JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Intelligent Machine Learning Framework for Automated Yawning and Fatigue Monitoring

1. Sandeep Bharti (Department of CSE Department) , Mtech scholar ,JP institute of Engineering &Technology, Meerut,sammrtmail@gmail.com
2.Ayan Rajput ,Assistant Professor ,Mtech Guide ,JP institute of Engineering & Technology, Meerut

Sign language is a rich, multidimensional form of communication that uses coordinated hand movements, facial expressions, spatial organization, and timing to convey meaning. Unlike spoken languages, it depends entirely on visual and physical cues. Indian Sign Language (ISL), widely used by the Deaf community in India, reflects substantial variation across regions and is shaped by diverse gestural expressions. While ISL is deeply embedded in the lives of Deaf individuals, communication with those unfamiliar with sign language often proves challenging, due to the stark contrast between gestural and verbal modes of interaction. Bridging this communication gap requires intelligent systems capable of interpreting ISL gestures in real-time. Despite the growth of computer vision technologies, most studies have focused on American Sign Language and isolated signs, leaving dynamic ISL sequences less examined. To address this, our research presents a ViViT-based framework, optimized with deeper transformer layers and fine-tuned attention mechanisms to better capture complex spatiotemporal features. The model was trained on VISL-PICT, a purpose-built dataset of 508 gesture videos. Our approach attained 96.69% overall accuracy and 99.55% top-5 accuracy, offering a robust solution for automated ISL recognition.

Index Terms— Recognition of Indian Sign Language Gestures, Analysis of Temporal and Spatial Hand Movements, Deep Learning Using Transformer Architectures, Automated Interpretation of Sign Language in Real-Time

I. Introduction

Effective communication is essential for sharing our thoughts and interacting with one another. For humans, speech is the primary mode of communication. A speaker produces sound vibrations through their vocal cords, which listeners hear and then process to derive meaning. However, challenges come when a speaker is unable to make a sound or a listener is unable to hear. Millions of individuals worldwide experience hearing or speech impairments, indicating the importance of implementing alternate communication strategies for their complete social integrationSign language is an important tool for individuals who have difficulties with hearing or speech impairments, enabling them to express their thoughts and feelings through gestures, hand movements, and facial expressions. Unfortunately, sign language is not widely understood outside of its community, causing difficulties in communication with individuals unfamiliar with it and potentially causing feelings of marginalization.Recent deep learning and NLP improvements have brought us tools for text and vocal language translation, such as Long Short-Term Memory (LSTM) [1] and Transformers [2]. However, there is a lack of computer vision frameworks that can recognize sign language, which can enable seamless communication between individuals with hearing or speech impairments and the broader society. This is quite challenging, especially because most computer vision systems become weak to handle out-of-distribution samples [3].

Sign language recognition has been a significant area of research, with many studies focusing on American Sign Language (ASL) [4]. However, sign language is highly diverse, with no universal standard; each region has its own variant. For example, there are German Sign Language, British Sign Language, Arabic Sign Language, and many others besides ASL. These sign languages can also differ because of regional languages and dialects. Other than that, every sign language has thousands of signs that look the same, with a slight difference in each. Computer vision systems face several challenges in accurately recognizing these signs due to other factors such as variations in lighting, background changes, facial expressions, and the speed of signing. Apart from this, some of these signs are static in nature, and some are dynamic, so we can't just use a simple image classification framework to recognize the signs; we also have to take the semantic information from all the previous frames to make sense of a sign. Addressing these challenges requires advanced recognition systems capable of effectively handling the diversity and nuances of different sign languages.

One of the key challenges in promoting ISL recognition is the non-availability of a standard high-quality dataset, which restricts research and sometimes even has to rely on datasets of ASL. There is an urgent need for a large-scale, high-quality dataset in order to enhance ISL recognition frameworks. 333, most current methods that use Convolutional Neural Networks (CNNs) are limited to static signs, making them less effective for real-world applications that involve dynamic signing. Recent transformer-based methods show promise with their ability to use attention mechanisms to analyze information temporally and semantically across video frames. This paper tries to fill these gaps, which leaves plenty of room for future research in ISL recognition.

