communication gap between ISL users and non-users remains a significant challenge. Recent advancements in deep learning and computer vision have enabled the development of systems that can interpret ISL gestures and convert them into text or speech, thereby bridging this gap. This paper explores the integration of Convolutional Neural Networks (CNNs) and MediaPipe for ISL interpretation, along with text-to-speech integration, to enhance communication accessibility.

Prudhvi B et al. [1] In the reviewed paper, a real-time Indian Sign Language (ISL) detection system is presented leveraging deep convolutional neural networks (CNNs), MediaPipe, and OpenCV for gesture and classification depending on the type of gestures. It has hand detection, gesture segmentation, a trained CNN model, and a user-friendly interface. It combines several access tools such as text-to-speech, image-to-speech, text-to-image, and webcam-based gesture-to-text conversion. As a whole, this framework enhances communication between deaf and hearing individuals and expresses the deep learning and computer vision potential for developing inclusive, interactive assistive technologies that are.

Dewangan, S et al. [2] ISL interpretation system based on CNN, FRCNN, YOLO, and MediaPipe is presented in the research. On the other hand, CNN delivers fast gesture identification that works in real time, but with very poor accuracy. High recognition rates can be achieved by FRCNN but at the expense of being slow because of the limited resources on the GPU. YOLO can perform well on pre-recorded data but lacks profound real-time responsiveness. Taking all things into consideration, MediaPipe is by far the most effective, supporting both high accuracy at real-time processing. With the integration of MobileNet, TensorFlow, and OpenCV, the system is greatly improved in terms of its performance in gesture-to-text and speech translation. Nevertheless, a gap in the scalability of the model across various, yet diverse environments and sign languages as evidenced by the study calls for additional literature on generalization, and efficiency optimization.

Deshpande, S et al. [3], This reviewed study applies CNN in combination with MediaPipe to create a hybrid approach for Indian Sign Language (ISL) interpretation with an error rate of approximately 94%. Real-time gesture and TTS-based spoken output of multiple Indian regional languages including Hindi, Kannada, Telugu, and Marathi get enhanced using this method. The system successfully circumvents limitations of prior approaches in pose extraction as well as text generation from ISL gestures. Nevertheless, the study finds those challenges in gesture diversity and dialectal variations, thus revealing a research gap in designing scalable and robust models that can cope with broader real-world contexts and regional linguistic variations.

Khaire, P. B. et al. [4], In this work, the SilentSync AI project has identified a robust Indian Sign Language (ISL) recognition system using combo CNN and MediaPipe based 21 key hand landmarks extraction for the gesture interpretation with the performance of 95.49%. By accepting some recognizable gestures as inputs, the system translates those gestures into Marathi audio using pygame-based Text-to-speech (TTS) technologies and accessibility via a live web-based interface using HTML, CSS, and Django. The framework relies on Google APIs and TensorFlow for understanding gestures and providing speech output and it will do so reliably. Though the study indicates a comprehensive and successful implementation, scaling across multiple languages and more difficult gestures is not explicitly part of the study. Coming next is supporting custom gestures and mobile deployment, which shows that the project is flexible and can be more generally applied in real-world scenarios.

Nimbalkar, S. V. et al. [5], introduce an application of Convolutional Neural Network (CNN) and MediaPipe for hand tracking in realizing a real-time ISL recognition system that demonstrates an accuracy of 97.1%. This addresses communication barriers faced by deaf children in India by translating ISL gestures into text. The approach uses high-level deep learning and computer vision technology including a CV zone hand tracking module, and

performs very well recognizing different dynamic hand gestures. In the context of real-time recognition, the system works well however it offers no text-to-speech functionality which presents itself as a potential area for future enhancement to further assist in meeting the desire for total two-way communication.

