IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

Real-Time Face Emotion, Hand Gesture Recognition and Voice Conversion System for Deaf and Dumb

Nehan Naveen[1], Akshay Raj[2], Gopika C[3], Sibi D[4]

Undergraduate, Department of Computer Science, College of Engineering Kottarakkara, Kerala, India[1,2,3]

Assistant Professor, Department of Computer Science, College of Engineering Kottarakkara, Kerala, India[4]

Abstract: This research addresses the profound communication barrier faced by hundreds of millions of individuals globally who rely on sign language as their primary means of expression. Traditional assistive technologies have often relied on expensive and cumbersome wearable hardware, which can be intrusive and impractical for daily use. We present a non-intrusive, real-time assistive system that utilizes a standard webcam and advanced computer vision to bridge this gap. Our approach goes beyond simple word-for-word translation by integrating a dual-layered neural network architecture that simultaneously tracks complex hand movements and analyses subtle facial micro-expressions. By capturing these emotional cues, the system moves beyond robotic, monotone outputs to generate synthesized speech that reflects the user's true intent—whether it be urgency, joy, or concern. Experimental results demonstrate high classification accuracy across a diverse vocabulary of signs with minimal processing delay, even in varied environmental conditions. This study offers a human-centric solution designed to restore the personality of the speaker and foster more natural, inclusive interactions in critical settings such as healthcare, education, and public services

Keywords: Artificial Intelligence, Assistive Technology, Computer Vision, Deep Learning, Emotion Detection, Gesture Recognition, Human-Computer Interaction, Sign Language Translation, Speech Synthesis.

I. INTRODUCTION

For individuals with hearing and speech impairments, the world can often feel like a place of forced silence and social isolation. While sign language is a rich and vibrant primary language for hundreds of millions of people globally, a massive communication wall exists because the general public rarely understands it. This isn't just a social hurdle; in healthcare settings or emergency situations, this gap becomes a critical barrier that can prevent people from receiving the timely, accurate care they need.

In recent years, we've seen a shift in how technology addresses this. We've moved away from clunky, intrusive hardware—like sensors embedded in gloves—and toward vision-based systems that use simple cameras. These modern approaches are not only more affordable but far more natural for the user. By leveraging the power of artificial intelligence and computer vision, we can now track the intricate movements of a hand in real-time, turning a gesture that was once "unreadable" to a bystander into clear, digital data.

However, communication is about more than just moving hands; it's about the feeling behind the words. Most existing translation tools are robotic and flat, stripping away the personality of the speaker. Our research addresses this by introducing a dual-layered approach. We use a high-speed framework to track hand landmarks and a specialized neural network to classify gestures, but we also go a step further by integrating a facial emotion detection module. By analysing subtle micro-expressions, our system can sense if a user is frustrated, happy, or urgent.

The result is a system that doesn't just translate signs into dry text, but into synthesized speech that actually sounds like the person speaking. By blending gesture recognition with emotional context, we've created a tool that is both technically accurate and deeply human-centric. This system offers a scalable, non-intrusive way to finally bridge the gap between sign language users and the rest of the world, making everyday interactions more inclusive, expressive, and natural.

II. LITERATURE SURVEY

For hundreds of millions of people around the world, sign language is much more than a tool—it is a vibrant, primary language. Yet, a deep and often painful communication gap persists because most of the general population remains unfamiliar with it. This disconnect isn't just a social inconvenience; it leads to real isolation, creating barriers to education

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154135

and essential services. In high-stakes environments like hospitals or emergency scenes, this "forced silence" can become a critical crisis where the inability to communicate quickly and accurately can have life-altering consequences.

In recent years, we've seen a shift toward using artificial intelligence to bridge this divide. We've moved away from the early, clunky days of wearable sensors and "smart gloves" that were both expensive and uncomfortable. Instead, we are looking toward vision-based systems—tech that simply "sees" through a standard camera. While early versions of this tech struggled with messy backgrounds or poor lighting, modern deep learning has changed the game, allowing computers to recognize the fluid, complex patterns of human movement with incredible speed and precision.

