Journal of Computer Science

Research Article

Addressing Emergency Communication Challenges: Deep Learning Solutions for the Speech and Hearing- Impaired

Poornima B V[1,2], Srinath S[2], Mustafa Basthikodi[1], Rashmi S[3] and Rakshitha R[2]

[1]*Department of CSE, Sahyadri College of Engineering & Management, Mangaluru, India*
[2]*Department of CSE, JSS Science and Technology University, Mysuru, India*
[3]*Government First Grade College, Nyamathi, India*

Article history
Received: 12-06-2024
Revised: 31-07-2024
Accepted: 05-08-2024

Corresponding Author:
Mustafa Basthikodi
Department of CSE, Sahyadri College of Engineering & Management, Mangaluru, India
Email: mbasthik@gmail.com

Abstract: Emergency communication plays a very important role in ensuring that help reaches people promptly and safely in case of any emergency. However, the biggest problem faced by those who cannot speak and hear properly is to convey their message clearly and understand others. In this regard, the proposed research work focuses on recognizing emergency gestures made in the Indian Sign Language. It recognizes 14 categories of emergency gestures for various medical-related words. Two types of novel deep learning methods are used in the process to increase recognition efficiency such as the hybrid architecture of 3D Convolutional Neural Networks and Long Short-Term Memory networks and TimeSformer with DenseNet pre-trained network. For the evaluation of both models, two specially developed benchmark datasets have been used such as ISL_CSLTR and INCLUDE. The average accuracy obtained in the experiment using the TimeSformer architecture is 97% while for the hybrid approach is 91%.

Keywords: Indian Sign Language (Isl), Emergency Gestures, Timesformer, Long ShortTerm Memory (Lstm), Sign Language Recognition (Slr), Densenet, 3dcnn

Introduction

Communication during emergency situations must be rapid, clear, and reliable to ensure timely assistance and safety. However, individuals who are deaf or hard of hearing face significant challenges in expressing distress signals and understanding critical information. Traditional assistive communication methods have often proven insufficient in such scenarios, emphasizing the need for advanced intelligent systems capable of interpreting gestures accurately. Deep learning-based approaches offer a promising solution, particularly for recognizing complex and dynamic sign language gestures. In sign language, gestures are broadly categorized into static and dynamic forms (Liang et al., 2023; Lahiani and Neji 2018). Dynamic gestures consist of continuous movements that evolve over time, unlike static gestures that represent a single pose. These dynamic gestures are inherently complex, as they require simultaneous understanding of spatial configurations and temporal transitions. Long-term temporal dependencies in such gestures play a crucial role in conveying meaning, where the sequence, duration, and progression of movements determine the semantic interpretation (Samal and Panda, 2021; Singha et al., 2018). Therefore, effective recognition of sign language requires models capable of capturing both spatial and extended temporal information. Recent advancements in deep learning have significantly contributed to the field of Sign Language Recognition (SLR) (Mohamed et al., 2021; Das et al., 2018). Among these, three-Dimensional Convolutional Neural Networks (3DCNNs) (Al-Hammadi et al 2020; Tatebe et al., 2018) have demonstrated superior performance over traditional two-dimensional CNNs (Rudregowda et al., 2022; Urabe et al., 2018), as they are capable of extracting both spatial and temporal features simultaneously. By performing convolutions across spatial and temporal dimensions, 3DCNNs effectively model motion patterns present in video sequences. Furthermore, Long Short-Term Memory (LSTM) networks (Mariappan et al., 2021; Staudemeyer et al., 2019) have been widely integrated with convolutional architectures to enhance temporal modelling by capturing sequential dependencies across frames. Despite these advancements, existing approaches exhibit several limitations. Many methods primarily focus on static gesture recognition or short-term temporal patterns, with limited emphasis on dynamic gestures involving long-term dependencies. Additionally, most

studies do not specifically address emergency communication scenarios, where rapid and accurate interpretation of gestures is critical. Prior works (Bhandary et al., 2021; Basthikodi and Poornima, 2025; Basthikodi et al., 2021; Salins et al., 2022) largely overlook dynamic Indian Sign Language gestures relevant to emergency situations, thereby limiting their practical applicability. To address these challenges, recent transformer-based architectures such as TimeSformer (Bertasius et al., 2021) have emerged as powerful alternatives for video understanding. By utilizing self-attention mechanisms (Basthikodi et al., 2024), TimeSformer effectively captures long-range temporal dependencies and global contextual relationships. In the proposed work, this capability is further enhanced by integrating DenseNet-based feature representations (Chen et al., 2024; Zhao and Du, 2016). DenseNet improves spatial feature extraction through dense connectivity and feature reuse, while also facilitating efficient gradient propagation during training (Pai et al.,2025; Basthikodi et al., 2019). This integration enables the model to leverage both strong spatial representations and advanced temporal modeling for improved dynamic gesture recognition. The proposed approach is evaluated on real-world datasets and benchmarked against existing methods. Unlike prior studies that predominantly focus on static gestures, this work specifically targets dynamic gesture recognition for emergency communication, addressing a critical research gap in assistive technologies (Bhandary et al., 2021; Basthikodi and Poornima, 2025; Basthikodi et al., 2021; Salins et al., 2022). By incorporating both spatial and long-term temporal modeling, the proposed framework aims to provide a more robust and practical solution for real-world sign language interpretation.

The main contributions of this work are as follows:

1. A novel hybrid framework that integrates TimeSformer with DenseNet-based feature representations to effectively capture both spatial and long-term temporal dependencies in dynamic gesture recognition
2. A dedicated focus on emergency gesture recognition, addressing a critical gap in existing sign language recognition systems that largely overlook time-sensitive communication scenarios
3. A comprehensive modeling of dynamic gestures, enabling improved understanding of continuous motion patterns essential for real-world sign language interpretation
4. Support for both word-level and sequence-level (sentence-level) gesture recognition, enhancing the practical applicability of the system in real communication settings
5. Extensive experimental evaluation and comparison with existing methods, demonstrating the effectiveness of the proposed approach on real-world datasets

Related Work

Liao et al. (2019) proposed a new multimodal algorithm for recognizing dynamic sign language named B3D ResNet, which is a combination of deep 3D residual ConvNet and bi-directional Long Short Term Memory (Bi-directional LSTM). This technique comprises of three key steps including hand object localization for reducing computational burden, automatic extraction of spatiotemporal features using B3D ResNet and finally generating an intermediate score of each action performed in the video clip. It is evident from the experiment results that this technique offers state-of-the-art performance with 89.8% on the DEVISIGN_D dataset and 86.9% on the SLR dataset. Venugopalan and Reghunadhan (2023) focused on the challenging problem of communication obstacles between deaf patients and healthcare professionals. This research work provides a new database with dynamic hand gestures depicting ISL words that are important for communication during emergencies with deaf patients suffering from COVID-19. For increasing the efficiency of gesture recognition, the researchers have adopted a hybrid framework by utilizing both deep CNN and LSTM networks together. The results obtained are promising with an average accuracy rate of 83.36%. A sign word recognition model that is resistant to changes in rotation, translation, and scale has been introduced by Saleh et al. (2023) with the use of CNN. First, the sign word dataset, which consists of 20 sign words, is transformed to generate an RTS dataset. The gesture segmentation technique involves the process of using otsu thresholding along with YCbCr color space and morphological operations followed by the use of the watershed algorithm. Next, the CNN model is trained and validated on both RTS and non-RS versions of the datasets. This model shows excellent results, scoring 99.30% on the original 20 sign word dataset, 99.10% on the RTS version of the dataset, 100% on the five sign word dataset, and 98.00% on the RTS version of the five sign word dataset. Venugopalan and Reghunadhan (2021) carried out research on gesture recognition for words in ISL that are frequently used by deaf farmers. Through a deep learning-based technique with a combination of convolutional LSTM network, the classification accuracy level is achieved to be 76.21% using a dataset of ISL words in agriculture-related contexts. The approach used by Sánchez Ruiz et al. (2023) for their SLR model has a more concentrated approach in terms of word level recognition based on an optimized subset of features. As opposed to the other methods that have been proposed before, their methodology emphasizes non-manual features. Once the ROI is identified and tracked, features