In this paper, we introduce a novel dataset designed for ISL recognition and present a new approach utilizing attention mechanisms via using a Video Vision Transformer. Our proposed method marks a new direction in dynamic sign language recognition, and we believe that both the dataset and the approach will significantly advance future research in ISL recognition. The key contributions of this paper are as follows:

- Novel VISL-PICT Dataset: We present a dataset comprising 42 dynamic word classes totaling 504 high-quality videos. These videos were recorded with six different signers against a green background to facilitate improved processing and augmentation. A few sample frames of the proposed dataset are shown in the Fig. 1.
- Initial Approach Using Attention Mechanisms: We propose a preliminary approach that applies the attention mechanism, using a Video Vision Transformer for ISL recognition. Given the recent advancements in generative AI, which have made transformers a top model for a number of tasks, we use this advanced framework to evaluate its performance for the ISL recognition task.

II. Related Works

Sign language recognition has experienced major breakthroughs in recent years, driven by advancements in computational resources and the evolution of sophisticated deep learning models in computer vision. These improvements have played a crucial role in enhancing communication with the deaf and hard-of-hearing population. Research in this field typically follows two key methodologies: sensor-based systems and vision-based systems. Sensor-Based Approaches: Initially, gesture recognition relied heavily on physical sensors, such as depth-sensing devices and motion-detecting gloves. Early studies, including those by Mehdi and Dipietro, incorporated gloves equipped with accelerometers and flex sensors to monitor intricate finger and hand movements. These sensors translated physical gestures into digital sign interpretations. Subsequently, researchers like Kumar developed systems that utilized multiple sensors to improve both the accuracy and speed of sign recognition. However, these solutions often involve expensive hardware, making large-scale or real-world deployment financially challenging. Vision-Based Approaches: Alternatively, vision-based methods utilize conventional 2D cameras to interpret hand gestures from images or video footage. These systems remove the need for specialized wearable devices, making them more practical and scalable. Their affordability and ease of integration with everyday technology—such as webcams and smartphones—have made vision-based models increasingly attractive for real-time, accessible sign language recognition.

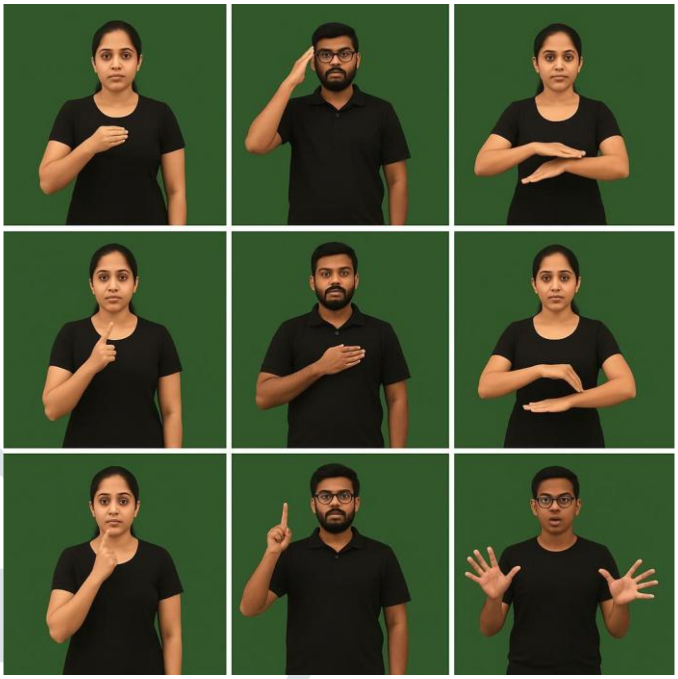


Fig. 1. frames from the proposed ISL recognition video dataset. These frames may come across as static signs, but they are snapshots of dynamic signs

The use of standard 2D camera technology has significantly simplified the process of data collection and deployment in sign language recognition systems. This cost-effective method has made the technology more accessible. To boost the performance and precision of these systems, researchers have increasingly started combining depth data with conventional 2D visuals. This integration has led to notable improvements in accuracy. Treating sign language recognition as a visual classification challenge has allowed the application of various deep learning models, driven by rapid advances in computer vision techniques. These innovations have made it possible to handle complex, continuous sign gestures more efficiently. In the area of American Sign Language (ASL) recognition, Obaid et al. developed a hybrid model that combines both spatial and temporal analysis. They used the Inception model to capture spatial features from video frames and a Recurrent Neural Network (RNN) to process motion and sequence information over time. The team also created a unique dataset to support their approach, enhancing the effectiveness of their system. Regarding Indian Sign Language (ISL), Sreemathy et al. carried out a comprehensive study exploring deep learning techniques used in recent ISL recognition efforts. In addition, Shridhar et al. built a large-scale ISL dataset featuring an extensive lexicon. Sharma et al. treated ISL as an object detection task and applied the YOLO framework. Based on these developments, this study puts forward a transformer-based method to further enhance ISL recognition performance.

