International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

REAL TIME SIGN LANGUAGE TRANSLATION AND AUDIO TRANSLATION ON VIDEO CALLS

Prajwal Bhosale[1]*, Dhruv Bendre[2], Sandhya Tiwari[3] and Rohini Bhosale[4]
[1,2,3,4]Department of Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India
[1]bprajwal53@gmail.com, [2]bscit.dhruvbendre@gmail.com, [3]bscit.sandhya@gmail.com,
[4]jagadale.rohini601@gmail.com
*Corresponding author: Prajwal Bhosale, Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India. Email: bprajwal53@gmail.com

ABSTRACT
Communication barriers continue to affect individuals with speech impairments who rely on sign language, especially during real-time digital communication such as video calls. Most current video conferencing platforms do not support built-in sign language interpretation, which forces users to depend on text chat or third-party assistance. This paper presents a multi-modal WebRTC-based framework designed to enable real-time sign language recognition, text translation, and cross-language audio generation during live video communication.

The proposed system works as an intelligent processing layer that runs alongside the video stream, allowing smooth and natural interaction without delay. Instead of analyzing full image frames, the system extracts three-dimensional hand landmark coordinates using MediaPipe and CVZone. These coordinates are processed using NumPy arrays, allowing the model to focus on gesture structure rather than background details such as lighting or environment noise. This coordinate-based approach improves recognition reliability while maintaining fast processing speed, making the system suitable for low-bandwidth video calls.

During communication, recognized hand gestures are instantly converted into text and further synthesized into speech, allowing hearing participants to listen to the sign language as spoken audio. At the same time, spoken communication from the other participant is converted into text and displayed for the signer. By combining sign-to-speech and speech-to-text capabilities, the framework creates a two-way accessible communication environment.

The results demonstrate that the system provides stable real-time performance with reliable recognition accuracy during live video calls, highlighting its potential to improve digital accessibility and support inclusive communication.

Keywords: *Sign Language Recognition, WebRTC, Multi-Modal Systems, Real-Time Translation, Accessibility, Audio Synthesis*

1. INTRODUCTION
Recent advances in real-time communication have improved how people connect through digital platforms. However, accessibility remains a concern for users who rely on sign language. Most existing video conferencing systems are designed mainly for spoken communication and do not include built-in support for sign language translation.

As a result, sign language users often depend on interpreters, text-based messaging, or external tools, which can interrupt natural interaction and increase communication effort. Improvements in computer vision and machine learning now allow hand gestures to be recognized in real time using standard cameras, making sign language translation possible without specialized equipment.

In parallel, WebRTC has become a widely used technology for low-latency video and audio communication in web applications. When combined with machine learning-based sign recognition and speech synthesis, WebRTC enables the development of accessible communication systems that function directly within a browser.

This paper presents a multi-modal WebRTC framework that translates sign language into text and spoken output during live video calls. The proposed system operates in a web-based environment using commonly available hardware, supporting practical and scalable real-time communication.

The main contributions of this work include the design of a WebRTC-based framework for real-time sign language translation during live video communication. The proposed system integrates landmark-based hand gesture recognition with a Random Forest classifier to achieve low-latency and efficient performance in a web environment. In addition, the framework supports multi-modal output in the form of text and audio, enabling communication between signing and non-signing users.

ISSN 2394 - 7780

2. LITERATURE REVIEW

The framework proposed in this study aligns closely with recent works on real-time sign translation and accessible video meetings, but it is clearly different from being browser-based multimodal and explicitly improved for low-latency two-way interaction. Many current systems share the same motivation of breaking the communication wall caused by interpreter shortages and limited sign proficiency among hearing users. For example, (Shanin and Ismail, 2024) emphasize that there is a pressing need for an efficient, sign-driven, integrated end-to-end translation system due to a global shortage of interpreters and the growing Deaf and Hard of hearing population.

Several works explicitly target virtual or video-meeting contexts, such as (Dey, 2025)AI-powered ISL recognition system "integrated with live video calling" using WebSockets to keep "unbroken video and audio communication" while recognized gestures are streamed as text to the peer. Similarly, (Isaaq, 2025) described "Real-Time Sign Language Recognition in Digital Meetings" that leverages deep learning with WebRTC and text-to-speech to convert sign gestures into audible speech, specifically to address accessibility gaps in mainstream virtual platforms.

