IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 12, December, 2025, pp. 3301-3311
Journal homepage: http://iieta.org/journals/isi

Hybrid CNN-Transformer for Dynamic Indian Sign Language Recognition with Non-Manual Gesture Analysis

Purva Badhe[1*], Vaishali Kulkarni[2]

[1] Department of AIML, D J Sanghvi College of Engineering, Vile Parle West, Mumbai 400056, India
[2] Department of AI, MPSTME, NMIMS University, Vile Parle West, Mumbai 400056, India

Corresponding Author Email: purva.pcb@gmail.com

https://doi.org/10.18280/isi.301222

Received: 3 November 2025
Revised: 23 December 2025
Accepted: 27 December 2025
Available online: 31 December 2025

Keywords:
sign language translation, gesture recognition, Indian Sign Language, video-based recognition, hybrid convolutional neural network, vision based attention transformer model

ABSTRACT

Sign language is a component element of communication between the mute and hearing-impaired communities that are indispensable to them, but it is mostly closed-off to the general population. To step into that gap, the current paper outlines the design of a hybrid Vision Transformer-Convolutional Neural Network system, officially focused on Indian Sign Language (ISL) gesture recognition, strong dynamic gestures, and face muscles. The edited database is 1 100 video samples in 22 different classes, which were recorded in the heterogeneous environmental conditions, to provide the robustness. The empirical findings indicate that the hybrid model has an exemplary training accuracy of 100, validation accuracy of 88.6, and a test accuracy of 82.14 and thus outperforms the state-of-the-art that provides accuracy of 88.7 to 92% of training accuracy. Proposed system thus achieves enhanced accuracy by 7-11% in case of continuous sign gestures. Through this, inclusivity and accessibility to the deaf community are thereby enhanced and future possibilities involve data enhancement as well as the integration of NLP-based text-to-speech synthesis.

1. INTRODUCTION

Sign language is a highly dense, visually based interaction, whose effectiveness is predetermined by a highly organized system of hand gestures in space and the concomitant use of musculature of the face, eye movements, articulatory gestures, and other non-manual signatures, which gives the speech of sign language an artistic shift of power [1]. There have arisen all over the world three hundred or so different sign systems, each of them diversifying iteratively in the specificity of the geographical region, the texture of the culture, and the subtlety of the language. Similar to the idea that is presented in spoken tongues of lingering dialects and accent, sign languages display substantial deviation when it comes to gesture construction, sequential correctness, idiomatic expression, and non-manual semiotic adornment. Especially, it is worth mentioning that Indian Sign Language (ISL) takes a prominent place in the South Asian linguistic landscape, which is the symbol of diversity and epistemic richness that defines the spectrum of sign languages [2]. Whereas sign languages can be coarsely categorized into the group of static gestures, where the hand shapes are created by one or both hands, and the group of dynamic gestures, where the temporal movement of the hands and the variations of the expression are added, the overwhelming part of the extant research has either overemphasized the former classification or oversimplified the current systems of gestures. This chauvinism poses a great gap in full identification of the broader spectrum of sign language that embraces facial expression and other more elaborate body languages [3]. Furthermore, the dominating paradigms are based on the recognition of transient manual movements or on those that are written in a speech form, thus overlooking the complete repertoire of sign-language communication and limiting the effectiveness of human-computer interaction systems that have been developed to correspond with the deaf community. There is an urgent need for powerful sign language recognition systems that go beyond the rigid gestures and written texts. We aim to develop a state-of-the-art solution that could understand dynamic sign language sufficiently well to include subtle facial expressions, among others, with the employment of state-of-the-art deep-learning algorithms. Precisely, this paper aims to:

1. Propose and combine a Vision Transformer (ViT) to enhance the Convolutional Neural Network (CNN) ability to identify sign language in video footage.
2. Use a sample of naturally deaf and trained Sign Language speakers so as to incubate the intricacy of sign languages as they are truly utilized.
3. A gap in the existing work of research will be resolved by focusing on dynamic, continuous signs with non-manual attributes, thus going beyond the static signs and crude gestures that have been prevalent in the literature.

The contribution to this paper is introducing a video-based sign-language recognition model which uses Vision Transformer to process dynamic signs, facial expressions of nuanced appearances and other non-manual societal readings. In this way, we get the field to a stage where it is not constrained by its own limitations, where more inclusive and effective solutions to human-computer interaction are possible that realize the richness and the complexity of the sign-

language interaction. After this introduction, Section 1 is the review of related work in sign-language recognition. Section-2 explains the suggested methodology and elaborates on how it will address the challenges mentioned. Section 3 describes the architecture of Vision Transformer and its video recognition application in detail. Section 4 describes and discusses the results of the experiment. Finally, Section 5 wraps up this research providing a summary of the main findings and future research avenues.