from body pose, facial expressions, hand region, head pose, and eye gaze estimation are obtained. Then, there is a process of data augmentation and dimensional reduction. Lastly, in the recognition stage, BiLSTM and Transformer architectures are employed for classification purposes. Das et al. (2023) have presented a novel method for the prediction of sign language by presenting the Expert System for Indian Sign Language Recognition (ESISLR). The system is developed to recognize the sign language and predicts isolated sign words effectively. It uses a key frame extraction algorithm to eliminate the redundancy of frames from large frame sequences. ESISLR integrates the CNN and handcrafted features, which are subsequently learned using a stack Bi-directional Long Short-Term Memory (BiLSTM). In order to extract the CNN features, the authors have used VGG-19, whereas for extracting handcrafted features, Hu Moments (HM) and Zernike Moments (ZM) are utilized. Masood et al. (2018) presented a method for overcoming the communication barrier through SLR by considering both the temporal and spatial information contained in video frames. Spatial information was extracted from the images through the Inception model, which is a CNN, whereas temporal information was extracted through an RNN. The images utilized in this work were collected from gestures of the LSA, which included 46 gesture types. This novel model has achieved a good level of accuracy, 95.2%, for images in the database. Bansal and Jain (2024) provides a complete technique for building an ISL recognition system that is aimed at dynamic words. This technique uses deep neural network models like VGG19, InceptionV3, DenseNet121, and ResNet50 to generate feature vectors from the input video images, which include important details about the hand positions and movements in ISL signs. Such extracted feature vectors are employed as representations for a robust classifier model of recurrent neural networks to detect ISL. Notably, the presented technique has achieved remarkable accuracies; namely, 99%, 97%, 98%, and 98% for VGG19, InceptionV3, DenseNet121, and ResNet50, respectively.

Also, recent advancements in ISL recognition have increasingly focused on improving dynamic gesture understanding through multimodal and deep learning-based approaches. In Geetha et al. (2025), a real-time continuous ISL recognition system is proposed using a multimodal framework that combines RGB data with pose estimation, enabling improved temporal and spatial understanding of gestures. Similarly, Bansal and Jain (2024) introduced a dual feature descriptor approach integrated with GMT-MASKRCNN for word-level gesture recognition in videos, enhancing feature representation and detection accuracy. In Nadaf et al. (2025), an efficient gesture recognition framework is developed using SENet-based fusion of multimodal data, emphasizing the importance of combining multiple input modalities to improve recognition performance. Furthermore, Agarwal et al. (2025) explored neural network-based approaches for ISL recognition, demonstrating the effectiveness of deep learning models in capturing gesture patterns from video data. These recent studies highlight a shift toward multimodal learning, real-time recognition, and advanced deep architectures.

Methods

In this section, a detailed information about the dataset employed in the experimentation phase and the implementation particulars of the proposed models are given. The general overview of the dynamic gesture's classification network is represented in Fig. 1. In subsequent sections, the model combining 3D Convolutional Neural Networks (3DCNN) with LSTM is denoted as Model-I, while Model-II refers to the implementation based on the Times former architecture. The overall pipeline of the proposed system follows a structured sequence consisting of: input video acquisition preprocessing spatial feature extraction temporal modeling classification output prediction. This structured flow ensures effective learning of both spatial and temporal characteristics of dynamic gestures.

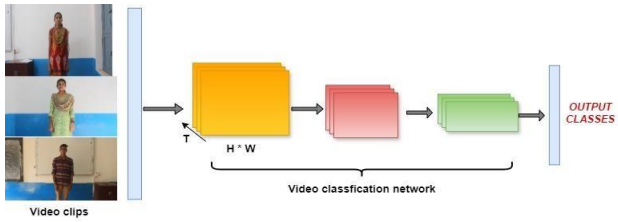


Fig. 1: Overview of dynamic sign language recognition system

Dataset Details

The experiment is done on the publicly available word and sentence level ISL gestures available at the repository: ISL_CSLTR_Corpus-Indian Sign Language Dataset for Continuous Sign Language Translation & Recognition (Elakkiya and Natarajan, 2021). and INCLUDE DATASET (Sridhar et al., 2020). Word-level gestures are sourced from the INCLUDE dataset repository, while sentence-level gestures are acquired from the ISL_CSL_TR_Corpus repository. The signs are performed by both male and female signers in cluttered and non-uniform background. Word level gesture, often referred to as "Isolated gesture," pertains to gestures that accompany individual words or phrases. On the other hand, sentence level gesture, known as "Continuous gestures," encompasses gestures that occur throughout an entire sentence providing continuous non-verbal communication. Information regarding the dataset at the word level is

represented in Table 1, while Table 2 presents details regarding the dataset at the sentence level. Sample frames of the input data is represented in Figs. 2 and 3.

The dataset consists of both word-level (isolated gestures) and sentence-level (continuous gestures), enabling the model to learn both short-term and long-term temporal dependencies. This dual-level dataset design supports comprehensive evaluation of the proposed models in realistic communication scenarios.

Proposed Methodologies

Algorithm: Dynamic Gesture Recognition Framework

Input: Video sequence $\boldsymbol{V}$
Output: Predicted gesture class $\boldsymbol{C}$

1. *Acquire input video sequence* $\boldsymbol{V}$
2. *Perform preprocessing:*
   - *Frame extraction*
   - *Resizing (e.g., 256×256)*
   - *Normalization*
3. *For each frame:*
   - *Extract spatial features using DenseNet / 3DCNN*
4. *Construct temporal sequence of feature vectors*
5. *Apply temporal modeling:*
   - *Model-I: LSTM layers*
   - *Model-II: TimeSformer (self-attention mechanism)*
6. *Learn spatio-temporal representations*
7. *Pass features to classification layer (Softmax)*
8. *Output predicted class label* $\boldsymbol{C}$

Model-I

The proposed 3DCNN integrated with LSTM captures the spatio-temporal (Al-Hammadi et al., 2020) data from the video input. The integration of 3D CNNs with LSTM networks in the context of ISL gesture recognition presents a comprehensive approach aimed at effectively capturing the spatial and temporal dynamics inherent in ISL gestures. ISL gestures are characterized by intricate hand movements and body postures, necessitating detailed understanding of both the spatial configurations and temporal sequences for accurate recognition.

The utilization of 3D CNNs enables the extraction of spatial features from video frames, encompassing critical aspects such as hand poses, movements and inter-segmental relationships. This spatial understanding forms the foundation for discerning the semantic content of gestures. Conversely, LSTM networks excel in capturing temporal dependencies within sequential data. Model architecture defined has two 3D convolutional layers (Conv3D). The first convolutional layer comprises 64 filters with a kernel size of 3x3x3 and ReLU activation. Subsequently, max-pooling layers (MaxPooling3D) with a pool size of 2x2x2 downsample spatial dimensions, followed by batch normalization to stabilize and accelerate the training process. Another convolutional layer with 128 filters and similar parameters is added for feature extraction. A TimeDistributed layer flattens the output of the convolutional layers while maintaining the temporal dimension. Two LSTM layers (LSTM) follow for sequential modeling, with 128 and 64 units respectively. A dropout layer is included to mitigate overfitting, randomly dropping 50% of the input units. Finally, a dense layer with softmax activation is used to classify the input into the respective number of classes. The proposed architecture of Model-I is shown in Fig. 4. The choice of 3DCNN is motivated by its ability to simultaneously capture spatial and short-term temporal features directly from video data, unlike 2DCNNs which process frames independently. The integration of LSTM further enhances the model by capturing long-term temporal dependencies, making it suitable for sequential gesture understanding. This hybrid approach provides a balance between local motion feature extraction and sequential modeling, which is essential for dynamic gesture recognition.