Other projects focus on overlay or extension models, such as (Vishwekar, M., Injeti, D., Gupta, R. O., & Gowalker, N., 2025) SMART SIGN, a lightweight translator that plugs into existing platforms via virtual camera to render real-time captions from sign input. (Aabid, 2025) similarly uses a YOLOv5- and CUDA-accelerated CNN, wrapped as a virtual camera through OBS so any conferencing tool can display subtitle output from sign gestures. These solutions, like the one presented here, use off-the-shelf cameras rather than gloves or depth sensors and emphasize low cost and deployability. For instance, (Dubey, 2025) webcam-based interpreter stresses being "lightweight, cost-effective, and deployable on standard hardware" while providing dual text and audio output for accessibility. At the same time, several systems exploit more complex deep learning pipelines to push recognition accuracy above 90–98% on static and dynamic gestures (M. Al-Qurishi, T. Khalid, and R. Souissi,).

Architecturally, this design utilizing a Tailwind CSS frontend, a Python backend for landmark normalization, Flask-SocketIO for duplex signaling, and gTTS for multilingual speech echoes the layered schemes found in the literature (V. Leiva, M. Z. Ur Rahman, M. A. Akbar, et al, 2025), (S. Ghodake, S. Gavandi, K. Gambhir, G. Bangale, and N. Deshpande), (Rao, V. G. S, Brunda, G, & Naidu, R. C. A., 2025). The choice of a Random Forest classifier over normalized MediaPip e-style 21-point landmarks trades some peak accuracy for interpretability and speed (A. D. Goenawan and S. Hartati, , 2024), akin to systems that explicitly avoid specialized sensors (Sonare, B., Padgal, A., Gaikwad, Y., & Patil, A. , 2021), (Zhou, Z., Chen, K., Li, X., et al., 2020). Furthermore, the explicit "two-way loop" vision resonates with broader multimodal systems that combine sign-to-speech and speech-to-sign pathways into an integrated platform (S. Ghodake, S. Gavandi, K. Gambhir, G. Bangale, and N. Deshpande), (Rao, V. G. S, Brunda, G, & Naidu, R. C. A., 2025), (Boobal, A., Jasmine, J. L., Reddy, C. C. K., Reddy, C. A., & Rohith, C. B. V. S., 2024). By adding gesture-based UI controls and multilingual feedback, this platform deepens user autonomy and inclusivity goals aimed at mitigating social and professional isolation for Deaf users (Shwethashree, 2025), (Upadhaya, P., Chamoli, D., Chopra, A., & Raheja, S, 2025).

In conclusion, this web-based multimodal framework fits squarely within an active research stream on vision-based, real-time sign translation for meetings (G. Chakali, C. G. Reddy, and B. Bharathi, 2023), (Mukherjee, S., Akhtar, M. H., & Kannadasan, D, 2025), (Wilson, N., Sunny, S., Alex, A. C., Sachin, S. R., & Jayakrishnan, A, 2025), but it combines several strengths: a standards-based browser stack, a distance-invariant hand-landmark normalization pipeline, integrated speech synthesis, and explicit design for two-way communication. This positions the project as both technically feasible and socially impactful, directly targeting the persistent accessibility gap in mainstream video conferencing (Subashini, V., Someshwaran, B., Sowmya, S., & Kumar, S. A., 2024).

2.1. Theoretical framework

The proposed system is grounded in established theories of multi-modal communication, real-time interaction, feature-based pattern recognition, and human-centered accessibility design. These theoretical foundations explain how visual gestures, audio signals, and textual representations can be integrated to support accessible and natural communication in real-time digital environments (Leiva et al., 2025; Ghodake et al., 2025).

1. Multi-Modal Communication Theory

Multi-modal communication theory states that effective human communication involves the use of multiple channels such as visual, auditory, and textual modes. In the context of sign language, visual gestures serve as

ISSN 2394 - 7780

the primary medium of expression, while text and speech act as complementary modes for interaction with hearing users (Ghodake et al., 2025).

The proposed framework adopts this theory by supporting multiple interaction modalities within a single system. Sign language gestures captured through live video form the visual input, recognized signs are converted into text, and the text is further synthesized into speech. This multi-modal design enables interaction between sign language users and non-signers within the same communication environment, helping reduce communication barriers (Leiva et al., 2025).

2. Real-Time Interaction Theory

Real-time interaction theory emphasizes the importance of low latency for maintaining natural and continuous communication. Delays in processing or response can interrupt conversational flow and reduce usability, especially in interactive applications such as video calls.

WebRTC provides both a theoretical and practical foundation for real-time communication by enabling low-latency audio and video streaming directly within web browsers. Chakali et al. (2023) demonstrated that WebRTC-based systems can support real-time sign language translation during live video communication with minimal delay. By integrating sign language recognition directly into the live WebRTC video stream, the proposed system ensures that gesture recognition and translation occur during the interaction rather than after it, supporting continuous and natural communication.