2. RELATED WORK

Early manual gesture recognition has been based on compositional and model-based methods. As an example, Heap and Hogg [4] built up a hand model that is deformable and in 3D, and considered the Principal Component Analysis (PCA) to align a dynamically scaled template to observed images, allowing tracking motion of a hand in real-time. Although this method served as a great base to have precision in tracking, it was unable to manage scale, rotation, and occlusions. Complementary compositional practices also employed principles of perceptual grouping in describing hand postures by combinations of hand parts but these too had pragmatic obstacles as complex scenes were dealt with. The later works dealt with the static recognition of gestures using contours and special hardware. Lee and You [5] made use of the wrist bands to segment the region of the hand and used an algorithm to match and classify. They were however sensitive in terms of background color and not robust in problematic environments. Chevtchenko et al. [6] also used a multi-objective evolutionary algorithm for the features sets and dimensions optimization, using Gabor filters and Zernike moments to reach accuracies up to 97.63% in 36 static gestures in a position. Huang et al. [7] target interpreting sign language into text or speech using a novel 3D CNN method automatically extracting discriminative spatial-temporal features from raw video streams. Deep learning models triggered the adoption of more advanced architectures as the transition was made to video-based recognition and dynamic gestures. Vision Transformers (ViT) was an attractive alternative to Conventional Neural Networks (CNNs), and sometimes they outperformed them in precision and speed. On the same note, Lai and Yanushkevich [8] used CNNs together with recurrent neural networks (RNNs) to use both the spatial and time data, and they obtained the highest accuracy of 85.46% with depth and skeleton data. Kamruzzaman [9] used ResNet50 and MobileNetV2 to do Arabic sign language achieving a combined accuracy of 98.2. Based on data augmentation mechanisms through CNNs, Zakariah et al. [10] and Zhang et al. [11] boosted the American sign language recognition and reached an accuracy of 99.52%. These strategies were further applied by other scientists to other sign languages and methods of feature extraction. Recently, De Coster et al. [12] incorporated OpenPose with a multi-head attention mechanism to get 74.7 percent accuracy on Flemish Sign Language. Vaswani et al. [13] designed a CNN that was used to identify hand gestures in small scale image begging mind hand gestures on a simple background, which achieved an accuracy of 97.1%. Shenoy et al. [14] performed skin color segmentation and grid-based feature extraction to identify the ISL gestures which were using k-nearest neighbors (KNN) and hidden Markov models (HMM) with remarkable success. Katoch et al. [15] used a Bag of Visual Words (BOVW) model that was coupled with CNNs and SVMs to recognize ISL letters and digits whereas Rokade and Jadav [16] used a combination of skin color-based segmentation with artificial neural networks (ANN) and SVMs to obtain robust fingerspelling recognition in ISL. Nanivadekar and Kulkarni [17] created an ISL database and proposed a hand tracking and segmentation based on three step algorithm. Badhe and Kulkarni [18] implemented an ISL gesture translator using hand tracking with combinational algorithm and recognition done using template matching. Badhe and Kulkarni [19] have proposed handcrafted feature extraction method for SL recognition where complex grammatical rules are captured with 98% accuracy. In addition to hand gestures, Kashika and Venkatapur [20] used deep learning as a way to detect objects on the panoramic video frame and Tran et al. [21] studied face recognition relying on SVM, but another team [22] proposed a new method of detecting objects. Sreemathy et al. [23] showed that deep learning could be used as an ICT method of identifying the signs of the ISL in English. Das et al [24] combined the handcrafted features and CNN-extracted features to counter the problem concerning the same hand orientations and different viewing angles. Sharma et al. [25] emphasized that transfer learning is efficient in the context of sign language recognition, and Liu et al. [26] explored the detection transformers which also has a feature extraction pyramid network in order to improve recognition performance. Al Essa et al. [27] proposed an approach of multi connect associative memory for recognition of American Signs. This approach solved a problem of misclassification of static signs which are too similar in gestures.

This changing arena of methodologies between early hand-modeling methodologies and more recent transformer-based architectures is indicative of a dynamic research domain. The limits of hand gesture and sign language recognition are constantly being extended with the integration of Vision Transformers, more advanced CNNs, and hybrid models, as more focus is placed on more advanced and sophisticated solutions. It is hoped that these developments can enhance the precision, flexibility, and applicability of sign language systems, which will favor more inclusive communication, and broaden the possibility of human-computer reaction under a variety and dynamic setting.

3. METHODOLOGY

Conventional CNN-based models which have historically been used in image and video tasks have good local feature extraction properties. Nevertheless, they in many cases use sequential feature summation (e.g., through RNNs or 3D convolutions) to do so, which can be costly and not always optimal to establish long-range correlations in videos. Whereas purely transformer-based schemes (ViT) can capture both the global and temporal context and find local differences, they can ignore small variations in local context that can distinguish similar gestures.

Through the incorporation of CNN layers into the ViT structure we have been able to maintain the local pattern recognition capabilities of CNNs whilst still exploiting the ViT capability to capture multi-frame complex temporal and contextual interactions. This synergy does enhance recognition and is especially accurate with complex ISL gestures which are based on both fine-grained hand configurations and complex temporal patterns.

CNN-Only Models:

CNN-based methods in pure form might be too restricted in terms of ability to realize the temporal global context unless other components, including RNNs or 3D convolutions, are added to them. In our initial experiments, we found that, because of the rich spatial feature retrieval capabilities of 3D CNNs, their performance in interpreting long sequences of behaviors was lower than that of the ViT-based counterparts.

RNN or LSTM-Based Models:

Though RNNs or LSTMs are capable of modeling temporal sequences, they can be more susceptible to such problems as vanishing gradients across long videos and even less efficient than attention-based models. Initial experiments with CNN-LSTM hybrids provided decent performance at the price of increased training time and reduced modeling capabilities of complicated time-dependent dependencies.

Pure ViT Models:

Pure ViT models are superior in modeling long-range dependencies. Nevertheless, they do not have any local feature extraction element and hence do not pick up finer details such as minute movements of the fingers or micro-expressions on faces. Experiments on our part showed that the general fine-grained recognition accuracy was enhanced with the addition of CNN layers, especially in harsh backgrounds.

Altogether, the hybrid ViT-CNN solution provides a middle ground solution based on global time modeling (transformers) and local space feature extraction (CNNs). Empirical studies demonstrate that this hybrid architecture outperforms purely CNN-based or purely ViT-based models and are comparable to CNN-RNN hybrids in accuracy, efficiency and generalization to various ISL gestures.

3.1 Flow of the study

The current research aims to develop an effective ISL recognition application that could decode dynamic hand gestures and facial expressions on the basis of video streams correctly. Where a large number of research focuses on gestural frames or static gestures, this research focuses on continuous gestural signs, including small video recordings of 1 to 3 seconds. In this direction, we choose a hybrid neural network combining the capabilities of global attention of a ViT with the potential of a local feature-extraction of a CNN. The whole processing chain is described in Figure 1 and summarized in the following parts:

3.1.1 Video capture

At the first stage, we obtain a rich set of video recordings of deaf subjects instructing a repertoire of predetermined ISL gestures in a range of lighting scenarios and background settings, which is used to strengthen the later system against environmental covariate influences.

3.1.2 Frame extraction

The videos that are captured are then divided into discrete frames; these temporal snapshots are the two elements of the manual component, i. e., the hand kinematics, and the non-manual component, i. e., the facial musculature, both of which cannot be done without reading the signs in any manner.

