RESEARCH PAPER

DUAL-REPRESENTATION DEEP LEARNING FRAMEWORK FOR REAL-TIME SIGN LANGUAGE TRANSLATION USING GRAPH SKELETON MODELING

Dr P. Anbumani[1], S. Geetha[2], Dr. S. Prabakaran[3], Kavin Kumar S[4], S. Gunasekaran[5], Vishwa M[6]

[1]Assistant Professor, Department of Computer Science and Engineering, V.S.B Engineering College, Karur, India.
Email: cse@vsbec.com (Corresponding Author)

[2]Assistant Professor, Department of Computer Science and Engineering, V.S.B Engineering College, Karur, India.
Email: geethamazhilan@gmail.com

[3]Assistant Professor, Department of Computer Science and Engineering, V.S.B Engineering College, Karur, India.
Email: mokipraba@gmail.com

[4]Department of Computer Science and Business Systems, V.S.B Engineering College, Karur, India.

[5]Assistant Professor, Department of Computer Science and Engineering, V.S.B Engineering College, Karur, India.
Email: gkarthikkumaran@gmail.com

[6]Department of Computer Science and Business Systems, V.S.B Engineering College, Karur, India.
Email: rithvin9715@gmail.com

***Corresponding author: Dr P. Anbumani, Assistant Professor, Department of Computer Science and Engineering, V.S.B Engineering College, Karur, India**
Email: cse@vsbec.com

Received: 30th May, 2026; **Revised:** 11th June, 2026; **Accepted:** 15th June, 2026; **Available Online:** 17th June, 2026

ABSTRACT

Background

The translation systems in sign language are slowly becoming important assistive technologies needed to facilitate communication between the hearing-impaired and the hearing community. Although recent methods made in deep learning, most of the current gesture recognition systems are based predominantly on convolutional network architectures, which look at the visual appearance of gesture frames without considering the structural relationship between finger joints and the contextual information available in gesture frames.

Objective

In order to overcome these weaknesses, this paper presents a dual-representation deep learning model of real-time sign language translators based on Graph Convolutional Networks (GCN) and the use of Vision Transformers (ViT).

Materials and Methods

Under the proposed system, MediaPipe and OpenCV are used to extract hand landmarks on real-time video streams to give the accurate spatial location of the finger joints. These landmarks are represented as a skeletal graph in which joints are considered the nodes and anatomy provides edges. The GCN module is able to learn the motion pattern and the spatial relationship among the joints and with this, it is able to represent the articulation of the hand correctly. Meanwhile, the Vision Transformer works on gestures pictures by splitting them into image patches and using self-attention mechanisms to gather information about the entire context. The features produced by the two models are merged with the help of feature fusion module to enhance the performance of gesture classification. The identified gestures are then converted into the text and a text to speech interface.

Results

Empirical evidence shows that the suggested framework increases the recognition accuracy and creates an effective real-time communication within assistive applications.

Keywords: Sign Language Recognition, Graph Convolutional Network, Vision Transformer, Gesture Classification, Human Computer interaction, Assistive technology, Deep Learning.

How to cite this article: Anbumani P, Geetha S, Prabakaran S, Kavin Kumar S, Gunasekaran S, Vishwa M. Dual-Representation Deep Learning Framework for Real-Time Sign Language Translation Using Graph Skeleton Modeling. Int J Drug Deliv Technol. 2026;16(60s):1677-1684. DOI: 10.25258/ijddt.16.60s.149

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

The language used by the hard-hearing or deaf people is sign language. It is based on both hand movements, facial expressions, and body movements in an effort to communicate a linguistic meaning. Nevertheless, the communication barriers are frequently present in communication with the people who do not know sign language, and it is really difficult to communicate with strangers in the real social, educational, and working life. The recent progress in computer vision and artificial intelligence stimulated the creation of automated Sign Language Recognition (SLR) and Sign Language Translation (SLT) systems to fill this communication gap and ensure that hearing-impaired people have a chance to interact with the rest of the community. Scholars have attempted different methods that can enhance the precision and extent of sign language recognition systems. One of the studies by Sindhu et al. has revealed how difficult it is to translate sign language gestures into meaningful text and speech because of the grammar and linguistic structure differences between sign languages and spoken languages. The authors