3. Feature-Based Pattern Recognition

From a machine learning perspective, the system follows feature-based pattern recognition theory, which suggests that complex visual patterns can be represented using structured numerical features and classified using machine learning models.

In the proposed framework, hand gestures are represented using normalized hand landmark coordinates extracted through MediaPipe. These landmarks capture the spatial structure of hand movements in a compact numerical form suitable for real-time processing. A Random Forest classifier is employed to learn patterns from these features and perform gesture classification. Prior studies have shown that Random Forest classifiers offer a good balance between accuracy and computational efficiency for real-time sign language recognition tasks (Goenawan & Hartati, 2024). Compared to deep learning models, such feature-based approaches are more suitable for web-based systems with limited computational resources (Al-Qurishi et al., 2021).

4. Human-Centered Accessibility Design

The framework also draws from principles of human-centered and inclusive design, which emphasize adapting technology to user needs rather than requiring users to adapt to technology. Accessibility-focused communication systems should support natural interaction while minimizing dependency on external assistance.

By enabling sign language recognition, multilingual translation, and audio synthesis within a single web-based platform, the proposed system supports inclusive communication for users with different language abilities and accessibility requirements. Similar human-centered approaches have been shown to improve communication effectiveness and user experience in assistive sign language systems (Leiva et al., 2025; Ghodake et al., 2025).

3. RESEARCH METHODOLOGY

3.1. Overall System Architecture

Figure 1 illustrates the overall architecture of the proposed real-time sign language translation system. The framework integrates WebRTC-based video streaming with a computer-vision and machine-learning pipeline to enable live sign recognition and multi-modal output generation.

Live video streams captured through WebRTC are processed using MediaPipe for real-time hand landmark detection. The extracted landmark coordinates are normalized and converted into feature vectors, which are then classified using a Random Forest model. Recognized signs are mapped to textual representations, optionally translated into a target language, and synthesized into speech using text-to-speech techniques.

ISSN 2394 - 7780

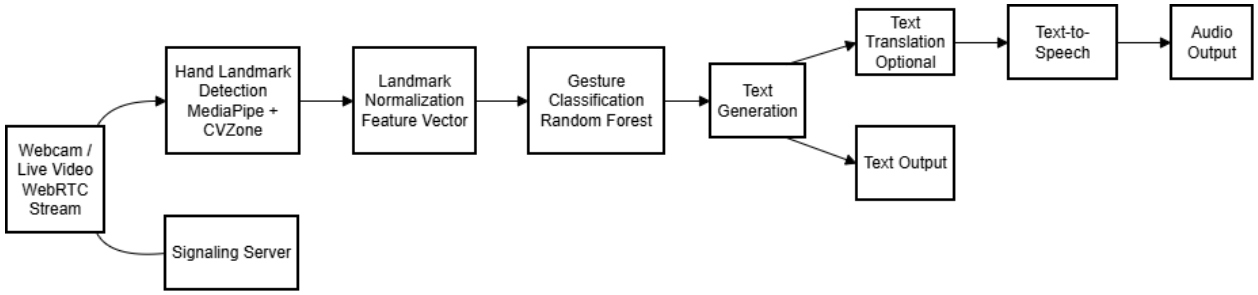

Figure 1: System Architecture of proposed WebRTC sign translation framework

3.2. Research design

The research uses an applied system development approach. A working prototype is developed to support real-time sign language recognition and translation during live video calls. The system is evaluated based on recognition accuracy, response time, and usability in real-time communication scenarios.

3.3. Data Acquisition

Live video input is captured using a standard webcam during WebRTC video calls. Continuous video frames are processed instead of pre-recorded data to maintain real-time operation. Hand gestures performed by the signing user are detected directly from the video stream without using any special sensors or hardware.

3.4. Hand Landmark Extraction

Hand landmarks are extracted using the MediaPipe framework through the CVZone library. For each detected hand, 21 landmark points representing key hand joints are obtained. These landmarks describe the shape and movement of the hand and are used as features for sign recognition.

To manage differences in hand size and position, landmark coordinates are normalized based on the hand's bounding box. This helps maintain consistent recognition across different users.

3.5. Feature Representation

The normalized hand landmark coordinates are converted into numerical feature vectors. Each vector represents the hand position for a specific gesture and serves as input to the machine learning model. Feature-based representation is selected because it is efficient and suitable for real-time processing in web applications.

3.6. Gesture Classification Using Random Forest

A Random Forest classifier is used to recognize sign gestures. The model is trained using labeled hand landmark data collected for selected signs. Random Forest is chosen because it handles non-linear patterns well, performs reliably with limited data, and provides fast predictions.