3.1.3. Frame counting and padding

We also come up with the frame tally of both records and the temporal maximum over the corpus; those videos that are shorter are electronically padded to that temporal maximum which normalizes the input dimension of all further processing phases.

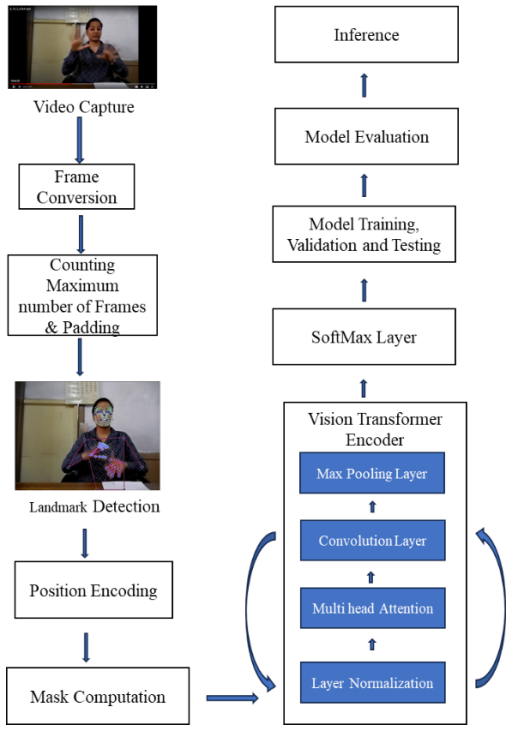

Figure 1. Block diagram

3.1.4 Landmark detection

A Boolean landmark-detection processes on each frame provide a list of salient features, most likely to occur around the hands and face of the signer, that provide a structured, fine-grained image of the gestures and subtle facial expressions. A similar approach is seen in the work of Jo et al. [28] for enhancing gestural interaction used in virtual and augmented reality with Media-Pipe based gesture recognition interface.

3.1.5 Position encoding

The network takes spatial context through positional encodings to the retrieved landmarks; this process embeds inter-point spatial relations and relative positions of anatomy parts into the model and provides it with a better understanding of gestural structure.

3.1.6 Mask computation

A saliency mask that gives more weight to the hand, face, and other regions of interest is synthesized by us and effectively reduces background clutter, sensor noise, and focal capacity of the model, focusing attention on the informative regions of the spatial map.

Development: ViT architecture: ViT is an architecture using artificial intelligence (AI) to recognize images as objects and extract information from them.

Motivation: ViT architecture: ViT is a type of architecture based on AI that identifies objects in images and derives information about objects contained within the image. Instead of defining images, or sequentially ordered frames, as a set of convolutional filters, Vision Transformers conceptualize images as a set of tokenized patches and use multi-head self-attention systems to encode long-range correlations and contextual interactions. We then subdivide the extracted frames (or their spatial encodings, e.g., landmark-based ones) into small patches, and in this manner, a linear projection of

the patch into a high-dimensional embedding space and positional embedding is generated.

3.1.7 Feature fusion

The unified approach of combining CNN with Transformer follows a sequential feature encoding and fusion mechanism. For each input video sequence, individual frames are first processed by a backbone of CNN, which extracts fine-grained spatial features corresponding to hand shape, finger articulation as well as non-manual cues like facial expressions and head orientation. The embeddings generated by CNN are then temporally arranged and sourced as token representations to the vision transformer where self-attention mechanisms model long-range sequential dependancies across frames. The transformer output is lastly fused with the CNN features by concatenating before classification.

Mathematical Formulation of the model:

Let

$$X = \{x1, x2, x3, \ldots., xT\} \quad (1)$$

be the video with T frames.

CNN feature extraction would be

$$f_t = CNN(x_t);\ where\ f_t \in \mathbb{R}^d \quad (2)$$

Stacked spatial matrix is represented by:

$$F = [\, f_1, f_2, f_3, \ldots\ldots, f_T] \quad (3)$$

Position encoding is depicted by

$$Z = F + P \quad (4)$$

where, P is Positional Encoding Matrix and Z is position aware feature representation. The Transformation attention spaces Query (Q), Key (K) and Value (V) are represented as

$$Q = ZW_Q, K = ZW_K \text{ and } V = ZW_V \quad (5)$$

The transformer output, Fusion representation and Classification equation is :

$$T_f = Transformer\ (Z) \quad (6)$$

$$F_{fusion} = [Pool(F)\ ;\ Pool(T_f)] \quad (7)$$

$$y = Softmax\ (WF_{fusion} + \mathrm{b}) \quad (8)$$

where, Pool represents Global maxpooling and [;] indicated concatination of feature vectors.

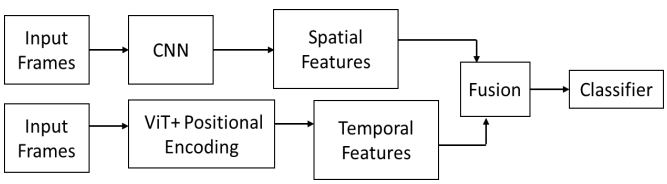

Figure 2. CNN- ViT fusion mechanism

The method of fusion is as depicted in Figure 2. This fusion design enables the model to preserve fine – graned spation discrimination through CNN feature extraction while simultaneously exploiting the Transformer's ability to capture temporal features. This synergy is particularly beneficial for dynamic ISL gesture recognition.

3.1.8 ViT and adapting to the Desired Task:

(a) Temporal Patching: Each frame or the set of landmarks that accompanies the frame is treated as a separate token, thus the transformer can focus on the spatial axis and simultaneously the time axis.

