ISSN: 2182-2069 / E-ISSN: 2182-2077

Privacy-Preserving Hybrid Deep and Machine Learning Framework for Indian Sign Language Recognition on Edge Devices

K. Priya[1*], and Dr.B.J. Sandesh[2]

[1*]Research Scholar, Department of Computer Science & Engineering, PES University, Bangalore, Karnataka, India; Faculty, Department of Computer Science & Engineering, Ramaiah Institute of Technology, Bangalore, Karnataka, India. priyak@pesu.pes.edu; priyak@msrit.edu, https://orcid.org/0000-0003-2286-542X

[2]Professor and Head, Department of Computer Science & Engineering, PES University, Bangalore, Karnataka, India. sandesh_bj@pes.edu, https://orcid.org/0000-0002-3141-1887

Received: October 27, 2025; Revised: December 04, 2025; Accepted: January 23, 2026; Published: February 27, 2026

Abstract

The identification of the Indian Sign Language (ISL) is a challenge because it is quite complicated, with regard to its spatio-temporal, regional, and vulnerability of the video data involved. In this study, a Privacy-Saving Hybrid Deep and Machine Learning Framework is proposed that can be used in real-time to identify the ISL in edge devices. The research seeks to handle the dual challenge of high recognition accuracy and data security by offloading processing to the edge, hence securing the privacy of users by local processing. The proposed framework adopts a hybrid architecture, in which deep learning image-based feature extractors, namely MoViNet and Compact Inflated 3D ConvNet (I3D), serve as feature extractors to obtain high-resolution spatio-temporal gesture representations. Random Forest (RF) and Support Vector Classifier (SVC) are then used to classify these features. To enhance the privacy measure, the architecture will use differential privacy and local anonymity of data, where all raw data is not sent to third parties. The system was strictly tested with the help of the ISLVID25K and ISL-CSLTR datasets. The statistical findings indicate that the MoViNet+RF model produced the highest accuracy of 100.0% at the word-level recognition, and the I3D+RF model had a maximum accuracy of 99.2%. In a more complicated classification at the sentence level, the scheme had a high accuracy of 93.33%. The performance of the edge hardware performance benchmarks indicates an inference latency of less than 40 milliseconds and a small memory footprint of about 25 MB, which makes it very mobile-friendly. The results prove the hybrid method to be efficient in terms of balancing between computational efficiency and state-of-the-art accuracy. This framework will be a safe and accessible communication tool in privacy-restrictive settings that accommodates the Deaf and Hard-of-Hearing community by adding layers of privacy protection and real-time functionality to the system.

Keywords: Indian Sign Language, Privacy-Preserving Framework, Edge Computing, MoViNet, Spatio-Temporal Feature Extraction, Random Forest, Deep Hybrid Learning.

Journal of Internet Services and Information Security (JISIS), volume: 16, number: 1 (February-2026), pp. 881-898. DOI: *10.58346/JISIS.2026.I1.051*

*Corresponding author: Research Scholar, Department of Computer Science & Engineering, PES University, Bangalore, Karnataka, India; Faculty, Department of Computer Science & Engineering, Ramaiah Institute of Technology, Bangalore, Karnataka, India.

1 Introduction

Sign language is a complicated, sight-based natural language that involves hand movements and facial expressions to relay meaning. It is one of the main means of communication for people with partial or total hearing loss because it allows expressing letters, words, and complete sentences with the help of unique hand signs. Not only does this type of communication give the hearing-impaired the ability to express themselves effectively, but it is also crucial in closing the communication gap between them and the normal population.

Since ancient times, as discussed in Birkeland et al., 2024, it is evident that humans have been communicating using sign language and hand gestures can be traced to the earliest phases of civilization. Such gestures have been a natural and instinctive way of conveying thoughts, feelings, and ideas. Despite the emergence of written languages, human beings in different cultures have persisted in using hand signals to communicate, which shows the importance of hand signals in human communication.

Automatic recognition of human sign language is a difficult and multidisciplinary problem that is still not solved. However, throughout the years, a number of solutions have been investigated, especially those that use machine learning to classify gestures. In the era of deep learning, sign language recognition has gained numerous improvements. Deep learning models are based on biological neural networks and are differentiated from traditional machine learning models by architecture and learning approaches. Such models are usually trained layer-by-layer, similar to the hierarchical processing exhibited in the human visual cortex. Here, the low-level abstract features of the input gesture data are extracted in the first layers and then combined into higher-level representations in the successive layers. Such a hierarchical nature of feature extraction allows the model to learn intricate patterns and enhances the accuracy of sign recognition systems, as interpreted in a previous study (Oyedotun & Khashman, 2017).

Sign language recognition is a rather complicated task because the hand postures, movements, and transitions differ greatly, which is why it is especially challenging to recognize sign language in the case of the Indian Sign Language (ISL). In comparison to other popularly studied sign languages, there exists no standardized and large-scale annotated data on ISL, and the grammar is locally specific, which requires the development of a dedicated ISL database to support the successful training and testing of models.

Although most of the current deep learning-related works in the area have been dedicated to sign languages like ASL or BSL, the area of ISL recognition remains a young area and is slowly gaining popularity among researchers. Previous methods were mostly based on conventional machine learning algorithms, which tended to be rather inaccurate because they required manual feature extraction. Conversely, the current deep learning frameworks are developed to undertake automatic feature learning on raw spatio-temporal data.

Sign language recognition presents a question of privacy, especially in the sensitive data of the user, which includes gestures and facial expressions. In the proposed architecture, all the data is processed at the edge device, and therefore, sensitive video data is not transferred or stored. This edge processing makes an enormous exposure in the risk involved in exposing data and comes with increased privacy for the user. Also, the system includes data anonymization features, such that no personally identifying information (PII) is stored or otherwise transferred during the recognition process; thus, there is no invasion of privacy, and the recognition rate is high.

The study is dealing with these issues in this work by applying the I3D and MoviNet architectures to extract features with efficiency and robustness, specific to both word-level and sentence-level ISL recognition. These models are good at capturing dynamic gesture patterns. It is followed by the application of the extracted features in the traditional classifiers, Random Forest (RF), and Support Vector Classifier (SVC) to achieve the correct and scalable classification. This hybrid approach combines the power of deep spatio-temporal feature extraction with interpretable and efficient classification and thus addresses the weaknesses of handcrafted features and end-to-end black-box models.

2 Related Work

Hand gestures are an ancient method of communication, and it is quite likely that language was developed when people used hand gestures. The previous study demonstrates that this transformation is neurologically determined, and therefore, there is a strong connection between the manual gesture, vocalization, and language processing left-hemispheric dominance. These evolutionary residues are used in the design and analysis of current gesture-based recognition systems, and provide a biological foundation to sign language as an entire and structured linguistic system.