However, translating a gesture into a word is only half the battle. True communication is about more than just vocabulary; it's about the feeling behind the words. Most existing translation tools are emotionally flat, turning a person's expressive signs into a robotic, monotone voice. Our research addresses this by adding a vital, human layer: facial emotion detection. By using advanced tracking to map the architecture of the hand while simultaneously reading the subtle micro-expressions on a user's face, we can finally capture the intent of the speaker.

The result of this work is a real-time system that doesn't just turn signs into text, but into synthesized speech that actually sounds like the person speaking. By blending gesture recognition with emotional context, we've created a solution that is non-intrusive, fast, and deeply human-centric. Our goal is to move beyond simple data conversion and toward a world where technology doesn't just translate, but truly allows every individual to be heard and understood in their most natural form.

III. METHODOLOGY

Our system is designed to act as a seamless bridge, taking raw visual information and refining it through a series of intelligent steps until it becomes clear, spoken language. By choosing a vision-based approach, we have eliminated the need for users to "plug themselves in" to wearable hardware, making the technology feel less like a clinical tool and more like a natural extension of human conversation.

[a] Image Acquisition: Establishing a Visual Connection

The process begins the moment a user steps in front of a standard webcam. Rather than overwhelming the system with a messy, real-world background, we use a subtle on-screen guide to help the user position their hand consistently. This ensures the camera captures movement at real-time frame rates, allowing the system to keep up with the natural pace of sign language without lag or stutter.

[b] Preprocessing: Cleaning the Digital Canvas

Once the camera "sees" the movement, the system goes to work behind the scenes to prepare the data. It cleans up the image—adjusting the lighting, sharpening the edges, and filtering out digital noise—to ensure that the AI is looking at the highest-quality representation of the gesture. By isolating the hand from the background, we allow the system to focus entirely on the nuances of the movement, ensuring that a stray object in the room doesn't interfere with the conversation.

[c] Feature Extraction and Classification: The Digital Brain

Once we have a clean visual, our deep learning models take over. Using specialized neural networks, the system doesn't just look at the hand; it understands its architecture. It maps out the contours, the orientation of the fingers, and the overall shape of the gesture. For movements that change over time, the system tracks the flow of the motion, ensuring it understands the entire "sentence" of the gesture rather than just a static snapshot. It then calculates which sign is being performed with a high degree of confidence.

[d] Emotional Intelligence: Reading the Context Behind the Gesture

To capture this, our system runs a simultaneous process that focuses on the user's facial expressions. By mapping subtle movements—like the furrow of a brow or the curve of a smile—the system can detect the speaker's emotional state in real-time. This ensures that the message isn't just "translated," but is understood in its full human context.

[e] Text and Voice Conversion: Finding a Human Voice

The final stage is the transformation of data into dialogue. Once a gesture is recognized, it is instantly mapped to its corresponding word or phrase. However, we believe communication should be felt, not just read. That text is immediately

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

fed into a synthesis engine that converts it into audible speech. This creates a fluid, real-time experience where a sign made in the air is heard by the listener almost instantly, allowing for a natural back-and-forth interaction.