III. Methodology

In this section, we detail our experimental approach by first discussing the dataset. We then describe the ViViT architecture that underpins our work. Afterward, we present our proposed model tailored for Indian Sign Language recognition. The section concludes with a brief summary of the methodology.

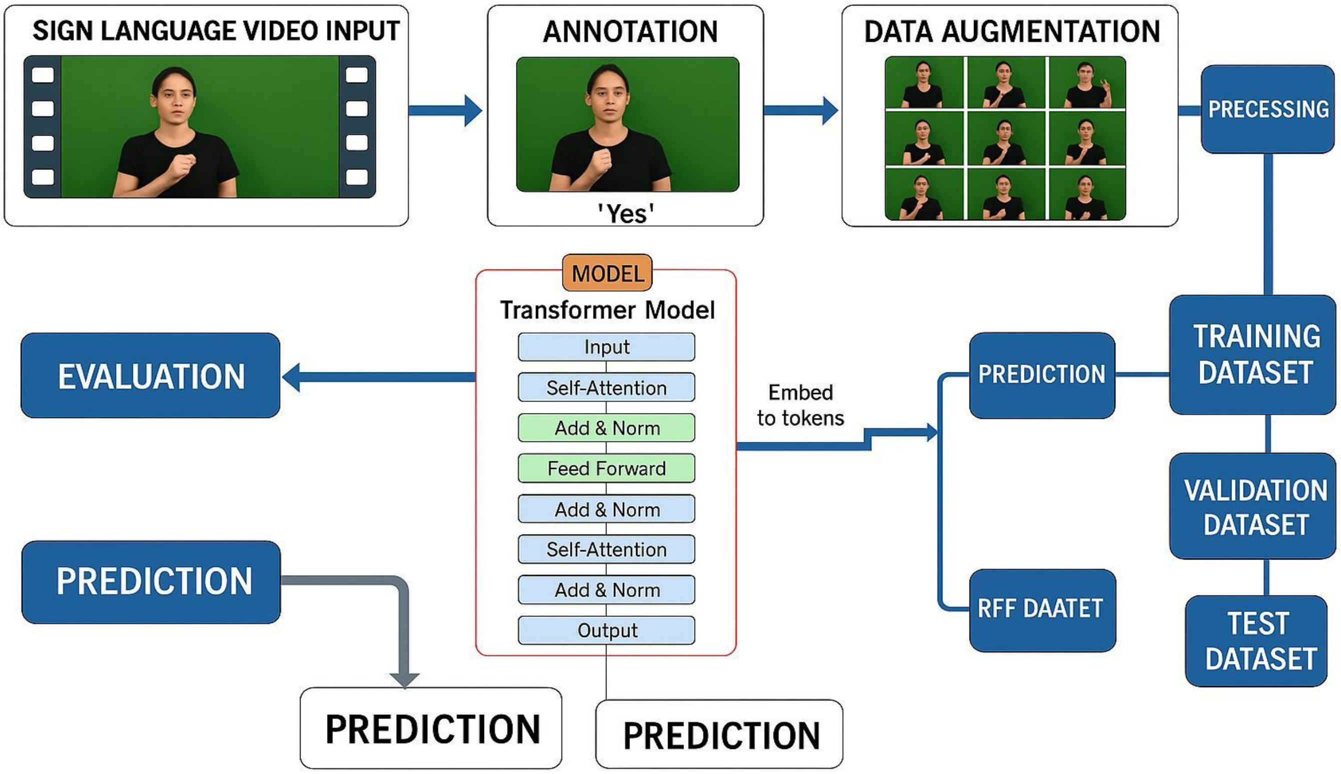


Fig. 2. Only training videos were augmented; test and validation videos underwent preprocessing and feature extraction before model input, ensuring fair evaluation without altering non-training data.

The Various performance indicators were used to evaluate the model. Figure 2 presents the structured framework applied throughout the model's training and evaluation process."