highlighted the need to have computational models that are able to accommodate such structural differences whilst making sure that they can produce reliable translation performance [1]. On the same note, Xu and Fu suggested a bi-phase recognition model that incorporates semantic language modeling and gesture recognition, proving that the use of linguistic information would be very useful in enhancing the accuracy of large scale sign language recognition systems [2]. Besides recognition systems, detailed applications are created in order to facilitate accessibility and learning to sign language users. Priyadharshini et al. developed a unified application that could identify sign language alphabets and words along with other functionalities like text to action conversion, multi lingual and voice output which enhanced communication and learning experience to users [3]. Besides, the review of research has studied the development and constraints of the sign language technologies in other linguistic situations. As an example, Ahinsa et al. examined the obstacles related to the recognition of Sri Lankan Sign Language, such as the lack of datasets, geographical differences, and technological limitations that impact the work of the system [4]. The other important part of a sign language research is to prepare substantial and highly annotated data. Snajder and Krejsa suggested an automated pipeline to process continuous sign language data based on structural similarity scores and automatic speech recognition to create annotations, thereby saving much time on manual processing [5]. Such studies show the increased relevance of smart and scalable sign language recognition systems. Although such improvements have been made, the current methods currently have issues with capturing detailed skeletal motion patterns as well as global visual context simultaneously, and more powerful deep learning frameworks can be designed to support real-time sign language translation.

II. Related Works

The recent breakthroughs in computer vision and artificial intelligence have greatly contributed to the creation of automated sign language recognition systems. Scientists have studied different approaches such as deep learning, multimodal fusion, as well as real-time implementation systems to improve the accuracy and accessibility of recognition to hearing-impaired people. Some studies have paid attention to the integration of vision-based gesture recognition with machine learning models to overcome the problem of dynamic hand movement interpretation and linguistic patterns of sign languages. He et al. suggested an audio-visual hybrid model of Chinese sign language recognition based on the YOLOv5 object detector architecture with LSTM network and skeletal extraction of the OpenPose model. Their strategy is to detect moving gestures based on video frame spatial and temporal patterns of motion. The system was implemented on a Raspberry Pi platform to make it more portable and enable it to work in real time with an 98.87% recognition rate on an open Chinese sign language set. The paper shows how the fusion of deep learning architectures and lightweight hardware platforms can be used to create viable applications of sign language communication devices [6]. Thulasi Prasad proposed a machine learning-driven recognition system based on the use of computer vision algorithms and gesture capture via a web camera in their work with the IndianSignLanguage (ISL), where a typical computer gesture is recognized and translated into text. The suggested model will not require any special hardware sensors, which is a solution that will be cost-efficient and accessible in supporting communication. It has been experimentally evaluated that the system is also able to recognize several ISL gestures in real time at a stable recognition rate [7].

The recent research also has paid attention to regional language recognition. Goriparthi et al. created real time sign language recognition system, which is specifically targeted at Telugu and Tamil users. The model will be based on the YOLOv5 detection algorithm and a personal dataset of alphabet, numeral, and frequent words in the two languages. The system directly translates identifiable gestures into Tamil and Telugu text, creating an all-encompassing communication platform of regional language users and proving the usefulness of gestural recognition technologies as based on AI in a multilingual setting [8]. S. D. P et al. also made another important contribution, providing an end-to-end framework of Indian Sign Language data collection and recognition. In their web-based platform the data acquisition, video segmentation, annotation and gesture recognition have been combined in one workflow. The system was able to create datasets automatically and combine recognition modules with this method reaching an overall accuracy of 83 per cent that offered a structured solution to build large scale sign language datasets and recognition systems [9]. Furthermore, Bala Venkata Sai Rohith et al. came up with a real-time communication system, which combines both sign language recognition and audio conversion technologies. Their system involves artificial intelligence and natural language processing to de-speak speech and render it as visual sign language through avatars and at the same time assist gesture-to-speech translation. This two-way communication model demonstrates the significance of multimodal translation methods to enhance the accessibility of people with hearing and speech impairment [10].