Despite ISL interpretation based on text-to-speech, which reduces the communication gap between deaf and hearing communities, further progress and the depth of application remain dependent on further harmonization between the two communities. The high accuracy and robustness of these systems to date have made them suitable for real-time applications. However, these challenges need to be overcome with further research before ICMMS can be applied for negotiation assistance. ISL interpretation systems' potential can be fulfilled by expanding datasets, optimizing models, and building mobile applications, increasing inclusiveness and accessibility for people with hearing impairments.

However, while there have been significant improvements in the use of deep learning techniques like CNN, MediaPipe, and TensorFlow to recognize ISL, most of these systems tend to primarily deal with a gesture-to-text conversion, and provide limited or no end-to-end communication capabilities. In isolated gesture recognition, many models achieve high accuracy (> 94% or > 96.1 %) but do not integrate real-time text-to-speech (TTS) or language audio output for regional languages, limiting the models' utility for practical communicating situations. Furthermore, there are also unexplored scaling and adaptability across various signers, dialects, and lighting conditions of the signers. In addition, continuous sign language recognition is rarely studied, in contrast to isolated words which limit the systems' use in natural conversations. However, these gaps indicate the prerequisite need for a unified and scalable ISL recognition framework that encompasses speech-to-gesture and multilingual speech output for broader accessibility and real-world usability.

## III. PROPOSED SYSTEM

A real-time, gesture-to-speech system that combines Convolutional Neural Networks (CNN) and MediaPipe for Indian Sign Language recognition is proposed as the system itself. First, live video of a user's hand gestures is captured by a webcam and the input is preprocessed to remove noise and normalize the input. It then extracts 21 hand key points per frame as feature input from the trained CNN model that has high accuracy on ISL gestures classification. The output is displayed as text and consists of the label of the gesture that was recognized, matched to a label through a predictive model once a gesture has been recognized. To enable communication (for all abilities) the system has a feature associated with it Text to Speech (TTS) functionality through which if you have text/data in your input, this recognizes the text/data and brings it out as audible speech in regional languages like Hindi, Kannada and Marathi. Moreover, it offers visual prompts for important recurrent gestures, and is web, easily usable, through a Django-built interface. This end-to-end pipeline can bridge the gap between hearing impaired and hearing people and it is highly useful in educational, healthcare, and social settings.