(b) Temporal Positional Embedding: It makes the network self-temporal: the network is given the ability to represent its dynamic properties of motion and gesture development over time by encoding order in vectors indicating frame chronology. The application of CNN Elements and ViT Behavior:

Although ViT is highly successful in observing the dependencies of the world, its peculiarity of using pure attention might not always adequately reflect local, subtle forms (such as fine grasping configurations or subtle facial micro-expressions). In order to address this weakness, we utilize CNN modules as part of the ViT pipeline. In everything hybrid, raw frames pass through a lightweight CNN that isolates salient primitives in space: edges and textures, contours, etc., before being subjected to the transformer. CNN feature maps such that result are in turn fused with the ViT embedding, allowing the transformer layers to have access to the enriched inputs that combine the benefits of global context framing with the benefits of distilled local detail.

3.1.9 SoftMax layer

The unprocessed output of the hybrid ViT-CNN processing is then fed into a SoftMax classifier to recede to an emergent representation, enabling categorical probabilities of every gesture classes repertoire of ISL to be produced, hence allowing decisive identification of signs with each input sequence. To identify the relationship, the model undertook and its strength, the tests involve training, validating, and testing the model.

The ViT-CNN architecture was trained and optimized on the curated dataset of ISL and the following hyper-parameters and regularization options were used:

o Learning Rate: Learning rate at the beginning will be 1e - 4 and as plateau in the validation accuracy is reached, it will be decreased by a factor of 0.1.

o Number of Epochs: 50 -100, based on convergence patterns.

o Batch Size: 8-16, which was selected according to the available memory of the GPUs and stability of the training.

o Regularization Techniques: To address overfitting, dropout layers (dropout rate of 0.3–0.5) are included not only in the CNN layers but also in the transformer ones. Also, early stopping and data augmentation are applied (e.g. random cropping, limited rotations) to guarantee improved generalization.

3.1.10 Model scoring

After the training process, a detailed assessment of the withheld test split is performed where we calculate conventional performance measures such as accuracy, precision, recall and the F1-score to objectively assess the efficacy of the system in both controlled laboratory and real world background circumstances. This test confirms that the model can be dependable in distinguishing between twenty-two different classes of ISL gestures.

3.1.11 Inference

When its validation with sufficient accuracy is achieved, the hybrid ViT-CNN pronounces a real-time inference feature. Once a new video stream is consumed, the system will run landmark extraction, positional encoding, and mask generation sequentially and run the content through the hybrid network, producing an estimation of a gesture and a confidence estimate in the end.

4. IMPLEMENTATION

4.1 Database

To work out the indispensable basis of further analyses, an ISL database was built strictly under a carefully curated one in the absence of a widely realized standardized corpus. A collection of 1,100 video recordings was made of the Ali Yavar Jung National Institute of Hearing Handicapped in Bandra, Mumbai, of 22 different ISL gestures executed by ten deaf signers. As our research objective demands to incubate the intricacy of sign language in its true sense, we ensured that the signers are naturally deaf and trained by an authentic ISL educator. The age group of the signers is 16 to 35 and it includes both male and female users. To make the system robust to variations like background, lighting conditions and signer bias, we recorded the videos in various backgrounds like – Classrooms, Personal desks or Cubicles. The lightning condition was not controlled. Also, sampling ensured to enclude left domninant as well as right-dominant users. Some recorded gestures are as shown in Figure 3.

Figure 3. Twenty-two Indian sign language gestures (Image format)

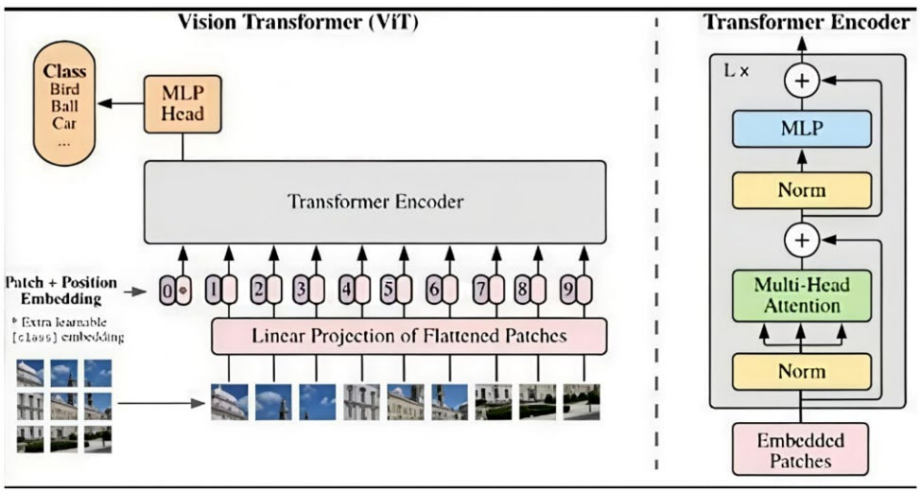

Figure 4. Vision Transformer (ViT) Architecture (Source: https://viso.ai/deep-learning/vision- transformer-vit/)

4.2 Vision transformer implementation

ViT as shown in Figure 4 was thematicized to accept video analysis by dividing each frame into temporal patches, projecting the patches into linear representations, and encoding the latent with positional encodings through which the temporal structure of the video can be identified in the footage. The resulting ViT encoder, which includes the Layer Normalisation, Multi-Head Self-Attention, and Multi-Layer Perceptrons, converts the tokenised embeddings into an overall representation that can be subsequently used by classification tasks occurring downstream.

The hybrid model provides the advantage of using both the global attentional attributes of ViT and fine-grained local feature extraction of CNN layers when combined with CNN features. The empirical analysis in the following passages indicates that this combined method is more accurate in recognition than the baseline CNN and CNN-RNN methods. When recovered with the local detail extraction by the CNN, the ViT would provide its ability to model long-range temporal dependencies, in the form of an architecture that is very sensitive to the complexities and subtleties of ISL gestures.

In order to come up with a very efficient hand-gesture recognition framework based on the Vision Transformer, a systematic approach was adopted that entailed careful data segmentation, intensive training, meticulous validation and extensive testing as well. The second part of the paper outlines the approach used to divide the available data, training, validation, and evaluation processes of the ViT recognition model.

