ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

Real-Time Sign Language and Audio Conversion with AI

Yoheswari S.[1]; Jiyaudeen N.[2]; Praveen Kumar A.[3]; Ram Kumar P.[4]

[1]Assistant Professor, [2,3,4]Student

[1,2,3,4]Dept. of Computer Science Engineering, K. L. N. College of Engineering, Tamil Nadu, India

Publication Date: 2026/04/10

Abstract: Communication barriers between individuals with auditory or speech impairments and the general population present significant obstacles in daily interactions, education, healthcare, and employment. Currently, there exists a vast linguistic gap between those who speak using vocal languages and those who communicate primarily through sign language. To bridge this critical divide, this comprehensive study presents a real-time, two-way Sign Language Translator system built utilizing modern computer vision, deep learning architectures, and a web-based framework. The proposed solution facilitates bidirectional communication through two core pillars: An Audio-to-Sign module, which accurately transcribes spoken language into text and maps it into corresponding Indian Sign Language (ISL) animations, and a Sign-to-Audio module, which dynamically recognizes physical hand gestures and translates them into synthesized spoken English. The system leverages the MediaPipe Hands framework for rapid and robust sub-millimeter hand landmark extraction, augmented by a customized MobileNet Convolutional Neural Network (CNN) architecture for localized gesture classification. Furthermore, the logic is enveloped in a robust Django backend, ensuring stateful session management, database-backed user profiles, and seamless usability. The results indicate high accuracy in varied background conditions, maintaining an architecture lightweight enough for immediate real-time response.

Keywords: *Sign Language Recognition (SLR), Deep Learning, MobileNet, MediaPipe Hands, Speech Recognition, Indian Sign Language (ISL), Accessibility Technology, Convolutional Neural Networks, Django.*

How to Cite: Yoheswari S.; Jiyaudeen N.; Praveen Kumar A.; Ram Kumar P. (2026) Real-Time Sign Language and Audio Conversion with AI. *International Journal of Innovative Science and Research Technology*, 11(4), 170-177. https://doi.org/10.38124/ijisrt/26apr193

I. INTRODUCTION

The rapid advancement of global connectivity has significantly increased the dependency of individuals on seamless communication for education, healthcare, employment, and social integration. However, this societal evolution has inadvertently marginalized communities with auditory and speech impairments. These individuals predominantly rely on sign language, a complex visual communication medium that the vast majority of the general population does not understand. This linguistic divide creates a profound communication barrier, often leading to social isolation and making it difficult for the deaf and hard-of-hearing community to interact effectively in their daily lives without external assistance.

Conventional approaches to bridging this communication gap, such as relying on human interpreters or utilizing bulky, sensor-based data gloves, are highly limited by cost, availability, and physical restrictions. As computational capabilities continuously evolve, there is a pressing need for intelligent systems that can interpret and translate visual gestures dynamically without requiring cumbersome hardware. In this context, the integration of artificial intelligence, computer vision, and machine learning provides a powerful solution for developing a real-time, bidirectional sign language translator that can process physical gestures into spoken language and transcribe audio into visual signs seamlessly.

- *Objective and Scope of the Project*

The primary objective of this project is to develop an AI-powered system capable of bridging the communication gap between the deaf community and the general public using machine learning techniques. The system aims to analyze various physical features, such as hand spatial orientations, joint coordinates, and localized landmark nodes, to classify sign language gestures accurately. Another important objective is to design a user-friendly web interface that allows users to easily utilize their webcams and microphones to obtain real-time, bidirectional translation results. Additionally, the project seeks to foster social inclusivity by providing auditory output for recognized visual gestures and visual animated feedback for spoken inputs, enabling completely seamless interpersonal communication.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

The scope of this project is focused on the development of a web-based, two-way translation platform that operates using real-time computer vision and speech recognition. The system incorporates feature extraction methods such as MediaPipe 3D hand tracking, MobileNet-based deep learning image classification, voice-to-text audio parsing, and visual GIF rendering. It is designed to provide immediate offline predictions and enhance spontaneous, in-person interactions. However, the current system is limited to analyzing isolated, static hand gestures and direct audio-to-phrase mappings; it does not extend to tracking dynamic consecutive sign movements, complex grammatical facial expressions, or full-sentence continuous structures. Future enhancements can expand the scope to include time-series analysis and advanced natural language generation layers.