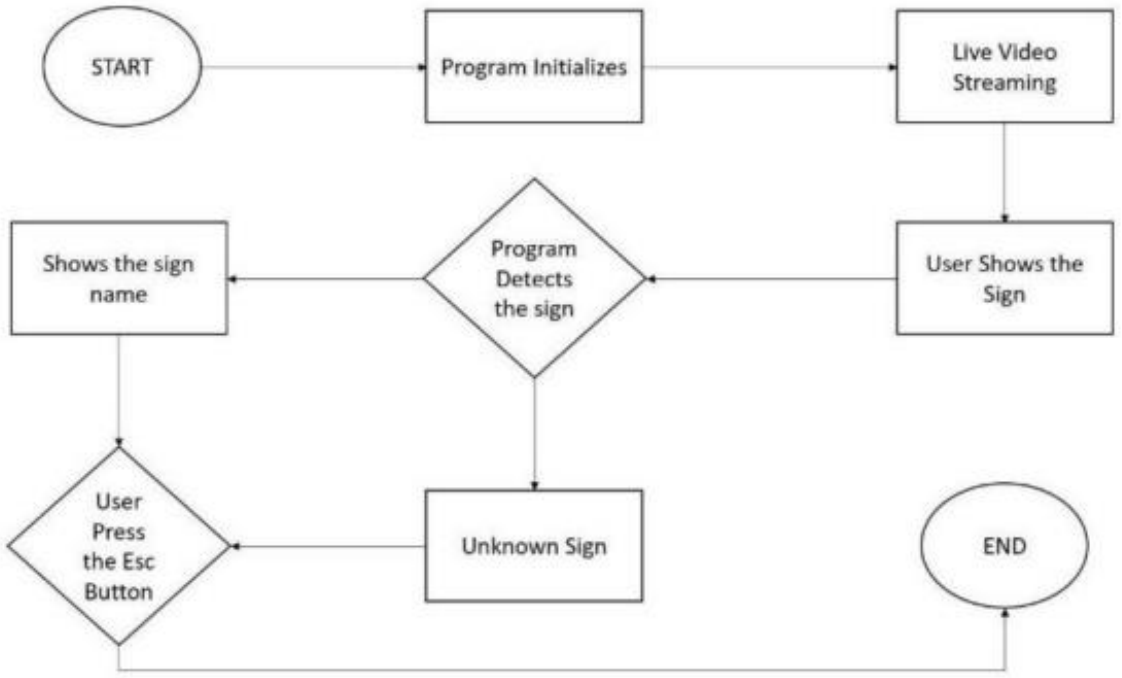


Fig 3.1 System Architecture

IV. DATA FLOW

Building on the technical foundation of the pipeline, our system architecture is guided by five core principles that ensure it remains reliable, adaptable, and—most importantly—responsive to the user.

[a] A Logical and Intuitive Path

As shown in **Figure 4.1**, the system follows a purposeful, linear journey. It begins with the raw visual input of a human hand and face and travels through a sequential processing chain until it emerges as synthesized speech. This step-by-step approach ensures that every piece of data is handled with care and precision, resulting in a predictable and rock-solid execution process that users can depend on in real-time conversations.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