Dataset Splitting

It is comprised of a corpus of 1,100 videos (portraying 22 different ISL gestures) that were recorded in varying conditions on ten deaf signers. To obtain a firm estimation of the model performance and ensure that there is generalisation to new data, the data was categorised into three mutually exclusive data subsets using subject wise split technique, including training data, validation data and testing data. Samples from the same subset were not shared across various subsets. The subjects used for training and validation were completely excluded from the testing subset ensuring that the model is evaluated on unseen subjects, reflecting real world deployment scenario.

Training Set (80%):

The training subset was provided with approximately 880 videos. This large assignment ensures that this model is left open to numerous background variations, signer idiosyncrasy and subtle gesture dynamics. This diversity makes it easy to extract meaningful spatial-temporal patterns, and the resulting learning of the complex hand movements and facial expressions that are characteristic of ISL recognition becomes resilient.

Validation Set (15%):

There was also a set of 165 videos that would be held back as a validation set. This data, used in isolation from the training phase, is used to monitor the performance of the model in an iterative way. The validation set provides feedback in time by evaluating the accuracy, loss and (where applicable) specialised metrics at the end of each training epoch. Whenever metrics become stagnant or worse off, then it is an indication that metrics require adaptations of hyperparameters, architectural parts or regularisation methods to reduce either over-fitting or under-fitting.

Testing Set (5%):

Finally, the testing set of 55 videos was left to be included in the final set and serve as a purely unseen control. This conclusive analysis establishes the capacity of the model to extrapolate under new cases and provides an approximate estimate of its effectiveness in the real world and practical situations. With the help of a small but representative test set, end metrics such as accuracy, precision, recall, and F1- score are exact measures of the performance of the model on unseen data.

The selected split ratios summarise a trade-off between maximizing training data to encourage robust learning and having adequate examples not seen to be validated and tested. Even though the fraction of the test can be viewed as small, the videos of the 55 types altogether are a headlong summation of the variegated nature of the dataset and still leave the testing phase as a strict and unbiased indicator of performance.

The Vision Transformerbert model was then trained by starting with the 880 training videos. All the videos were pre-processed into homogenous temporal patches and positional embeddings and then fed to the ViT. The CNN modules used were linked together to obtain local spatial aspects but the attention mechanisms of the ViT extracted it alongside the long-range association and time connection between frames.

Parameters and Procedures of training.

- Epochs: 50100 most common, and early stopping occurred when validation measures stopped improving over a specified patience (e.g., ten epochs).
- Learning rate: To start with, the learning rate is initialized to approximately 1 -10-4 times and decreased by the same factor each time a plateau is reached in the validation accuracy.
- Batch Size: Eight to sixteen, with a compromise between speed and stability of training, due to the limitation of using the GPU memory.

Regularisation

- Dropout: CNN and ViT layers were applied with a rate of 0.3-0.5 to prevent over-fitting through the elimination of co-adaptation of features.
- Data Augmentation: Mild randomly spaced spatial transformations (e.g., cropping and small rotations) were applied as mechanisms to improve robustness and generalisation.
- Early Stopping: Training was terminated when validation metrics stopped improving with the increase of the number of epochs and did not lead to needless over-training and wasted computations.

Validation Process

The model prediction on the 165 validation videos after each epoch was determined. Accuracy, validation loss, and, when it is applicable, precision-recall metrics were considered core metrics since they must identify core issues in class imbalance or particular difficulties in gestures. The differences in these metrics were used to perform hyper-parameter optimization and architecture-level changes, including learning rate schedule or dropout rate modulation.

Model Testing

After training and validation had been done, the model was tested on the 55-video test subset. These samples that were not observed during training and validation provided a true measure of generalisation. Gesture predictions of the model were compared with ground-truth gestures and the ultimate performance indicators, accuracy, precision, recall, and F1-score, were calculated. These outcomes have been compared to the existing practices and reported to exemplify the effectiveness of the model based on ViT.

Outcome and Significance

The study provides credible evidence of the validity and applicability of the trained ISL recognition model by adopting the comprehensive approach, including a reasonable division of data, extensive hyper-parameter optimization through the use of validation, and a strict final analysis of the test on unknown data. The attained output highlights the potential of Vision Transformers especially with CNN components, to drive sign language recognition systems to the next stage of being more inclusive and accessible communicative technologies.

Model Evaluation

An integrated assessment plan was used to effectively evaluate the performance of the Vision Transformer-based sign language recognition model. The evaluation involved a set of measures, such as accuracy, as well as precision, recall, and F1-score, thus providing detailed information about the model effectiveness and efficacy.

Accuracy: Accuracy is a fundamental metric that measures the proportion of correctly predicted instances from the total instances in the testing dataset. Itis a primary indicator of the model's overall correctness in recognizing hand gestures [21]. Mathematically, accuracy is defined as the ratio of true positive (TP) and true negative (TN) predictions to the total number of predictions:

$$Accuracy = \frac{TP+TN}{TOTAL\ PREDICTIONS} \quad (9)$$

Precision: Precision gauges the model's ability to correctly identify positive instances (correctly recognizing a specific hand gesture) among all instances predicted as positive. It focuses on the model's propensity to avoid false positives, i.e., instances wrongly classified as positive. Precision is calculated using the formula:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Recall (Sensitivity): Recall, also known as sensitivity or the true positive rate, quantifies the model's capacity to correctly identify positive instances from all actual positive instances [21]. This metric highlights the model's ability to capture all relevant occurrences of a particular hand gesture. A recall is calculated as:

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

F1 Score: The F1 score is a harmonic mean of precision and recall, providing a balanced assessment of the model's performance by considering false positives and false negatives. It offers a single metric considering Type I (false positive) and Type II (false negative) errors. The F1 score is calculated as follows:

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (12)$$

The capabilities of the ViT based sign language recognition model can be holistically evaluated with the help of these metrics. High accuracy indicates a strong overall performance and precision identifies the model accuracy in performing a positive prediction. As mentioned, recall highlights the effectiveness of the model in capturing all the positive instances, whereas F1 score provides a balanced trade-off of precision and recall to rightfully consider that between false positive and false negative.

The evaluation measures presented in the model based on the testing data allow a deep insight into its shortcomings and advantages. Moreover, these measures allow making strict parallels with the current methods and standards, which allows obtaining useful information about the possible practical implementation of the ViT-based hand gesture recognition system into practice related to the interaction through the ISL.