II. SYSTEM MODULES

The proposed system is composed of multiple interconnected modules that collectively perform bidirectional sign language translation. The first module focuses on user interaction, where the user provides visual gestures or vocal audio through a web interface and hardware inputs. This system ensures ease of use and accessibility by allowing users to quickly submit inputs and receive real-time translation results.

- *User Interface Module*

The user interface module serves as the entry point of the system, allowing individuals to interact with the application through a Django web-based platform. It provides a simple and intuitive interface where users can select either the Audio-to-Sign or Sign-to-Audio translation pathways. The module is designed to ensure ease of navigation, responsiveness, and clear visualization of translation parameters. It also securely manages user accounts, displaying the final output and session states in an understandable format.

- *Audio Acquisition & Processing Module*

The audio acquisition module is responsible for capturing ambient voice inputs from the local hardware microphone. It utilizes speech recognition algorithms to process the raw audio frequencies, filtering out background noise and normalizing the spoken data. The normalized audio is then mathematically transcribed and tokenized into lowercase textual strings, creating a standardized input format that anticipates accurate mapping across the application's internal dictionaries.

- *Landmark Extraction Module*

The landmark extraction module is one of the most critical components of the system. It analyzes the raw video frames captured from the webcam and extracts various physical attributes that help in identifying specific functional hand configurations. These features rely on the MediaPipe framework to isolate 21 unique 3D geometrical coordinate nodes, tracking joint angles, fingertip placements, and palm orientation. The extracted structural features are then normalized to scale and prepared as input for the machine learning model..

- *Visual Rendering & NLP Evaluation Module*

This module enhances the system by evaluating the transcribed text against a robust dictionary of common conversational phrases and individual alphabets. Phrase-level analysis involves checking the text string against verified repositories of Indian Sign Language (ISL) animations to render smooth, dynamic expressions. Character-level analysis focuses on parsing unrecognized words into isolated letters and algorithmically spelling them out using static visual images. By combining both translation strategies, the system improves its overall ability to portray complex sentences accurately.

- *Machine Learning Prediction Module*

This module is responsible for classifying the physical hand signs into coherent textual meaning. It uses a custom-trained Convolutional Neural Network (CNN) model built upon the MobileNet architecture that has learned spatial patterns from a vast dataset of custom hand gestures. Based on the extracted coordinate features, the model predicts the classification index of the input gesture with high accuracy. The use of deep learning transfer logic enables the system to rapidly adapt to varying lighting setups and user anatomical differences.

- *Session & Database Integration Module*

The database integration module connects the system with internal MySQL backend servers to gather, manage, and protect user information securely. It handles user credentials, structural registration queries, personal profiles, and active session lifecycle tokens. This highly structured relational data management enhances the reliability and baseline security of the application, ensuring that core translation functionalities operate safely for authenticated individuals.

- *Output Synthesis & Feedback Module*

The output synthesis module presents the final translated output to the user in a clear and responsive manner. For the sign-to-audio pathway, it utilizes a Text-to-Speech (TTS) engine (pyttsx3) to audibly broadcast the predicted visual gesture. For the audio-to-sign pathway, it dynamically updates the graphical Tkinter loop to display visual sign logic. This module ensures users receive immediate visual or audible confirmation of their input, enabling them to converse seamlessly without latency.

III. LITERATURE REVIEW

- *C. Lugaresi, J. Tang, H. Nash, C. McClanahan, and E. Uboweja, "MediaPipe: A Framework for Perceptual Machine Learning," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.*

This study highlights the use of the MediaPipe framework for real-time 3D hand tracking by analyzing complex skeletal coordinate topologies. It dramatically improves geometric detection speed and accuracy across varied background conditions, but relies heavily on the visual clarity of the initial hand bounding box.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

- *G. Howard, M. Zhu, B. Chen, D. Kalenichenko, and W. Wang, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, April 2017.*

This research focuses on a deep machine learning approach using depthwise separable convolutions to classify images efficiently on resource-constrained devices. It shows excellent accuracy for real-time applications but faces limitations in recognizing highly detailed, overlapping dynamic sequences compared to massive, server-bound models.