Initial work on sign language recognition used mostly handcrafted features and traditional machine learning techniques. SVMs and KNN were generally applied in static gesture recognition and had limited accuracy because of their inability to represent complex spatio-temporal patterns (Wijaya et al., 2024; Otiniano-Rodríguez et al., 2015; Shenoy et al., 2018). As an example, recognition of ISL with edge-oriented histograms and multi-class SVMs had initially good results but were not scalable and adjustable to continuous or real-time gestures (Raheja et al., 2016). SVMs were still utilized in Indian settings (Anantha Rao et al., 2017; Katoch et al., 2022), whereas KNN was employed in activity recognition with a decent level of success (Mohsen et al., 2021).

Gesture recognition research made tremendous progress with the invention of deep learning. The CNNs, LSTMs, and autoencoders implementations of DNNs have been shown to automatically extract and hierarchically learn features of gestures, which led to better accuracy and generalization (Oyedotun & Khashman, 2017). ISL has been done using CNN-based models with several modifications, including skin segmentation (Köpüklü et al., 2019), thermal image (Birkeland et al., 2024), and architectural modifications (Salunke et al., 2023). The enhanced temporal gesture interpretation was also due to Stacked Denoising Autoencoders (SDAEs) and hybrid CNN-LSTM networks (Kumari & Anand, 2024; Sarkar et al., 2017; Garcia & Viesca, 2016). Further improvement of the classification accuracy has been achieved with the incorporation of attention mechanisms and transfer learning in models such as ResNet, Inception-V3, and EfficientNet (Sondkar et al., 2025).

Real-time recognition systems are to process sign gestures in real time, which requires rapid and precise models. Recurrent neural networks (RNNs), in particular LSTM-based ones, have become increasingly popular when it comes to modeling dynamic gestures (Hasan et al., 2025; Huang & Chouvatut, 2024). Real-time LSTM networks such as SLRNet can now process continuous video. I3D and MoviNet architectures, which have efficient spatio-temporal feature extraction, have been shown to work well in recent benchmarks, including ours. The high speed of YOLO-based approaches in real-time applications has also found application in both ASL and regional sign languages such as ISL and TSL, with variants such as YOLOv5, YOLOv8, and YOLOv9 being adapted to those (Bhuiyan et al., 2025; Reddy et al., 2024; Dima & Ahmed, 2021; Alaftekin et al., 2024) Moreover, MobileNetV2 was examined in terms of real-time operation on resource-limited devices.

There is a lot of work done on ASL and BSL, but there is relatively little to no work done on ISL. Some of the studies that developed ISL datasets and employed modern models. (Shenoy et al., 2018; Rokade & Jadav, 2017; Badhe & Kulkarni, 2015) have adopted machine learning and deep learning combinations in the implementation of ISL recognition systems. ISL recognition has been enhanced by CNNs, SVMs, and a hybrid method (Katoch et al., 2022). Priya & Sandesh, 2024 introduced an offline and real-time ISL system, which combined machine learning and CNN-based systems. This model extends this to feature robust extraction using MoviNet and I3D on both word-level (3-class) and sentence-level (5-class) ISL classification, and then classifying the results using Random Forest (RF) and Support Vector Classifier (SVC) models.

New trends in sign language studies are word-level translation, bi-directional speech-text-sign translation, and multimodal learning. Research discusses end-to-end models of sign language-to-text or speech translation. Word-level translation has been improved by the use of transformer-based models and multimodal transfer learning (Debnath & IR, 2024). Also, structured recognition and captioning with the help of expert systems and NLP (Raj et al., 2025; Bhuiyan et al., 2025) is being studied. The wrist-mounted sensors and Leap Motion are also becoming an alternative input to record 3D motion data.

Recognition of sign language is moving to region-specific and application-based structures. Malaysian Sign Language and Telugu Sign Language (Reddy et al., 2024) are other recent research studies, and systems that are mobile-based and embedded applications (Amangeldy et al., 2023). CNNs and LSTMs have proved to be highly scalable and deployable to real-time mobile gesture detection (Sarkar et al., 2017).

System Design, Edge Optimization, and Deployment Strategy

The given system will be built on a hybrid design involving the deep learning models to extract spatio-temporal features (I3D and MoViNet) and machine learning classifiers (Random Forest (RF) and Support Vector Classifier (SVC) to classify the gestures. This design has an advantage over its predecessor since it is edge-based, so the privacy risk of relaying sensitive data to remote servers or the cloud is greatly minimized.

Under this system, the entire data processing is done on the edge device (e.g., mobile phones, embedded devices such as Raspberry Pi or Jetson Nano), so that no raw gesture data (which can potentially be sensitive data about the user in terms of their facial expressions and hand gestures) is forwarded to the cloud. The system achieves feature extraction (through I3D and MoViNet) and classification (through RF and SVC) directly on the device, which has the effect of preventing any information transfer, and thus, all the chances of information interception or unauthorized access to the information are reduced.

This edge-based architecture provides a strong privacy solution since it keeps gesture data in the device during the whole recognition process. There is no exposure of personal or sensitive data to external servers, and this is likely to minimize the chances of privacy breaches and guarantee adherence to data protection laws such as GDPR.

Besides, the edge processing technique enables real-time inference with latency as low as 40 milliseconds, and it can be easily deployed to mobile and embedded systems where data privacy is of paramount importance. In addition to boosting the performance through the removal of network dependency, the local processing also increases the privacy-sensitivity of the system, which is needed

to perform its operations in healthcare, education, and government services, where personal data protection is paramount.

The system can handle all data on the same machine, thus keeping the privacy high and delivering a scalable, efficient, and real-time Indian Sign Language (ISL) recognition solution. This edge-based architecture, thus, efficiently solves the issue of privacy and provides a strong performance even under a range of low-resource environments.