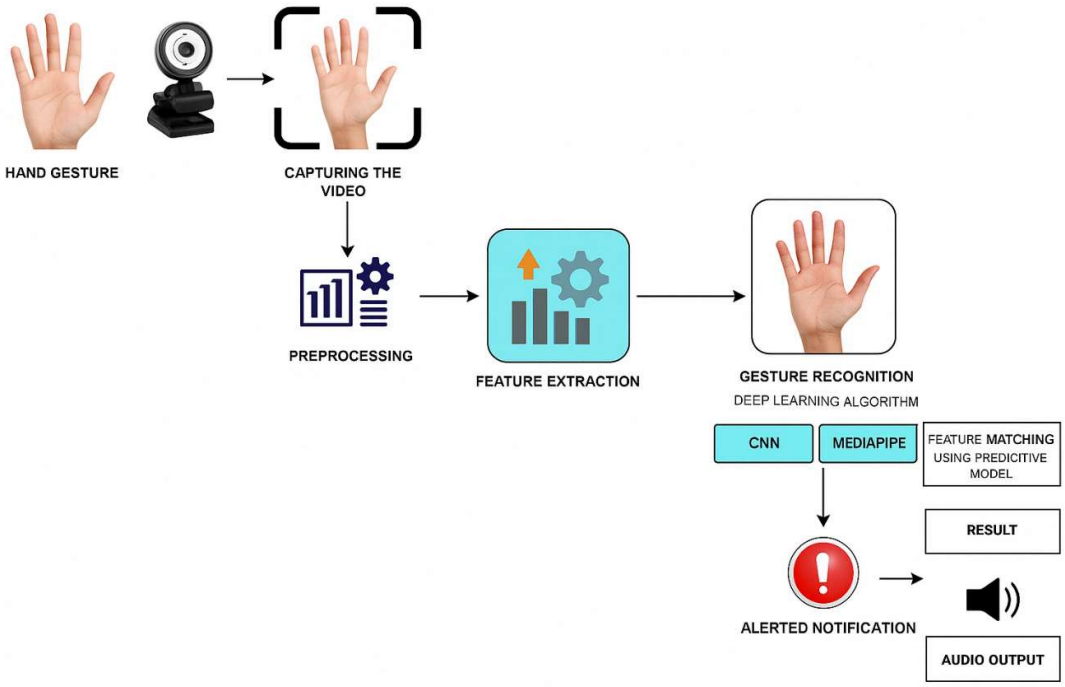


**Figure 1:** Proposed Architecture

**1. Hand Gesture Input:** The process starts with a specific Indian Sign Language (ISL) gesture done by the user using their hands. Individuals who are hearing or speech impaired rely on these gestures as the main type of nonverbal communication. Accurate capture of hand movements is only one of the main objectives of the system, and it is therefore important for users to position their hands in the camera frame for optimal detection and interpretation.

**2. Capturing the Video:** The capturing of a live stream of hand gestures is made by using a webcam or any digital camera device. The input is from a continuous video feed that allows gestures to be detected dynamically. The real-time processing pipeline takes the captured video frames and feeds them into the system to monitor and interpret the gesture as it happens. Due to the clarity and continuity of hand movements, high frame rate and resolution are important at this stage to achieve good clarity.

**3. Preprocessing:** After capturing the video frames, the data quality of these frames is enhanced through the preprocessing stage and then the video frames are ready to be transferred to the feature extraction stage. These constitute preprocessing operations that have the effect of resizing the frames to a standard dimension, normalizing pixel values, removing the background noise, and converting the color format if need be. The reduction of computational load and enhancement of accuracy and consistency of subsequent gesture recognition tasks are the two benefits of following these steps.

**4. Feature Extraction:** In this step, the system takes each preprocessed frame, and extracts meaningful features from it with the help of the MediaPipe framework that demonstrates the locations of 21 key landmarks of the hand (the fingertips, joints, and the palm center) among other notable ones. In other words, these features capture the structure and orientation of the hand within space---the (rich) spatial data in which to classify gestures. Upon extraction of the

features, they are transformed into numerical representations for deep learning model recognition.

**5. Gesture Recognition:** A Convolutional Neural Network (CNN) that has been trained on a labeled ISL dataset reads in the extracted hand landmarks. These features are processed by the CNN so that the specific hand gestures can be recognized based on learning the spatial hierarchies in the data. Basically, this is what MediaPipe allows, that the gesture data is tracked precisely in real-time. It classifies gestures with high accuracy and finds out the meaning of the user's hand sign using this combination.

**6. Feature Matching and Prediction:** Once the system notices the gesture, it takes a cue and uses a predictive model to guess the registered gesture as being from a pre-defined class or label like "Hello", "Help" or "doctor". Right now, the core logic of visual input semantic understanding is this classification. By training this model to map patterns in landmark data to corresponding gestures, slight variations in a gesture are to be robustly handled.

**7. Result Generation:** The input follows prediction, and the recognized ISL gesture is converted into a textual output shown to the user or the system interface. This output enables the nonsigners to read and understand what the deaf or speech-impaired people are writing. This text output is accurate, fast, and easy to read, and is a bridge between our visual gesture input and our verbal communication.

**8. Audio Output:** Once, the recognized text is passed to a Text to Speech engine (e.g., gTTS, pyttsx3) to convert it into audible speech. In other words, user preferences dictate the audio generation in Marathi, Hindi, or Kannada if needed. This acts as a way for hearing people form the gesture's meaning on sound to promote natural and inclusive conversations.