Table 1: Word level dataset details

Class	No. of video clips	Duration of the video	No. of frames per video
Doctor	100	4s	240
Hospital	100	4s	240
Medicine	100	4s	240
Sick	100	4s	240
Healthy	100	4s	240
Hungry	100	4s	240

Table 2: Sentence level dataset details

Class	No. of video clips	Duration of the video	No. of frames pervideo
Bring water for me	100	6s	240
I am feeling cold	100	6s	240
I am fine, thank you	100	6s	240
I am hungry	100	6s	240
I am suffering from fever	100	6s	240
I am tired	100	5s	240
I need water	100	5s	240
Serve the food	100	5s	240

Fig. 2: Sample frames for the gesture “I AM FINE”

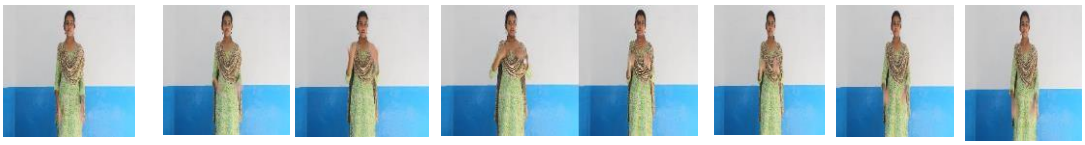


Fig. 3: Sample frames for the gesture “HEALTHY"

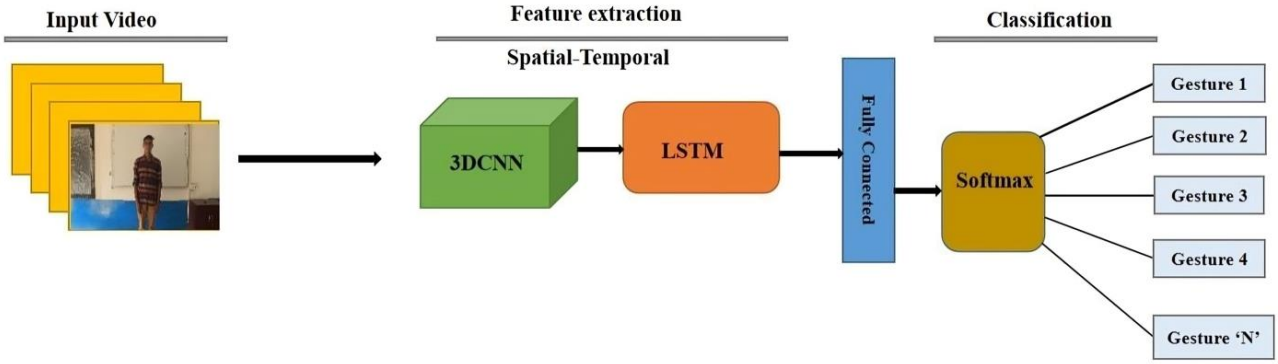


Fig. 4: 3DCNN with LSTM architecture

Model-II

The proposed Model-II introduces an innovative method for enhancing sentence and word level gesture recognition by combining DenseNet based (Altaf et al., 2023) feature extraction with the TimeSformer model. DenseNet121, a renowned pre-trained CNN, is employed to extract intricate spatial features from individual frames of ISL gesture videos. The utilization of DenseNet allows for the extraction of rich spatial information, crucial for discerning complex hand movements and body postures within ISL gestures, especially against challenging backgrounds. These spatial features are then integrated into the TimeSformer model, an attention-based transformer architecture renowned for its ability to capture both spatial and temporal dependencies within sequential data. This integration enables the model to accurately recognize and interpret complex gestures at both sentence and word levels, even amidst complex backgrounds. These frames are pre-processed by resizing them to 256x256 pixels. Subsequently, DenseNet121 extracts features from each frame, resulting in a feature vector of length 1024 that encapsulates the spatial characteristics of the gesture. In addition to dense net’s capability to extract spatial features, the TimeSformer model is used to extract both spatial and temporal dependencies from gesture videos. The architecture of the TimeSformer model consists of positional embedding layers and a Transformer Encoder layer. The positional embedding layer improves the capabilities of the model in capturing temporal sequences by adding positions into the input features. Conversely, the Transformer encoder layer utilizes self-attentive models to extract both spatial and temporal dependencies. In terms of our architecture, the pre-trained dense net model will be considered a feature extractor model. When an image frame is passed into the Dense Net model, each of the layers of the network will extract different levels of spatial feature representations ranging from low-level features such as edges and texture to high-level features such as object part and shapes. The feature extraction process results in the generation of high-level abstract feature representation of the input image. Fig. 5 shows the proposed architecture of Model-II. The next step involves feeding the input frames to the TimeSformer model. The input clip is decomposed into smaller patches or segments. This allows the model to process the video data more efficiently by focusing on localized information. Each patch is then linearly embedded into a lower-dimensional representation using a linear transformation. This can be represented as shown in the Equation (1):

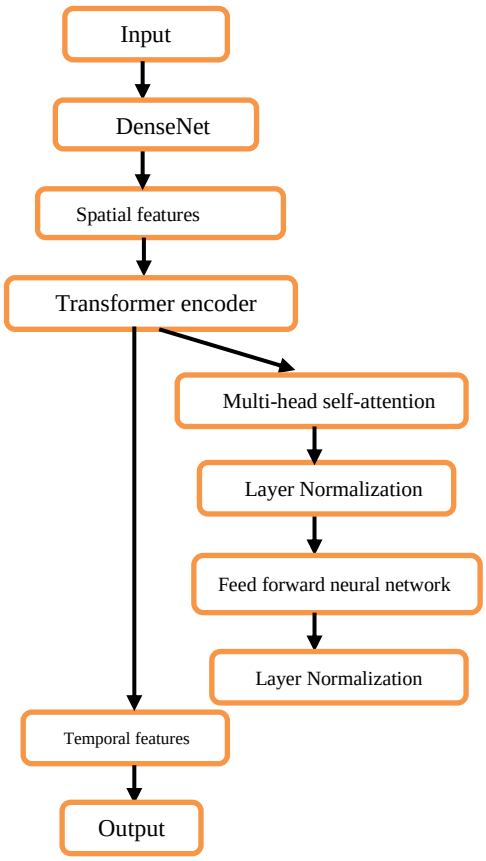


Fig. 5: Work flow of the Model-II, second box after Input Frames should be DenseNet

$$Encoder\,output = LayerNorm.(Xemb + Attention(Q,K,V) \quad (1)$$

Where X*emb* is the embedded representation of the input patch *X*. W*emb* is the weight matrix and *b emb* is the bias term. Further, Query-key-value computation involves computing query, key and value vectors for each embedded patch. These vectors are used in the self-attention mechanism to determine the importance of different patches relative to each other.

This computation can be represented as shown in Equation (2):

$$Q = XembWQ,.K = X\,embWK,.V = XembWV \quad (2)$$

With *Q*, *K*, and *V* being the query, key, and value matrices, while *WQ*, *WK*, and *WV* represent the respective weight matrices. Self-Attention generates a weighted sum of the values based on how compatible the queries are with the keys.

This can be represented as shown in Equation (3):

$$Attention(Q,.K,.V) = Soft\max\frac{(QK^T)}{(\sqrt{dk})}V \quad (3)$$

Where *dk* represents the dimensions of key vectors, and the softmax function is applied row-wise to calculate the attention weights. The result of the self-attention process is added to the initial patch embeddings through the residual connection and layer normalization processes.

This can be represented as shown in Equation (4):

$$Encoder\,output = LayerNorm.(Xemb + Attention(Q,K,V) \quad (4)$$

Finally, the output from the encoding step is used for classification tasks. Proposed TimeSformer is shown in Fig. 6.