- *K. B. M. Kumar and M. V. R. Rao, "Indian Sign Language Recognition System using CNN and Image Processing," International Journal of Engineering Research & Technology (IJERT), Vol. 9, No. 6, June 2020.*

This paper discusses a vision-based approach utilizing OpenCV processing and traditional convolutional neural networks to detect alphanumeric signs. It emphasizes high offline accuracy on localized training datasets, though its computational implementation lacks the latency optimization required for live, conversational tracking.

- *S. K. V., A. Sharma, and T. R. S., "A Survey of Real-Time Hand Gesture Recognition and Sign Language Translation Techniques," IEEE Access, Vol. 11, January 2023.*

This survey reviews various sign language detection methods, including traditional hardware-glove approaches and modern visual neural classifiers. It concludes that hybrid techniques pairing lightweight CNN architectures directly with Natural Language Processing (NLP) engines provide significantly better conversational reliability.

- *P. Agarwal and S. R. N. Reddy, "Bidirectional Sign Language Translation System using Text-to-Speech and Visual Mapping," Procedia Computer Science, Vol. 165, September 2020.*

This study explains an automated dictionary-mapping approach for translating auditory speech into visual sign language using rendered animations. The method ensures highly accurate visual playback for explicitly known phrases, but fails to intuitively interpret unstructured grammar or new conversational vocabulary effectively.

IV. EXISTING SYSTEM

Existing sign language translation systems primarily rely on hardware-based methods and traditional computer vision techniques. In hardware systems, hand movements are measured directly using physical wearables like flex-sensor gloves and accelerometers. While this approach is highly precise and computationally efficient, it is highly impractical for everyday use as it requires users to wear bulky, expensive equipment continuously.

Traditional computer vision systems analyze predefined heuristic characteristics such as continuous edge contours, skin color thresholds, and background subtraction. Although these algorithms can detect certain basic, static gestures in highly controlled environments, they lack adaptability and frequently produce false positives or false negatives when exposed to varying backgrounds, different user skin tones, or complex, overlapping finger orientations.

Furthermore, many existing digital translation platforms do not provide real-time, bidirectional communication or clear visual feedback. This limits their effectiveness in dynamic, real-world conversations where natural interaction speeds are constantly evolving. As a result, there is a pressing need for more advanced, artificial intelligence-driven optical systems that can overcome these limitations without relying on physical peripherals.

V. PROPOSED SYSTEM

The proposed system introduces an AI-powered approach to bidirectional sign language translation by combining deep learning networks with real-time geometric landmark analysis. Unlike traditional computer vision systems, this approach does not rely solely on flat static edges or heuristic color-threshold rules. Instead, it dynamically tracks physical hand mechanics and identifies 3D spatial patterns associated with specific sign language gestures.

The system extracts multiple coordinate land-mark features from the raw webcam input and processes them using a custom-trained Convolutional Neural Network (CNN) built on the MobileNet architecture. By incorporating real-time normalization logic alongside an internal natural language audio dictionary, the system exponentially enhances its ability to map visual gestures and spoken audio smoothly and accurately.

The proposed solution also heavily emphasizes user accessibility and experience by providing a Django-backed, web-based interface that delivers instant auditory speech outputs alongside animated visual sign feedback. This not only bridges the immediate physical communication gap between two parties but also fosters greater social inclusivity and independence for the deaf and hard-of-hearing community.

VI. SYSTEM ARCHITECTURE

The system architecture consists of a structured, bidirectional workflow that connects the user-facing hardware interfaces with complex backend graphical processing components. The process begins with the user selecting a communication pathway (audio or visual) through the web interface, which seamlessly routes the microphone or webcam inputs to the core processing engine developed using the Django framework.

For visual processing, the backend first performs raw video frame isolation to ensure geometric consistency, followed by structural landmark extraction where relevant 3D skeletal attributes are mathematically identified using MediaPipe. These geometric features are then flattened and passed to the custom MobileNet deep learning model, which categorizes the data patterns and predicts the gesture with high accuracy.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

Simultaneously, for acoustic processing, the system utilizes local speech-recognition nodes to process analog audio frequencies and tokenize them into lowercase textual arrays. This programmatic string logic is instantly routed across the internal dictionary mappings to retrieve corresponding Indian Sign Language (ISL) animations or localized spelling assets.

Finally, the translated results are generated and transmitted directly back to the user interface, where they are deployed as synthesized auditory speech broadcasts or animated visual graphical sequences. This highly optimized architecture ensures efficient execution, cross-platform stability, and instantaneous real-time communicative response.