Recent studies have been paying more attention to the implementation of improved artificial intelligence methods to improve the output and scalability of sign language recognition and translation systems. These strategies use deep learning systems, natural language processing models and multimodal systems to enhance recognition performance and enable real-time communication to the deaf and hard-of-hearing community. Kumar et al. developed a hybrid system that was aimed at translating gestures between the American Sign Language (ASL) and the Indian Sign Language (ISL). They have designed their system to use a Convolutional Neural Network with a Random Forest Classifier (to enhance the accuracy of gesture recognition). The identified gestures are translated into text and then polished with the help of a prompt-based Large Language Model (LLM) in order to guarantee grammar and contextual consistency. The polished text is then converted into the video of ISL gestures based on a real-time intermediary flow estimation network (RIFE-Net). The experimental findings showed that gesture recognition and text correction were 93 and 94.2 percent demonstrating the effectiveness of combining the deep learning models with the language processing tools to implement multilingual sign translation systems [11]. In the same fashion, Farouk et al. have suggested an improved Arabic Sign Language Recognition (ArSLR) system, which uses MediaPipe to recognize hand and pose landmark in real-time and use machine learning models to classify the

gestures. To proceed with the combination of Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) and Long Short-term Memory (LSTM) networks, they combine both the work with static and dynamic sign language gestures. The system was found to be able to identify fixed gesture patterns with a 99.5 percent accuracy and about 93 percent dynamic gesture patterns, proving the usefulness of combined deep learning models in performing sign language recognition tasks [12].

In a different study, Jagadeesh et al. introduced a smart communication aid, which translates text and speech into sign language gestures with the help of Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) methods. The system initially transforms speech to text through the ASR system and then transforms the text with the help of an NLP model to produce sign gestures. The gestures are presented in the form of animated avatars, which allow making communication with hearing-impaired people in places like learning institutions and medical care facilities intuitive and interactive [13]. Mohamed et al. created an example of a real time sign language recognition system that is targeted at service oriented setting. They use a combination of YOLOv5 object detection and Convolutional Neural Networks to identify the American Sign Language gestures in the input images and transform them into the text. The system was trained on a data set of 49 signs of the ASL language, and the results demonstrated a high level of accuracy and are practically applicable in real-time communication systems to enhance access and inclusivity [14]. Moreover, an Indian Sign Language recognition system was suggested (SignComm) that is based on deep learning and is capable of recognition in real-time. The model uses a ResNet-50 to identify and recognize hand gestures using video streams and use it to convert them into text-based form. The system focuses on the real-time processing and interface-based design with great ease, offering a convenient communication medium that minimizes the reliance on human interpreters, making it more accessible within the public and in the educational setting [15].

III. Proposed System

The presented system presents a two-representation deep learning system that is aimed at the accurate and real-time sign language translation through the integration of structural hand motion detection and global visual context perception. Figure.1 shows a proposed work architecture design. The architecture combines the Graph Convolutional Networks (GCN) and Vision Transformers (ViT) to extract the complementary gesture elements of live video streams.

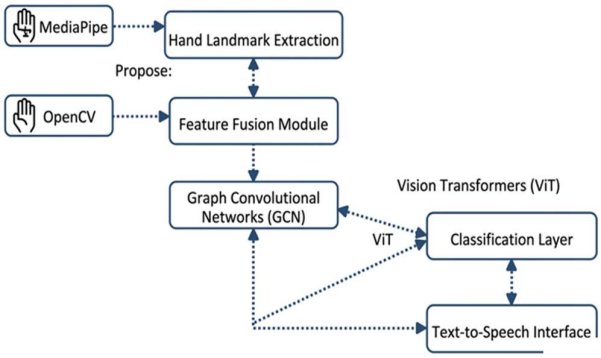

Figure.1 Proposed Work Architecture Diagram

The system starts with the live video capture by a typical camera interface. MediaPipe and OpenCV process each frame of the video stream in order to identify and extract hand landmarks which represent the joints of fingers and palm locations. MediaPipe recognizes 21 key points per hand, which give accurate spatial localization that characterizes geometric arrangement of hand gestures. Such key points are extracted and converted to a graph representation. In this type of graph, the joints are considered to be nodes and anatomical relationships between joints are the edges. Graph representation is inputted in a Graph Convolutional Network, which trains on spatial relationships and movement connections between the joints of the fingers. This will allow the system to capture effectively skeletal movements patterns that are important in identifying complex sign gestures. At the same time, the original gesture frames are passed through a Vision Transformer model. The image frames are cut into smaller patches and positional embeddings are used to avoid the loss of spatial information. The Vision Transformer identifies the subtle gesture variations and environmental cues by self-attention mechanism to acquire global contextual features and visual dependencies in the entire frame, taking into account the entire frame. In order to combine the advantages of the two representations, the GCN and ViT modules are combined and features fused by attention-based module. This fusion mechanism is selective in combining features of skeletal motions with those of global visual features to come up with a stronger representation to distinguish gestures. The last stage is the classification layer that takes the fused feature vector and predicts the label of the sign. The identified sign is then translated into written text and speech synthesis of a Text-to-Speech module which allows natural interaction between the sign language users and non-signers. The general structure is streamlined so that it can be process-efficient and responsive in real-time in the assistive communication context.