5. RESULTS

This paper is dedicated to strict analysis and subtle understanding of naturally pre-established ISL frameworks, and the major aim to outline the gap in communication between the disabled population and the rest of the community.

The study combines ViT modules with traditional CNN models by building an indigenous ISL dataset by recording participants who are deaf and therefore enhancing the recognition accuracy.

The sensitive application of encoder transformer avoids the use of complicated data preprocessing, and the discriminating addition of the ViT highlights the strength and performance measure of the model.

Model Summary:

Table 1 provides a concise overview of the model architecture and the cumulative trainable parameters.

Table 1. Model summary

Layer (Type)	Output Shape	Parameters #
Input_1 (Input Layer)	(None,144,258)	0
Frame_position_embed ding	(None,144,258)	37152
Transformer_layer	(None,144,258)	270646
Global_max_pooling1d	(None,258)	0
Dropout	(None,258)	0
Dense_2(None,22)	(None, 22)	5698
Total Parameters: 313496		
Trainable parameters: 313496		
Non- trainable parameters: 0		

Training and Validation Performance:

During the course of training, significant success is found. Training accuracy reaches amazing 100ʿ, which shows the ability of the model to generalize the training data. The training accuracy is 95, and the recall is 92 with the resulting F1 score of 0.95 in 1,393 epochs. These measures highlight the capability of the model to pick the positive instances correctly and have a balance between the precision and recall.

During validation stage, the model maintains a healthy performance. The agreement of validation stabilizes to 88.60 per cent, and the precision is 87 and the recall is 86. Validation F1 score: 0.89, realized in 2,987 epochs. These statistics show that the model is a good generalization, which can still maintain a good performance when it is applied to data that is not seen.

Construction: testing Performance and Comparative Analysis:

The model when rigorously tested on another set of 55 videos gives a testing accuracy of 82.14. Accuracy is 81.89, and recall is 81.36 with an F1 score of 0.81 in 3,000 epochs. Such findings support the effectiveness of the model in identifying ISL gestures in real world situations.

These performance metrics point out the consistency of proficiency of the model in training, validation and testing phases.

Table 2. Model performance

Phase	Training (880 Videos)	Validation (165 Videos)	Testing (55 Videos)
Accuracy (%)	100	88.60	82.14
Precision (%)	95	87	81.89
Recall (%)	92	86	81.36
F1 Score	0.95	0.89	0.81
Epochs	1393	2987	3000

Class -Level Performance and Error Analysis:

To gain more insights, the performance of the model was analyzed according to the classes, presented in Table 2. And Figure 5 can be seen as the confusion matrix. The results evidently show strong validation of the system's suitability to correctly classify signs like Namaste, Hello, Danger, Help Me, and I am Hungry. This proves the system's ability to perform practical ISL translation as the signs included are for: Greeting, Emergency Communication and conveying basic needs. Classification summary indicated that the features which are visually and dynamically distinct lead to near- perfect separability. The most contributing features to classification are Hand Shape, Orientation, Motion Pattern and Semantic uniqueness of the gestures. Although the majority of classes have high F1 scores, some of the classes such as "Ten" have relatively lower F1. In order to gain more insight into these discrepancies, a confusion matrix was obtained, as shown in Figure 5. The confusion chart demonstrates that gestures with similar hand shapes or orientations were wrongly classified quite often. The confusion matrix has strong diagonal dominance. An example is that the model was more likely to confuse the signs that have similar fingers arrangement or indicate slight rotations in their hands.

Some of the difficult classes include, but are not limited to, classes: Ten and Nine which exhibit a significant level of inter-class confusion as a result of the fact that these two classes are similar in their hand shape and position. Equally, those signs that require rapid changes or faint body expressions were also mistaken. The model may not be able to differentiate between gestures that are differentiated by slight variance in finger positioning or slight movement of the wrist and such therefore will be misclassified. The classification results can be grouped in 3 distinct categories. Group A: Perfectly Classified Classes, Group B: Moderately Strong Classes, Group C: Weak Classes as seen in Table 3. We can see that the gestures where distinctive sand shapes and motions are dominant have been precisely classified with a high F1 Score. The gestures where partial feature overlap is possible within the neighboring frames are moderately low in F1 score. The most misclassified signs are too similar in nature. The hand posture, orientation and visual similarity between the signs One, Ten and Hundred are too close. The analysis indicated that the error pattern is not random, however semantically more meaningful. Visual Resemblances of Gestures here stand out as a possible cause of misclassification. Little movements in bending fingers or the position of thumbs can be very hard to detect by the model. The results demonstrate that the proposed framework achieved excellent recognition for emergency, basic need and other conversational gestures, while giving challenges in some numeric sign gestures with similar visual patterns due to high inter-class similarity and subtle articulation differences.

Even specific refinements (like a more careful data augmentation, better landmark detection, or using other cues, like depth or skeleton data) can be made by looking at cases of misclassification and understanding their causes.

