**Educational Administration: Theory and Practice**
2024, 30(6),4399-4407
ISSN: 2148-2403
https://kuey.net/

**Research Article**

Educational Administration Theory and Practice

# Emotion-Aware Indian Sign Language Recognition: A Multimodal Approach With Sign-Expression Correlation Analysis

Hriday Ranka[1], Darshit Sarda[1], Haardhik Kunder[1], Aaditya Rajesh[1], Aniket Kore[1]

[1]Department of Computer Engineering, Dwarkadas J. Sanghvi College Of Engineering, Mumbai, India.
hridayr1234@gmail.com; darshitsarda10@gmail.com; kunderhaardhik@gmail.com;
aadityarajesh04@gmail.com; aniket.kore@djsce.ac.in;

**Citation:** Hriday Ranka et al. (2024), Emotion-Aware Indian Sign Language Recognition: A Multimodal Approach With Sign-Expression Correlation Analysis, Educational Administration: Theory and Practice, 30(6), 4399-4407
Doi: 10.53555/kuey.v30i6.7043

**ARTICLE INFO**

## ABSTRACT

This research introduces a novel approach to developing a real-time sign language recognition system designed for Indian Sign Language (ISL) and the emotional expressions of signers. The system aims to assist individuals with hearing impairments and those unfamiliar with sign language. It employs Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and Natural Language Processing (NLP) techniques to enable efficient and accurate interpretation of both signs and emotions. By combining hand sign recognition with facial expression analysis, the system enhances communication by providing more natural interactions and offering additional context regarding the signer's emotional state. The emotion detection component uses advanced computational models to analyze facial expressions, improving the understanding of emotional nuances during communication. A notable innovation in this research is the development of a custom dataset specifically for Indian Sign Language, rather than relying on existing datasets, ensuring the system is tailored to local needs. Furthermore, the integration of sign language recognition and emotion detection into a single system, along with correlation analysis, distinguishes this project from existing solutions.

To further enhance accessibility, the system incorporates a feature that converts written text into spoken words using Google Text-to-Speech, supporting five different Indian regional languages. This functionality facilitates seamless communication across diverse linguistic groups and settings, such as hospitals, schools, courts, and training institutions.

**Keywords:** LSTM, CNN, Indian Sign Language(ISL), Correlation Analysis

## 1 Introduction

In the realm of computer-assisted communication, there is a growing trend to incorporate machine learning and sign language recognition to enhance inclusivity and accessibility for individuals with hearing impairments. The inclusion of sign language as a standard subject in schools is particularly significant, as it benefits individuals with hearing challenges and promotes wider social understanding. Developing an efficient sign language recognition system is crucial due to the challenges faced by the deaf community in their daily social interactions. Such a system eliminates the reliance on time-consuming methods like tapping for attention, the necessity of learning sign language, and dependence on interpretation services across various fields, including healthcare, education, and training environments.

The proposed research presents a novel system capable of providing real-time Indian Sign Language (ISL) interpretation and emotion recognition within a single digital solution. This innovation addresses the growing demand for effective communication tools by integrating sign language recognition with emotion detection. This dual capability represents a significant breakthrough in assistive communication technologies, enabling users to express both language and emotions simultaneously.

This system has the potential to overcome existing communication barriers in assistive technologies. It aims to transcend traditional paradigms by seamlessly integrating advanced intelligent systems with the intricate expressive capabilities of human language. By bridging the communication gap faced by individuals with hearing impairments, this solution promotes inclusivity and accessibility. Unlike routine assistive systems, this approach provides the hearing-impaired community with a comprehensive tool for interpreting ISL gestures while simultaneously recognizing emotional context.

Furthermore, this research introduces a novel feature: the integration of Google Text-to-Speech [1][2]. This feature transforms written text into spoken words, significantly enhancing the system's communication capabilities. By addressing both linguistic and emotional dimensions, this research contributes to the broader goal of creating a more unified and connected world, offering innovative tools to empower the impaired community.