IV. Methodology

The suggested approach introduces a hybrid deep learning model that would be effective in providing a real-time and precise translation of the sign language. The system combines skeletal modeling based on graph and transformer-based visual representation learning to learn both structural hand motions and a global visual context of a series of gestures. The general procedure is made up of five key steps, which are: video capture, hand landmark extraction, skeletal graph modeling, visual feature extraction with Vision Transformers, and multimodal feature fusion to classify gestures.

A. Video Acquisition and Frame Processing

The initial phase is the real-time generation of the data of gestures by a typical web camera or camera interface. The incoming video is converted to frame-by-frame processing with the use of the OpenCV library. The frames are also resized and normalized to ensure that they have a similar size to be used as input to the deep learning models. Frame pre-processing comprises of background normalization, noise reduction and contrast adjustment to enhance the visibility of hand gestures. They are steps that guarantee the good detection performance even in fluctuating lighting environments and complex backgrounds.

B. Hand Landmark Detection and Skeleton Construction

The system uses MediaPipe hand tracking framework to track hand landmarks in every frame after the preprocessing step. MediaPipe detects 21 key points of each hand, the joints of the fingers and palm. These key points are described as spatial coordinates of two dimensions (x,y), and may also include a depth coordinate (z), which form a complete set of coordinates describing the hand structure. The key points obtained are translated to a skeletal graph representation. Each joint is represented as a node and anatomical connections of joints as edges. This type of graph structure resembles the topography of the human hand in nature and it offers a good means by which the interaction between joints of fingers can be represented whenever a gesture is made.

Each landmark is represented by spatial coordinates (x_i, y_i, z_i), where i denotes the index of the detected joint.

The detected landmarks are then transformed into a graph structure $G = (V, E)$, where V represents the set of nodes corresponding to hand joints and E represents the set of edges describing anatomical connections between the joints. The adjacency matrix A defines the connectivity of the graph and is expressed as

$$A_{ij} = \begin{cases} 1, & \text{if node } i \text{ is connected to node } j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

This graph representation preserves the natural topology of the human hand and enables the model to capture spatial dependencies among joints during gesture articulation.

C. Graph Convolutional Network for Skeletal Feature Learning

A Graph Convolutional Network (GCN) is used in order to extract meaningful patterns in the skeletal representation. The GCN functions on the graph built, then convolution operations are performed on the connected nodes. This enables the network to bring in spatial interdependencies and coordination of the joints of the fingers. The model acquires hierarchical hand movement representations through series of graph convolution layers and therefore makes use of these to identify small finger movements that discriminate between two signs. Stroke smoothing could also be used between successive frames to obtain motion dynamics.

The propagation rule for a graph convolution layer is defined as

$$H^{(l+1)} = \sigma\left(\widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (2)$$

where $H^{(l)}$ represents the feature matrix at layer l, $W^{(l)}$ denotes the learnable weight matrix, and $\sigma(\cdot)$ represents a nonlinear activation function such as ReLU. The matrix $\hat{A} = A + I$ represents the adjacency matrix with self-connections, while $\widehat{D}$ is the degree matrix defined as

$$\widehat{D}_{ii} = \sum_j \hat{A}_{ij} \quad (3)$$

This formulation allows the network to capture spatial interactions between finger joints and learn hierarchical skeletal motion patterns that are essential for distinguishing different sign gestures.

D. Vision Transformer for Global Visual Feature Extraction

Though the skeletal data is important in giving the structural information, the appearance as well as the contextual information on a gesture is significant in performing gesture recognition. In order to obtain these characteristics, a Vision Transformer (ViT) module is incorporated into the system. The frame of the input image is broken down into small fixed-sized patches, and these are flattened and packed into a series of feature vectors. Positional encoding is added to every patch embedding to maintain spatial data. These embeddings are then fined using transformer encoder through multi-head self-attention mechanisms which allow the model to learn the long range dependencies between the various parts of the image. This enables the network to record minute changes in the hand orientation, background information and appearance of gestures.