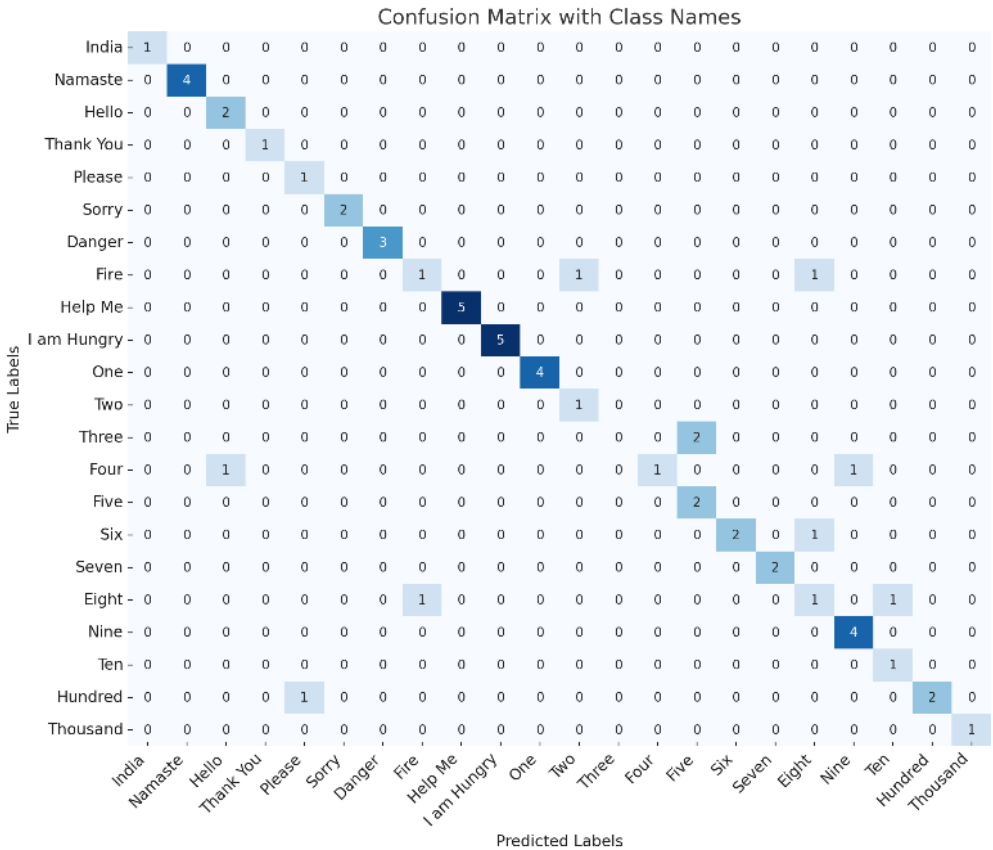

Figure 5. Confusion matrix

Table 3. Grouping of classes as per performance of the recognition

Group	F1 Score	Gestures Included	Gestural Characteristics
A (Perfectly Classified)	Approx 100%	India Namaste Hello Sorry Danger Help Me I am Hungry	Distinct Hand Shapes and Motions Low Intra Class Variations Low Intra Class Similarities
B (Moderately Strong)	65-85%	Two Five Eight Thank You Fire Three Four Six Seven Nine	Partial visual overlap with neighboring frame Similar finger counts or transitions Variation in signing speed and orientation
C (Weak Classification)	Less than 50%	Thousand One Ten Hundred	Single Finger Count Similar hand position/ posture Visual Similarity

Table 4. Class wise performance

Class Name	Accuracy %	Precision %	Recall %	F1 Score
India	100	100	100	100
Namaste	100	100	100	100
Hello	100	100	100	100
Thank you	96	80	100	89
Please	100	100	33	50
Sorry	100	100	100	100
Danger	100	100	100	100
Fire	95	67	100	80
Help Me	100	100	100	100
I am Hungry	100	100	100	100
One	70	50	62	50
Two	100	100	100	100
Three	100	100	67	80
Four	93	50	100	67
Five	100	100	100	100
Six	100	100	67	80
Seven	94	50	100	67
Eight	100	100	100	100
Nine	60	50	100	67
Ten	50	33	33	33
Hundred	50	50	33	40
Thousand	92	50	100	67

By examining misclassifications as depicted in Table 4 and understanding their underlying causes, targeted improvements—such as more elaborate data augmentation, refined landmark detection, or incorporating additional cues (like depth or skeleton data)—can be implemented.

Consequences of the Real-World Usage:

The noted matter of confusion has an important implication for the usage of the sign recognition system in the life environment. This is a must in the daily communicative interactions of a system where slight differences in gestures must be dealt with high accuracy. In other settings, including educational institutions, clinical facilities, or customer service touchpoints, misclassification of a particular gesture could trigger the occurrence of an expensive misunderstanding. This may lead to a necessity on the part of practitioners to embrace more vivid signing conventions, or the training corpus may be expanded by engineers to a wider cohort of signer heterogeneity and ambient environmental situations.

Future work could be done by increasing the size of the corpus to include a wider range of difficult gestures examples.

Adding multi-modal sensors of sensory data (depth sensors or skeletal tracking) to provide a more detailed context system.

Optimizing the hybrid ViT-CNN like in a cross-head, i.e., adding to the model, to reinforce its local element extraction capabilities, especially on gestures of nearly the same shape.

Implementation of domain adaptation measures so that there is robustness to the context environment interactions and variations in the background.

6. CONCLUSIONS

This research predicts the existence of a major gap in current academic literature. The literature corpus on the subject has concentrated mostly on the static sign, which places minimal interest in the dynamic sign on which the associated facial expressions have a significant role. Our question, on the other hand, takes the holistic approach to ISL recognition, at the same time looking to the changing gestures and the fine nature of the role of facial expressions.

Based on the idea of augmenting existing CNN techniques with the ViT, our methodology provides efficient classification of a wide range of gestures, which does not require large-scale data augmentation or transfer learning. The efficiency results in a decrease in training time and computational complexity hence curbing problems that are common in recurrent architectures.

Our proposed framework is efficacious as well as evidenced to have attained a validation accuracy of 88.60 and a test accuracy of 82.14 which are performance metrics exceeding the present-day state-of-the-art. A case study that was done by ablation proves that convolutional encoding shows significant improvement on accuracy in the recognition of ISL. Going forward, we will explore an expanded range of pre-trained ViT frameworks by increasing recognition accuracy further. Furthermore, we would increase the dataset to cover a more diverse range of dynamic signs and facial expressions. In addition, the introduction of Natural Language Processing

(NLP) as a text-to-speech processing solution and the creation of a user-friendly graphical user interface (GUI) are inscribed as some of the major goals, which will expand the application and availability of the suggested method.

By focusing on dynamic gestures and facial expressions, this study essentially draws attention to a significant gap in the field of sign language recognition. The effective application of ViT methodologies will not only find the way to create superiority over traditional CNN methods, but also set the path to make significant developments in sign language interpretation, thus contributing to more inclusive communication between various communities.