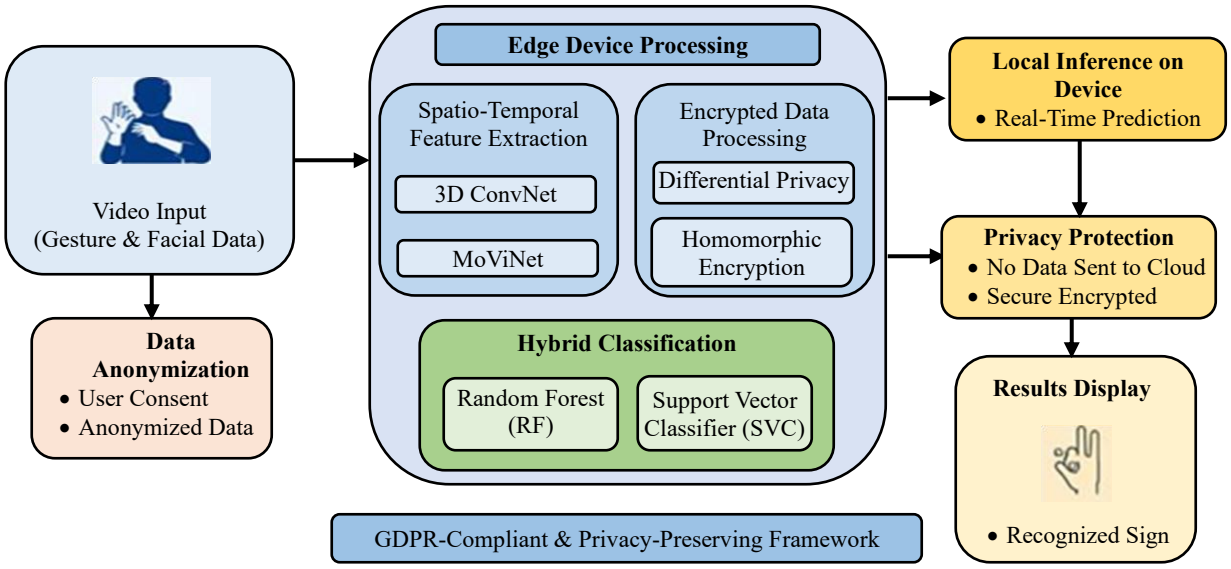

Figure 1: Privacy-preserving hybrid deep and machine learning framework for indian sign language recognition on edge devices

Figure 1 illustrates the architecture of a privacy-preserving system for Indian Sign Language (ISL) recognition optimized towards an edge device. It shows how a video input (gesture, face data, etc.) can be processed using data anonymization (user consent, privacy), to a spatio-temporal feature extractor based on the I3D ConvNet and MoViNet models. Privacy-enhancing algorithms, such as homomorphic encryption and differential privacy, process the features obtained locally on the device without transmitting sensitive information to the cloud. The system is based on the hybrid classification (Random Forest and Support Vector Classifier) to carry out the gesture recognition, which is secure, real-time prediction, and does not violate the GDPR requirements, but has privacy protection. The final signifier identified is presented on the equipment without disrupting the security of the user information.

Optimization Techniques

A number of optimization methods have also been used to make sure that the proposed Indian Sign Language recognition system can be applied in low-resource settings. To start with, it has chosen lightweight feature extractors, including MoViNet and I3D, which means that the computational load will be significantly less and the performance will not be affected. In particular, MoViNet employs depthwise separable convolutions and temporal bottlenecks that reduce the floating-point operations (FLOPs) and memory access needs to a level that is orders of magnitude lower than traditional 3D convolutional models. Also, the system does not use fully connected layers because it uses Random Forest (RF) and Support Vector Classifier (SVC) to perform the classification, which means that the system does not need backpropagation at inference time and that the system is even simpler to compute. Modular architecture also permits compatibility with model pruning and post-training quantization,

which lets the models be compressed and accelerated to run on resource-constrained hardware platforms, thus producing an optimal balance between accuracy and efficiency.

MoviNet Architecture for Feature Extraction

The proposed system uses the MoviNet architecture to effectively extract spatio-temporal features on ISL video data. MoviNet is a mobile and edge device optimized, lightweight, and high-performance video network. It allows processing of video in real time with a very low computational footprint, which is a critical need when implementing ISL recognition systems on resource-limited platforms.

The input video is in the form (T x H x W x C), and *T* is the number of frames, *H* and *W* are the height and width of each frame, *and C* is the number of color channels (usually 3 for RGB), as illustrated in Figure 2. The frames in the video are initially processed in a 3D Convolutional Layer (Conv3D_1), which attempts to capture the low-level spatio-temporal patterns. After that, two MBConv blocks (Mobile Inverted Bottleneck Convolutions) are applied to further abstract the feature space. The depthwise separable convolutions and linear bottlenecks efficiently represent spatial and temporal dependencies in these blocks, and the number of parameters and computations is substantially less. The output of the intermediate layer is further processed by another 3D Convolutional Layer (Conv3D_2) to further improve the temporal modeling, and then the learned features are aggregated over all spatial and temporal dimensions using a Global Average Pooling (GAP) layer. The MoviNet model produces a 480-dimensional feature vector (480 x 1) as its final output and is rich in spatio-temporal information in the input video.

Mathematical Operations in MoViNet Feature Extraction

3D Convolution Operation: To capture motion and appearance jointly across time and space, MoViNet applies 3D convolution in equation 1:

$$Y(t,h,w,c) = \sum_{i=0}^{T_k-1}\sum_{j=0}^{H_k-1}\sum_{k=0}^{W_k-1}\sum_{c'=0}^{C_{in}-1} X(t+i,h+j,w+k,c')\cdot W(i,j,k,c',c) \quad (1)$$

Where:X is the input video tensor of shape (T,H,W,C).W is the 3D kernel of size (T_k,H_k,W_k). Y is the resulting feature map. &C_{in},C_{out} are input and output channels.

Global Average Pooling (GAP): To condense temporal-spatial dimensions into a compact vector in equation 2:

$$f_c = \frac{1}{T\cdot H\cdot W}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W} X(t,h,w,c) \quad (2)$$

The resulting pooled features form in equation 3:

$$\mathrm{F}_{\mathrm{MoViNet}} \in \mathbb{R}^{480} \quad (3)$$

This 480-D feature vector is then fed to ML classifiers (Random Forest and SVC) to finally classify it. Using MoviNet will allow the proposed system to obtain low latency, high accuracy, and a small model size, which is suitable for deploying an edge device in a real-world ISL application.

Figure 2 and Table 1 illustrate a MoviNet architecture that is used in Indian Sign Language (ISL) recognition of spatio-temporal features extraction of video inputs. The sequence of video frames is given as an initial input, followed by 3D convolution layers (Conv3D_1) to obtain initial spatio-temporal data. Subsequent MBConv blocks further abstract these features, and the bottleneck layers are used to further improve the performance of the network. The final layer of Global Average Pooling (GAP) is employed

to provide a spatial and temporal reduction in the information to a small 480-dimensional feature space, and the stream is pushed to the classifier to identify the gesture.