A. VISL-PICT Dataset Description

The A significant challenge faced during the development of an Indian Sign Language (ISL) recognition system was the lack of publicly available benchmark video datasets. As a result, we undertook the task of creating a new dataset tailored to our research objectives. The data collection process was conducted in a controlled lab setting using a Sony a7III DSLR camera to ensure high-quality video capture. All recordings were made in front of a green screen backdrop, which allowed for consistent backgrounds and minimized visual noise, enhancing the clarity of the hand gestures and facial expressions. Six individuals with speech and hearing impairments participated voluntarily in the recording sessions. Their contributions were essential in producing a diverse and authentic dataset. The dataset encompasses 42 unique ISL word classes, with each word performed multiple times. Specifically, every word was recorded in 12 different video samples — two from each participant — using three distinct camera angles to capture varied visual perspectives. This yielded a total of 504 RGB video clip Each video is recorded in high resolution at 1921x1800 pixels and a frame rate of 25 frames per second, ensuring smooth motion representation. Due to the differing complexity and duration of individual signs, the video lengths vary; the shortest video contains 22 frames, while the longest extends to 77 frames. This variability reflects the real-world diversity in sign duration and execution styles among individuals.Overall, the dataset lays an essential foundation for future advancements in ISL gesture recognition. It represents a step forward in making sign language technologies more accessible and effective, especially for aiding communication for the speech and hearing-impaired community. This dataset aims to facilitate progress in dynamic video ISL recognition by offering a dependable basis for building and assessing recognition models. Figure 1 illustrates example frames that highlight the dataset's diversity and structural consistency.

Augmentation and Preprocessing

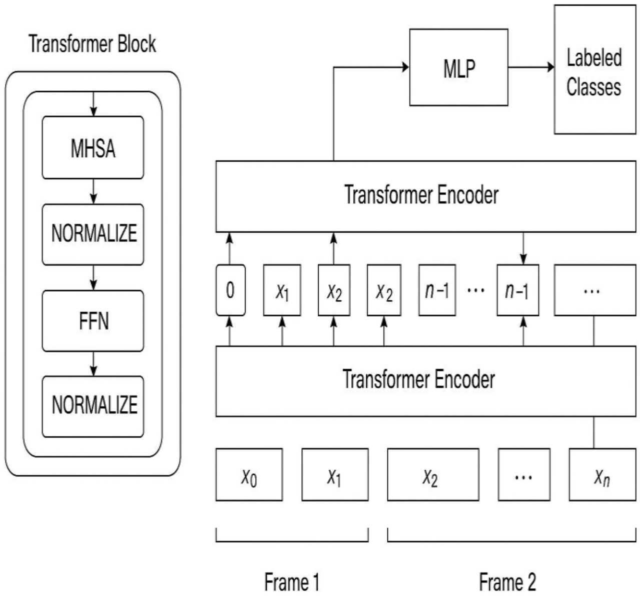
Transformer-based Models based on the transformer architecture, such as ViViT, generally demand large-scale datasets to achieve optimal accuracy. However, the VISL-PICT dataset includes only 504 videos, which limits its capacity to train deep learning models effectively. To overcome this limitation, an automated augmentation tool was developed to expand the dataset. This tool introduced variations by randomly adjusting brightness, rotating frames, flipping them horizontally, zooming, and altering color properties. Each original video was used to generate nine augmented versions, substantially increasing the total dataset size.

After augmentation, the dataset underwent preprocessing to prepare it for use with the ViViT model. Every video was broken down into individual frames and saved as NumPy arrays. These frames were then converted to floating-point numbers and reshaped to match the format expected by 3D convolution layers. The associated labels were also converted into float format. To improve training efficiency and generalization, the dataset was shuffled before being fed into the model for learning.