REFERENCES

[1] Campbell, L., Grondona, V. (2008). Ethnologue: Languages of the world. Language, 84(3): 636-641.

[2] Zeshan, U. (2000). Sign Language in Indo-Pakistan. John Benjamins Publishing Company.

[3] Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.B., Corchado, J.M. (2022). Deepsign: Sign language detection and recognition using deep learning. Electronics, 11(11): 1780. https://doi.org/10.3390/electronics11111780

[4] Heap, T., Hogg, D. (1996). Towards 3D hand tracking using a deformable model. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, pp. 140-145. https://doi.org/10.1109/AFGR.1996.557255

[5] Lee, D.L., You, W.S. (2018). Recognition of complex static hand gestures by using wristband-based contour features. IET Image Processing, 12(1): 80-87. https://doi.org/10.1049/iet-ipr.2016.1139

[6] Chevtchenko, S.F., Vale, R.F., Macario, V. (2018). Multi-objective optimization for hand posture recognition. Expert Systems with Applications, 92: 170-181. https://doi.org/10.1016/j.eswa.2017.09.046

[7] Huang, J., Zhou, W., Li, H., Li, W. (2015). Sign language recognition using 3D convolutional neural networks. In 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, pp. 1-6. https://doi.org/10.1109/ICME.2015.7177428

[8] Lai, K., Yanushkevich, S.N. (2018). CNN+RNN depth and skeleton-based dynamic hand gesture recognition. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, pp. 3451-3456. https://doi.org/10.1109/ICPR.2018.8545718

[9] Kamruzzaman, M.M. (2020). Arabic sign language recognition and generating Arabic speech using a convolutional neural network. Wireless Communications and Mobile Computing, 2020(1): 3685614. https://doi.org/10.1155/2020/3685614

[10] Zakariah, M., Alotaibi, Y.A., Koundal, D., Guo, Y., Elahi, M.M. (2022). Sign language recognition for Arabic alphabets using transfer learning technique. Computational Intelligence and Neuroscience, 2022(1): 4567989. https://doi.org/10.1155/2022/4567989

[11] Zhang, J., Bu, X., Wang, Y., Dong, H., Zhang, Y., Wu, H. (2024). Sign language recognition based on dual-path background erasure convolutional neural network. Scientific Reports, 14(1): 11360. https://doi.org/10.1038/s41598-024-62008-z

[12] De Coster, M., Van Herreweghe, M., Dambre, J. (2021). Isolated sign recognition from RGB video using pose flow and self-attention. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, pp. 3436-3445. https://doi.org/10.1109/CVPRW53098.2021.00383

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), pp. 6000-6010.

[14] Shenoy, K., Dastane, T., Rao, V., Vyavaharkar, D. (2018). Real-time Indian sign language (ISL) recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, pp. 1-9. https://doi.org/10.1109/ICCCNT.2018.8493808

[15] Katoch, S., Singh, V., Tiwary, U.S. (2022). Indian sign language recognition system using SURF with SVM and CNN. Array, 14: 100141. https://doi.org/10.1016/j.array.2022.100141

[16] Rokade, Y.I., Jadav, P.M. (2017). Indian sign language recognition system. International Journal of Engineering and Technology, 9(3): 189-196. https://doi.org/10.21817/ijet/2017/v9i3/170903S030

[17] Nanivadekar, P.A., Kulkarni, V. (2014). Indian sign language recognition: Database creation, hand tracking and segmentation. In 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, India, pp. 358-363. https://doi.org/10.1109/CSCITA.2014.6839287

[18] Badhe, P.C., Kulkarni, V. (2015). Indian sign language translator using gesture recognition algorithm. In 2015 IEEE International Conference On Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, India, pp. 195-200. https://doi.org/10.1109/CGVIS.2015.7449921

[19] Badhe, P.C., Kulkarni, V. (2020). Artificial neural network based indian sign language recognition using hand crafted features. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1-6. https://doi.org/10.1109/ICCCNT49239.2020.9225294

[20] Kashika, P.H., Venkatapur, R.B. (2022). Deep learning technique for object detection from panoramic video frames. International Journal of Computer Theory and Engineering, 14(1): 20-26. https://doi.org/10.7763/IJCTE.2022.V14.1306

[21] Tran, C.K., Ngo, T.H., Nguyen, C.N., Nguyen, L.A. (2021). SVM-based face recognition through difference of Gaussians and local phase quantization. International Journal of Computer Theory and Engineering, 13(1): 1-8. https://doi.org/10.7763/IJCTE.2021.V13.1282

[22] Ye, N., Wang, R., Li, N. (2021). A novel active object detection network based on historical scenes and movements. International Journal of Computer Theory and Engineering, 13(3): 79-83. https://doi.org/10.7763/ijcte.2021.v13.1293

[23] Sreemathy, R., Turuk, M., Kulkarni, I., Mohanty, S. (2023). Sign language recognition using artificial intelligence. Education and Information Technologies, 28(5): 5259-5278. https://doi.org/10.1007/s10639-022-11391-z

[24] Das, S., Biswas, S.K., Purkayastha, B. (2023). Automated Indian sign language recognition system by fusing deep and handcrafted features. Multimedia Tools and Applications, 82(11): 16905-16927. https://doi.org/10.1007/s11042-022-14084-4

[25] 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

[26] Liu, Y., Nand, P., Hossain, M.A., Nguyen, M., Yan, W.Q. (2023). Sign language recognition from digital videos using feature pyramid network with detection transformer. Multimedia Tools and Applications, 82(14): 21673-21685. https://doi.org/10.1007/s11042-023-14646-0

[27] Al Essa, H.A., Hanon, W., Wotaifi, T.A., Raheem, A.K.A. (2025). Associative memory for recognition and translating American sign language. Ingénierie des Systèmes d'Information, 30(3): 703-711. https://doi.org/10.18280/isi.300314

[28] Jo, B.J., Kim, S.K., Kim, S. (2023). Enhancing virtual and augmented reality interactions with a MediaPipe-based hand gesture recognition user interface. Ingénierie des Systèmes d'Information, 28(3): 633-638. https://doi.org/10.18280/isi.280311