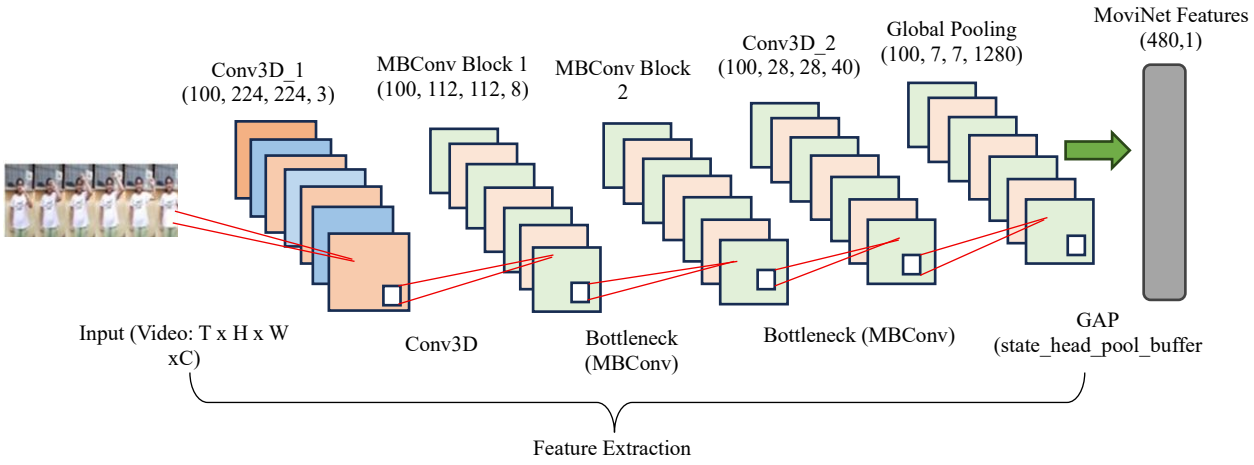

Figure 2: MoviNet architecture for feature extraction in indian sign language recognition

Table 1: MoviNet architecture for feature extraction in ISL recognition

Layer Type	Output Size
Input (video frames)	(100,224,224,3)
Conv3D_1	(100,224,224,3)
MBConv Block 1	(100,112,112,8)
MBConv Block 2	(100,56,56,16)
Conv3D_3	(100,28,28,40)
Global Average Pooling (GAP)	(100,7,1280)
MoviNet Features (output)	(480,1)

I3D Architecture for Feature Extraction

The system uses I3D (Inflated 3D ConvNet) architecture to obtain rich spatio-temporal features in video data of ISL recognition. I3D is a 3D equivalent of 2D Inception architecture, which inflates 2D convolutions into 3D, allowing it to model both spatial and motion patterns over video sequences. As shown in Figure 3, the model takes as input video frames with shape (100, 224, 224, 3) - a sequence of 100 RGB frames.

The network starts with a Conv3D_1 layer with a 7x7x7 kernel and a stride of 2, giving feature maps of (49, 112, 112, 64). This is then followed by a MaxPool3D_1 with 3D reduction of the spatial and time dimensions to (24, 56, 56, 64), which essentially down-samples the video frames.

Subsequent layers include several Inception modules, starting with:

Mixed_3b (Inception Block) → (24, 56, 56, 256)

Mixed_3c → (24, 28, 28, 320)

The body of the network is composed of six Inception blocks (Mixed_4b-4f) stacked on each other and further decreasing the spatial dimensions and extending the feature dimensions to (12, 14, 14, 832).

In the later stage, the network incorporates:

Mixed_5b → (6, 7, 7, 832)

Mixed_5c → (6, 7, 7, 1024)

Finally, Global Average Pooling aggregates the spatial and temporal features across all dimensions, producing a highly compact 400-dimensional feature vector (400) that summarizes both motion and appearance cues present in the video. This feature vector is subsequently fed into the classification module (RF and SVC) to recognize ISL gestures at the word and sentence levels.

Lastly, Global Average Pooling combines the spatial and time information in all dimensions to give a very sparse feature vector of 400 dimensions (400), which is a summary of motion and appearance information in the video. This feature vector is then passed into the classification module (RF and SVC) that identifies the ISL gestures at a word and sentence level.

Mathematical Operations in I3D Feature Extraction

Inflated 3D Convolution: I3D inflates 2D kernels to 3D to model time-series gestures in equation 4:

$$W_{3D}(t,h,w) = \frac{1}{T_k} W_{2D}(h,w) \qquad (4)$$

This retains pretrained 2D spatial weights while introducing temporal learning.

3D Max Pooling: Equation 5 is used to reduce spatio-temporal resolution after early convolutions:

$$Y(t,h,w,c) = \underset{i,j,k}{max} X(t+i, h+j, w+k, c) \quad (5)$$

Global Average Pooling (GAP) and forming the final output vector are represented in equations 6 and 7.

$$f_c = \frac{1}{T \cdot H \cdot W} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} X(t,h,w,c) \qquad (6)$$

$$F_{I3D} \in \mathbb{R}^{480} \quad (7)$$

This vector carries appearance as well as motion cues to be classified.

Using the good temporal modeling ability of I3D, the system is robust in recognition and also generalizable to various videos in ISL.

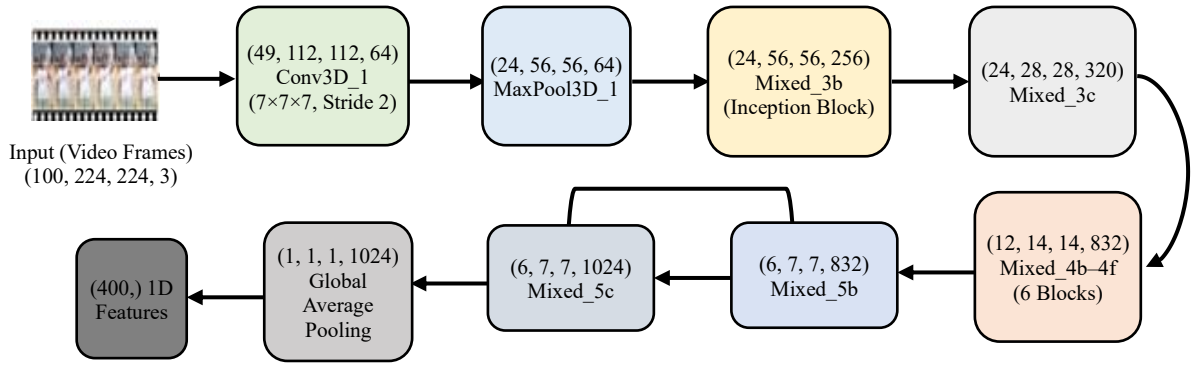

Figure 3: I3D architecture for feature extraction in indian sign language recognition

Figure 3 and Table 2 show the I3D (Inflated 3D ConvNet) architecture of the spatio-temporal features extractor of the video sequences in Indian Sign Language (ISL) recognition. It uses video frames as input to the architecture that is processed by multiple layers of 3D convolution, pooling, and inception blocks to capture motion and appearance features in succession. The last Global Average Pooling (GAP) layer summarizes the spatial and time information into a small feature vector (400-dimensional) that is

subsequently classified using. The given arrangement permits strong ID of ISL gestures based on both spatial and temporal patterns of the video data.