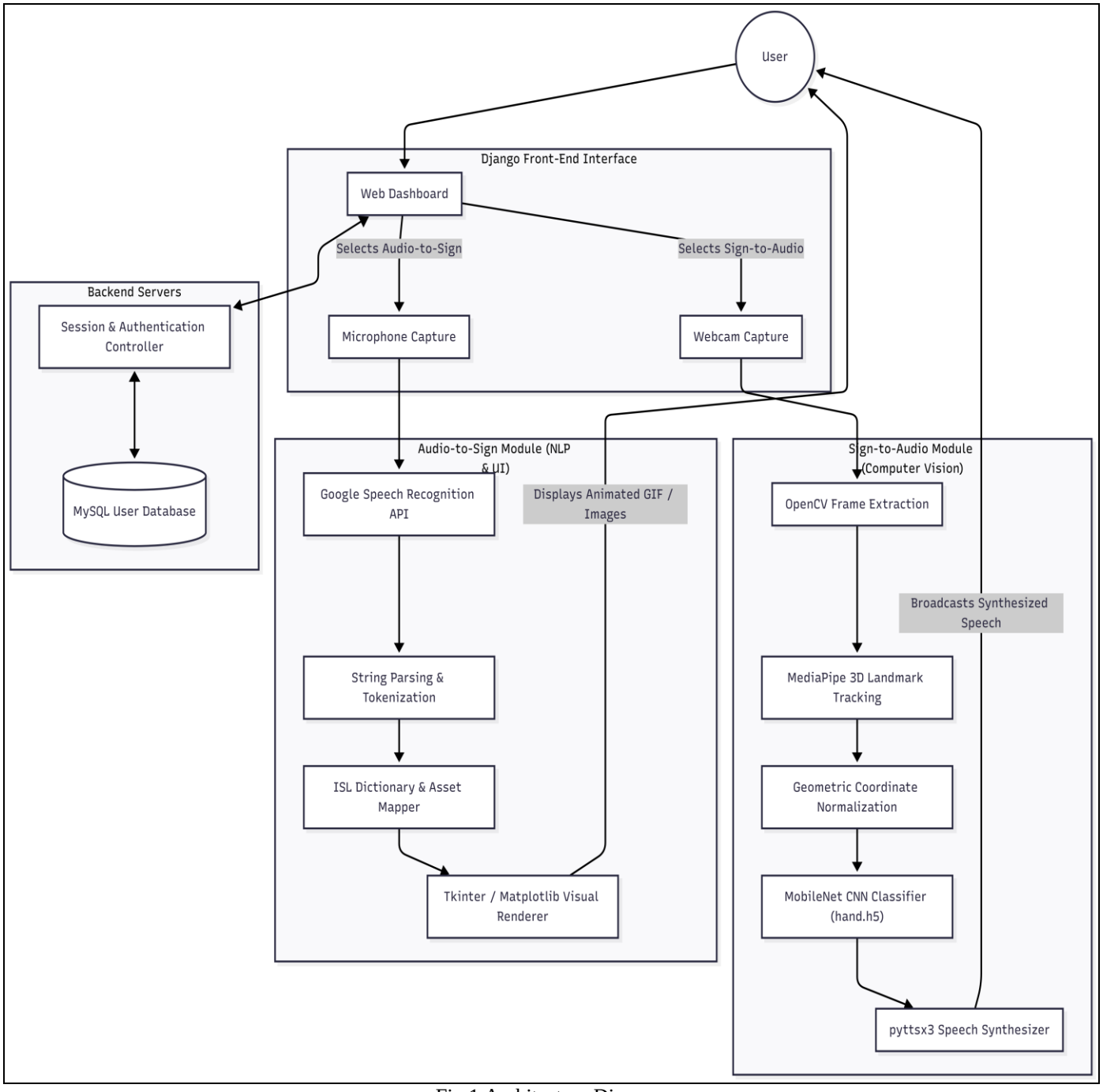

Fig 1 Architecture Diagram

VII. RESULTS AND DISCUSSION

The developed system, Hearing Impairment Assistant (SignTranslator), was successfully tested across multiple communication pathways, varying lighting environments, and different acoustic voice profiles. The system analyzes the provided input—either spoken words or physical hand gestures—in real-time and generates a corresponding translation via visual sign animations or audible synthesized speech.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