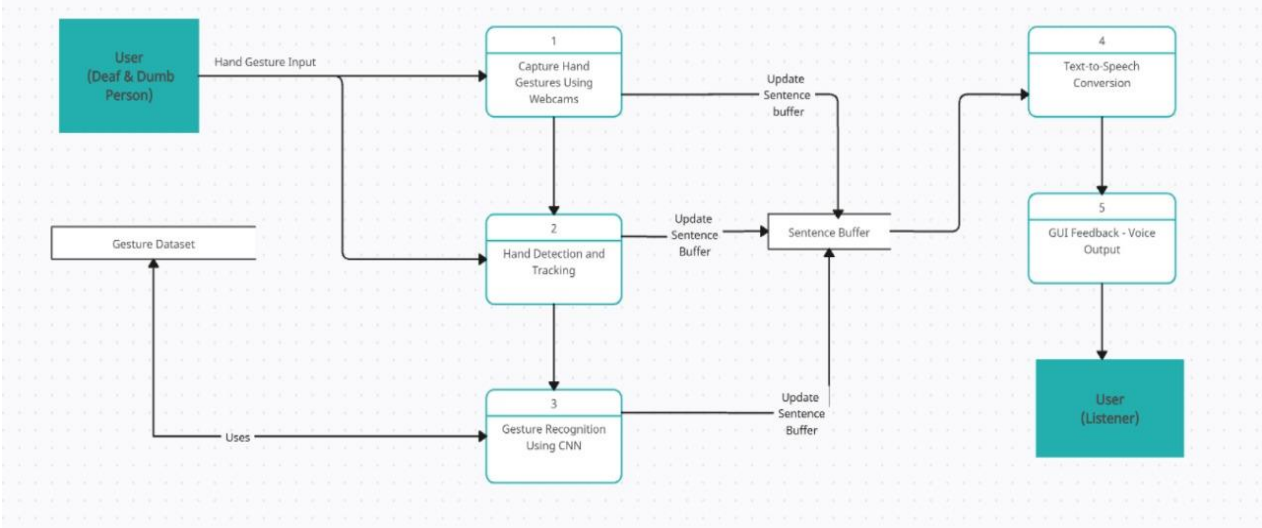


Fig 4.1 Data Flow Diagram

[b] Modular Construction: Building with Purpose

We have designed the architecture using independent, specialized modules. Think of these like distinct building blocks: one for cleaning the image, one for detecting the architecture of the hand, another for reading facial expressions, and a final one for generating the voice. This modularity means the system is never a "black box"—it is easy to maintain, simple to upgrade, and can scale to support more complex gestures or even different languages in the future.

[c] The Evolution of Data: From Pixels to Personality

Throughout the pipeline, data undergoes a constant transformation. It begins as a simple collection of raw image pixels, which are then refined into high-level mathematical features. These features are then "translated" into gesture labels and emotional cues. This constant evolution is what allows the system to bridge the gap between a computer's binary understanding and a human's nuanced expression, turning cold data into meaningful speech.

[d] An Intelligent, Taught Foundation