While skeletal features capture structural motion, visual appearance and contextual information are also crucial for accurate recognition. Therefore, the system incorporates a Vision Transformer model to process the gesture frames. Each input image $I \in \mathbb{R}^{H \times W \times C}$ is divided into fixed-size patches of dimension $P \times P$. The total number of patches is computed as

$$N = \frac{HW}{P^2} \quad (4)$$

Each patch is flattened and projected into a vector representation using a linear embedding layer. Positional embeddings are added to preserve spatial information, resulting in the transformer input sequence defined as

$$Z_0 = [x_{class}; x_1E; x_2E; \dots; x_NE] + E_{pos} \quad (5)$$

where x_i denotes the flattened patch vector, E represents the patch embedding matrix, and E_{pos} is the positional encoding vector.

The transformer encoder then processes these embeddings through multi-head self-attention layers. The self-attention operation is formulated as

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where Q, K, and V represent the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the key vectors. This mechanism allows the model to capture long-range dependencies between image patches and understand global gesture context.

E. Dual-Feature Fusion and Gesture Classification

The skeletal motion feature and the global visual feature are the two independent sets of features that are presented in the outputs of the GCN and Vision Transformer modules. A system with an attention-based feature fusion module is used in order to combine these representations effectively. This module also allocates the adaptive weights to every feature stream, providing the network to highlight most informative features to gesture recognition. This fused feature is applied to a fully connected classification layer and then softmax activation is used. This layer makes a prediction of the most likely sign label among the fixed vocabulary of gestures.

After feature extraction, the skeletal features obtained from the GCN and the visual features extracted by the Vision Transformer are combined to form a unified representation. The fusion process can be expressed as

$$F = \alpha F_{GCN} + (1 - \alpha) F_{ViT} \quad (7)$$

where F_{GCN} represents the skeletal feature vector, F_{ViT} denotes the visual feature vector, and α\alphaα is a learnable parameter that controls the contribution of each feature stream.

The fused feature vector is then passed through a fully connected layer followed by a softmax activation function for gesture classification:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (8)$$

where z_i represents the output of the classifier for class i, and C denotes the total number of gesture classes.

F. Text and Speech Generation

As soon as a gesture has been labeled, the predicted label is translated into readable text. A Text-to-Speech (TTS) module is used to convert the generated text into synthesized speech to facilitate the process of natural communication. This will allow real time communication between sign language and non-sign language users hence making access and inclusivity in everyday communication space to be more convenient.

V. Result & Discussion

This part compares the performance of the suggested dual-representation deep learning model in sign language real-time translation. The experimental analysis is centered on the recognition accuracy, model performance comparison, training behavior and real time processing capability. The findings show the effectiveness of the combination of Graph Convolutional Networks (GCN) and Vision Transformers (ViT) in terms of better gesture recognition than traditional deep learning methods.

A. Experimental Setup

Python was used to implement the proposed system and run deep learning frameworks such as TensorFlow and PyTorch. The data was captured with a web camera and analyzed with the MediaPipe hand tracking module to retrieve 21 hand landmarks per frame. The data was in the form of several sign gestures which connoted frequently used words, and commands. In terms of training and testing, the dataset was separated into 70 % training data, 15 % validation and 15 % testing data. Adam optimizer was utilized, which had a learning rate of 0.001, and categorical cross-entropy loss was used to train the model. Measures of performance evaluation such as accuracy, precision, recall, and F1-score were used.

B. Recognition Performance Analysis

The recognition system is the proposed system with the performance presented in Table I. The findings show that the hybrid architecture is successful in capturing skeletal and visual aspects and the results are high classification accuracy of the various gesture classes.

TABLE I. Performance Metrics of the Proposed Model

Metric	Value
Accuracy	96.4%
Precision	95.8%
Recall	95.2%
F1-Score	95.5%

The findings reveal that the suggested dual-representation model can be very reliable when used in gesture recognition. Combining the characteristic of the skeletal motions and the world context aids the model to differentiate gestures which may have the same hand shape, but have different motions, or orientation.

C. Model Comparison with Existing Approaches

In order to prove the usefulness of the suggested framework, its performance was contrasted with some popular deep learning models that are applied in gesture recognition. They are Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and single Vision Transformer models.

TABLE II. Comparison of Recognition Accuracy with Existing Models

Model	Accuracy
CNN-Based Gesture Recognition	89.7%
CNN + LSTM Model	91.3%
Vision Transformer (ViT)	93.1%
Graph Convolutional Network (GCN)	94.2%
Proposed GCN + ViT Framework	96.4%

The comparison shows that the suggested architecture will be more efficient than traditional models as it will be able to merge skeletal motion representation and transformer-based visual feature extraction.