Video Vision Transformer

Video The Video Vision Transformer (ViViT), presented by Arnab et al. [16], extends the capabilities of the Vision Transformer (ViT) [17] to accommodate the unique challenges posed by video-based tasks. Unlike image models, which handle static spatial information, video analysis requires understanding both spatial and temporal aspects. ViViT achieves this by incorporating spatiotemporal modeling into a single transformer framework. Inspired by transformer models originally crafted for natural language processing [2], ViViT adapts these architectures to process sequential video frames. It leverages the self-attention mechanism to extract spatial details from individual frames while also learning the temporal relationships between frames. This dual capability allows the model to interpret both static and dynamic elements present in videos. In our approach, we utilized ViViT as the core structure and made specific adjustments to suit the needs of our dataset, as visualized in Fig. 3.To prepare the input, ViViT employs uniform sampling, extracting a fixed number of frames from each video to maintain a consistent input size. This method ensures an optimal trade-off between preserving motion continuity and keeping computation manageable. Once sampled, frames are divided into non-overlapping patches. Rather than processing patches from a single frame, ViViT stacks patches from successive frames into what are known as "tublets," which incorporate spatial and temporal cues into a unified format. These tublets are then processed by transformer encoders, where multi-head self-attention identifies relationships across both frames and spatial locations. Positional encoding is applied to help the model retain the sequence of patches and frame order, ensuring that the temporal structure is preserved.ViViT supports multiple configurations tailored for different tasks. One can apply attention separately to spatial and temporal dimensions or fuse them in a joint attention scheme. This design flexibility allows ViViT to scale efficiently across various video analysis problems.Training ViViT typically involves optimization using AdamW, along with regularization strategies such as stochastic depth and layer normalization, both of which aid in reducing overfitting. The integration of uniform sampling and tublet-based embedding gives the model a strong foundation to learn intricate spatiotemporal patterns, making it effective for large-scale video understanding tasks

Setups experiment and Network Perimeters :



In this This section details the methodology and architectural settings utilized in building the proposed model. The approach is composed of essential elements aimed at improving recognition of sign language from video sequences. Initially, the data undergoes preprocessing to standardize the input format. As part of data enhancement, the videos are augmented through various techniques such as rotation, zoom alterations, adjustments in brightness and contrast, and application of color-based filters to diversify the training set.To maintain uniform input dimensions, videos are capped at 22 frames, determined by the shortest video in the dataset. Each frame is then resized to 320×180 pixels to reduce memory and computation requirements. A pixel threshold of 10 is applied to filter out low-quality visual information. The dataset is then split into training and validation segments, with 20% of the data reserved for validation. This setup yields an input array with a shape of (16,