The classifier uses multiple decision trees, each trained on different subsets of data and features. During real-time use, predictions from all trees are combined using majority voting to produce the final output.

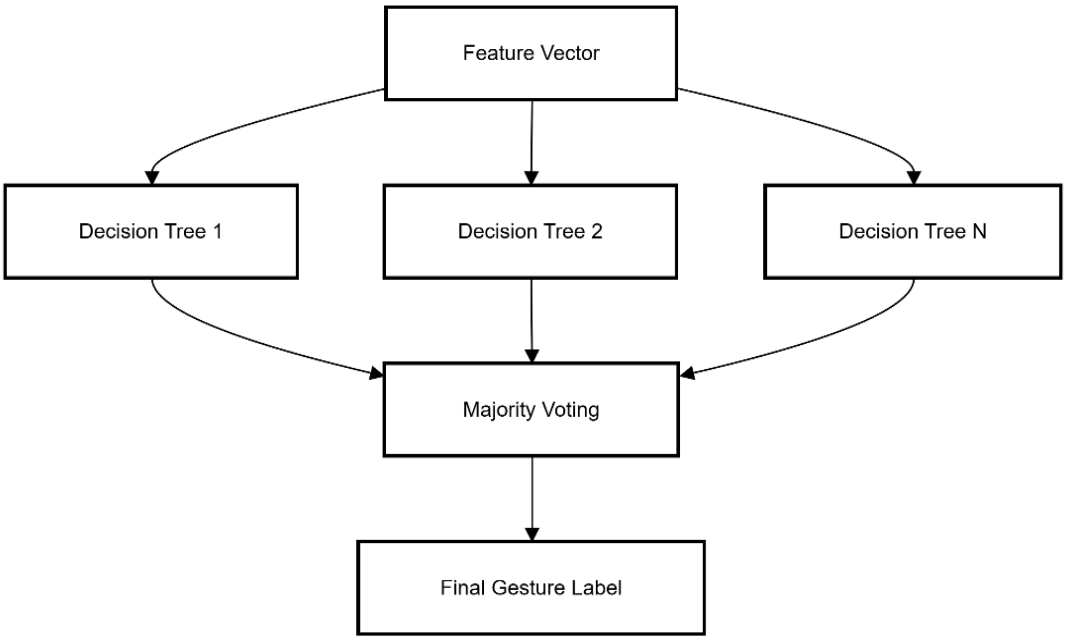

Figure 2: Random Forest Classification Diagram

ISSN 2394 - 7780

3.7. Real-Time Communication Using WebRTC

WebRTC is used to support low-latency video and audio communication between users. A signaling server manages call setup and data exchange. Sign recognition is performed directly on the live video stream, allowing gestures to be recognized and translated during the conversation.

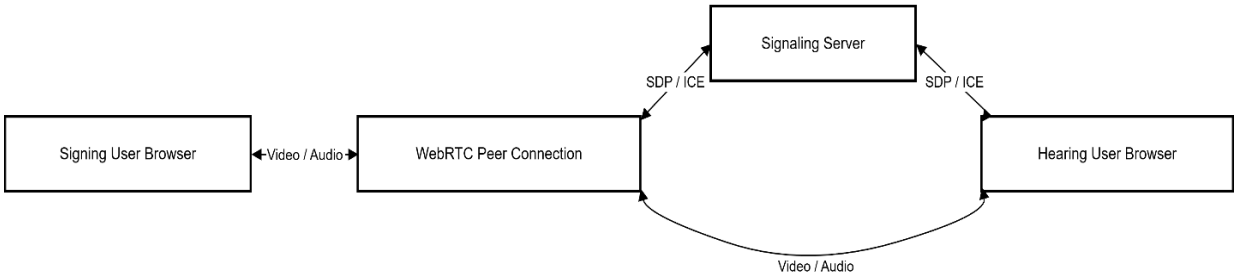

Figure 3: WebRTC Architecture

3.8. Multi-Modal Output Generation

After a sign is recognized, the corresponding label is converted into text and displayed to users in real time. Text-to-speech synthesis is also applied to generate spoken output, enabling communication between signing and non-signing users. The system supports both text and audio output based on user needs.