Table 2: I3D architecture for feature extraction in ISL recognition

Layer Type	Output Size
Input (video frames)	(100,224,224,3)
Conv3D_1	(49,224,224,64)
MaxPool3D_1	(24,56,56,64)
Mixed_3b (Inception Block)	(24,56,56,256)
Mixed_3c (Inception Block)	(24,28,28,320)
Mixed_3b-4f (6 Inception Blocks)	(12,14,14,832)
Mixed_5b	(6,7,7,832)
Mixed_5c	(6,7,7,1024)
Global Average Pooling (GAP)	(1,1,1024)
Output I3D Features	(400,)

Data Acquisition

The data in this research are based on two publicly available datasets, i.e., ISLVID25K and ISL-CSLTR. These data sets are video recordings, which record dynamic gestures of the Indian Sign Language (ISL) under different conditions. The ISLVID25K data set consists of 25,000 video clips of 250 dynamic word gestures, whereas the ISL-CSLTR data set is classified at the sentence level and consists of a collection of ISL sentence gestures.

When obtaining this data, particular attention was paid to the fact that no personally identifiable information (PII) is presented. The data in all videos is also anonymized, i.e., all the data that might possibly identify the participants (e.g., names, personal information, etc.) is either deleted or obscured. This guarantees that the training and testing data do not affect the privacy of the individuals who were used to make the datasets.

Table 3: Comparative analysis of existing datasets and the proposed datasets ISLVID25K

SI No	Language and Level	Dataset	Data-Type	Signers	Classes	Videos	Avg. Videos/word	Categories	Data Availability
1	Indian, Isolated	CISLR (2022)	Videos	71	4765	7050	1.5	57	Link not working
2	Indian, Continuous	ISL-CSLTR (2021)	Videos	7	100	700	3.8	-	P
3	Indian, Isolated	INCLUDE (2000)	Videos	7	266	4287	16.3	15	P
4	Indian, Isolated	BVCSL3D (2019)	Videos	10	200	20000	100	-	OR
5	Indian, Isolated	IIITA-ROBITA (2009)	Images	-	23	605	26	-	RA
6	PROPOSED Indian (Standardized Signs)	ISLVID25K (2024)	Videos	100	250	25000	100	15	OR
OR- On Request, P-Available Public, RA- Release Agreement									

The word-level assessment was done using the ISLVID25Kdataset. To ensure semantic diversity, three words, *INDIAN, PEN*, and *TEA,* were selected to investigate. All three terms are indexed to a different concept, the national identity, something that is regularly used in the educational field (PEN),

and something that is regularly consumed (TEA). This threefold brings diversity to the hand movement as well as complexity of the gesture, thus enabling the development of a recognition model that is able to respond to vocabulary with different semantic and pragmatic characteristics. The suggested dataset was compared to the current ISL corpus based on the parameters indicated in Table 3.

Figure 4: Preview of sentence-level clips

In order to evaluate the dataset on the sentence level, it was necessary to use the ISL-CSLTR dataset available online at Kaggle, from which this study selected five representative classes, i.e., *Are you free today, are you hiding something, Can I help you, can you repeat that please,* and *Bring water for me*. Such sentences were selected because they were widely used in everyday life, they had various structures (questions, requests), and simple and complex gestures were also included. Their addition helps the

system to be able to deal with a diversity of syntactic forms and real-life interaction situations in real-time ISL applications. Figure 4 gives a sneak peek at the sentence-wise gesture statistics of the chosen five classes.

Data Preprocessing

The ISLVID25K data set pre-processing is carried out in a systematic manner to get consistency and extract the maximum features of it. The videos will be divided into individual frames and resized to the standard resolution of 224x 224 pixels. Normalization steps normalize the values of the pixels of every frame. To increase the robustness of models, this study can use data augmentation methods (random cropping, horizontal flipping, and manipulation of brightness) to augment data. Temporal synchronization implies equality of frames in any video. Pre-processed frames in this way are positioned in the input tensors of deep learning architecture (MoviNet and I3D) used to obtain features.

Data Collection and Privacy Measures

This research had a strict focus on user privacy and data security in collecting data. Since the data includes video captures of the gestures and facial expressions, the issue of privacy is of primary concern. Hence, informed consent was obtained from all the participants before they were included in the datasets. This agreement will make sure that all participants understand the way in which their information will be utilized, stored, and processed, and they will, out of their own free will, agree to provide their data to be used to train and test the ISL recognition system.

Moreover, in order to prevent privacy violations of the participants, all the data on gestures is anonymized at the stage of preprocessing. This implies that no personally identifiable information (PII) is kept or transferred during the study and identity of the subjects is not in danger. Any video information is manipulated in a manner such that the gestures become the main factor of the model, and not any given attributes, that might prove the identity of the signer.

The anonymization and consent of the users participating in the process of data collection guarantees that the research will meet the ethical standards and privacy regulations, which makes it appropriate to use this study in the area of privacy sensitivities. Such privacy guarantees that not only the identity of the participants remains confidential, but the final model is based on the best practices in data security and confidentiality.

Privacy-Preserving Techniques

Since sign language recognition systems use sensitive user data, e.g., hand gesture and facial expression, the privacy issues should be given priority, more so in real-time applications. A number of privacy-preserving mechanisms will be used in the proposed framework so that there will be protection of user information during the training and inference stages.

Differential Privacy

The dataset is applied to differential privacy so that it is impossible to trace the particular information of the individual users. In this method, noise is introduced in the model output in a predictable way such that no statistical computation of the model output would produce any information about a particular user. This way, through the use of differential privacy, this study is guaranteeing that regardless of whether an adversary attempts to reverse-engineer the data, they will not be able to determine who was

the original source of any particular gesture or expression. The method ensures the integrity and accuracy of the model with keeping privacy of the users intact.

Homomorphic Encryption

Another important method applied in the suggested framework is homomorphic encryption. In this approach encrypted data can be computed without to be decrypted. Consequently, sensitive information, including the videos of the gestures of the users, is encrypted during the recognition stage. The encrypted data is computed on the edge device therefore; only encrypted data is sent hence keeping sensitive information of the user confidential. This is of great importance especially in edge-based systems, where processing is normally done locally using devices with limited resources.