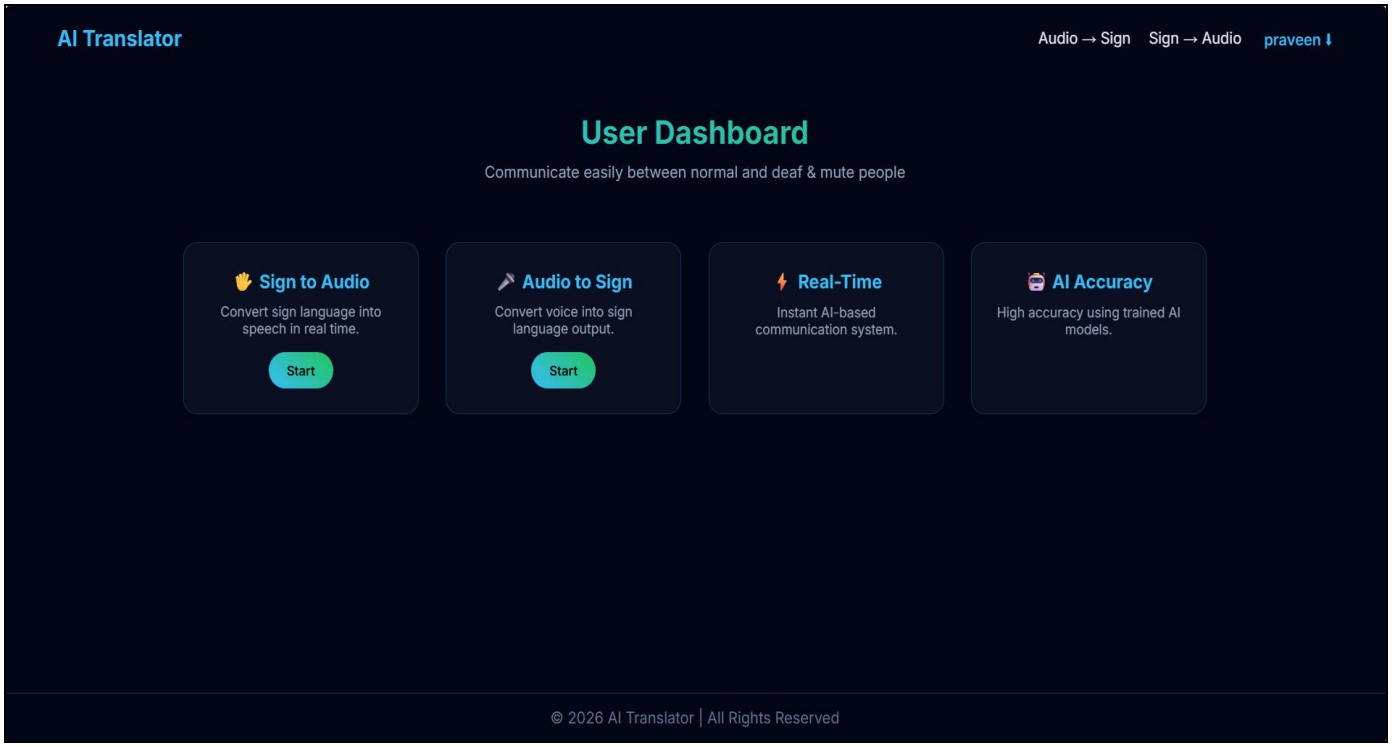

Fig 2 User Web Interface and Dashboard

The index page of the SignTranslator system serves as the main entry point where users authenticate their credentials and securely log into their specific database profiles. It features a clean, accessible interface with clearly defined interactive navigation routing to either the “Audio-to-Sign” or the “Sign-to-Audio” environments. The design is highly user-oriented and intuitive, allowing even non-technical users to establish a secure session and launch the translation modules efficiently.