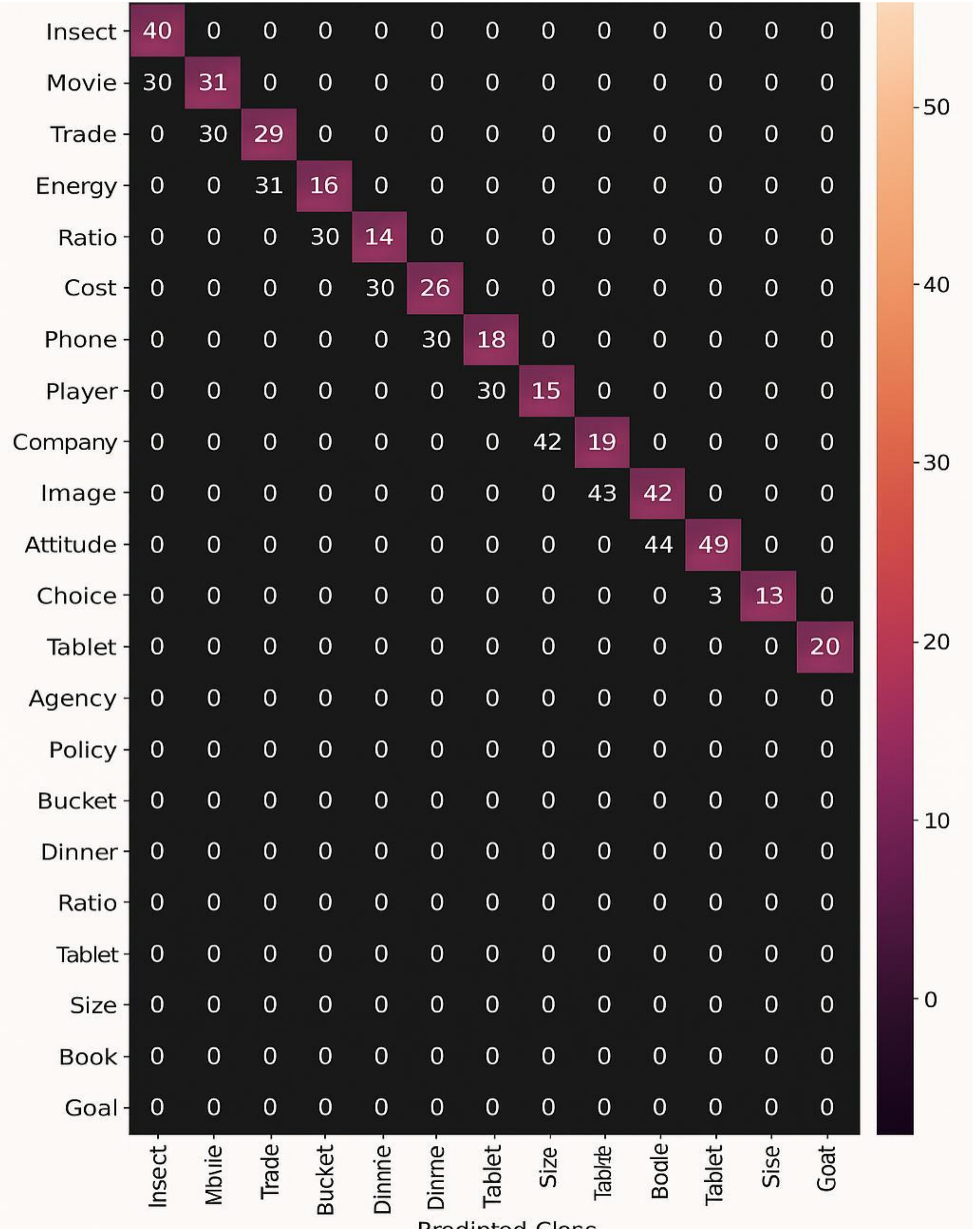


Figure 4. Evaluation matrix: model output against expected labels on the validation set.

180, 320, 3), structured for optimal processing. For feature extraction across time and space, a tubelet embedding method is adopted. The patch size is set to (16, 8, 8), resulting in 256 distinct patches for each input sequence. Multiple configurations of transformer blocks were tested, altering the number of layers and attention heads, while keeping the output feature dimension at 128.Training was conducted over 85 full passes through the dataset (epochs), utilizing a batch size of 2 due to hardware limitations. The learning rate was fixed at 0.0001, with a weight decay factor of 0.00001 applied to reduce overfitting risks. Overall, this pipeline combines effective data preparation with a tailored transformer-based model, making it highly suitable for Indian Sign Language (ISL) gesture recognition. A step-by-step flow of this system is depicted in Figure 2

Assessment Criteria :

To 1To gauge the effectiveness of our system, we used two primary metrics: classification accuracy and top-5 accuracy. These indicators help measure how often the model assigns the correct label to each video clip.

- of To quantify how well the model performs, we rely on the metric known as accuracy, which reflects the proportion of total predictions that were correctly classified. The formula used to compute accuracy is as follows

$$\text{Accuracy} = \frac{\text{Number of Accurate Predictions}}{\text{Total Predictions Made}} \qquad (1)$$

In This metric is central to our evaluation process. It helps us understand the frequency with which the model's top predicted class aligns with the actual label. Since our dataset has a balanced distribution across classes, accuracy serves as a fair and dependable measure of overall performance. However, in scenarios involving closely related classes—like various signs in Indian Sign Language (ISL)—accuracy alone may not capture near-correct predictions. To address this, we also use Top-5 Accuracy, a metric that examines whether the true label is found within the five predictions with the highest confidence scores. The formula is expressed as:

$$\text{Top-5 Accuracy} = \frac{\text{Distance where True Label in Top five}}{\text{Total Predictions}}$$

(2

) This approach is especially relevant when dealing with similar gesture patterns, as it provides a broader picture of how close the model gets to the correct result even if it's not ranked first.

IV. PERFORMANCE OUTCOME AND OBSERVATIONS

In This portion outlines the insights gained from our experimental results. Among all tested variations, the architecture comprising 10 layers and 10 attention heads per layer outperformed others. This setup reached a validation accuracy of 96.69%, as illustrated in Figure 5. The corresponding confusion matrix is presented in Figure 4, while Table I compares the performance with alternative designs. We recorded a Top-5 Accuracy of 99.55% at epoch 85. Both the loss and accuracy curves stabilized around epoch 65, signaling convergence. Training was executed on an NVIDIA RTX 4090 GPU and took roughly one hour using a batch size of two. The accuracy over time is shown in Figure 6, while Figure 7 tracks the Top-5 accuracy across epochs.