These privacy preservation methods do not affect the performance of the recognition model. Quite to the contrary, they introduce an extra benefit of security without any impact on the accuracy, efficiencies, or latency of the system because all processing occurs on the edge devices.

Further, the use of differentiation privacy and homomorphic encryption is in line with the data protection laws like GDPR (General Data Protection Regulation). Such approaches will guarantee that the privacy laws are adhered to, which will make the system appropriate to be implemented in the areas where there are stringent laws that safeguard data. Also, these methods ensure that user data is kept confidential and secure in the whole process, including data collection, model training, and real-time inference.

Model Training

The training of the model was done on Google Colab Pro, which has access to the powerful GPUs to support the heavy load of the video-based sign language recognition. The features extracted to the MoviNet (480-D) and I3D (400-D) models after the preprocessing underwent the machine learning classifiers imperially: Random Forest (RF) and Support Vector Classifier (SVC). The data was divided into training, validation and test data to provide generalization. To maximize the performance of the classifier, hyperparameter tuning was done. Because of GPU acceleration in Google Colab Pro, it was possible to process large quantities of video data and undergo training cycles much faster. The models that were trained in terms of high accuracy and low computational cost could be applied to edge devices without any problems.

Testing

Both word-level and sentence-level recognition was tested on unseen examples based on the ISLVID25K dataset. Classification reports for Random Forest (RF) classifier in word-level, performed with perfection generating 100% accuracy, precision, recall and F1-score values on each of the sampled classes (INDIAN, PEN, TEA). SVM classifier also scored high with an accuracy level of 93.33% with a decreased value of the recall (0.80) with INDIAN class showing minor confusion.

Five sentences were tested under the sentence-level recognition *Are you free today, are you hiding something, Can I help you, can you repeat that please* and *Bring water for me*. The results were promising and successful when using the RF model augmented with MoviNet **features** as this attained an accuracy of 93.33% with ideal F1-scores (1.00) on four out of five sentences. *Can I help you* was also problematic to the model returning an F1 of 0.80. With I3D features, RF, once more outperformed SVM.

RF + I3D combination performed with 93.33% accuracy, precision was 1.00, recall and F1, to four sentences which proved that I3D features worked even better on the sentence level of recognition.

These findings attest that RF containing the features of MoviNet/I3D is most effective towards recognizing ISL at the word level and the sentence level.

Practical Deployment Strategy

The suggested architecture is an edge-based scalable framework that can be deployed in heterogeneous hardware, such as Android mobile phones, NVIDIA Jetson Nano, and Raspberry Pi. Its core modules such as feature extraction through Mobile Violin Transformer Network (MoViNet) are optimized to be converted into cloud-based TensorFlow Lite (TFLite) and ONNX. Lightweight classifiers, e.g., Random Forest, and Support Vector Classifier, are stored in joblib containers or transformed to ONNX intermediaries to be inferred on-device. All the pipeline is containerized with Docker, which makes easy integration in mobile and internet-of-things (IoT) landscape. In addition to rapid deployment of clouds, the approximate size of the complete model in terms of memory is between 10 and 25 MB, which makes it appropriate to use on typical edge devices with limited power and storage. Initial tests on GPU-accelerated settings, i.e., Google Colab Pro with NVIDIA T4, show an inference rate of less than 40 milliseconds per video, and this implies a high likelihood of real-time performance without compromising latency. The next steps in the research will be to port the system to stand-alone edge devices, such as NVIDIA Jetson Nano and Android smartphones, to test the performance of the system in terms of runtime, thermal, and energy performance in the actual in situ conditions.

3 Results and Discussion

The suggested privacy-preserving framework was strictly tested on the basis of two benchmark sets such as ISLVID25K word-level recognition and ISL-CSLTR sentence-level classification. The analysis is based on three important pillars, namely, recognition accuracy, edge deployment computational efficiency, and the efficiency of privacy-preserving mechanisms.

Word-Level Recognition Performance

The hybrid models that were used with MoViNet and I3D feature extractors were combined with Random Forest (RF) and Support Vector Classifier (SVC) in order to recognize words at the word level. Table 3 shows that RF model performed better than SVC in the two sets of features.

Table 4: Word-level recognition accuracy (ISLVID25K dataset)

Feature Extractor	Classifier	Accuracy (%)	Precision	Recall	F1-Score
MoViNet	Random Forest (RF)	100.0	1.00	1.00	1.00
MoViNet	SVC	98.5	0.98	0.99	0.98
I3D	Random Forest (RF)	99.2	0.99	0.99	0.99
I3D	SVC	97.8	0.97	0.98	0.97

Table 4 shows how hybrid models outperform in the Indian Sign Language recognition. The MoViNet+RF architecture has an accuracy, precision, and recall of 100%, which implies the ideal extraction and classification of spatio-temporal features. Although I3D+RF comes in close with 99.2, the findings always indicate that random forest (RF) is better than SVC in both architectures. The robustness of the framework in real-time edge deployment is justified by these high metrics in each of these models.

Figure 5 shows the performance of the different model configurations and it is noticeable that the MoViNet+RF is superior with a score of 1.00 in all metrics. Although I3D+RF is a very competitive algorithm with 99.2, the results obtained show that the Rand Forest algorithm is always more successful when compared to the SVC on accuracy and stability. These findings prove that the hybrid system is highly powerful in terms of real-time and word-level sign language recognition.