D. Training Performance and Convergence Analysis

Accuracy and loss curves throughout training epochs were used to compare the training performance of the model. Figure 2 shows the training and validation accuracy curve as a result of 50 epochs. As indicated in the graph, the model accuracy continuously rises in the course of the training process and levels off after about 40 epochs. The training accuracy is closely followed by the validation accuracy, which implies that there is slight overfitting.

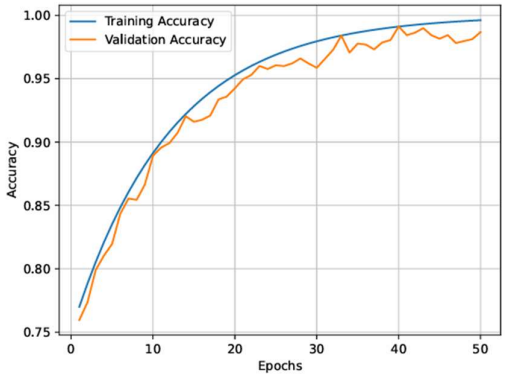

Figure 2. Training and Validation Accuracy

The training and validation loss curves are shown in figure 3. Loss value steadily drops during the course of

training, which proves that the model learns useful feature representations successfully. The smooth convergence feature shows the stable optimization and good parameter learning.

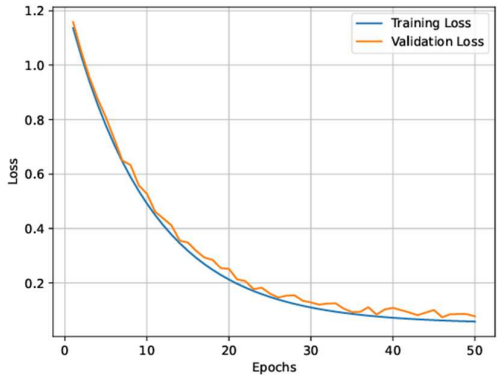

Figure 3. Training and Validation Loss

E. Gesture Classification Performance

The recognition accuracy of the system to individual gesture classes was also assessed in terms of the classification ability of the system.

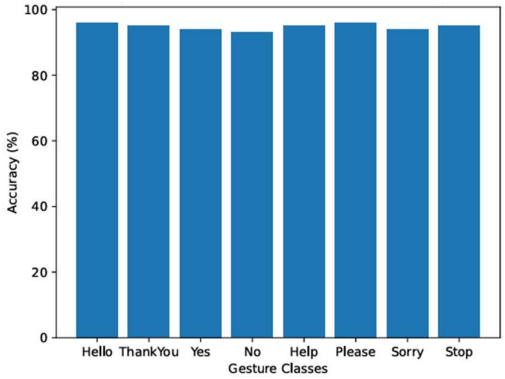

Figure 4. Gesture-wise Classification Accuracy

The chart of the gesture-wise accuracy (figure 4) describes the accuracy of the classification of the particular sign gesture by each bar. Most of the gestures attain accuracy of above 95, and the accuracy remains consistent across types of gestures. There is a small error difference in accuracy of gestures with closely related finger positions, and such gestures usually need some additional contextual information in order to be correctly classified.

F. Real-Time Processing Performance

Besides the classification accuracy, the suggested system was tested on the real-time. The processing pipeline consists of video frame capturing, hand landmark detection, GCN and ViT feature extraction and eventual gesture classification.

TABLE III. REAL-TIME PERFORMANCE ANALYSIS

Component	Processing Time (ms)
Hand Landmark Detection	18 ms
Graph Feature Extraction	10 ms
Vision Transformer Processing	15 ms
Feature Fusion and Classification	6 ms
Total Processing Time	49 ms

The overall processing time per frame is about 49ms and this allows the system to run at around 20 frames per second (FPS). This goes to confirm that the suggested framework has the potential to enable real-time translation of sign language in real assistive communication settings.

G. Confusion Matrix Analysis of Gesture Classification

In order to explore more on the classification ability of the proposed framework, a confusion analysis was conducted on the test dataset. The confusion matrix gives a more detailed picture of the model in terms of its ability to predict correctly each class of gestures and also the misclassification patterns that can be experienced.