## 2 Literature Review

In recent years, the advancement of Facial Emotion Recognition (FER) techniques has been driven by the widespread adoption of Convolutional Neural Networks (CNNs) [3]. Achieving an impressive 93% accuracy in real-time, CNN-based FER systems have surpassed previous benchmarks, marking a significant milestone in the field. According to research presented at the SMART–2021 conference [3], FER plays a critical role in enhancing microsocial interactions and supporting advanced technologies. Utilizing datasets such as LFW and the Extended Yale Face Database B, the study advocated incorporating face processing techniques to further improve FER, offering promising directions for future research and development. Similarly, [4] conducted an extensive review of real-time facial emotion detection and classification, leveraging various deep learning models, including CNNs, deep learning (DL), and Haar Cascade [5]. Their proposed classification model achieved 97% accuracy across seven emotion classes using the Fer2013 dataset. By integrating histogram equalization and background subtraction techniques, their research enhanced classification precision, establishing a robust foundation for real-world FER applications. On the other hand, significant progress has also been made in sign language recognition, particularly within the Indian context. The study by [6] proposed a CNN-based approach for Indian Sign Language (ISL) recognition using visual signals. Their system aimed to bridge communication barriers by converting ISL gestures into text or audio through techniques like Canny Edge Filtering and text-to-audio conversion. With an accuracy rate of 95.31%, their research addressed challenges in ISL gesture recognition and underscored the importance of integrating audio resources to facilitate inclusive communication for individuals with hearing impairments. Expanding on this, [7] explored real-time ISL recognition using deep learning techniques. By employing CNNs, MediaPipe, and OpenCV, they developed an efficient system for detecting and interpreting ISL gestures in real time. Their findings emphasized the adaptability of such technologies in reducing communication barriers and fostering greater inclusion for the Deaf community. In addition, [8] focused on advancing American Sign Language (ASL) recognition and real-time conversion into text alongside emotion recognition. By combining CNN, SVM, HAAR, K-means clustering, and LSTM algorithms, their system achieved approximately 80% accuracy. Their research highlighted the significance of dynamic hand gesture analysis and emotion recognition in reducing communication gaps between signers and non-signers, thereby promoting accessibility and inclusivity for hearing-impaired individuals. This literature review underscores the substantial progress in sign language and emotion recognition through machine learning applications. The reviewed studies collectively advocate for incorporating sensory and audio components into sign language recognition systems. Such advancements empower the hearing-impaired community by enabling more effective communication and fostering inclusive interactive media experiences across diverse social contexts.

## 3 Proposed Methodology

Our innovative approach provides an all-inclusive solution to improve communication accessibility for individuals with hearing impairments, with a particular focus on the Indian Deaf community. The system comprises four key components: a hand sign recognition module, an emotion recognition module, a text-to-speech converter supporting five regional Indian languages, and a correlation NLP module.

### 3.1 Palm Detection Model using MediaPipe

The Palm Detection model leverages a pre-trained MediaPipe pipeline to recognize and track hand gestures in a video stream. The process begins with capturing frames from the video feed (30 frames at a time) and analyzing them for hand detection. MediaPipe's hand-tracking model identifies and tracks the keypoints of the palm, fingers, and wrist within each frame. The detected keypoints are organized into a NumPy array, which serves as input for further analysis. Each keypoint position is normalized and converted into floating-point arrays using NumPy for consistency and precision. An LSTM model is trained on these keypoints to learn and interpret complex hand gestures. This step is repeated for every 30 frames in the video stream, enabling real-time feedback on the hand signs performed by individuals.

When no hands are detected, a corresponding message is displayed on the output screen, ensuring uninterrupted monitoring and interpretation of hand gestures. This continuous feedback mechanism enhances the system's reliability and usability for real-world applications.
gestures.