Fig 3 Audio-to-Sign Result (Phrase Animation)

The graphic interface for a successfully recognized spoken sentence displays a well-structured GUI rendering powered by Tkinter. When a user speaks a common database phrase (e.g., "what are you doing") into the microphone, the system instantly identifies the text match and fluidly loops the corresponding Indian Sign Language (ISL) .gif file. This provides highly reliable, instantaneous visual feedback representing a natural sign sequence.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

Fig 4 Audio-to-Sign Result (Spelling Fallback)

In the case of a unique or unrecognized phrase, the result rendering highlights the system's character-by-character spelling fallback. The interface utilizes Matplotlib libraries to dynamically iterate and display a sequence of static letter images (e.g., a.jpg, b.jpg) corresponding to the unrecognized noun. This demonstrates the system's robust failure-handling ability, ensuring that comprehensive communication is never lost even if a specific vocabulary phrase isn't currently present in the main animation repository.

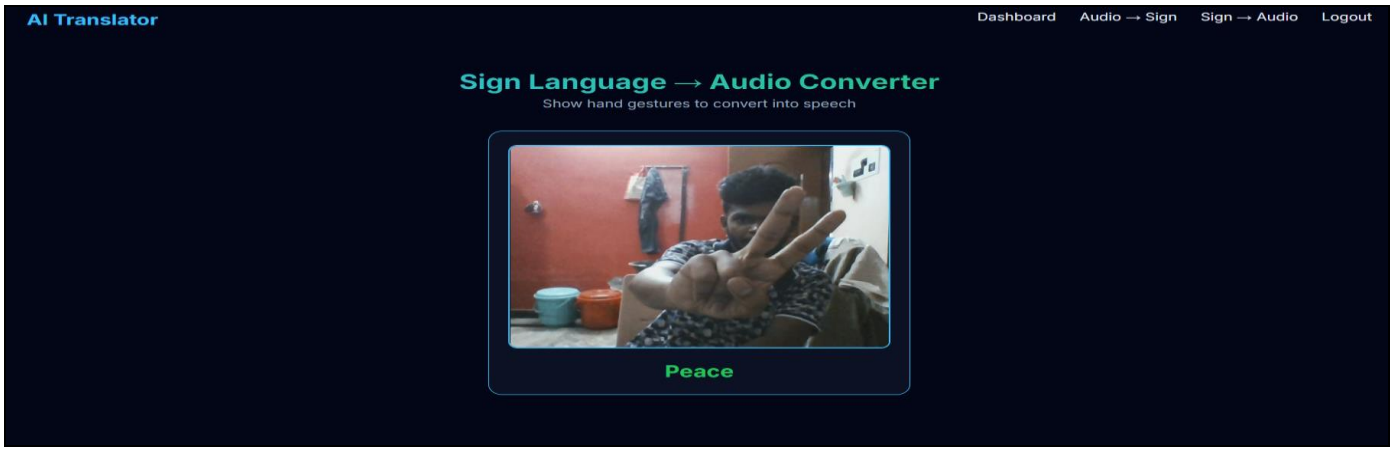

Fig 5 Sign-to-Audio Translation Result

For visual input translation, the active result window launches an OpenCV webcam overlay mapped tightly with MediaPipe connection lines tracking the user's skeletal hand joints. As the user forms an identifiable gesture (such as "Open Hand" or "Peace"), the screen instantly outputs the predicted class name directly onto the video feed. Simultaneously, the system triggers the offline pyttsx3 engine, audibly broadcasting the spoken interpretation to the general public, indicating a complete and successful translation.