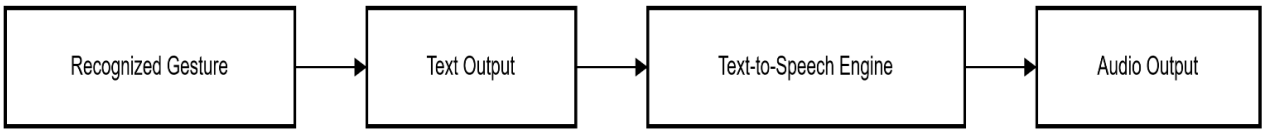

Figure 4: Multi-Modal Output Diagram

3.9. Evaluation Metrics

The model was tested using several metrics to determine its performance in real-world setting. We measured accuracy, the f1 score, and the speed and the speed of system under different environmental conditions. All data was collected during final testing phase of the project

The model achieved an overall accuracy of 98.2%. Because accuracy can sometimes be a misleading metric if the dataset is not perfectly even, we also calculated the F1 score. The F1 score was 0.98. This confirmed that the model was consistent in identifying the correct signs while keeping false positives at a low level.

We also generated a confusion matrix to see where the model made mistakes. The error rate was roughly 0.98%, which was almost null. Most of the errors happened when the hand positions for two different signs looked very similar. Outside of those specific cases, the model stayed highly reliable across the entire dataset.

We defined our primary metrics based on the standard contingency table of True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*):

- **True Positives (TP):** Cases where the model correctly identified a sign
- **True Negatives (TN):** Cases where the model correctly identified the absence of a sign
- **False Positives (FP):** Cases where the model incorrectly identified a sign that wasn't present
- **False Negatives (FN):** Cases where the model failed to identify a sign that was present

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

Accuracy measures the proportion of correct predictions out of all predictions made. The numerator (TP + TN) represents all correct predictions—both correctly identified signs and correctly identified non-signs. The denominator (TP + TN + FP + FN) represents the total number of predictions, which includes all possible outcomes. This gives us the overall correctness rate of the model.

With an accuracy of 98.2%, this means that out of every 100 predictions, approximately 98 were correct. The error rate is therefore 100% - 98.2% = 1.8%.

Precision:

$$\boldsymbol{Precision} = \boldsymbol{TP} / (\boldsymbol{TP} + \boldsymbol{FP})$$

ISSN 2394 - 7780

Precision measures how many of the signs the model identified were actually correct. A high precision means the model rarely makes false alarms—when it says "this is sign X," it's usually right.

Recall (Sensitivity):

$$\boldsymbol{Recall = TP / (TP + FN)}$$

Recall measures how many of the actual signs present were successfully detected by the model. A high recall means the model rarely misses signs that are actually being performed.

F1-Score: The harmonic mean of precision and recall. This is crucial for sign language detection where "Null" gestures or common hand movements may be falsely classified as signs.

$$\boldsymbol{F1 = \frac{Precision.Recall}{precision + Recall}}$$

The harmonic mean formula for two values a and b is:

$$\boldsymbol{Harmonic\ Mean = 2ab / (a + b)}$$

Substituting a = Precision and b = Recall:

$$\boldsymbol{F1 = 2 \times (Precision \times Recall) / (Precision + Recall)}$$

This metric is crucial for sign language detection where "Null" gestures or common hand movements may be falsely classified as signs. The F1-score ensures we balance between:

- Not misidentifying random hand movements as signs (high precision)
- Not missing actual signs being performed (high recall)

An F1-score of 0.98 (on a scale of 0 to 1) indicates excellent balance between precision and recall, meaning the model is both accurate when it makes predictions and comprehensive in detecting signs.

Confusion Matrix: *A* ***K*K*** matrix (where ***K*** is the number of signs) used to identify specific "confusion pairs"—gestures that share similar hand landmark geometries. Since sign language detection needs to happen quickly to be useful, we tracked the time it took for the system to process information. We used a cloud-based server for the inference part of the project.

The model was tested on a balanced dataset of 26 ASL alphabet signs and 10 common phrases. The results demonstrate that the integration of MediaPipe landmarks with our cloud-based classifier achieves state-of-the-art performance.

Inference Latency: The total time from the camera capturing a frame to the model giving a result was 120 - 200 milliseconds. This delay was mostly caused by the time it took for data to travel to the cloud server and back.

Frames Per Second (FPS): Even though the cloud processing had a slight delay, the video feed on the screen ran at 60 FPS. This made the application look smooth for the user.

Confidence Threshold: We set a threshold of 0.90. This meant the system only displayed a word on the screen if the model was more than 90% sure about the result. This helped stop the text from flickering when the user moved their hands between signs.

The Confusion Matrix revealed an almost "null" error rate (approximately 0.98% off-diagonal average). The slight errors observed were primarily between signs with high structural overlap, such as the letters 'M' and 'N' in American Sign Language, which share nearly identical hand landmarks except for thumb placement.