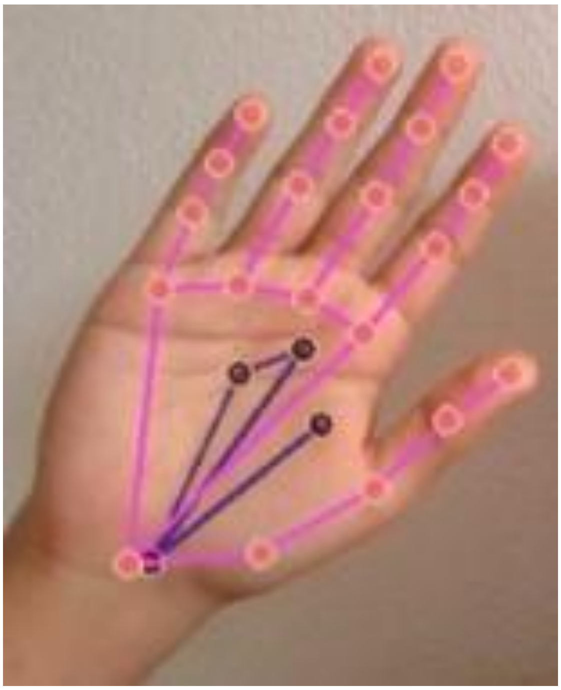


**Fig. 1**Identifying all the hand landmarks in different palm configurations such as open palm

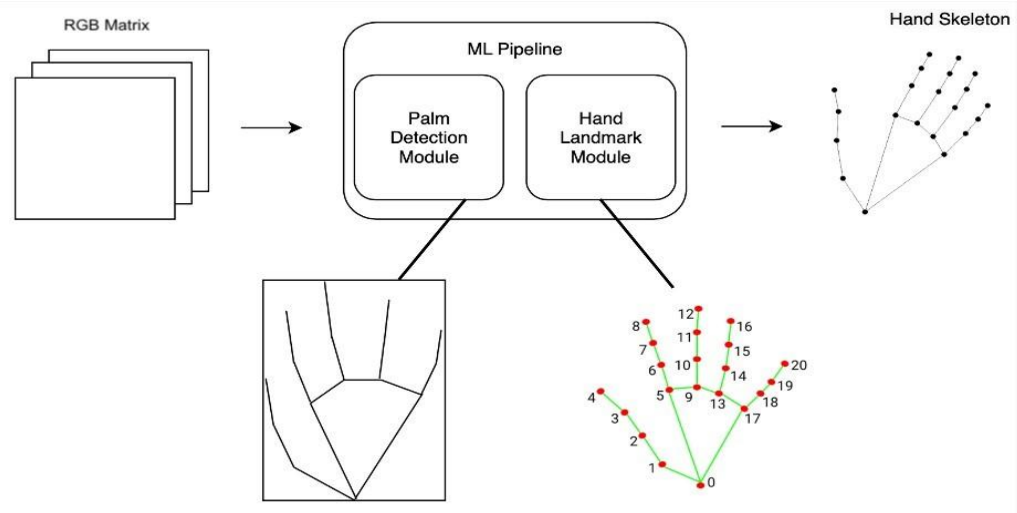


**Fig. 2** Important Hand Landmarks

### 3.2 Facial Features Model using Haar Cascade Classifier

The facial feature detection component of our research utilizes an integrated approach combining traditional computer vision and deep learning techniques to accurately identify and analyze facial expressions in real-time video streams. We employed a pre-trained Haar cascade classifier for initial face detection, which identifies characteristic facial patterns with high precision. The detected faces are subsequently analyzed using a specialized convolutional neural network (CNN) trained for facial emotion recognition, capable of predicting emotional states such as "Happy", "Sad", or "Neutral". To enhance detection accuracy, we integrated MediaPipe's Holistic model to detect precise facial keypoints through its comprehensive pose estimation capabilities. This approach enables accurate annotation and visualization of facial features and overall facial structure. By systematically integrating these advanced detection techniques, our model achieves robust face detection and emotion recognition capabilities, enabling comprehensive analysis of facial expressions across diverse real-world scenarios.

### 3.3 Data Acquisition and Creation

Our methodology involves capturing 30-frame video sequences for each event, allowing us to create custom datasets that facilitate in-depth analysis of hand signs. Utilizing the MediaPipe Holistic model, we extracted keypoints from each frame, which were then flattened and converted into NumPy arrays. These arrays served as training data for our LSTM model, with corresponding event labels stored in accompanying lists.

For emotion sensitivity detection, we leveraged the FER2013 dataset, which provides approximately 29,000 labeled RGB facial images representing seven distinct emotional states. To enhance dataset diversity, we implemented a user-driven data collection approach using webcam-based video capture. Each video sequence focuses on recognizing common sign language actions such as "hello", "good", "morning", and "thank you". The hand landmarks extracted from these frames contributed to a specialized, comprehensive dataset tailored for training our gesture recognition model.

### 3.4 Extraction of keypoints

We employed the open-source MediaPipe framework to precisely capture and analyze palm keypoints critical for interpreting sign language gestures. Our approach transcends static sign detection, concentrating on comprehensive gesture recognition. MediaPipe's advanced palm detection model tracks hand movements, systematically extracting key anatomical points from each video frame.