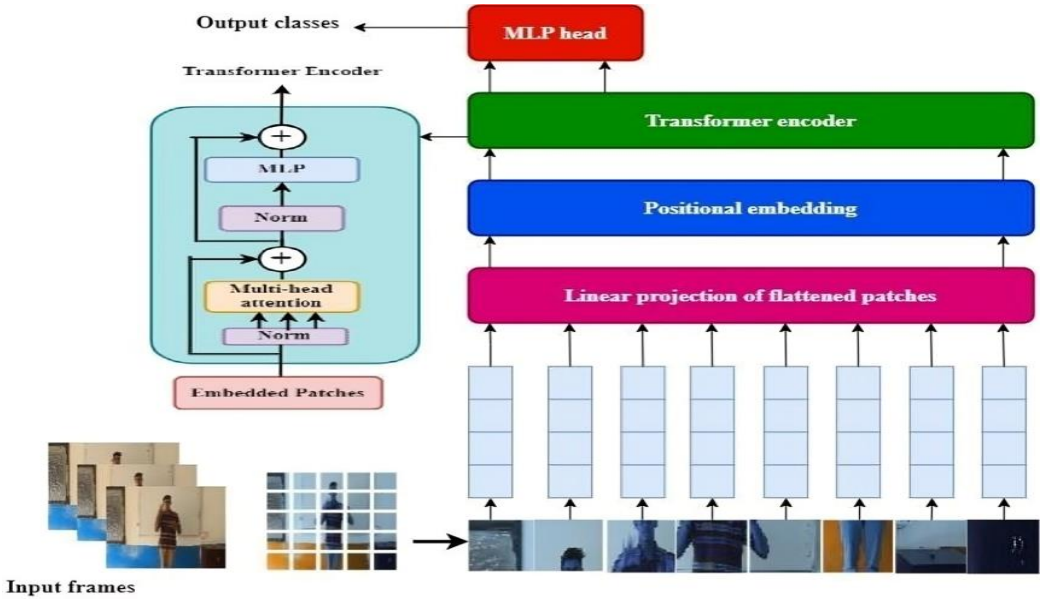


Fig. 6: TimeSformer architecture for dynamic SLR

To combine both spatial features obtained from the DenseNet model and the temporal features provided by the TimeSformer model, the DenseNet121 model is first imported with the help of Keras using weights set to "imagenet," which shows that the pretrained model will be based on the weights obtained from the ImageNet dataset. The parameter include_top is set to False in order to remove the fully connected layers present at the end of the DenseNet model architecture, keeping only the convolutions. When performing gesture recognition, each frame of the video is first processed through the pretrained DenseNet model to extract features before being passed onto the TimeSformer model.

The selection of the TimeSformer architecture is driven by its superior capability to model global temporal dependencies using self-attention mechanisms, which overcome the limitations of recurrent models like LSTM. Unlike traditional CNN-LSTM approaches, TimeSformer processes the entire video sequence holistically, enabling better understanding of long-range interactions between frames. DenseNet is integrated as a spatial feature extractor due to its dense connectivity, which promotes feature reuse, improves gradient flow, and reduces redundancy. Compared to conventional CNN architectures, DenseNet provides more efficient and discriminative feature representations, particularly in complex backgrounds. The combination of DenseNet and TimeSformer results in a hybrid architecture that effectively captures both fine-grained spatial features and long-range temporal relationships, making it more suitable for dynamic gesture recognition than standalone CNN or LSTM-based models.

Results and Discussion

The experimental results depict that Model-II demonstrated superior performance compared to Model-I, achieving the good accuracy while minimizing loss and requiring fewer epochs. Initially, frames extracted are resized to 256x256 pixels. Video clips consist of word gestures lasting 4 seconds and sentence-level gestures lasting 6 seconds. For experimentation, 240 frames are extracted per video clip. Standardizing the duration, all videos are uniformly set to 4 seconds for word- level gestures and 6 seconds for sentence-level gestures to mitigate data variability. In both Model I and Model II, they have been trained to recognize six word-level gestures and eight sentence-level gestures, where all gestures contain an equal number of videos per gesture type. In particular, there were 100 videos per word-level gesture, and also 100 videos for each sentence-level gesture. The experiment is performed on Google Colab, running on an NVIDIA RTX 3090 GPU.

Performance of Model-I

Model-I is trained over 100 epochs, with the training process entailing the partitioning of the dataset into training and testing sets utilizing an 80-20 split. Accuracy fluctuated as the number of epochs increased, stabilizing notably around epochs 90-98, characterized by minimal loss. On average, Model-I achieved an accuracy of 92.3% for word- level gestures and 90.9% for sentence-level gestures. However, accuracy varied across different gestures. The parameters used for Model-I are detailed in Table 3.

Performance of Model II

Model-II underwent a training spanning 50 epochs, with the dataset partitioned into an 80:20 training and testing ratio. The TimeSformer model is configured with specific hyperparameters. Remarkably, Model-II exhibited exceptional overall accuracy, achieving 98.4% for word-level gestures and 94.96% for sentence-level gestures. The performance improvement can be attributed to the self-attention mechanism inherent in the TimeSformer architecture. Unlike traditional recurrent or convolutional models, TimeSformer excels in capturing long-range dependencies and intricate temporal details more efficiently. Comparatively, Model-II achieved superior results with fewer epochs than Model-I. The self-attention mechanism enables TimeSformer to effectively capture and integrate information from distant time steps within the sequence, facilitating more efficient learning and convergence. Consequently, fewer training epochs are required to achieve optimal performance, as the model can quickly adapt and incorporate relevant temporal dependencies during training. The parameters used for Model-II are detailed in Table 4.

The results showing the actual and predicted classes when tested on unseen data is depicted in Fig. 7.

Table 3: Parameters employed for Model-I

Parameters	Values
Learning Rate (3DCNN/LSTM)	0.001
Optimizer (3DCNN/LSTM)	SGD-stochastic gradient descent
Momentum	9
Activation (3DCNN)	Leaky ReLU
Activation (LSTM)	Sigmoid
Filter Size (3DCNN)	3x3x3
Number of Convolutional Layers	2
Number of LSTM Units	64
Time Steps	240 frames
Sequence Length	Fixed, 4s and 6s
Batch Size	32
Dropout Rate (3DCNN/LSTM)	0.5
Combination Method	Concatenation
Loss Function	Categorical cross-entropy

The confusion matrices for Model-I and Model-II as represented in Figs. 8a-b and 9a-b, provide a clear view of class-wise performance for both word-level and sentence-level gestures. For Model-I (3DCNN +

LSTM), the word-level gestures achieved a high overall accuracy of approximately 92.3%, with only minor misclassifications occurring between visually similar gestures such as "Doctor" and "Hospital." In contrast, the sentence-level gestures, which involve longer temporal sequences, showed slightly lower accuracy at around 90.9%. Misclassifications in sentence-level gestures were primarily observed for sentences with overlapping or similar gesture segments, highlighting the limitations of Model-I in capturing long-term temporal dependencies. Model-II (DenseNet + TimeSformer) demonstrated superior performance across both word-level and sentence-level gestures. Word-level gestures were recognized almost perfectly, achieving an accuracy of 99%, with very few errors. This performance can be attributed to DenseNet's effective spatial feature extraction combined with TimeSformer's self-attention mechanism, which efficiently captures temporal relationships. For sentence-level gestures, the model achieved an accuracy of 94.9%, outperforming Model-I. The remaining misclassifications were mostly limited to sentences with similar beginnings or endings, indicating that the model successfully manages complex temporal patterns across longer sequences.

Table 4: Parameters used for Model-II

Parameters	Values
Learning Rate	0.001
Optimizer	Adam
Activation	GeLU
Number of Transformer Layers	3
Number of attention heads	4
Sequence Length	240 frames
Batch Size	64
Loss function	Sparse categoricalcross-entropy
Dropout Rate	0.5
Combination Method	Concatenation