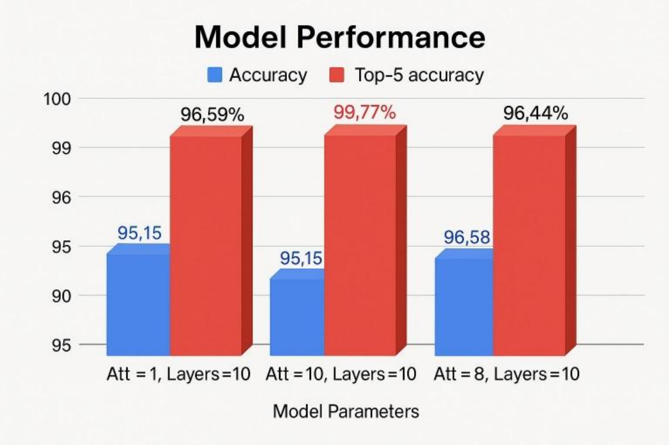


Fig. 5. This chart illustrates the impact of altering model architecture on accuracy and Top-5 accuracy. "Att" denotes the count of attention heads, while "layers" signify the network's depth applied during model assessment.

Performance was evaluated by modifying the count of attention heads and layers. The table highlights the most effective results obtained through careful tuning of these architectural components during experimentation.

F

Attention Heads	Layers	Accuracy	Top5 Accuracy
1	10	95.15	96.59
10	8	95.15	99.77
8	10	96.58	96.44

Results show that attention-based architectures have strong potential for advancing this research area. Future developments may involve redesigning the model to better process longer videos while maintaining frame-wise semantic and temporal continuity throughout the sequence.

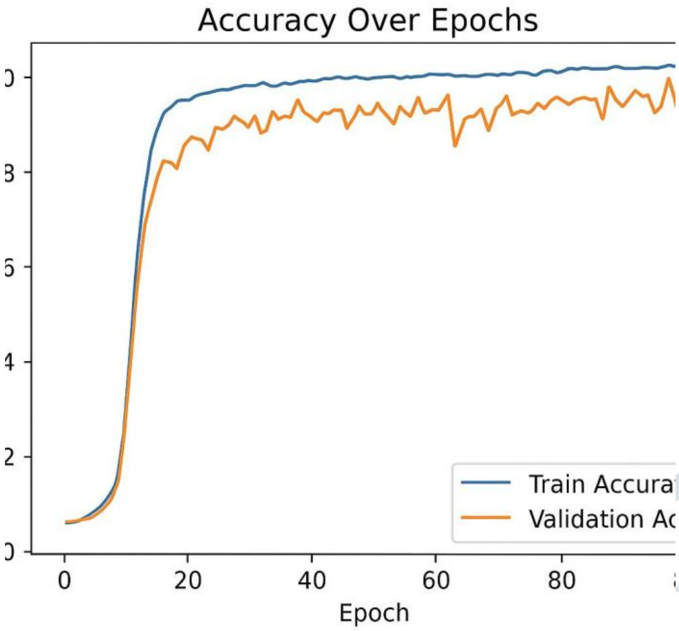
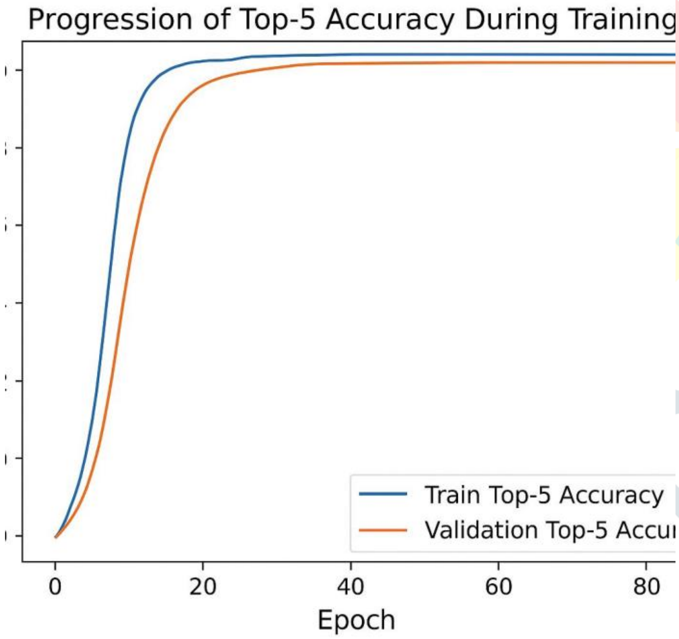


Fig. 6. “Top accuracy reached by model in each epoch.”