The 30-frame video capture methodology ensures comprehensive gesture representation. These extracted points are methodically organized into NumPy arrays, serving as input for our LSTM model training. This approach enables effective learning of complex, dynamic palm gestures. Additionally, MediaPipe's integrated face detection capabilities allow for precise identification and isolation of facial regions of interest, including eyes, nose, mouth, and overall facial structure. This holistic approach ensures a thorough analysis, enabling our model to interpret both hand gestures and facial expressions with high accuracy.
accurately

### 3.5 Deep Learning Models

In predicting sign language gestures, we use an LSTM (Long Short-Term Memory) model designed to process temporal sequences of gesture data. The model architecture includes multiple layers to capture the complex patterns and relationships inherent in sign language. The first LSTM layer has 64 units with ReLU activation and is configured to return sequences. The second LSTM layer contains 128 units, also returning sequences. A final LSTM layer with 64 units does not return sequences, effectively summarizing the learned information.

Following these LSTM layers, two dense layers with 64 and 32 units, respectively, use the ReLU activation function to enhance learning. The output layer, with a softmax activation function, classifies gestures into predefined categories. The model is compiled with the Adam optimizer and uses categorical cross-entropy as the loss function. It is trained over 2000 epochs to ensure robust performance.

MediaPipe is used for real-time hand gesture detection and landmark extraction. We process the input frame, converting it to RGB for the model’s prediction and then revert it back to BGR. Combining LSTM and Dense layers allows the model to handle the complexity and variability of sign language gestures, providing accurate real-time predictions. The LSTM layers handle the sequence dynamics, while the Dense layers fine-tune features for classification.

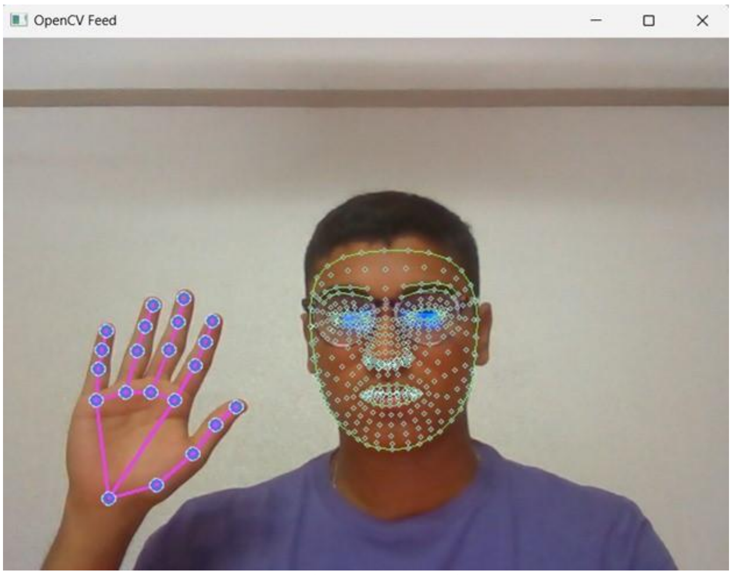


**Fig. 3** Extraction of key points for sign language

### 3.6 Emotion Detection

The emotion recognition component of our system uses OpenCV techniques combined with deep learning methods to analyze and interpret facial expressions detected in every frame. The model is based on a fine-tuned convolutional neural network (CNN), trained on large facial emotion datasets.

The Haar Cascade classifier detects faces in each frame of the real-time video feed. Detected faces are cropped, resized to a fixed 48x48 pixel size, and fed into the CNN model, which accurately predicts emotions such as happiness, sadness, and anger. The CNN analyzes facial features, generating a probability distribution of emotions for each detected face.

The real-time video feed is then overlaid with the label of the most likely predicted emotion. The Haar Cascade classifier continuously detects faces, refining regions of interest to ensure accurate emotion detection even in dynamic environments. It also maintains a history of detected emotions by collecting predictions from each frame, providing valuable insights into emotional content.