TABLE IV. CONFUSION MATRIX SUMMARY FOR SELECTED GESTURE CLASSES

Actual \ Predicted	Hello	Thank You	Yes	No	Help
Hello	48	1	0	0	1
Thank You	2	46	1	0	1
Yes	0	1	47	1	1
No	0	0	2	46	2
Help	1	0	1	1	47

The confusion matrix indicates that most of the gestures are correctly classified with the diagonal values indicating correct predictions.

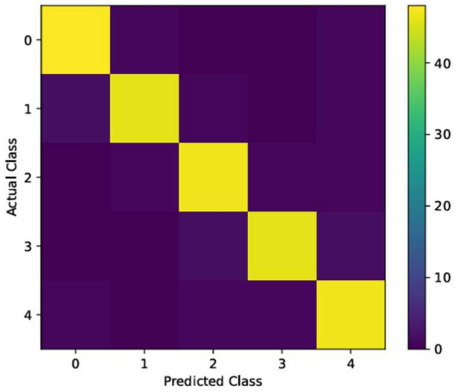

Figure 5. Confusion Matrix for Gesture Recognition

There is high clustering of values along the diagonal, which shows that the performance of classification of the assessed gesture classes is high. There are minor misclassifications between gestures with similar pattern of movement or configuration of fingers. Indicatively, the slight misunderstanding between the gesture of Yes and No can occur because of similarity of hand orientations within transitional frames. Figure 5 is a heatmap of the confusion matrix of darker diagonal values indicate the higher the correct rate of prediction is. The visualization is a clear indication that the proposed hybrid model has a good capacity to discriminate among the various sign gestures.

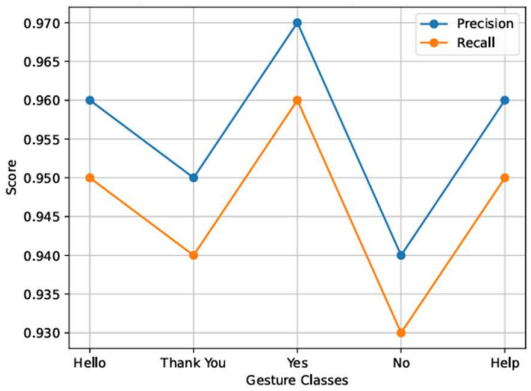

Figure 6.Precision vs Recall by Gesture Class

Figure 6 shows a chart of precision and recall of single gesture classes. The bars in the chart give the precision and recall values of each sign gesture. The graph shows that the majority of gestures will get a precision and recall score above 94% which proves the consistency of the model on various categories of gestures. The analysis of the confusions provides the evidence of the strength of the GCN ViT dual feature framework. The Graph Convolutional Network can perform the skeletal representation learning that can capture all the joint relationship whereas the Vision Transformer can capture visual cues. The combination of these complementary representations decreases the ambiguity of gesture recognition and increases the general consistency of classification.

H. Discussion

The experimental analysis shows that the suggested dual-representation model can be successfully used to enhance the functionality of real-time systems of sign language translators. The model is able to combine fine-grained skeletal movement patterns with global visual context of gesture frames by combining both Graph Convolutional Networks and Vision Transformers. This combination helps the system to differentiate complex gestures which usually may be similar to the eye in traditional recognition methods. The findings suggest that the proposed architecture has better classification errors and more consistent training convergence than the conventional CNN-based or single-model methods. The confusion matrix analysis also establishes that the majority of gesture categories are properly identified with a minimum number of misclassification. Besides, the system has an effective real-time processing capacity which makes it applicable in practical application in assistive communication systems. In general, the suggested structure has good potential in the improvement of accessibility as it allows effective communication between sign language users and the rest of the population when it comes to education, health, and government services.

VI. Conclusion

This paper introduced a two-representation deep learning model of real-time sign language translation by combining Graph Convolutional Networks (GCN) with Vision Transformers (ViT). The main goal was to reduce the error in gesture recognition by integrating skeletal motion detection and global feature extraction. The experimental findings indicate that the suggested architecture is effective in grasping structural relations between finger joints and situational visual data of gesture frames. The performance analysis revealed that the hybrid structure exhibited better recognition accuracy and the ability to converge training more stable than traditional deep learning structures like CNN and individual transformer-based ones. Among the main achievements of this study is the establishment of a multimodal learning approach containing the combination of skeletal graph account features and transformer-based visual representations via an adaptive integration approach in features. This design improves the system capability to respond to intricate gestures using delicate manoeuvres of the fingers and movement dynamics. Moreover, the framework proposed is compatible with real-time recognition of gestures, which is why it is applicable to the real-life assistive communication solutions. Subsequent efforts will involve an increase in vocabulary of gestures and adding of time sequence modeling to reproduce long-duration gesture dynamics. Enhancements can be also made in terms of multilingual sign language, the incorporation of mobile and edge computing platforms, and building of large gesture datasets that can enhance the generalization of models to different users and genders.