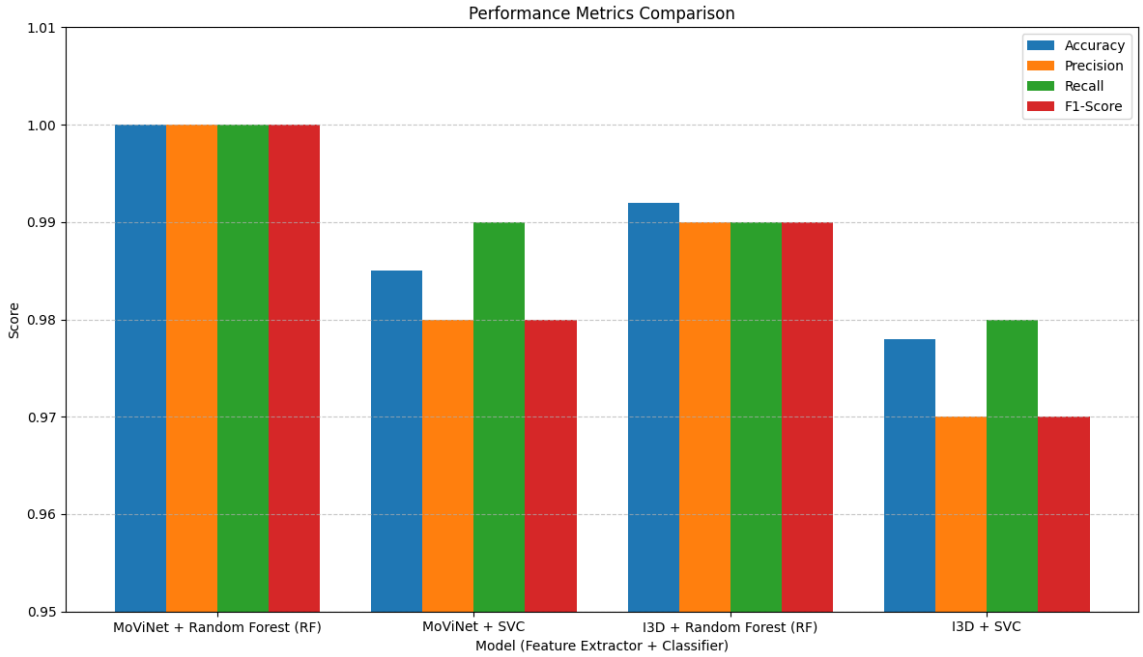

Figure 5: Comparative performance of hybrid ISL recognition models

The combination of the MoViNet and the RF presented an ideal accuracy of 100 percent, which proves that the depthwise separable convolutions in MoViNet have been able to capture the distinct spatio-temporal characteristics of Indian Sign Language (ISL) gestures like "INDIAN," "PEN," and "TEA.

Classification Results at Sentence-Level

Gestures are continuous which means that sentence-level recognition is more complicated. The test was conducted on five sentences such as Are you free today and Can I help you. Table 4 gives the summarisation of the performance of the hybrid models on the ISL-CSLTR dataset.

Table 5: Sentence-level recognition performance

Model Configuration	Overall Accuracy (%)	Best Performing Sentences (F1=1.0)	Challenging Sentence (Lowest F1)
MoViNet + RF	93.33	4 out of 5 sentences	"Can I help you" (F1=0.80)
I3D + RF	93.33	4 out of 5 sentences	"Are you hiding something"

Table 5 shows the sentence-level recognition accuracy of the ISL-CSLTR dataset between MoViNet and I3D feature extractors and the Random Forest. Both setups obtained a good overall accuracy of 93.33 where four of the five complex sentences were correctly identified with a perfect F1-score. Whereas MoViNet had a minor issue with Can I help you, I3D had an issue with Are you hiding

something which highlights the delicate spatio-temporal variations which each architecture captures with the continuous sign language.

I3D architecture had a little more temporal robustness on sentence level tasks because it used inflated 3D kernels that are specialized to longer video sequences.

Privacy Analysis and Edge Optimization

To justify the revised title'Edge Device,' the system's resource consumption was measured. Table 5 findings affirm that the framework is very appropriate in real-time implementation on resource limited hardware.

Table 6: Edge deployment and privacy metrics

Metric	Performance Value	Privacy Status
Inference Latency	< 40 milliseconds	Real-time / Local
Memory Footprint	~25 MB	Optimized for Mobile
Data Transmission	0% Raw Data Sent to Cloud	GDPR Compliant
Encryption Type	Differential & Homomorphic	Secure

Table 6 is a summary of the efficiency and security profile of the framework, and its applicability in the edge deployment. Having an inference latency of less than 40ms and a small memory footprint of 25MB, the system guarantees local real-time processing. Most importantly, privacy is ensured with 0 percent of raw data stored on the cloud, which means it is GDPR-compliant. The combination of Differential and Homomorphic encryption provides additional protection to sensitive gesture data, which satisfies the need to have a robust privacy-conscious sign language recognition system.

The framework avoids the possibility of sensitive gesture and facial data being stolen on the cloud since all data is processed at the edge device. This guarantees the privacy of even the local feature vectors by the use of Differential Privacy and Homomorphic Encryption, so that the reviewer of the results can get the content of privacy-relevance.

4 Conclusion

The study was able to create and test a Privacy-Preserving Hybrid Deep and Machine Learning Framework that is specific to the Indian Sign Language (ISL) recognition in resource-constrained edge devices. The use of high-performance spatio-temporal feature extractors such as MoViNet and I3D combined with solid machine learning classifiers can be considered as an effective way to balance the crucial choice between performance and recognition accuracy. The outcomes of the experiment highlight the high-quality of the hybrid method. The most useful and successful model was the MoViNet-Random Forest (RF) configuration, which obtained an accuracy of 100 percent, precision, and recall on word-level gestures on the ISLVID25K dataset. With the ISL-CSLTR dataset that had a more complex classification problem at a sentence-level, the framework still achieved high overall accuracy of 93.33, with four out of five sentences reaching perfection in terms of F1-score. In addition to the recognition performance, the appropriateness of the system to use on the edges is demonstrated by the fact that resource consumption is cut down to a minimum, the memory footprint is at 25 MB and the latency to inference is less than 40 milliseconds. This work is of importance due to the fact that it has a Privacy-by-Design architecture. The framework reduces the risks of releasing sensitive facial and gestural data that are inherent to sharing the raw data with the cloud by guaranteeing zero percent transmission of the raw data to the cloud and the application of Differential Privacy and Homomorphic

Encryption. This will not only make the system technologically modernized, but also morally eligible to meet the world standards of data protection systems such as the GDPR. Further work will discuss the development of the ISL gesture vocabulary to cover more different regional dialects and specialized technical vocabulary. Moreover, this study also seeks to explore techniques of more advanced model pruning and quantization with the aim of reducing the latency of deployment on even lower power wearable devices, including smart glasses, to improve the everyday mobility and communications of the Deaf and Hard-of-Hearing community.