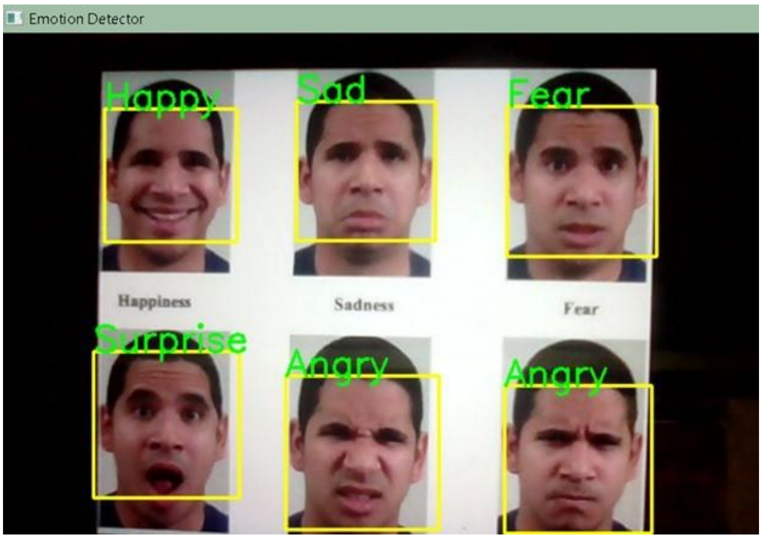


**Fig. 4** Emotion Detection Model

### 3.7 Merging Models

The real-time recognition of both facial expressions and sign language gestures represents a significant advancement in human-computer interaction. The MediaPipe framework supports detailed interpretation of sign language and facial expressions, capturing the subtleties of emotion.

Through careful dataset creation and video image processing, our model reads both hand and facial landmarks, capturing intricate details of the human hand and face. The hand gesture recognizer, built using an LSTM model, goes beyond static signs to interpret dynamic gestures, analyzing sequences of hand formations over multiple frames. Simultaneously, a CNN-based facial expression recognition model analyzes facial landmarks to categorize various emotional states with high precision.

This integrated approach allows the system to evaluate both sign language and emotional expressions in real time, making it an advanced tool for a wide range of applications. It provides instant sign language interpretation and represents an empathetic and inclusive communication tool.

### 3.8 Correlation of the models

The system's core functionality processes sequences of hand gestures to predict their meanings within the context of sign language. Simultaneously, a sentiment analysis module evaluates the emotional content of these recognized gestures. This is achieved through the use of semantic libraries, such as spaCy, which generates meaningful sentences from the recognized gestures while ensuring grammatical correctness. This forms the foundation for semantic understanding, which allows the system to interpret the contextual meaning of gestures.

The sentiment analysis is performed using the NLTK Sentiment Intensity Analyzer. It classifies the sentences derived from the gestures into three categories: positive, negative, or neutral. This step provides valuable insight into the emotional tone of the communication. It's crucial for addressing potential discrepancies where the gesture analysis suggests one emotional state, but the user's facial expression conveys another. This mismatch could be due to several factors, including suppressed emotions, misinterpretation of gestures, or varying cultural interpretations of signs.

By integrating sentiment analysis with the facial emotion detection data, the system ensures the emotional content conveyed through sign language is consistent with the user's facial expression. This feature improves the accuracy and reliability of the communication process, making it more empathetic and contextually aware. Furthermore, the system’s output includes both the recognized gesture and the corresponding emotional state of the user, providing a comprehensive understanding of the communication. This integration ensures that the intended emotions of the user are effectively conveyed, fostering more natural and meaningful interactions. Thus, the correlation between the gesture analysis and facial emotion detection strengthens the system's ability to recognize and convey both the verbal and emotional intent of the user, improving the consistency and empathy of the communication tool.

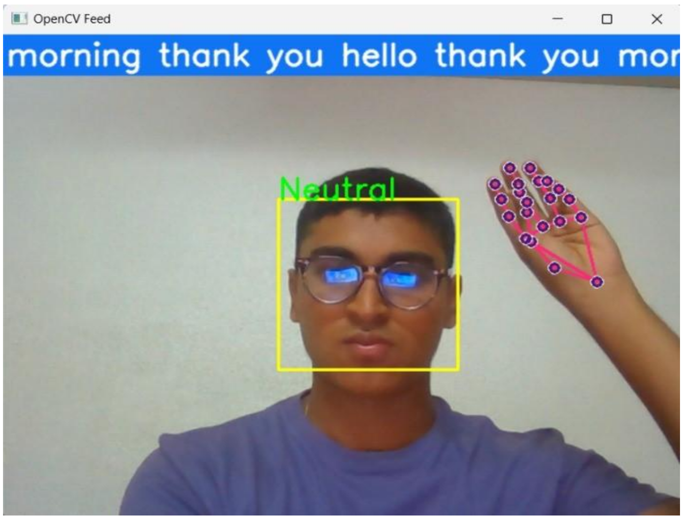
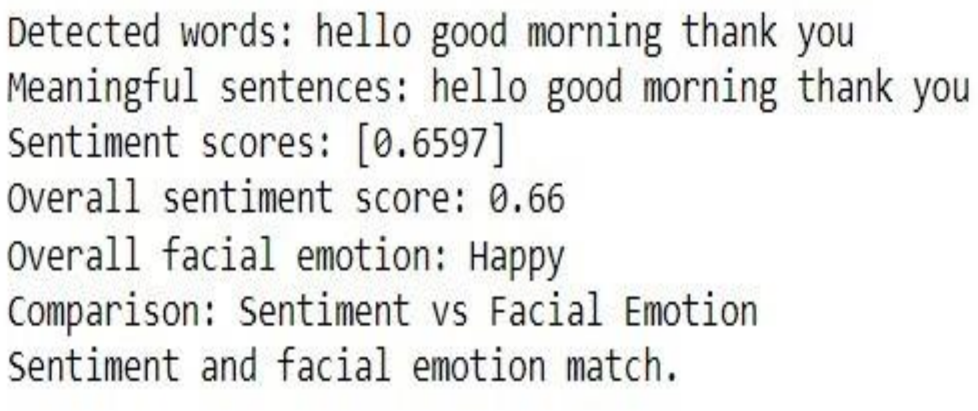