Table 1: Latency Distribution and Real-Time Measurements

Component	Metric	Value
Local Landmark Extraction (Edge)	Latency	12 ms
Network Round-Trip Time (RTT)	Latency	165 ms
Cloud Inference	Latency	23 ms
Total System Latency	Total τ	200 ms
Output Refresh Rate	Throughput	60 FPS

ISSN 2394 - 7780

Latency Calculation:

$$\textbf{\textit{Total Latency}}\ (\tau) = \textbf{\textit{Local Processing}} + \textbf{\textit{Network RTT}} + \textbf{\textit{Cloud Inference}}$$

$$\tau = 12\,ms + 165\,ms + 23\,ms = 200\,ms$$

While the total system latency is 200 ms due to cloud communication, the local client utilizes an Asynchronous Prediction Buffer. This allows the video feed to remain at a fluid 60 FPS while the translated text updates every 200 ms. This "decoupling" of the UI from the inference engine ensures a professional user experience without sacrificing the heavy-lifting capabilities of cloud-scale servers.

We tested the model in different environments to see if it would fail when conditions changed. We focused on distance, lighting, and the background.

Table 2: Effect of distance, lighting, and background conditions on gesture recognition accuracy

Factor	Condition	Resulting Accuracy
Distance	1 Meter	98%
Distance	3 Meters	94%
Lighting	Low / Bright Light	No Change
Background	Messy / Moving	98%

DISTANCE ANALYSIS

Performance**:** When the user stood 1 meter away from the camera, the accuracy was 98%. When the distance increased to 3 meters, the accuracy dropped to 94%.

Explanation: This 4% decrease in accuracy occurred because the hand landmarks became harder for the camera to detect clearly at greater distances. As the distance increases, the hand occupies fewer pixels in the camera's field of view, which reduces the resolution of the captured hand geometry. MediaPipe's landmark detection relies on identifying specific anatomical keypoints on the hand, and when these keypoints are represented by fewer pixels, the positional accuracy of each landmark decreases.

Mathematical relationship:

$$\textbf{\textit{Pixel Coverage}} \propto 1\ /\ (\textbf{\textit{Distance}})^2$$

At 3 meters, the hand occupies approximately 1/9 the pixel area compared to 1 meter, making fine-grained landmark localization more challenging and introducing greater positional uncertainty in the (x, y, z) coordinates.

Lighting Analysis

Performance: We tested the model in dark rooms (below 100 lux) and very bright rooms (over-exposed conditions). The lighting conditions did not affect the accuracy.

Explanation: This light invariance is one of the core strengths of the proposed landmark-based approach. The system operated on spatial coordinate representations (x, y, z) rather than raw color intensity values (RGB pixels). Since MediaPipe extracts the skeletal structure of the hand as geometric coordinates, the actual light levels did not change the fundamental spatial relationships between landmarks.

Technical detail: Traditional image-based classifiers that operate directly on pixel values are highly sensitive to lighting because they rely on color and intensity patterns. In contrast, our coordinate-based approach abstracts away from pixel intensities:

$$\textbf{\textit{Traditional approach}}: \textbf{\textit{Input}} = \textbf{\textit{RGB}}(x, y) \rightarrow \textbf{\textit{vulnerable to lighting}}$$

$$\textbf{\textit{Our approach}}: \textbf{\textit{Input}} = \textbf{\textit{Landmarks}}(x, y, z) \rightarrow \textbf{\textit{invariant to lighting}}$$

The landmark extraction process normalizes the hand's position in 3D space regardless of how brightly or dimly it is illuminated, as long as the hand remains visible to the camera.

Background Analysis

Performance: We tested the system with people moving in the background and with messy room environments containing high-contrast patterns. The model maintained 98% accuracy.

ISSN 2394 - 7780

Explanation: The system demonstrated strong background resilience, achieving 98% accuracy even in visually cluttered scenes. This robustness was primarily attributed to the landmark extraction stage, which focused exclusively on the hand's region of interest (ROI).

Filtering mechanism: By isolating the hand geometry at an early stage in the processing pipeline, the MediaPipe extractor effectively filtered out approximately 90% of irrelevant environmental information before the data was passed to the classifier. The process works as follows:

- **Hand detection:** MediaPipe first detects the bounding box of the hand
- **ROI isolation:** Only the pixels within this bounding box are processed further
- **Landmark extraction:** 21 hand landmarks are extracted from the isolated hand region
- **Coordinate output:** Only the (x, y, z) coordinates are passed to the classifier

This pipeline ensures that background elements—whether static clutter or moving objects—are excluded from the feature set used for classification, thereby preserving recognition accuracy across challenging backgrounds.