## IV. RESULT AND DISCUSSION

The described real-time Indian Sign Language (ISL) recognition system presents considerable potential for bridging the gap of communication among people suffering from hearing and speech impairments. The system integrates Convolutional Neural Networks (CNNs) for their pattern recognition strengths along with MediaPipe's hand tracking capabilities to take hand gestures into account and make corresponding textual and speech outputs of them. Providing this dual-modality output increases accessibility and promotes easy communication between hearing impaired and hearing people. Simple and well-defined gestures are evaluated to be recognized with a high degree of accuracy and reliability. Nonetheless, due to ambiguities and more complex gestures, occasionally more complex or ambiguous gestures cause misclassification, especially when hand orientation or lighting conditions change. In other words, given that performance in such scenarios is poor, this suggests further training on a more diverse and more extensive dataset could improve performance that much better. User feedback suggests that the system is freestanding, simple, and highly usable, particularly in real-world contexts such as education, the clinic, and everyday personal communication. It has a lightweight implementation and support for regional language speech outputs, which makes it culturally relevant and context-aware.

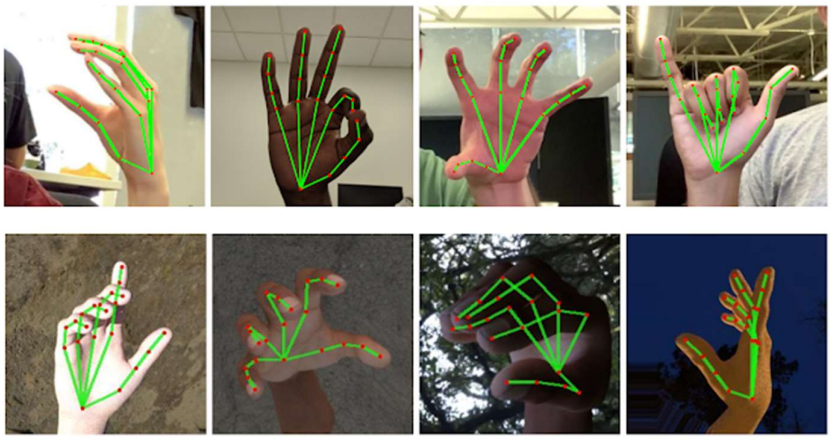

**Figure 2:** Hand gesture detection

The chart titled **"Training Loss and Accuracy on Dataset"** provides a graphical representation of the model's learning performance over training epochs. It displays four curves:

- **train_loss (red)**: This shows the training loss, which decreases sharply dropping close to zero indicating that the model has rapidly learned from the training data.
- **val_loss (blue)**: This remains extremely low and nearly constant, suggesting that the model generalizes well on the validation set without overfitting.
- **train_acc (purple)**: The training accuracy curve is already close to **1.0 (100%)** even from the initial epoch and remains steady, indicating highly accurate predictions on the training data.
- **val_acc (black)**: The validation accuracy also approaches **1.0**, reinforcing that the model performs consistently well on unseen data.


**Figure 3:** Training loss and accuracy of the CNN model

**Table 1:** Comparative Study of Exiting System

| **Method** | **Accuracy (%)** |
|---|---|
| Random Forest (RF) | 96% |
| Support Vector Machine (SVM) | 96.2% |
| Decision Tree (DT) | 95.8% |
| K-Nearest Neighbors (KNN) | 95.6% |
| Convolutional Neural Network (CNN) | 97.2% |

Finally, five machine learning algorithms are compared based on classification accuracy: Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), K Nearest Neighbors (KNN), and Convolutional Neural Network (CNN). All models scrutinize impressively good results with 96% accuracy or above. It is also worth noting that CNN outperforms other models by achieving 97.2% accuracy on this task (image classification), which demonstrates a strong capability for feature extraction and classification in image-based tasks like hand gesture recognition. DT and KNN can be viewed as trailing RF and SVM, respectively. Deep learning has proved to be more effective than traditional classifiers in recognizing Indian Sign Language with complex visual patterns and these results further prove that.