**Fig. 5** Demonstration of how a sentence can be formed with the help of recognized words

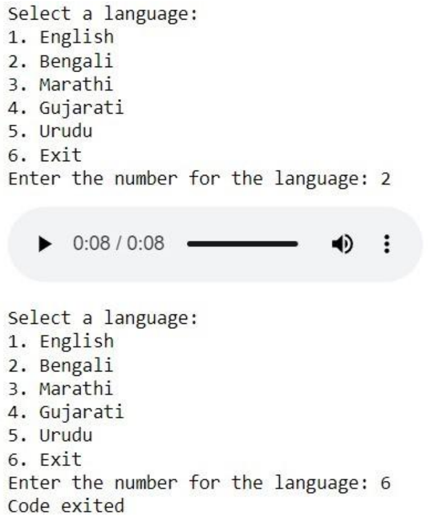


```
Detected words: hello good morning thank you
Meaningful sentences: hello good morning thank you
Sentiment scores: [0.6597]
Overall sentiment score: 0.66
Overall facial emotion: Happy
Comparison: Sentiment vs Facial Emotion
Sentiment and facial emotion match.
```

**Fig. 6** Correlation analysis

### 3.9 Audio Integration

The text-to-speech (TTS) functionality is a crucial component for enhancing accessibility and user engagement. This feature converts text into speech, enabling users to interact with information through their sense of hearing. The TTS system operates through a set of functions that manage language preferences and handle program termination. After translating a sentence, the core module generates synthesized speech asynchronously using external libraries like "gtts" and "translate." After translation, memory-handling techniques integrate the synthesized speech into an audio stream. The system allows users to select multiple languages in the same session, using threading to manage this process. The speech is then remembered, providing a more inclusive and accessible user experience.



```
Select a language:
1. English
2. Bengali
3. Marathi
4. Gujarati
5. Urudu
6. Exit
Enter the number for the language: 2
```

```
Select a language:
1. English
2. Bengali
3. Marathi
4. Gujarati
5. Urudu
6. Exit
Enter the number for the language: 6
Code exited
```

**Fig. 7** Text to Speech feature

## 4 Results

The proposed solution offers a comprehensive approach to enhancing communication for individuals who rely on Indian Sign Language (ISL) by providing real-time sign language interpretation and emotion analysis. By leveraging LSTM (Long Short-Term Memory) models and keypoint extraction techniques, the system is capable of interpreting a wide array of ISL gestures, ensuring accurate communication within the sign language community. This capability serves as the core of the model, enabling uninterrupted interactions for users dependent on sign language.

In addition to gesture recognition, the model integrates facial emotion recognition, offering insights into the emotional cues associated with the signs being made. This emotional analysis is crucial for understanding the emotional context behind the communicated message. By combining the gesture interpretation with emotion detection, the system adapts and responds according to both the linguistic content and the emotional state of the user, enhancing the overall interaction experience.

Correlation analysis further refines the system’s performance by ensuring consistency between the gestures and the emotional context. As depicted in Figures 8 and 9, the model accurately links the gestures to the corresponding emotions. In Figure 8, for example, the word "good" is accompanied by the emotion "happy," while Figure 9 illustrates how a sentence is formed after detecting multiple signs (e.g., "morning," "thank you," "hello," "thank you"), along with the associated emotional cues.