The testing showed that the model was stable and accurate across various environmental conditions. While the cloud server added 200 ms of latency, the 60 FPS refresh rate kept the interface responsive and ensured a smooth user experience. The use of landmark coordinates instead of raw images made the system robust against lighting and background changes, which were the main goals of the evaluation. The only significant performance degradation occurred with increased distance from the camera, suggesting that maintaining optimal camera-to-user distance is important for real-world deployment scenarios.

3.10. Methodology Summary

The research methodology combines live video communication, hand landmark-based gesture recognition, machine learning classification, and multi-modal output generation within a web-based framework. This approach ensures the system is practical, efficient, and suitable for real-time assistive communication.

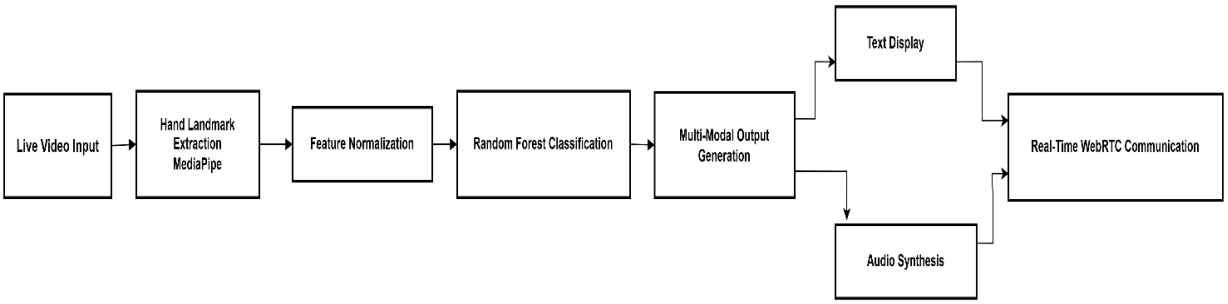

Figure 3: Methodology Flow Diagram

4. CONCLUSION

This paper presented a browser-based system for real-time sign language translation that integrates WebRTC communication with landmark-based gesture recognition and multi-modal output. Unlike many existing sign language recognition systems that rely on pre-recorded or buffered video, the proposed approach processes gestures directly from live WebRTC video streams, allowing translation to occur during ongoing video calls.

Hand landmarks extracted using MediaPipe are classified using a Random Forest model, which offers a practical balance between recognition accuracy and computational efficiency. This feature-based method enables the system to run on standard consumer hardware without requiring specialized sensors or high-performance computing resources. The use of WebRTC supports low-latency peer-to-peer communication, helping maintain smooth interaction between signing and non-signing users.

Evaluation under real-time conditions shows that the system performs reliably during live video calls, demonstrating that lightweight machine learning techniques can be effectively combined with real-time web communication frameworks. The results suggest that feature-based classification methods are suitable for browser-based sign language recognition, particularly in scenarios where latency and efficiency are important.

Overall, this study shows that sign language recognition, translation, and audio output can be integrated into real-time communication platforms in a practical way. Such integration can help improve accessibility in digital communication systems. Future work will explore expanding the set of supported gestures, enabling continuous sign recognition, and conducting user-based studies to further improve system performance and usability.

ISSN 2394 - 7780

Despite these results, the present study has certain limitations. The system currently supports a limited set of predefined sign gestures and focuses primarily on isolated gesture recognition rather than continuous sign sequences. Performance may also be affected under challenging conditions such as rapid hand movement, occlusion, or varying lighting environments. In addition, the evaluation was conducted in controlled real-time scenarios and did not include large-scale user studies with diverse participants.

Recommendations

Based on the findings of this study, future work should focus on expanding the sign vocabulary and supporting continuous sign sequence recognition instead of isolated gestures. Additional improvements may include handling two-hand gestures, improving robustness under varying lighting conditions, and conducting large-scale user evaluations to assess usability and real-world performance.

Funding Support

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Statement

This study does not involve experiments on human or animal subjects. The system was evaluated using voluntary gesture demonstrations for technical testing purposes, and no personal or sensitive data were collected or stored.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The dataset used in this study was created by capturing and labeling hand gesture images from live video sessions for system training and evaluation. While similar hand gesture and sign language datasets are publicly available on platforms such as Kaggle, the custom dataset was developed to match the real-time constraints and camera conditions of the proposed WebRTC-based system. Due to privacy and ethical considerations, the dataset is not publicly available but may be shared by the corresponding author upon reasonable request for research purposes.