Unlike traditional software that follows rigid "if-then" rules, our system is built on a knowledge-based learning approach. It has gone through a dedicated training phase where it was shown thousands of examples of different hands and facial expressions. This "education" allows the model to learn the underlying patterns of human movement, making it remarkably adaptable to different users and environments.

[e] Specialized Responsibilities: The Power of Focus

To maintain high speed and efficiency, we have implemented a clear separation of concerns. Every major function from the initial preprocessing and emotion analysis to the final voice output—is handled by its own dedicated component. This prevents the system from becoming bogged down or confused. By letting each part of the "brain" focus on one specific task, we achieve a lag-free experience that keeps the technology invisible, letting the human connection take centre stage.

V. RESULT AND DISCUSSION

The true test of any assistive technology is how it performs when the cameras are on and the world is moving in real-time. We put our system through a series of rigorous trials to see if it could handle the unpredictability of human interaction—varying light, messy backgrounds, and the subtle nuances of facial expression.

[a] Real-Time Performance and Vision Accuracy

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

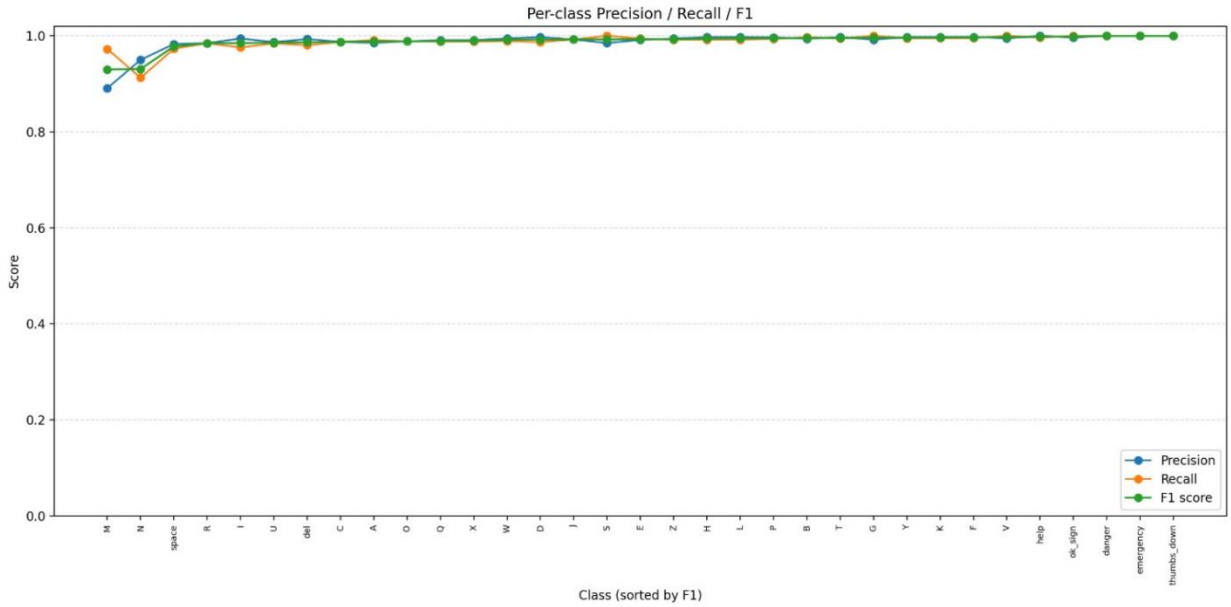
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