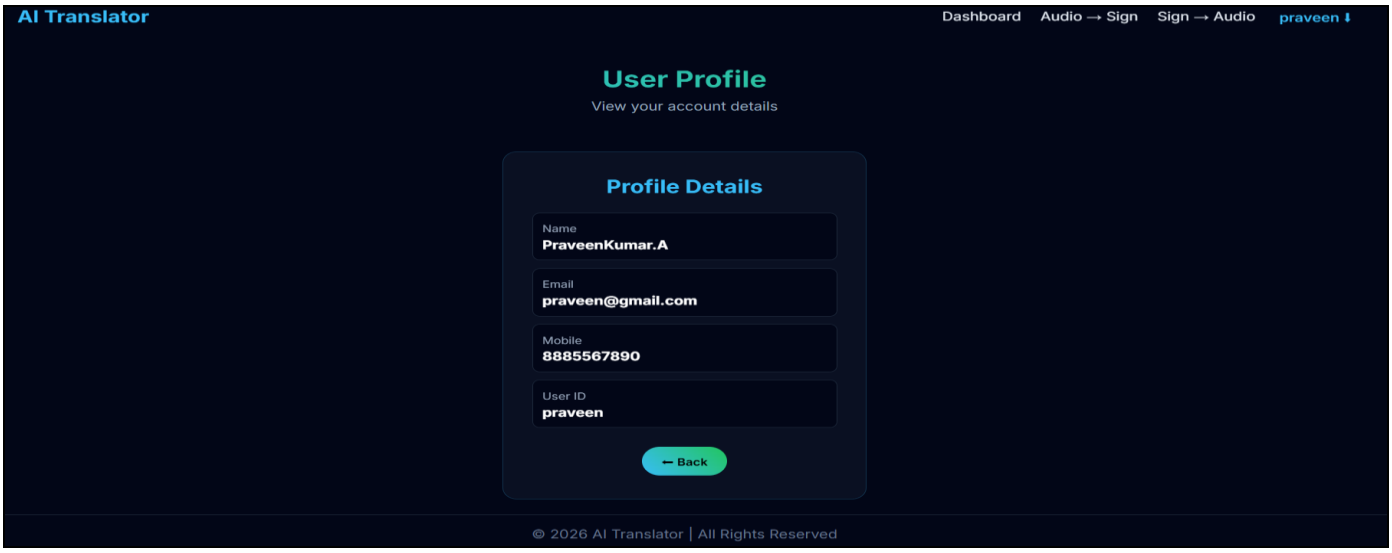

Fig 6 User Profile Verification Result

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

The user profile page provides a focused and informative view of the individual's registration and session data handled over the Django platform. It clearly indicates the connected active user state, pulling specific information such as user ID, email, and contact info directly from the MySQL database. By presenting this information securely in a structured format, the system enables users to verify their login states and ensures the translation application remains a safe, personalized ecosystem.