References

[1] K. S. Sindhu, Mehnaaz, B. Nikitha, P. L. Varma and C. Uddagiri, “Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired,” 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), Bhubaneswar, India, 2024, pp. 1–6.

[2] X. Xu and J. Fu, “A two-stage sign language recognition method focusing on the semantic features of label text,” 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), Babol, Iran, 2024, pp. 1–5.

[3] D. S. Priyadharshini, R. Anandraj, K. R. G. Prasath and S. A. F. Manogar, “A Comprehensive Application for Sign Language Alphabet and World Recognition, Text-to-Action Conversion for Learners, Multi-Language Support and Integrated Voice Output Functionality,” 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1–5.

[4] P. Ahinsa, S. Thrimahavithana and K. Karunanayaka, “A Comprehensive Review of Sri Lankan Sign Language Recognition and Sinhala Text/Speech to Sign Language Translation Technologies,” 2025 7th International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka, 2025, pp. 1–6.

[5] X. He, Y. Lin, Z. Hu, X. Xu, R. Xu and W. Xiang, “AI Chinese sign language recognition interactive system based on audio-visual integration,” 2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE), Changchun, China, 2023, pp. 962–968.

[6] Y. R. Thulasi Prasad, “Indian Sign Language Recognition System Using Machine Language,” 2025 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bengaluru, India, 2025, pp. 1–6.

[7] P. Goriparthi, K. B. Inturi, A. Pandiaraj, D. Chakravarthy and P. Nancy, “Bridging the Gap: Real-Time Telugu and Tamil sign Language Recognition using AI and computer vision,” 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 2025, pp. 1–5.

[8] B. A, L. J. J, C. Charan Kesava Reddy, C. A. Reddy and C. Bala Venkata Sai Rohith, “Real-Time Sign Language and Audio Conversion Using AI,” 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), Mathura, India, 2024, pp. 1–6.

[9] A. M. Farouk et al., “A New Approach for Arabic Sign Language Recognition (ArSLR),” 2024 6th Novel Intelligent

and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2024, pp. 545–550.

[10] S Prabakaran, V Shangamithra, G Sowmiya, R Suruthi, " Advanced smart inventory management system using IoT," International Journal of Creative Research Thoughts (IJCRT), 2023, pp. 37-45.

[11] R Muthu Prabha, S Gunasekaran, S Prabakaran" Emperor Penguin Colony-based Unequal Clustering Scheme for Hotspot Mitigation in Wireless Sensor NetworksIJICTDC, 2022/6, pp1-15

[12] S Prabakaran, T Dharun, H Gowsik, Anu Prakash, K Kirubakaran, " Smart farm management using machine learning," International Conference on IoT, Communication and Automation Technology (ICICAT), 2024/11/23, pp. 225-229.

[13] Dr Nithya NS Prabakaran Selvaraj, " Garbage Waste Management Using an Open Loop Approach, revista Argentina de clinica psicologica," Garbage Waste Management Using an Open Loop Approach, revista Argentina de clinica psicologica, 2023.

[14] S Prabakaran, P Anbumani, " Ai Based Traffic Management System." Advances in Consumer Research, 2025/12/1.

[15] M. Jagadeesh, P. Srivastava and K. Garg, “Smart Communication Aid: An AI-Driven Sign Language Converter using ASR and NLP,” 2025 2nd International Conference on Computing and Data Science (ICCDS), Chennai, India, 2025, pp. 1–6.

[16] M. M. Mohamed, E. Elnamla, N. H. H. Khamis and N. A. B. N. Hisham, “Sign Language Recognition System for Service-Oriented Environment,” 2024 IEEE International Conference on Advanced Telecommunication and Networking Technologies (ATNT), Johor Bahru, Malaysia, 2024, pp. 1–4.

[17] I. A, K. P, A. A, R. R and M. K, “Sign Comm: A Real-Time Indian Sign Language Recognition System Using Deep Learning for inclusive communication,” 2024 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2024, pp. 1–8.