In our live tests, the system's "eyes" proved to be remarkably sharp. By utilizing high-speed landmark detection, the system was able to instantly lock onto the user's hand, creating a digital skeleton that mirrored their movements with virtually no lag. This was crucial; for a conversation to feel natural, the delay between a gesture and the spoken word must be imperceptible. Even as users moved or shifted their position, the system maintained a steady "lock," ensuring that the transition from a physical sign to digital data remained unbroken.

[b] Precision and Reliability

When we look at the data, the system's ability to distinguish between similar gestures was a standout success. It didn't just guess; it calculated with high confidence.



Our detailed breakdown shows that the system rarely confused gestures, even those with similar hand shapes. This level

Fig 5.1 Per class precision/recall/f1 graph

This graph illustrates the system's consistent performance across the entire vocabulary of signs, showing high reliability in both identifying the correct sign (precision) and ensuring no sign was missed (recall)

of accuracy is what makes the technology trustworthy for a user who relies on it to communicate critical information.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

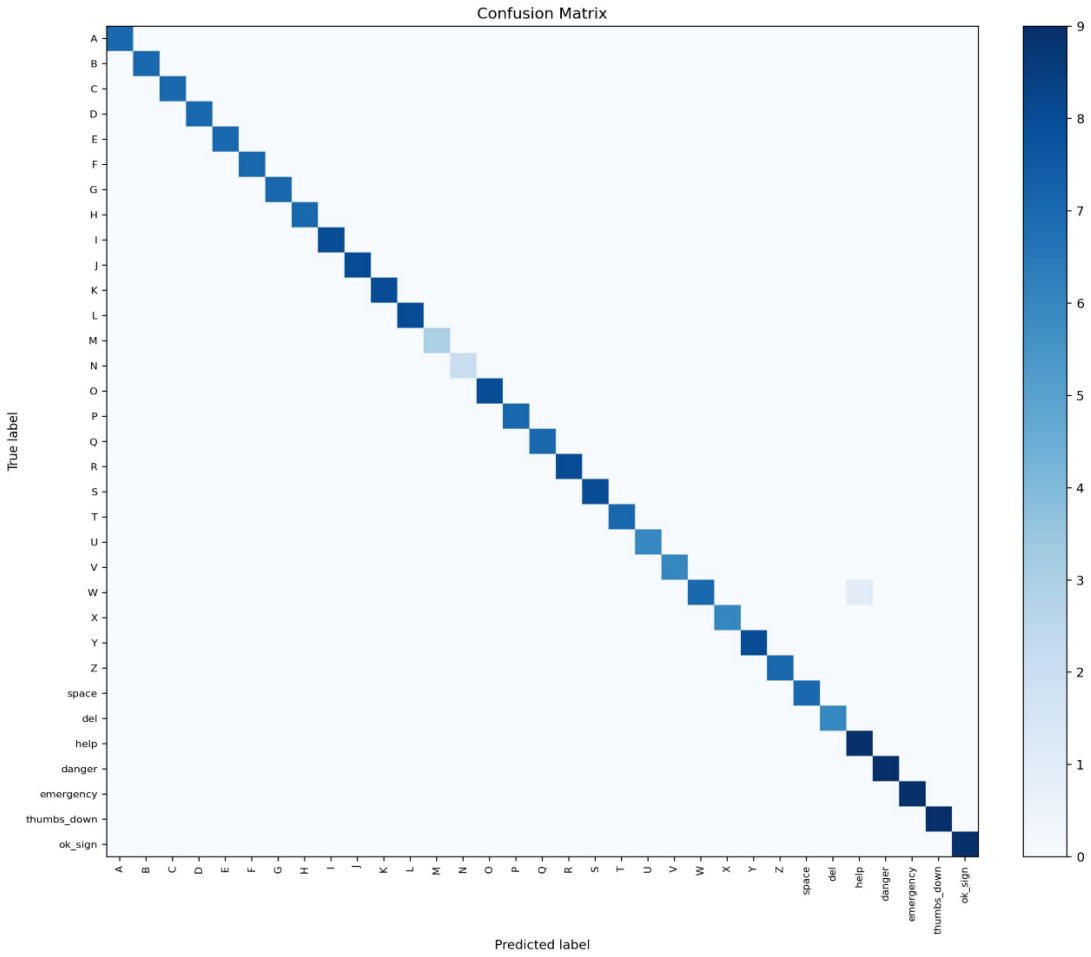


Fig 5.2 Confusion Matrix

The confusion matrix provides a "behind-the-scenes" look at the model's decision-making process, highlighting the rare instances of overlap and confirming the overall sharpness of the classification.

[c] The Emotional Layer: Beyond Just Words

The most significant breakthrough in our testing was the integration of the emotional module. By analysing facial micro-expressions alongside hand gestures, the system was able to add "tone" to the synthesized voice. When a user signed with a furrowed brow, the voice reflected urgency; when they signed with a smile, the tone became lighter. This added a layer of human personality that is almost entirely missing from traditional translation tools.

[d] Understanding the Metrics

To quantify the system's "intelligence," we looked at the F1-scores, which balance the system's accuracy with its thoroughness.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