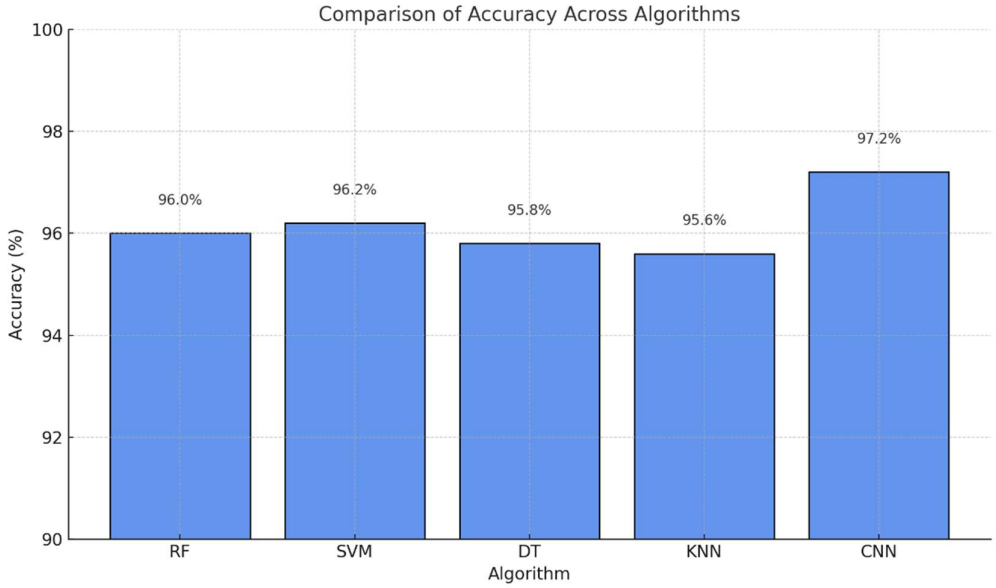


**Figure 3:** Accuracy, Precision, Recall, and F1-Score

## 5. CONCLUSION

A robust and efficient system for real-time Indian Sign Language (ISL) recognition based on Convolutional Neural Network (CNN), MediaPipe for detailed hand gesture tracking and classification in this research. Inclusive, two-way communication between hearing and speech-impaired individuals is facilitated via the combination of gesture-to-text and text-to-speech (TTS) capabilities. High performance is achieved while also being highly usable in multiple real-world scenarios including classrooms, healthcare environments, and daily communication, as demonstrated by an accuracy of 96.1–97.2%. Finally, a comparative analysis is made with classical machine learning algorithms such as SVM, RF, KNN, and DT to demonstrate how CNN is more competent at recognizing complex gestures than the other classifiers. In addition, the embedded lightweight design and web-based user interface provide portability and accessibility of the system across devices. Although it offers high performance, future work will focus on the scalability of continuous gesture recognition, the number of regional ISL variations, and deployment to mobile platforms. More broadly, this work shows the way for the use of deep learning and computer vision to create practical and real-time assistive technologies that help empower the deaf family to become socially included.