Model Performance:

- LSTM Model for Sign Prediction: The system achieves an impressive accuracy of 98.12% on test data, ensuring reliable sign language recognition.
- CNN Model for Emotion Detection: The emotion detection model performs with an accuracy of 79.22% on test data, showcasing its ability to detect a wide range of emotional expressions based on facial features.

This integration of sign language recognition with facial emotion analysis not only facilitates clear communication but also ensures that emotional nuances are conveyed, making the interaction more natural and contextually aware.

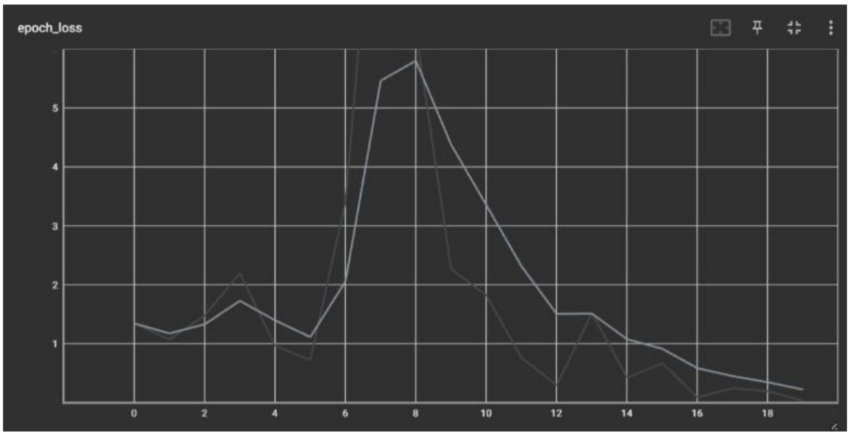


**Fig. 8** Epoch loss during Training of Sign Language model

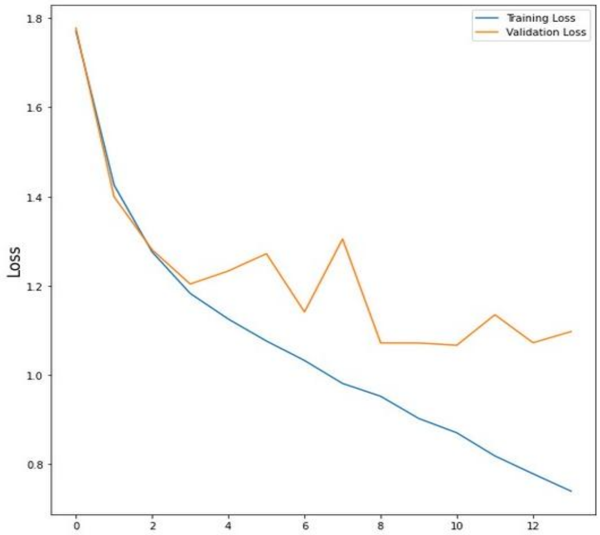


**Fig. 9** Epoch loss during Training of Emotional recognition model

## 5 Comparison with other models

| Papers | SL | Methods | Dataset | Size | Tr:Tt | Acc. (%) |
|---|---|---|---|---|---|---|
| [9] (2020) | ASL | RF, NB, SVM, LR, KNN, MLP | A-Z, 3-Spec. Char. | 3000 | 75:25 | 72.23 - 98.31 |
| [10] (2018) | ASL | Inception, LSTM | 100 signs | 2400 | 75:25 | 93.00 |
| [1] (2020) | ASL | VGG-16 | A-H | 5000 | 75:25 | 99.62 |
| [11] (2020) | ASL | SVM | A-Z, 'space' | 3000 | 80:20 | 98.89 |
| [12] (2020) | ISL | KNN | A-Z | 220 DH, 800 SH | NA | 95.84 DH, 94.88 SH |
| [13] (2014) | ISL | Haar Cascade | 5 phrases | 1000 | 90:10 | 92.68 |
| Note: NA - Not Available, Tr:Tt - Training: Testing, SL - Sign Language | | | | | | |

## 6 Conclusion

In conclusion, the integration of LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks) presents a groundbreaking solution to the challenges faced in recognizing Indian Sign Language (ISL). This system offers real-time sign language interpretation and emotional context analysis, making it an invaluable tool for individuals who are hearing impaired. By breaking the communication barriers between hearing and deaf individuals, this system fosters inclusivity and enhances mutual understanding.