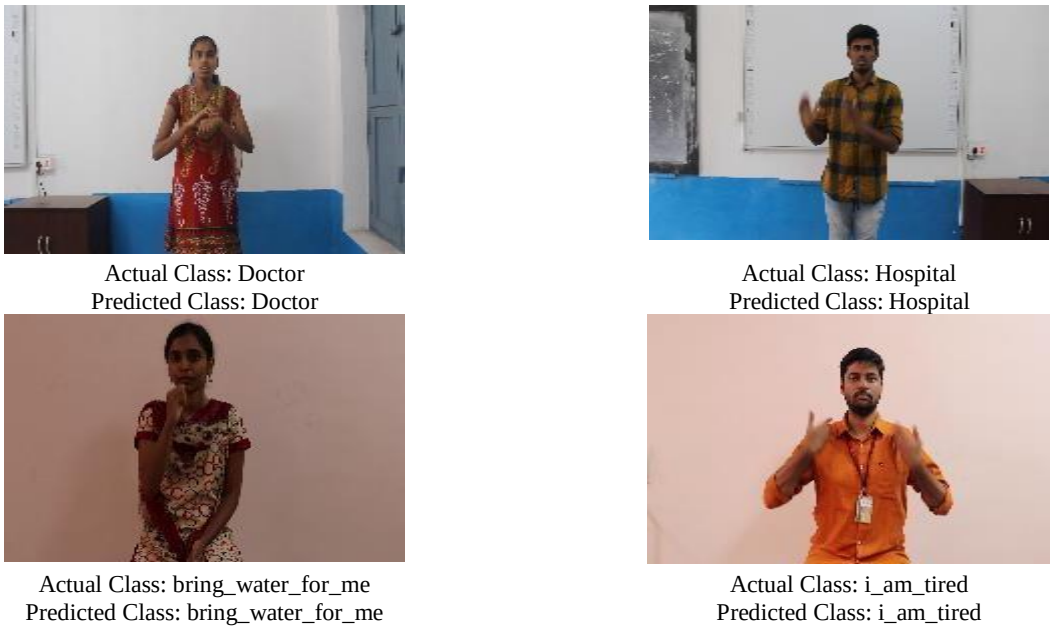
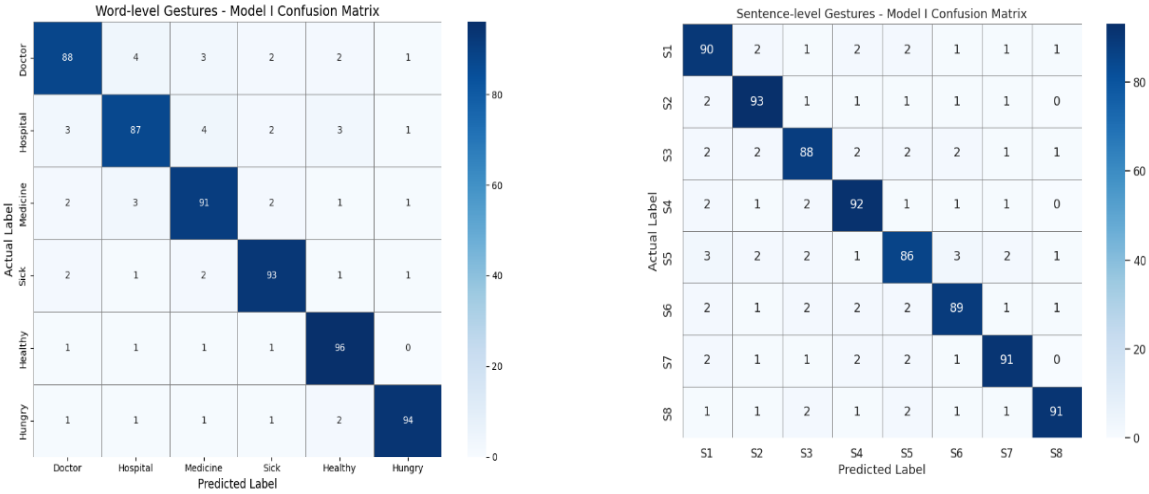


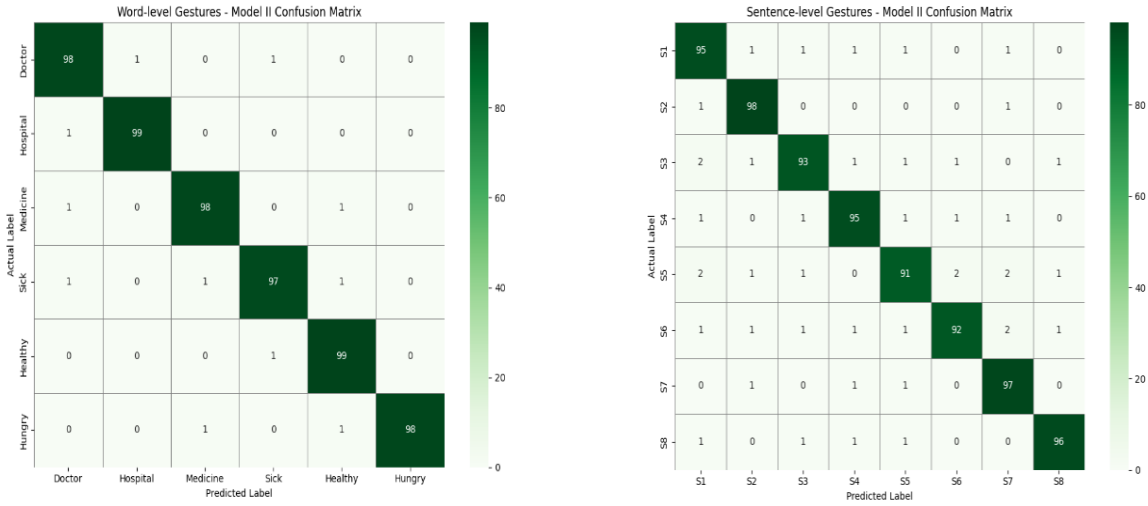
Fig. 7: Screenshots showing the actual and predicted classes

The training behavior of both models is illustrated in Figures 10 and 11 through accuracy and loss curves for word-level and sentence-level gestures. For Model-I, the word-level gesture training demonstrates a gradual increase in accuracy, converging around 92.3%, while the sentence-level accuracy stabilizes near 90.9%. A slight dip towards the end of training indicates minor overfitting, which is also reflected in the corresponding loss curves that steadily decrease but exhibit small fluctuations typical of real-world training. In contrast, Model-II exhibits faster convergence due to the self-attention mechanism of TimeSformer, achieving 99% accuracy for word-level gestures and 94.9% for sentence-level gestures within fewer epochs. Both word and sentence-level loss curves for Model-II show a smooth decrease, indicating stable learning with minimal oscillations. Overall, these plots highlight the superior convergence efficiency and robustness of Model-II over Model-I, while also revealing the effect of temporal complexity on sentence-level gestures compared to simpler word-level gestures.