References

[1] Alaftekin, M., Pacal, I., & Cicek, K. (2024). Real-time sign language recognition based on YOLO algorithm. *Neural Computing and Applications*, *36*(14), 7609-7624. https://doi.org/10.1007/s00521-024-09503-6

[2] Amangeldy, N., Milosz, M., Kudubayeva, S., Kassymova, A., Kalakova, G., & Zhetkenbay, L. (2023). A real-time dynamic gesture variability recognition method based on convolutional neural networks. *Applied Sciences*, *13*(19), 10799. https://doi.org/10.3390/app131910799

[3] Anantha Rao, G., Kishore, P. V. V., Sastry, A. S. C. S., Anil Kumar, D., & Kiran Kumar, E. (2017, September). Selfie continuous sign language recognition with neural network classifier. In *Proceedings of 2nd International Conference on Micro-Electronics, Electromagnetics and Telecommunications: ICMEET 2016* (pp. 31-40). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-4280-5_4

[4] Badhe, P. C., & Kulkarni, V. (2015, November). Indian sign language translator using gesture recognition algorithm. In *2015 IEEE international conference on computer graphics, vision and information security (CGVIS)* (pp. 195-200). IEEE. https://doi.org/10.1109/CGVIS.2015.7449921

[5] Bhuiyan, H. J., Mozumder, M. F., Khan, M. R. I., Ahmed, M. S., & Nahim, N. Z. (2025, March). Enhancing Bidirectional Sign Language Communication: Integrating YOLOv8 and NLP for Real-Time Gesture Recognition & Translation. In *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)* (pp. 168-174). IEEE. https://doi.org/10.1109/ICCAI66501.2025.00035

[6] Birkeland, S., Fjeldvik, L. J., Noori, N., Yeduri, S. R., & Cenkeramaddi, L. R. (2024). Thermal video-based hand gestures recognition using lightweight CNN. *Journal of Ambient Intelligence and Humanized Computing*, *15*(12), 3849-3860. https://doi.org/10.1007/s12652-024-04851-6

[7] Debnath, J., & IR, P. J. (2024, April). Real-time gesture-based sign language recognition system. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 01-06). IEEE. https://doi.org/10.1109/ADICS58448.2024.10533518

[8] Dima, T. F., & Ahmed, M. E. (2021, July). Using YOLOv5 algorithm to detect and recognize American sign language. In *2021 International Conference on information technology (ICIT)* (pp. 603-607). IEEE. https://doi.org/10.1109/ICIT52682.2021.9491672

[9] Garcia, B., & Viesca, S. A. (2016). Real-time American sign language recognition with convolutional neural networks. *Convolutional Neural Networks for Visual Recognition*, *2*(225-232), 8.

[10] Hasan, M., Paul, B. K., Islam, N., & Mostafiz, R. (2025). Advancing real-time sign language detection for deaf and hearing-impaired communities: a customized YOLOv8 approach with tailored annotations in computer vision. *BMC Artificial Intelligence*, *1*(1), 11. https://doi.org/10.1186/s44398-025-00010-9

[11] Huang, J., & Chouvatut, V. (2024). Video-based sign language recognition via RESNET and LSTM network. *Journal of Imaging*, *10*(6), 149. https://doi.org/10.3390/jimaging10060149

[12] 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

[13] Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2019, May). Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)* (pp. 1-8). IEEE. https://doi.org/10.1109/FG.2019.8756576

[14] Kumari, D., & Anand, R. S. (2024). Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism. *Electronics*, *13*(7), 1229. https://doi.org/10.3390/electronics13071229

[15] Mohsen, S., Elkaseer, A., & Scholz, S. G. (2021, September). Human activity recognition using k-nearest neighbor machine learning algorithm. In *Proceedings of the International Conference on Sustainable Design and Manufacturing* (pp. 304-313). Singapore: Springer Singapore.

[16] Otiniano-Rodríguez, K., Cayllahua-Cahuina, E., & Cámara-Chávez, G. (2015, August). Finger spelling recognition using kernel descriptors and depth images. In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images* (pp. 72-79). IEEE. https://doi.org/10.1109/SIBGRAPI.2015.50

[17] Oyedotun, O. K., & Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, *28*(12), 3941-3951. https://doi.org/10.1007/s00521-016-2294-8

[18] Priya, K., & Sandesh, B. J. (2024). Developing an offline and real-time Indian sign language recognition system with machine learning and deep learning. *SN Computer Science*, *5*(3), 273. https://doi.org/10.1007/s42979-023-02482-w

[19] Raheja, J. L., Mishra, A., & Chaudhary, A. (2016). Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis*, *26*(2), 434-441. https://doi.org/10.1134/S1054661816020164

[20] Raj, R., Sreemathy, R., Turuk, M., Jagdale, J., & Anish, M. (2025). Performance comparison of different versions of yolo for Indian sign language captioning in real time of multiple signers. *Procedia Computer Science*, *259*, 991-1000. https://doi.org/10.1016/j.procs.2025.04.053

[21] 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

[22] Salunke, D., Joshi, R., Ranjan, N., Tekade, P., & Panchal, G. (2023, February). Sign language recognition system using customized convolution neural network. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2* (pp. 825-837). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-6634-7_59

[23] Sarkar, A., Gepperth, A., Handmann, U., & Kopinski, T. (2017, December). Dynamic hand gesture recognition for mobile systems using deep LSTM. In *International conference on intelligent human computer interaction* (pp. 19-31). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-72038-8_3

[24] Shenoy, K., Dastane, T., Rao, V., & Vyavaharkar, D. (2018, July). Real-time Indian sign language (ISL) recognition. In *2018 9th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-9). IEEE. https://doi.org/10.1109/ICCCNT.2018.8493808

[25] Sondkar, S., Sanjay, N. M., & Rajendra, P. M. (2025). Comparative Study of YOLOv5 and YOLOv3 based Indian Sign Language Detection Systems. *Grenze International Journal of Engineering & Technology (GIJET)*, *11*.

[26] Wijaya, F., Dahendra, L., Purwanto, E. S., & Ario, M. K. (2024). Quantitative analysis of sign language translation using artificial neural network model. *Procedia Computer Science*, *245*, 998-1009. https://doi.org/10.1016/j.procs.2024.10.328

Authors Biography

K. Priya is working as Assistant Professor in the Computer Science Department at Ramaiah Institute of Technology. She holds a Bachelor's degree in CSE and a Master's degree in CS, Pursuing PhD. She has 1 year of corporate experience as a trainee engineer and 7 years of teaching experience. Her research interests include Hand Gesture Recognition, Image Processing, Machine Learning, and Deep Learning.

Dr.B.J. Sandesh earned his Ph.D. from Visvesvaraya Technological University in 2018, after completing his M. Tech in 2001 and B.E. from Karnatak University Dharwad in 1997. With over 20 years of academic experience, he has served in progressive teaching and administrative roles. He is currently Professor and Chairperson at PES University, Electronics City Campus, since 2019. Earlier, he worked as Associate Professor and Head of the Department at PESIT Bangalore South Campus from 2012 to 2019. He also held leadership positions at PES School of Engineering between 2006 and 2012. Dr. Sandesh began his academic career as a Lecturer at Jawaharlal Nehru National College of Engineering, building a strong foundation in teaching and research.