V. Conclusion

The Communication between individuals who are speech or hearing impaired and those unfamiliar with sign language remains a significant hurdle due to the limited usage and variation of regional sign languages. To bridge this gap, there is a growing need for an intelligent system capable of translating sign gestures into understandable formats for the general population. Our study addresses this issue by employing a video-based approach using vision transformers that utilize attention mechanisms to learn and interpret gesture sequences effectively. As part of this work, we introduce a dedicated dataset titled VISL-PICT, consisting of 508 high-resolution video clips that represent 42 commonly used words in Indian Sign Language (ISL). This curated dataset aims to support future research efforts in ISL recognition. Our model demonstrates an impressive performance, achieving 96.69% accuracy and 99.55% top-5 accuracy on the validation set, highlighting its robustness and potential for real-world deployment. Nevertheless, the approach faces a potential limitation in the form of high computational requirements, especially when handling extended video sequences with both spatial and temporal complexity. Despite this, the results are encouraging and mark a step forward in creating practical ISL translation tools.

References

[1] In 1997, S. Hochreiter introduced the Long Short-Term Memory (LSTM) architecture, which has become foundational in handling sequential data within neural networks, especially for overcoming the vanishing gradient issue common in earlier recurrent models.

[2] 2. A. Vaswani and collaborators revolutionized natural language processing by introducing the Transformer model in their 2017 arXiv publication titled “Attention Is All You Need”, emphasizing self-attention mechanisms for handling long-range dependencies.

[3] 3. In 2023, O. Ojaswee, A. Agarwal, and N. Ratha explored how image classifiers perform under physical alterations and out-of-distribution scenarios, presenting their findings at the ICCV Workshops to highlight robustness challenges in computer vision models.

[4] 4. K. Bantupalli and Y. Xie explored deep learning techniques in combination with computer vision to interpret American Sign Language, presenting their solution at the IEEE International Conference on Big Data in 2018.

[5] 5. Z. Zafrulla and colleagues demonstrated the potential of using the Kinect sensor to recognize American Sign Language, with their work showcased at the 13th International Conference on Multimodal Interfaces in 2011.

[6] 6. S. A. Mehdi and Y. N. Khan focused on recognizing gestures using gloves equipped with motion sensors. Their 2002 work was presented at the International Conference on Neural Information Processing (ICONIP).

[7] 7. L. Dipietro, A. M. Sabatini, and P. Dario provided an in-depth overview of glove-based input systems for gesture capture in their 2008 paper published in IEEE Transactions on Systems, Man, and Cybernetics, highlighting multiple practical applications.

[8] 8. In a 2017 study published in Pattern Recognition Letters, P. Kumar and his team introduced a hybrid method for sign language detection by integrating data from various sensors through a coupled HMM-based model.

[9] 9. R. Cui, H. Liu, and C. Zhang applied a recurrent convolutional neural network architecture to handle continuous gesture recognition using an iterative optimization approach, presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] 10. V. Kumar, R. Sreemathy, and their colleagues developed a real-time Indian Sign Language recognition technique based on skeletal features, which they presented at the 2023 ICCCNT conference.

[11] 11. In a 2023 survey at the OCIT conference, R. Sreemathy and co-authors reviewed state-of-the-art methods in sign language recognition, particularly emphasizing the effectiveness of deep learning in this domain.

[12] 12. R. Sabeenian, S. S. Bharathwaj, and M. M. Aadhil detailed the integration of deep learning and vision-based systems to recognize sign gestures in their 2020 article in a special issue of JARDCS.

[13] 13. In a 2020 publication in Applied Computer Systems, F. Obaid and his team utilized a combination of convolutional and recurrent neural networks for gesture recognition from video footage.

[14] 14. A. Sridhar and collaborators presented Include, a large-scale dataset for Indian Sign Language, during the 28th ACM Multimedia Conference in 2020, aimed at expanding resources for gesture recognition tasks. DOI: https://doi.org/10.1145/3394171.3413528

[15] 15. S. Sharma and team proposed a technique for word-level ISL recognition in real-time using the YOLOv4 architecture, presented at the INCOFT conference in 2022.

[16] 16. A. Arnab and co-authors developed ViViT, a video-specific transformer architecture aimed at extracting temporal and spatial patterns, presented at the IEEE/CVF International Conference on Computer Vision in 2021.

[17] 17. Lastly, A. Dosovitskiy and collaborators explored the application of transformer-based models to visual data in their 2021 preprint titled “An Image Is Worth 16x16 Words”, highlighting scalability in image classification. Available at https://arxiv.org/abs/2010.11929.