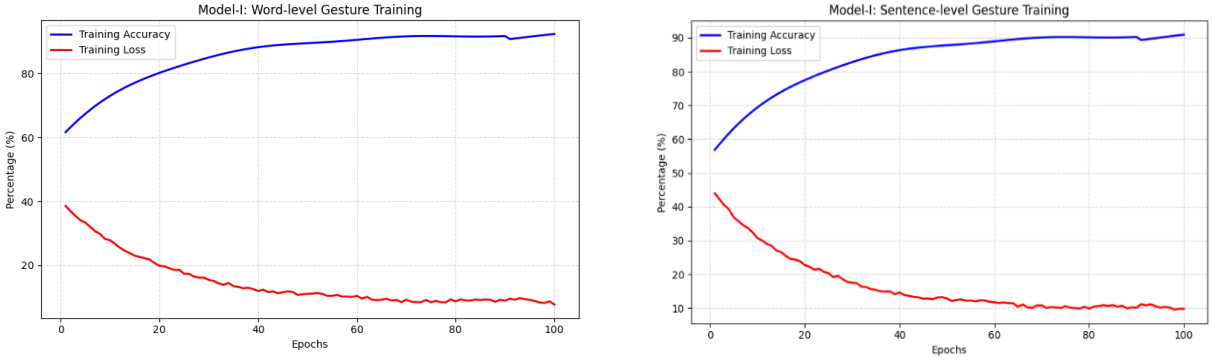
a: Word level b. Sentence level

Fig. 8a-b: Confusion Matrix - Model I



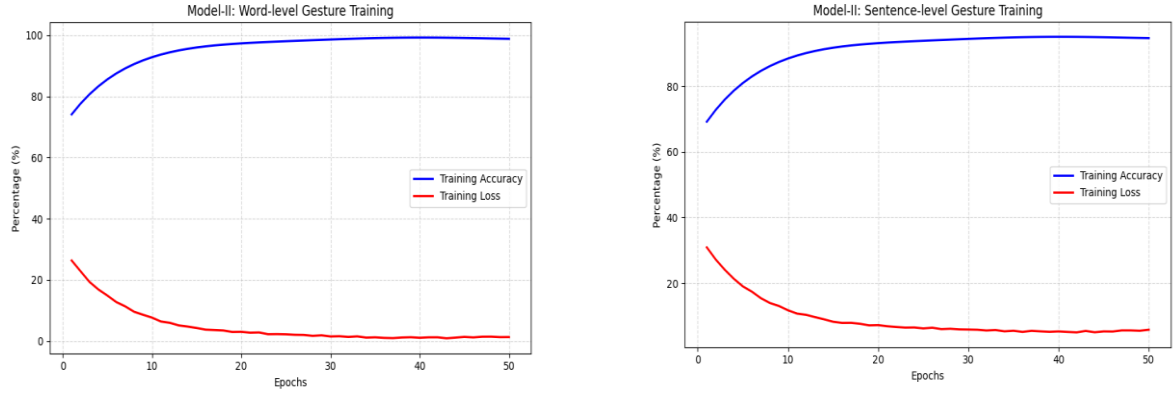
a. Word level b. Sentence level

Fig. 9a-b: Confusion Matrix - Model II



a. Word level b. Sentence level

Fig. 10a-b: Training plots for Model I



a. Word level b. Sentence level

Fig. 11a-b: Training plots for Model II

The comparative analysis made with existing approaches for certain dynamic gestures is shown in Table 5. Existing methods such as the 2-layer Vision Transformer and CNN + LSTM with MediaPipe, often struggle with capturing the full complexity of temporal dynamics and detailed gesture sequences due to their limited architectural depth and integration constraints. Models combining CNN, RNN, and LSTM or CNN with BiLSTM and CTC may face challenges in effectively managing long-term dependencies and continuous gesture transitions. The comprehensive training on diverse datasets ensures that the proposed models are not only accurate but also generalize well across various real-world scenarios, further enhancing their effectiveness over traditional methods.

Table 6 details the state of art methods for INCLUDE dataset. Existing models such as Mobilenetv2 with BiLSTM, BERT Transformer, and Inceptionv3 with LSTM RNN exhibit certain limitations in capturing the full complexity of dynamic gestures. Mobilenetv2 with BiLSTM may struggle with intricate temporal dynamics, BERT Transformer, while powerful for text, is less effective with continuous gesture sequences, and Inceptionv3 combined with LSTM RNNs, despite robust feature extraction, may fall short in managing long-term dependencies and subtle gesture transitions. In contrast, the proposed methods demonstrate superior performance on the same benchmark dataset, achieving higher accuracy in word level gesture recognition.

Table 5: Proposed system Vs. Existing systems

Work done	Classes	Gesture level	Architecture	Accuracy
Agarwal et al. (2023)	72	Word	2-layer ViT	99.56%
Mishra et al. (2023)	8	Word	CNN+LSTM, MediaPipe	97.53%
Jayanthi et al. (2023)	15	Word	CNN, RNN, LSTM	89.99%
Sharma et al. (2023)	90	Sentence &Word	CNN, BiLSTM, CTC	96.6%
Proposed	14	Sentence &Word	Model-I	92% for word
				90% for sentence
			Model-II	99% for word
				94.9 % for sentence

The state of art methods for ISL_CSLTR sentence level gestures is given in Table 7. The existing methods have notable limitations. Dynamic GAN may struggle with the variability and intricacy of dynamic gestures due to its focus on generative modeling rather than temporal sequence analysis. Media Pipe combined with CNN+BiLSTM, while effective, can have difficulty handling long-term dependencies and subtle nuances in gesture transitions due to the limited scope of its temporal modeling.

Graph showing the class level accuracy for both the models for sentence level gestures is represented in Fig. 12 and 13 represents the accuracy for each word level class. S1 to S8 represents 8 sentence gestures and W1 to W6 represents 6 words gestures considered for the experiment.

Table 6: State of art methods for “INCLUDE” dataset

Work done	Gesture level	Architecture	Accuracy
Sridhar et al. (2020)	Word	Mobilenetv2, BiLSTM	85.6%
Prathap et al. (2023)	Word	BERT Transformer	89.5%
Katti et al. (2023)	Word	Inceptionv3+LSTM RNN	79%
Proposed	Word	MODEL-I MODEL-II	92%
			99%

Table 7: State of art methods for "ISL_CSLTR" dataset

Work done	Gesture level	Architecture	Accuracy
Natarajan, and Elakkiya, (2022)	Sentence	Dynamic GAN	93.7%
(Rajalakshmi et al., 2022)	Sentence	MediaPipe, CNN+BiLSTM	95%
Elakkiya et al. (2023)	Sentence	Hybrid NMT	-------
Proposed	Sentence	MODEL-I MODEL-II	90% 94.9

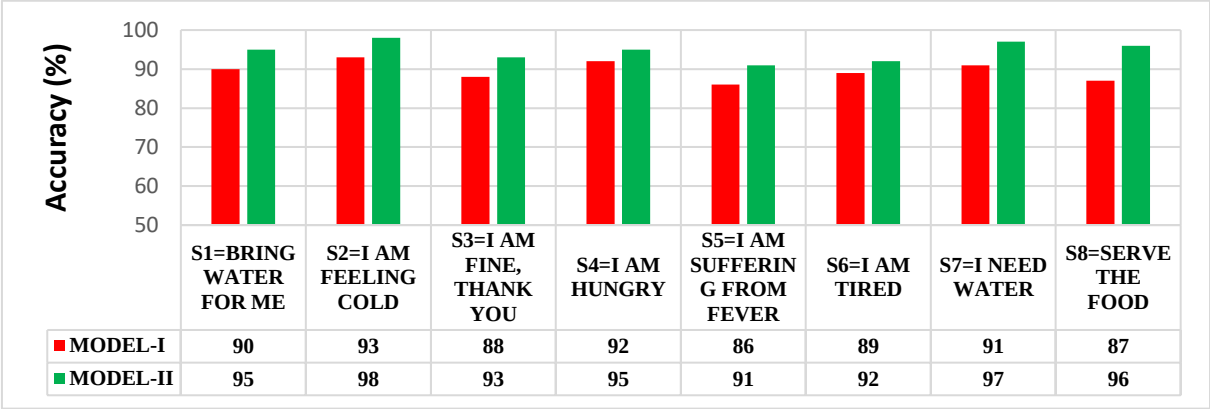


Fig. 12: Recognition rate obtained for sentence gestures

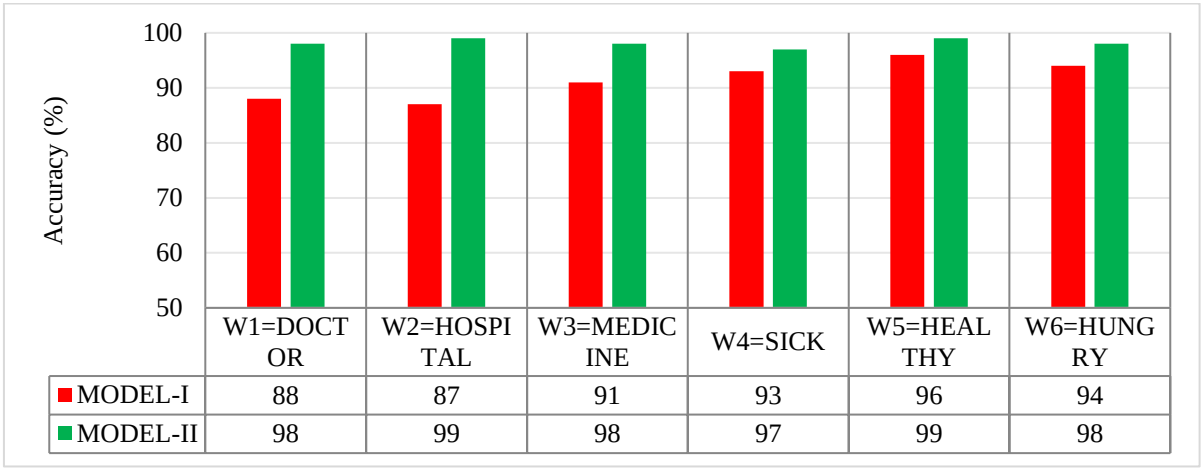


Fig. 13: Recognition rate obtained for word gestures

In the experiment, each gesture video is standardized to 240 frames per clip. For the word-level gestures, the video duration is fixed at 4 seconds, which corresponds to an approximate frame rate of 60 frames per second (fps). For the sentence-level gestures, the video duration varies between 5 to 6 seconds, which implies an approximate frame rate of 40–48 fps. Specifically, a 6-second gesture clip with 240 frames corresponds to 40 fps, while a 5-second clip with 240 frames corresponds to 48 fps. In Model-I, the LSTM component processes the sequence of spatial features extracted by 3DCNN, modelling temporal dependencies between consecutive frames by maintaining a hidden state. This ensures the incorporation of information from previous frames into subsequent analysis, effectively capturing temporal patterns and dependencies. In Model-II, the TimeSformer operates on the spatial feature sequence from Dense Net, utilizing self-attention mechanisms to capture long-range dependencies and temporal relationships across frames. This allows the model to maintain connections between frames, capturing intricate temporal patterns and dependencies within the video sequence, even across distant time steps. The disparity in accuracy between word-level and sentence-level gestures is primarily attributed to the temporal complexity inherent in each gesture type. Word-level gestures involve shorter and simpler temporal sequences, corresponding to individual words or short phrases. In contrast, sentence-level