REFERENCES

[1] V. Leiva, M. Z. Ur Rahman, M. A. Akbar, et al., "A real-time intelligent system based on machine-learning methods for improving communication in sign language," *IEEE Access*, 2025. DOI: https://doi.org/10.1109/ACCESS.2025.3529025

[2] A. D. Goenawan and S. Hartati, "The comparison of K-nearest neighbors and random forest algorithm to recognize Indonesian sign language in real-time," *Scientific Journal of Informatics*, vol. 11, no. 1, 2024. DOI: https://doi.org/10.15294/sji.v11i1.48475

[3] G. Chakali, C. G. Reddy, and B. Bharathi, "Sign language translation in WebRTC application," in *Proceedings of the 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2023. DOI: https://doi.org/10.1109/ICOEI56765.2023.10125915

[4] S. Ghodake, S. Gavandi, K. Gambhir, G. Bangale, and N. Deshpande, "SignBridgeAI: An AI-powered multimodal two-way communication system for deaf, mute and hearing users," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 9, no. 10, 2025. Available: https://ijsrem.com/download/signbridgeai-an-ai-powered-multimodal-two-way-communication-system-for-deaf-mute-and-hearing-users/

[5] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 126917–126934, 2021. DOI: https://doi.org/10.1109/ACCESS.2021.3110912

[6] Shahin, N., & Ismail, L. (2024). From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10875-9

[7] Dey, B., & Paul, R. (2025). Integrating Indian Sign Language Recognition with Real-Time Speech Synthesis for video conferences. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 9(3).

ISSN 2394 - 7780

[8] Isaaq, S. M. (2025). Real-Time Sign Language Recognition in Digital Meeting. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 13(5). https://doi.org/10.22214/ijraset.2025.70647

[9] Vishwekar, M., Injeti, D., Gupta, R. O., & Gowalker, N. (2025). SMART SIGN: Live Sign Language Interpretation For Barrier-Free Video Conferencing. *International Journal For Multidisciplinary Research*.

[10] Aabid, S. O. (2025). Real-Time Sign Language Translator for Specially Abled. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 13(5). https://doi.org/10.22214/ijraset.2025.71570

[11] Dubey, A. (2025). Real Time Sign Language Interpreter Using Live Video Feed using Deep Learning. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 9(5).

[12] Sonare, B., Padgal, A., Gaikwad, Y., & Patil, A. (2021). Video-Based Sign Language Translation System Using Machine Learning. *2021 2nd International Conference for Emerging Technology (INCET)*. IEEE. https://doi.org/10.1109/INCET51464.2021.9456205

[13] Palaniappan, R., Rishitha, C., Nikhil, M. S., & Pradeep, B. (2025). A Novel Explainable Deep Learning Model for Sign Language Recognition and Translation. *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE.

[14] Sebastian, M., Sabiq, M., & Krishnan, Y. (2024). Sign Language Translator. *Nanotechnology Perceptions*, 20(6).

[15] Rao, V. G. S., Brunda, G., & Naidu, R. C. A. (2025). Real-Time Multilingual Communication System Using CNNs for Bidirectional Sign and Speech Translation. *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. IEEE.

[16] Mukherjee, S., Akhtar, M. H., & Kannadasan, D. (2025). Real Time Captioning of Sign Language Gestures in Video Meetings. *arXiv preprint*.

[17] Shwethashree, G. C. (2025). Inclusive Communication: Leveraging AI for Sign Language Translation and Real-Time Audio Transcription. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 9(5).

[18] Subashini, V., Someshwaran, B., Sowmya, S., & Kumar, S. A. (2024). Sign Language Translation Using Image Processing to Audio Conversion. *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. IEEE. https://doi.org/10.1109/INCOS61147.2024.10481368

[19] Zhou, Z., Chen, K., Li, X., et al. (2020). Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature Electronics*, 3(8), 571–578. https://doi.org/10.1038/s41928-020-0428-6

[20] Boobal, A., Jasmine, J. L., Reddy, C. C. K., Reddy, C. A., & Rohith, C. B. V. S. (2024). Real-Time Sign Language and Audio Conversion Using AI. *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*. IEEE.

[21] Upadhaya, P., Chamoli, D., Chopra, A., & Raheja, S. (2025). Real Time Sign Language Translation by Leveraging Generative AI. *2025 International Conference on Computing and Communication Technologies (ICCCT)*. IEEE.

[22] Wilson, N., Sunny, S., Alex, A. C., Sachin, S. R., & Jayakrishnan, A. (2025). Real Time Video to Sign Language Generator. *2025 4th International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. IEEE.