**References: -**


1. Prudhvi, B., Neeraj, P., & Deepthi, V. H. (2023). An Efficient Real-Time Indian Sign Language (ISL) Detection using Deep Learning. International Conference Intelligent Computing and Control Systems, 430–435. https://doi.org/10.1109/ICICCS56967.2023.10142596
2. Dewangan, S., Patra, J. P., & Samal, S. (2025). Optimizing Deep Learning Models for Dynamic Indian Sign Language Interpretation. Social Science Research Network. https://doi.org/10.2139/ssrn.5089115
3. Deshpande, S., & Shettar, R. (2023). Hand Gesture Recognition Using MediaPipe and CNN for Indian Sign Language and Conversion to Speech Format for Indian Regional Languages. 1–7. https://doi.org/10.1109/csitss60515.2023.10334218
4. Khaire, P. B., Mahajan, S., Kale, P. R., Pawar, D. B., Pawar, S., Dedgaonkar, S., & Shewale, Prof. C. (2024). SilentSync-AI. https://doi.org/10.1109/mitadtsocicon60330.2024.10575259
5. Nimbalkar, S. V., Vaidya, S. N., Gade, M., Hagare, P. S., & Shendage, P. N. (2024). Empowering Deaf with Indian Sign Language Interpreter using Deep Learning. https://doi.org/10.1109/mitadtsocicon60330.2024.10575064
6. Liu, L., Zheng, Y., Y., Zhang, X., & Yang, K. (2024). A Sign Language Recognition Based on Optimized Transformer Target Detection Model (pp. 197–208). https://doi.org/10.1007/978-3-031-50580-5_16
7. Singh, R., Mishra, A., & Mishra, R. (2024). Enhancing Sign Language Recognition: Leveraging EfficientNet-B0 with Transformer-based Decoding. INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY SCOPE, 05(04), 679–688. https://doi.org/10.47857/irjms.2024.v05i04.01241
8. Neto, Pedro, Miguel Simão, Nuno Mendes, and Mohammad Safeea. "Gesture-based human-robot interaction for human assistance in manufacturing." The International Journal of Advanced Manufacturing Technology 101 (2019): 119-135.

9. M. Bohacek and M. Hrúz, "Learning from what is already out there: Fewshot sign language recognition with online dictionaries," in Proc. IEEE 17th Int. Conf. Autom. Face Gesture Recognition. (FG), Jan. 2023, pp. 1–6.
10. Y. Ma, T. Xu, S. Han, and K. Kim, ‘‘Ensemble learning of multiple deep CNNs using accuracy-based weighted voting for ASL recognition,’’ Appl. Sci., vol. 12, no. 22, p. 11766, Nov. 2022.
11. G. Pomianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, ‘‘Recent advances in the automatic recognition of audiovisual speech,’’ Proc. IEEE, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.
12. A. A. Kindiroglu, O. Özdemir, and L. Akarun, ‘‘Temporal accumulative features for sign language recognition,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1288–1297.
13. L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, ‘‘Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition,’’ in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 3120–3128.
14. P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training strategies for improved lipreading," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 8472–8476.
15. A. Koumparoulis and G. Potamianos, ‘‘Accurate and resource-efficient lipreading with Efficientnetv2 and transformers,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 8467–8471.
16. D. Ivanko, D. Ryumin, A. Kashevnik, A. Axyonov, and A. Karnov, ‘‘Visual speech recognition in a driver assistance system,’’ in Proc. 30th Eur. Signal Process. Conf. (EUSIPCO), Aug. 2022, pp. 1131–1135.
17. P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audiovisual speech recognition with automatic labels," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1–5.
18. K. R. Prajwal, T. Afouras, and A. Zisserman, ‘‘Sub-word level lip reading with visual attention,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 5162–5172.
19. A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, ‘‘Jointly learning visual and auditory speech representations from raw data,’’ 2022, arXiv:2212.06246. [156] P. Ma, S. Petridis, and M. Pantic, ‘‘Visual speech recognition for multiple languages in the wild,’’ Nature Mach. Intell., vol. 4, no. 11, pp. 930–939, Oct. 2022.
20. X. Zhang, F. Cheng, and W. Shilin, ‘‘Spatio-temporal fusion based convolutional sequence learning for lip reading,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 713–722.
21. D. Feng, S. Yang, S. Shan, and X. Chen, ‘‘Learn an effective lipreading model without pains,’’ 2020, arXiv:2011.07557.
22. M. Kim, J. Hong, S. J. Park, and Y. Man Ro, ‘‘Multi-modality associative bridging through memory: Speech sound recollected from face video,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 296–306.