gestures encompass longer and more complex temporal patterns, representing complete sentences composed of 3 to 5 words. This increased temporal variability poses challenges for models like 3DCNN + LSTM and DenseNet + TimeSformer, impacting their ability to effectively capture and generalize from the richer temporal dynamics of sentence-level gestures. As a result, while word-level gestures are recognized more accurately due to their simpler temporal nature, sentence-level gestures experience lower accuracy despite representing concise combinations of words. Also, the variation in the accuracy could be because of the complexity in the gesture movement based on the spatial and temporal features. Pre-trained DenseNet weights provide a strong starting point for feature extraction, but they may not be optimized specifically for word and sentence-level gestures. Fine-tuning the DenseNet weights along with training the TimeSformer model on a dataset specifically focused on word and sentence-level gestures can help adapt the features to the task at hand. 3DCNNs extract low-level spatiotemporal features from raw video frames, capturing basic visual patterns and motion information. LSTM networks process the sequence of features extracted by 3DCNNs, learning higher- level representations and temporal dependencies among these features. This hierarchical approach enables the model to capture both fine-grained details and long-term temporal relationships inherent in gestures.

Conclusion and Future Scope

This study presents novel deep learning frameworks for dynamic Indian Sign Language recognition, targeting emergency communication for individuals with speech and hearing impairments. By integrating 3DCNN with LSTM networks and leveraging a DenseNet-augmented TimeSformer model, the proposed approaches capture both spatial and long-term temporal dependencies, enabling accurate recognition of 14 word- and sentence-level emergency gestures. Comparative experiments on benchmark datasets (ISL_CSLTR and INCLUDE) demonstrate that the TimeSformer model significantly outperforms the 3DCNN + LSTM approach, achieving faster convergence and superior classification performance. These findings highlight the effectiveness of combining attention-based temporal modeling with robust spatial feature extraction for emergency-focused gesture recognition. Despite these advances, the study has limitations: the gesture set is restricted to 14 classes, dataset diversity is limited, real-time deployment has not been tested, and the computational requirements may challenge resource-constrained applications. Addressing these limitations will be critical for practical implementation in assistive technologies. Future work will extend the gesture vocabulary across diverse emergency and everyday contexts, implement real-time recognition on mobile and edge devices, and explore multimodal learning that combines video with sensor or audio cues to enhance robustness under challenging conditions. By providing accurate, timely, and context-aware communication support, this research contributes a meaningful step toward real-world assistive systems for the speech and hearing impaired, with significant potential for improving emergency response and accessibility.

Acknowledgment

The authors would like to thank their respective institutions for providing the necessary support and resources to carry out this research work.

Authors Contributions

Poornima B V: Conceptualized the study, carried out the implementation, performed experiments, and prepared the original draft of the manuscript.

Srinath S: Contributed to methodology design, supervision, and manuscript review.

Mustafa Basthikodi: Assisted in data collection, preprocessing, and experimental validation.

Rashmi S: Contributed to analysis, result interpretation, and manuscript edited. provided guidance in validation, review, and overall supervision of the research work.

Rakshitha R: Provided guidance in validation, review, and overall supervision of the research work.

All authors have read and approved the final manuscript.

Disclosure Statement

The authors declare that there is no conflict of interest regarding the publication of this paper. All authors meet the authorship criteria and have approved the final version of the manuscript. The addition of the new author has been made to accurately reflect contributions to the work.

References

Agarwal, A., Gupta, A., Devadas, D., Duduskar, A., & Patil, G. (2025). Indian Sign Language Recognition using Neural Networks. *Proceedings of the 2025 5th International Conference on Artificial Intelligence and Signal Processing (AISP)*, 1–5. https://doi.org/10.1109/aisp68263.2025.11396150

Agarwal, A., Sreemathy, R., Turuk, M., Jagdale, J., & Kumar, V. (2023). Indian Sign Language Recognition using Skin Segmentation and Vision Transformer. *Proceedings of the 2023 IEEE 20th India Council International Conference (INDICON)*, 857–862. https://doi.org/10.1109/indicon59947.2023.10440818

Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). Hand Gesture Recognition for Sign Language Using 3DCNN. *IEEE Access*, *8*, 79491–79509. https://doi.org/10.1109/access.2020.2990434

Altaf, Y., Wahid, A., & Kirmani, M. M. (2023). Deep Learning Approach for Sign Language Recognition Using DenseNet201 with Transfer Learning. *Proceedings of the 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–6. https://doi.org/10.1109/sceecs57921.2023.10063044

Bansal, N., & Jain, A. (2024). Word recognition from Indian Sign Language using Transfer Learning Models and RNN Classifier. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(9s), 182–189.

Bansal, N., & Jain, A. (2024). Word recognition from Indian Sign Language in videos using dual feature descriptor and GMT-MASKRCNN recognition technique. *Multimedia Tools and Applications*, *84*(5), 2565–2597. https://doi.org/10.1007/s11042-024-20384-8

Basthikodi, M., & Poornima, B. V. (2025). Developing an explainable human action recognition system for academic environments: Enhancing educational interaction. *Results in Engineering*, *26*, 105014. https://doi.org/10.1016/j.rineng.2025.105014

Basthikodi, M., Chaithrashree, M., Ahamed Shafeeq, B. M., & Gurpur, A. P. (2024). Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques. *Scientific Reports*, *14*(1), 26023. https://doi.org/10.1038/s41598-024-77243-7

Basthikodi, M., Faizabadi, A. R., & Ahmed, W. (2019). HPC Based Algorithmic Species Extraction Tool for Automatic Parallelization of Program Code. *International Journal of Recent Technology and Engineering*, *8*(2S3), 1004–1009. https://doi.org/10.35940/ijrte.b1188.0782s319

Basthikodi, M., Prabhu G, A., & Bekal, A. (2021). Performance Analysis of Network Attack Detection Framework using Machine Learning. *Sparklinglight Transactions on Artificial Intelligence and Quantum Computing*, *01*(01), 11–22. https://doi.org/10.55011/staiqc.2021.1102

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *Computer Vision and Pattern Recognition*, 8132–8145. https://doi.org/10.48550/arXiv.2102.05095

Bhandary, A., Gurpur, A. P., Basthikodi, M., & Chaitra, K. M. (2021). Early Diagnosis of Lung Cancer Using Computer Aided Detection via Lung Segmentation Approach. *International Journal of Engineering Trends and Technology*, *69*(5), 85–93. https://doi.org/10.14445/22315381/ijett-v69i5p213

Chen, Z., Wang, S., Yan, D., & Li, Y. (2024). A Spatio-Temporl Deepfake Video Detection Method Based on TimeSformer-CNN. *Proceedings of the 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 1–6. https://doi.org/10.1109/icdcece60827.2024.10549278

Das, A., Gawde, S., Suratwala, K., & Kalbande, D. (2018). *Sign Language Recognition Using Deep Learning on Custom Processed Static Gesture Images*. Proceedings of the 2018 International Conference on Smart City and Emerging Technology (ICSCET). https://doi.org/10.1109/icscet.2018.8537248

Das, S., Biswas, S. Kr., & Purkayastha, B. (2023). Indian Sign Language Recognition System for Emergency Words by Using Shape and Deep Features. *Processing of the 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, 1–6. https://doi.org/10.1109/iemecon56962.2023.10092312

Elakkiya, R., & Natarajan, B. (2021). ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition". *Mendeley Data*.