The system’s ability to accurately interpret sign language gestures and their emotional nuances significantly improves communication quality. It promotes better understanding regardless of context, ensuring that both the linguistic content and the emotional undertones are conveyed effectively. This not only aids in day-to-day communication but also contributes to building deeper, more empathetic interactions. Furthermore, the application of advanced machine learning algorithms opens the door to new opportunities across various domains, from everyday life to scientific research. The "translation" of sign language, coupled with emotional recognition, adds an additional layer to communication that goes beyond simple information transmission, enriching the quality of interaction.

The system's ability to assist individuals with hearing impairments has the potential to transform societal interactions, helping them integrate seamlessly into an inclusive society. It demonstrates that hearing loss can be leveraged as a force for positive change, facilitating greater accessibility and inclusion for those who rely on sign language.

## 7 Future Scope

Looking ahead, this research can be expanded by developing a user-centered mobile application for individuals with hearing loss. The app would provide real-time, accurate sign language interpretation and emotion detection, enabling seamless communication and enhancing users' independence in daily interactions. Key features could include customizable gesture sensitivity, support for various sign language dialects, and options for verbal and non-verbal communication based on user preferences. Extending this solution to multiple platforms—such as desktops, tablets, and wearables—would maximize accessibility and broaden its reach globally. Additionally, integrating the system into smart environments, like smart homes or public spaces, could further enhance real-time communication and social integration. Ultimately, this research envisions a unified, inclusive communication ecosystem that bridges gaps, fosters connection, and empowers individuals with hearing impairments to interact more freely and independently in diverse settings.

## References

1. Tiku, K., Maloo, J., Ramesh, A., R., I.G.: Real-time conversion of sign language to text and speech. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 346–351 (2020)
2. Kumar, M., Conversion of sign language into text. (2018). https://api.semanticscholar.org/ CorpusID:201682290
3. Gill, R., Singh, J.: A deep learning approach for real time facial emotion recognition. In: 2021 10th International Conference on System Modeling Advancement in Research Trends (SMART), pp. 497–501 (2021). https://doi.org/10.1109/ SMART52563.2021.9676202
4. Winyangkun, T., Vanitchanant, N., Chouvatut, V., Panyangam, B.: Real-time detection and classification of facial emotions. In: 2023 15th International Conference on Knowledge and Smart Technology (KST), pp. 1–6 (2023). https: //doi.org/10.1109/KST57286.2023.10086866
5. Riyantoko, P., Sugiarto, Maulida Hindrayani, K.: Facial emotion detection using haar-cascade classifier and convolutional neural networks. Journal of Physics: Conference Series **1844**, 012004 (2021) https://doi.org/10.1088/1742-6596/1844/ 1/012004
6. Unkule, P., Shinde, C., Saurkar, P., Agarkar, S., Verma, U.: Cnn based approach for sign recognition in the indian sign language. In: 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), pp. 92–97 (2022). https://doi.org/10.1109/ICAISS55157.2022.10010911
7. Surya, B., Suresh Krishna, N.V., SankarReddy, A.S., Prudhvi, B.V., Neeraj, P., Deepthi, V.H.: An efficient real-time indian sign language (isl) detection using deep learning. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 430–435 (2023). https://doi.org/10.1109/ ICICCS56967.2023.10142596
8. Jamwal, A., Vasukidevi, G., Malleswari, T.N., Vijayakumar, T., Reddy, L.C.S., Gupta, A.S.A.L.G.G.: Real time conversion of american sign language to text with emotion using machine learning. In: 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 603–609 (2022). https://doi.org/10.1109/I-SMAC55078.2022.9987362
9. Sharma, A., Mittal, A., Singh, S., Awatramani, V.: Hand gesture recognition using image processing and feature extraction techniques. Procedia Computer Science **173**, 181–190 (2020) https://doi.org/10.1016/j.procs.2020.06.022 . International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020
10. Bantupalli, K., Xie, Y.: American sign language recognition using deep learning and computer vision. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 4896–4899 (2018). https://doi.org/10.1109/BigData.2018.8622141
11. Dutta, K.K., Bellary, S.A.S.: Machine learning techniques for indian sign language recognition. In: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 333–336 (2017). https://doi.org/10.1109/CTCEEC.2017.8454988
12. Dabre, K., Dholay, S.: Machine learning model for sign language interpretation using webcam images. In: 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 317– 321 (2014). https://doi.org/10.1109/CSCITA.2014.6839279
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017) https://doi.org/10.1145/3065386