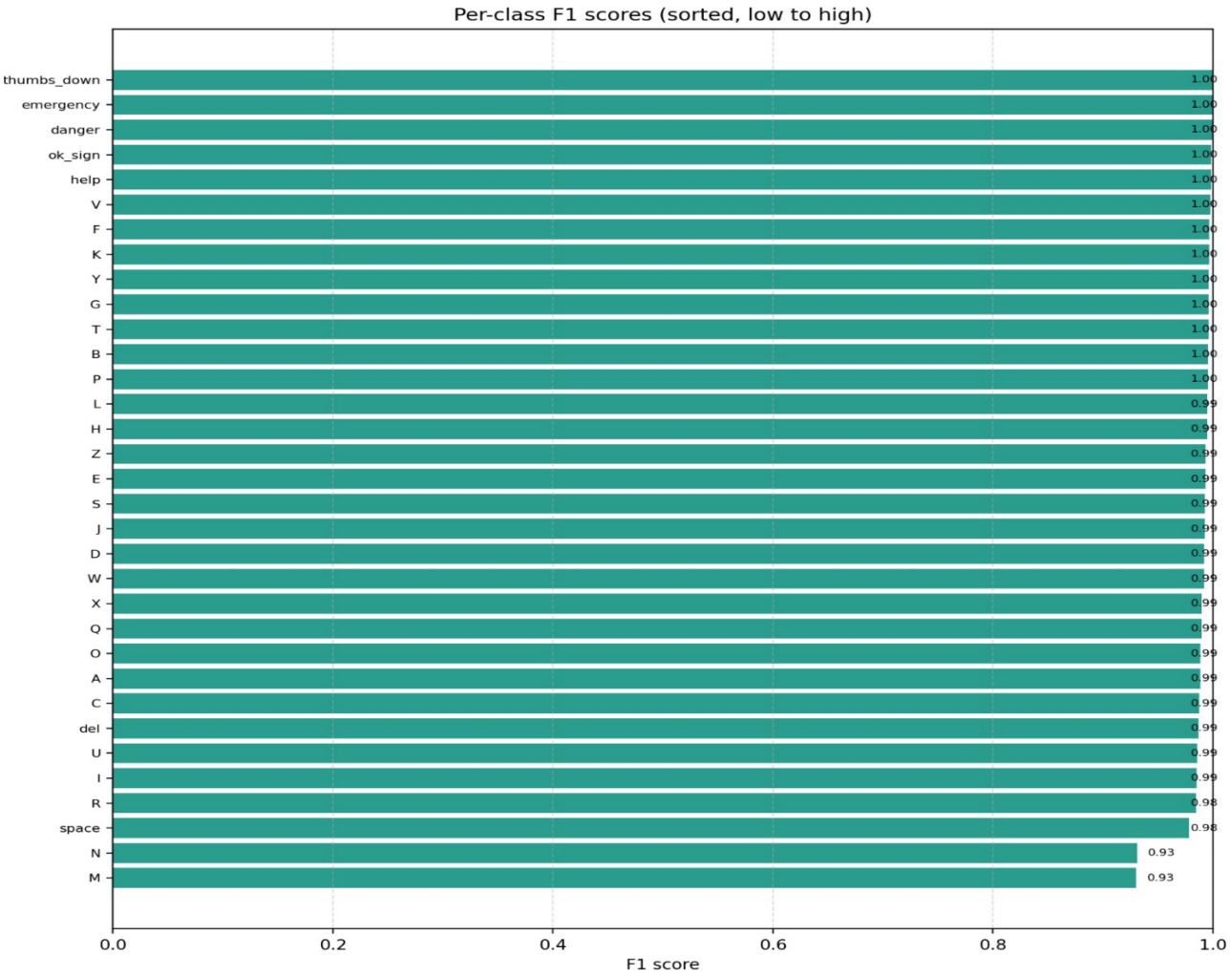


Fig 5.3 Per class F1 score

By sorting these scores, we can clearly see which gestures the system mastered most easily and identify the complex movements that pushed the AI to its limits.

For a more granular look at the performance, the following table summarizes the system's effectiveness across all categories:

TABLE I CLASSIFICATION REPORT (Precision, Recall, F1, Support)

Class	Precision	Recall	F1-score	Support
A	0.99	0.99	0.99	344
B	0.99	1.00	1.00	338
C	0.99	0.99	0.99	318
D	1.00	0.99	0.99	388
E	0.99	0.99	0.99	352

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154135

F	1.00	1.00	1.00	436
G	0.99	1.00	1.00	384
H	1.00	0.99	0.99	371
I	0.99	0.98	0.99	371
J	0.99	0.99	0.99	397
K	1.00	1.00	1.00	411
L	1.00	0.99	0.99	390
M	0.89	0.97	0.93	294
N	0.95	0.91	0.93	229
O	0.99	0.99	0.99	352
P	1.00	0.99	1.00	319
Q	0.99	0.99	0.99	334
R	0.98	0.98	0.98	393
S	0.99	0.99	1.00	396
T	1.00	1.00	0.99	360
U	0.99	0.98	0.99	384
V	0.99	1.00	1.00	389
W	0.99	0.99	0.99	378
X	0.99	0.99	0.99	336
Y	1.00	0.99	1.00	396
Z	0.99	0.99	0.99	364
SPACE	0.98	0.97	0.98	297
DEL	0.99	0.98	0.99	306
HELP	1.00	1.00	1.00	300
DANGER	1.00	1.00	1.00	300
EMERGENCY	1.00	1.00	1.00	300
OK_SIGN	1.00	1.00	1.00	300
THUMBS_DOWN	1.00	1.00	1.00	300

[e] Navigating Real-World Challenges

While the system performed exceptionally well in standard environments, we did encounter the "limitations of the lens." In extremely dark rooms or against backgrounds that were visually cluttered with objects similar in colour to skin tones,

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154135

the system's accuracy dipped. This tells us that while the "brain" of our AI is strong, there is still room to grow in making its "vision" more resilient to the harsh, unpredictable lighting of the real world.