Geetha, M., Aloysius, N., Somasundaran, D. A., Raghunath, A., & Nedungadi, P. (2025). Toward Real-Time Recognition of Continuous Indian Sign Language: A Multi-Modal Approach Using RGB and Pose. *IEEE Access*, *13*, 60270–60283. https://doi.org/10.1109/access.2025.3554618

Jayanthi, P., Sathia Bhama, P. R. K., Bhama, S. B., & Madhubalasri, B. (2023). Sign Language Recognition using Deep CNN with Normalised Keyframe Extraction and Prediction using LSTM. *Journal of Scientific & Industrial Research*, *82*(07), 623–632.

Katti, R. K., Chiplunkar, S., Desai, P., & Gopalan, S. (2023). Character and Word Level Gesture Recognition of Indian Sign Language. *Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/i2ct57861.2023.10126314

Lahiani, H., & Neji, M. (2018). *Hand gesture recognition method based on HOG-LBP features for mobile devices*. Procedia Computer Science. https://doi.org/10.1016/j.procs.2018.07.259

Liang, Z., Li, H., & Chai, J. (2023). *Sign Language Translation: A Survey of Approaches and Techniques*. Electronics. https://doi.org/10.3390/electronics12122678

Liao, Y., Xiong, P., Min, W., Min, W., & Lu, J. (2019). Dynamic Sign Language Recognition Based on Video Sequence With BLSTM-3D Residual Networks. *IEEE Access*, *7*, 38044–38054. https://doi.org/10.1109/access.2019.2904749

Mariappan Hariharan, M., & V, G. (2021). Indian Sign Language Recognition through Hybrid ConvNet-LSTM Networks. *EMITTER International Journal of Engineering Technology*, *9*(1), 182–203. https://doi.org/10.24003/emitter.v9i1.613

Masood, S., Srivastava, A., Thuwal, H. C., & Ahmad, M. (2018). Real-Time Sign Language Gesture (Word) Recognition from Video Sequences Using CNN and RNN. In A. Abraham, V. Snášel, A. E. Hassanien, J.-S. Pan, & S. Kozielski (Eds.), *Proceedings of International Conference on Intelligent Systems Design and Applications (ISDA 2018)* (1st ed., Vol. 941, pp. 623–632). Springer Singapore. https://doi.org/10.1007/978-981-10-7566-7_63

Mishra, A., Gupta, S., Goel, D., & Tiwari, V. (2023). ISL Recognition of Emergency Words Using MediaPipe, CNN and LSTM. *Proceedings of the 2023 International Conference on Power Energy, Environment & Intelligent Control (PEEIC)*, 322–325. https://doi.org/10.1109/peeic59336.2023.10450425

Mohamed, N., Mustafa, M. B., & Jomhari, N. (2021). A Review of the Hand Gesture Recognition System: Current Progress and Future Directions. *IEEE Access*, *9*, 157422–157436. https://doi.org/10.1109/ACCESS.2021.3129650

Nadaf, A. I., Pardeshi, S., & Gupta, R. (2025). Efficient gesture recognition in Indian sign language using SENet fusion of multimodal data. *Journal of Integrated Science and Technology*, *13*(6), 1145. https://doi.org/10.62110/sciencein.jist.2025.v13.1145

Natarajan, B., & Elakkiya, R. (2022). Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks. *Soft Computing*, *26*(23), 13153–13175. https://doi.org/10.1007/s00500-022-07014-x

Natarajan, B., Elakkiya, R., & Prasad, M. L. (2023). Sentence2SignGesture: a hybrid neural machine translation network for sign language video generation. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 9807–9821. https://doi.org/10.1007/s12652-021-03640-9

Natarajan, B., Rajalakshmi, E., Elakkiya, R., Kotecha, K., Abraham, A., Gabralla, L. A., & Subramaniyaswamy, V. (2022). Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation. *IEEE Access*, *10*, 104358–104374. https://doi.org/10.1109/access.2022.3210543

Pai, P., Amutha, S., Basthikodi, M., Ahamed Shafeeq, B. M., Chaitra, K. M., & Gurpur, A. P. (2025). A twin CNN-based framework for optimized rice leaf disease classification with feature fusion. *Journal of Big Data*, *12*(1), 89. https://doi.org/10.1186/s40537-025-01148-z

Prathap, K. B., Swaroop, G. D., Kumar, B. P., Kamble, V., & Parate, M. (2023). ISLR: Indian Sign Language Recognition. *Proceedings of the 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, 1–6. https://doi.org/10.1109/pcems58491.2023.10136062

Rudregowda, S., Arunakumari, B. N., Manjunath, A. S., & Manibyrappa, R. (2022). Indian Sign Language Recognition Using 2-D Convolution Neural Network and Graphical User Interface. *International Journal of Image, Graphics and Signal Processing*, *14*(2), 61–73. https://doi.org/10.5815/ijigsp.2022.02.06

Saleh Musa Miah, A., Shin, J., Hasan, A. M., Rahim, A., & Okuyama, Y. (2023). Rotation, Translation and Scale Invariant Sign Word Recognition Using Deep Learning. *Computer Systems Science and Engineering*, *44*(3), 2521–2536. https://doi.org/10.32604/csse.2023.029336

Salins, R. D., Ashwin, T. S., Prabhu, G. A., Basthikodi, M., & Mallikarjun, C. K. (2022). Person identification from arm's hair patterns using CT-twofold Siamese network in forensic psychiatric hospitals. *Complex & Intelligent Systems*, *8*(4), 3185–3197. https://doi.org/10.1007/s40747-022-00771-0

Samal, B., & Panda, M. (2021). Integrative Review on Vision-Based Dynamic Indian Sign Language Recognition Systems. *Proceedings of the 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, 1–6. https://doi.org/10.1109/ODICON50556.2021.9429002

Sánchez Ruiz, D., Olvera-López, J. A., & Olmos-Pineda, I. (2023). Word Level Sign Language Recognition via Handcrafted Features. *IEEE Latin America Transactions*, *21*(7), 839–848. https://doi.org/10.1109/tla.2023.10244183

Sharma, S., Gupta, R., & Kumar, A. (2023). Continuous sign language recognition using isolated signs data and deep transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 1531–1542. https://doi.org/10.1007/s12652-021-03418-z

Singha, J., Roy, A., & Laskar, R. H. (2018). *Dynamic hand gesture recognition using vision-based approach for human–computer interaction*. Neural Computing and Applications. https://doi.org/10.1007/s00521-016-2525-z

Sridhar, A., Ganesan, R. G., Kumar, P., & Khapra, M. (2020). INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. *Proceedings of the 28th ACM International Conference on Multimedia*, 1366–1375. https://doi.org/10.1145/3394171.3413528

Sridhar, A., Ganesan, R. G., Kumar, P., & Khapra, M. (2020). INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. *Proceedings of the 28th ACM International Conference on Multimedia*, 1366–1375. https://doi.org/10.1145/3394171.3413528

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *Neural and Evolutionary Computing*. https://doi.org/10.48550/arXiv.1909.09586

Tatebe, Y., Deguchi, D., Kawanishi, Y., Ide, I., Murase, H., & Sakai, U. (2018). Pedestrian detection from sparse point-cloud using 3DCNN. *Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT)*, 1–4. https://doi.org/10.1109/iwait.2018.8369680

Urabe, S., Inoue, K., & Yoshioka, M. (2018). Cooking activities recognition in egocentric videos using combining 2DCNN and 3DCNN. *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*, 1–8. https://doi.org/10.1145/3230519.3230584

Venugopalan, A., & Reghunadhan, R. (2021). Applying deep neural networks for the automatic recognition of sign language words: A communication aid to deaf agriculturists. *Expert Systems with Applications*, *185*, 115601. https://doi.org/10.1016/j.eswa.2021.115601

Venugopalan, A., & Reghunadhan, R. (2023). Applying Hybrid Deep Neural Network for the Recognition of Sign Language Words Used by the Deaf COVID-19 Patients. *Arabian Journal for Science and Engineering*, *48*(2), 1349–1362. https://doi.org/10.1007/s13369-022-06843-0

Zhao, W., & Du, S. (2016). Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4544–4554. https://doi.org/10.1109/tgrs.2016.2543748