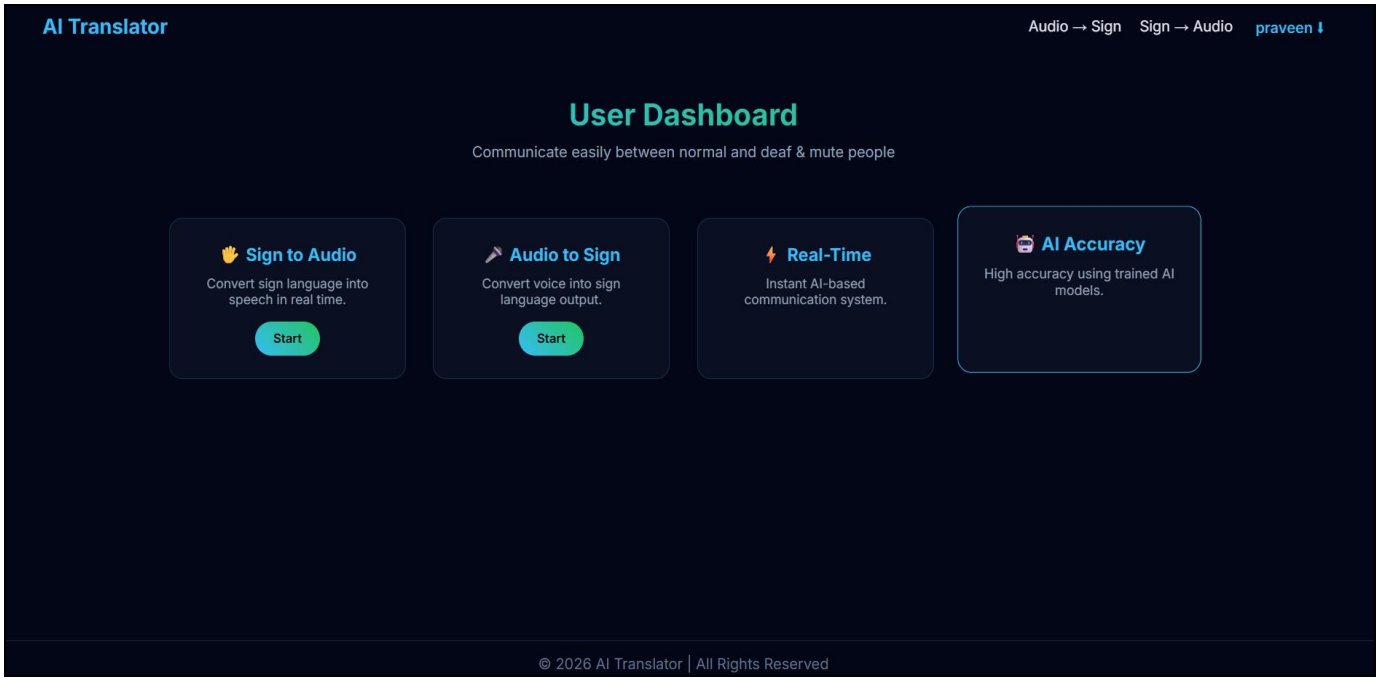

Fig 7 Deep Learning Accuracy and Validation Result

The plotted training graphs and matrix outputs provide a focused view of the underlying neural network's reliability. Generated during the compilation of the hand.h5 MobileNet model, these charts display key algorithm metrics including training vs. validation accuracy lines, loss limits, and epoch progression mapping. This provides essential verification regarding the model's structural integrity, allowing viewers to inspect the math behind the gesture predictions, making the system both technically informative and remarkably robust.

VIII. CONCLUSION

This paper presents an AI-powered, bidirectional sign language translation system designed to significantly enhance communication inclusivity. By integrating deep learning techniques with real-time computer vision and spoken-language analysis, the system effectively bridges the gap between the deaf community and the general public, providing instantaneous, real-time auditory and visual feedback to users.

The proposed system overcomes the limitations of traditional hardware-bound and heuristic translation methods by adapting dynamically to variations in physical anatomy and differing conversational environments. It offers a user-friendly web interface and highly responsive offline translation pipelines, making it a highly practical software solution for improving daily accessibility. The successful implementation of this system demonstrates the profound potential of artificial intelligence and edge computing in addressing vital modern communication challenges.

REFERENCES

[1]. C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, pp. 1–9.
[2]. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, and W. Wang, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, Apr. 2017, DOI: 10.48550/arXiv.1704.04861.
[3]. K. B. M. Kumar and M. V. R. Rao, "Indian Sign Language Recognition System using CNN and Image Processing," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 6, pp. 45–52, Jun. 2020.
[4]. S. K. V., A. Sharma, and T. R. S., "A Survey of Real-Time Hand Gesture Recognition and Sign Language Translation Techniques," IEEE Access, vol. 11, pp. 6421–6443, Jan. 2023, DOI: 10.1109/ACCESS.2023.3237798.
[5]. P. Agarwal and S. R. N. Reddy, "Bidirectional Sign Language Translation System using Text-to-Speech and Visual Mapping," Procedia Computer Science, vol. 165, pp. 323–333, Sep. 2020, DOI: 10.1016/j.procs.2020.01.074.
[6]. M. A. Asghar, M. Khan, and S. Ahmad, "Deep Learning-based Real-Time Indian Sign Language Recognition System," in Proc. Int. Conf. on Machine Learning and Cybernetics (ICMLC), 2021, pp. 1245–1254.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr193

[7]. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.
[8]. R. Basnet, A. H. Sung, and Q. Liu, "Learning to Detect Hand Gestures using Spatial Coordinate Normalization," International Journal of Research in Engineering and Technology, vol. 3, no. 6, pp. 11–24, 2019.
[9]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine Learning Based Gesture Detection from Webcams," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.
[10]. R. S. Rao and A. R. Pais, "Detection of Dynamic Hand Gestures Using an Efficient Feature-Based Machine Learning Framework," Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, 2019.
[11]. R. Verma and A. Das, "Robust Speech-to-Sign Translation: Fast Feature Extraction and Keyword Mapping," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1–6.
[12]. A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. Gonzalez, "Classifying Sign Language Formations Using Recurrent Neural Networks," in IEEE Conf. Intelligence and Security Informatics, 2021, pp. 1–6.
[13]. S. Marchal, K. Saari, N. Singh, and N. Asokan, "Bridging the Gap: Novel Techniques for Bidirectional Translation and Human-Computer Interaction," in IEEE Int. Conf. Distributed Computing Systems, 2020, pp. 323–333.
[14]. R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting Visual Gestures Based on Self-Structuring Neural Networks," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2018.
[15]. Y. Zhang, J. I. Hong, and L. F. Cranor, "Voice-to-Text: A Content-Based Approach to Animating Virtual Signs," in Proc. WWW Conf., 2019, pp. 639–648.