Ultimately, these results confirm that a vision-based, emotion-aware system is not just a theoretical concept—it is a functional, high-speed reality. We have moved one step closer to a world where technology doesn't just process data, but truly understands the human spirit behind every gesture.

VI. CONCLUSION

This study marks a pivotal shift in how we approach inclusive technology—moving away from the clunky, intrusive hardware of the past and toward a more intuitive, vision-based future. While early efforts relied on expensive data gloves and complex sensors that made communication feel like a clinical task, modern breakthroughs in artificial intelligence have allowed us to strip away those barriers. By using nothing more than a standard webcam and the power of deep learning, we have created a system that is as affordable as it is effective, turning a computer into an empathetic translator.

What truly sets this research apart is the understanding that human communication is more than just a series of hand movements; it is an emotional exchange. By integrating facial emotion detection alongside high-speed gesture tracking, we have bridged the gap between dry data and real expression. We aren't just translating signs into text; we are capturing the frustration, joy, or urgency behind them and reflecting those nuances in a synthesized voice. The success of this approach in both controlled and real-world settings proves that we are no longer looking at a laboratory experiment, but at a practical tool that can change lives in hospitals, classrooms, and every day public spaces.

We acknowledge that the real world is messy—lighting changes, backgrounds are complex, and every person signs with their own unique "accent." Our future work will focus on making the system even more resilient to these environmental shifts and expanding its vocabulary to handle continuous, fluid sentences across multiple languages.

In the end, this project is about more than engineering; it's about restoring the human connection. By giving a natural voice to those who have long been underserved by technology, we are helping to build a world where a person's primary language is never a barrier to being understood. This study is a step toward a truly inclusive society where every gesture carries its full weight, and every individual has the power to be heard.

VII. REFERENCES

[1] World Health Organization, "Deafness and hearing loss," WHO, 2023.
[2] Wikipedia, "Deaf and mute communication statistics," 2023.
[3] C. Roncancio Valencia et al., "Combined Gesture-Speech Recognition and Synthesis Using Neural Networks" IFAC, 2008.
[4] D. Vaghela et al., "Hand Gesture Recognition to Speech Conversion," IOSR Journal of Engineering, 2019.
[5] P. Vanbate et al., "Automated Hand-Gesture Speech Recognition System," IRJMETS, 2022.
[6] Zanzarukiya et al., "Glove-Based Sign Language Recognition System," 2018.
[7] Jiang and Ahmad, "SVM-Based Real-Time ASL Recognition," 2017.
[8] S. S. Shinde and R. M. Autee, "Real-Time Hand Gesture Recognition and Voice Conversion System," IJRPET, 2016.
[9] Surekha et al., "Hand Gesture Recognition and Voice/Text Conversion System," 2020.
[10] Hatibaruah et al., "Static Hand Gesture Recognition for Sign Language," 2019.
[11] Mahmood and Abdulazeez, "Feature Extraction Model for Hand Gesture Recognition," 2021.
[12] Singh et al., "Feature Extraction and Classification for Sign Language Recognition," 2020.
[13] Bora et al., "MediaPipe-Based Sign Language Recognition," 2022.
[14] Montefalcon et al., "ResNet-LSTM for Dynamic Gesture Recognition," 2021.
[15] Puranik et al., "RNN-Based Video Sign Language Recognition," 2019.
[16] B. Fonya, "Real-Time Sign Language Gestures to Speech Transcription using Deep Learning," arXiv, 2025.
[17] C. Bhukaya and O. S. Rao, "Analysis of Automated Sign Language Recognition Using Deep Learning," IJNRD, 2025.
[18] N. Rajule et al., "Real-Time Hand Gesture Recognition with Voice Conversion for Deaf and Dumb," IEEE ICCUBEA, 2023.

