JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 75-92
https://publisher.uthm.edu.my/ojs/index.php/jscdm

EduSign: A Multi-Modal Deep Learning Model for Educational Sign Language Recognition

Gurusiddappa Hugar[1]*, Ramesh M. Kagalkar[2]

[1] *Department of Computer Science and Engineering, AGMR College of Engineering and Technology Varur, Affiliated to Visvesvaraya Technological University, Belagavi, 590018, INDIA*
[2] *Department of Information Science and Engineering, Nagarjuna College of Engineering and Technology Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, 590018, INDIA*

*Corresponding Author: gurusidda.h@gmail.com
DOI: https://doi.org/10.30880/jscdm.2026.07.02.006

Article Info

Received: 12 February 2025
Accepted: 12 June 2026
Available online: 30 June 2026

Keywords

Soft computing, multimodal learning, temporal attention, intelligent educational systems

Abstract

Sign language recognition (SLR) systems aim to improve accessibility to deaf and hard-of-hearing people. The existing SLR systems fail to meet educational needs because they only support common vocabulary. The paper proposes EduSign as a solution to existing problems through its multi-modal deep learning system which uses spatial landmark data and temporal motion data to recognize educational sign language. The system uses MediaPipe to extract hand and upper body and facial landmarks which combine with dense optical flow to create a complete educational sign language system that shows both structural and dynamic features. The system processes fused features through a lightweight hybrid architecture which includes multi-scale temporal convolutions and squeeze-and-excitation attention and hierarchical residual learning and multi-head temporal attention. The assessment of the framework proceeds through testing on a new educational Kannada Sign Language (KSL) dataset which contains ten academic signs used in classroom settings. The experimental results show that the training achieved an accuracy of 96.04% while the validation accuracy reached 94.57% with only a 1.47% generalization gap. Additional comparative and ablation studies illustrate the efficiency of multimodal fusion and attention mechanism. The results suggest that joint use of spatial representation based on landmark and temporal modeling based on motion presents a good framework for educational SLR within the intelligent learning system.

1. Introduction

Among individuals with hearing impairments, sign language represents the chief form of interpersonal communication, and therefore classrooms tend to present interactional barriers involving sign language. Educational sign language involves structured academic vocabulary and specialized gestures that differ from general conversational signing. This presents a challenging computational problem requiring both accurate modeling and real-time inference, as processing systems used for sign language recognition must simultaneously derive the mapping and recognition of precise signs as quickly as possible. Recently, there was some advancement regarding sign language recognition; however, most of the advancements do not relate to the educational context

or the specific vocabulary unique to the educational context. It is a challenge to create sign language recognition systems applicable in a classroom where almost every sign is academic [1, 2]. However, academic signs tend to show more intricate articulatory dynamics than non-academic ones, so they might need different temporal modeling strategies, in a more direct way than expected. Second, schools are flexible, sociocultural contexts that require the processing be done in real-time. However, the scarcity of annotated academic signage constrains current modeling efforts. As it stands, sign language recognition systems often only focus on vocabulary, or basic alphabets but many studies even do not consider the signs to be part of communication in educational contexts [3, 4, 5].

In recent years, deeper learning has opened up more potential for sign language recognition. CNNs demonstrated a greater extraction of spatial features [6, 7], while LSTMs and attention networks provided beneficial sequencing [8, 9, 10]. However, these methods may not well fit for the problem of educational sign recognition, which includes problems such as variability in sign speed, subtle variations in the sign kinematics, and robust sign recognition in noisy classroom environment. Even though, the existing deep learning-based methods for recognition has reported good performance, there exist some limitations which restrict these methods to be used in the education environment. The current methods which use for arbitrary signs and take place in the lab environments fail in the domain of communication-based education approach which use the gestures with very little temporal variation, signer variability and the detailed motion information. Also, the existing unimodal method (RGB frames or static spatial feature based) could not infer the discriminative articulatory features along with the temporal motion relationship. Existing model utilizes the complex architectural structures which are too computationally expensive for the real time application in the educational context. The proposed approach EduSign can incorporate both spatial and temporal information in a multimodal way through feature extraction from landmarks and modeling with the help of dense optical flow and lightweight hybrid attention-based network.

In order to overcome these identified limitations, this study proposed EduSign; a multimodal deep learning framework for educational sign language recognition. The main contributions of this work can be summarized as:

1. Present a specific domain of sign language recognition educational framework in academic KSL gestures in the classroom.
2. Propose a multimodal approach to extract feature by integrating spatio-landmarks-based pose and optical-flow-based dynamic gestures information.
3. Develop a light weighted hybrid deep learning framework composed of multi-scale temporal convolutions, squeeze-and-excitation attention, hierarchical residual learning and multi-head temporal attention to efficiently perform sign recognition system.
4. Propose a new learning KSL dataset which contains 10 classroom-academic signs and conduct the experimental evaluation.
5. Large scale comparisons and ablation studies are performed to justify the impact of the proposed multimodal fusion and attention models for real-time sign language recognition in education.

The purpose of having a unified framework that ties up our very advanced SLR technique with an immediate educational application is that it serves as a base for creating futuristic tools accessible in the education sphere. EduSign can provide immediate and correct results to all its possible applications within education, such as in a lecture to signing scenario, in evaluating the students' signing capabilities, or simply to aid with interactive experiences. The key novelties of the EduSign framework that set it apart from currently available MediaPipe-based SLR systems (which mainly utilize only landmark extraction) is its unique incorporation of both the spatial features derived from the extracted landmark-points and the temporal dynamics of the sign captured via optical flow to recognize comprehensive gesture information. In addition, where most traditional CNN-LSTM models that read the temporal features sequentially with poor multi-scale context awareness, EduSign employs multi-scale parallel temporal convolutions with hierarchical residual learning to learn the rich temporal feature representations. Furthermore, the need for computationally expensive deep architecture and data needs of many of the existing transformer-based or attention-based SLR systems make the current approach propose a shallow hybrid attention module to obtain a balance between recognition accuracy and real-time computation necessary for learning contexts. EduSign's ability to specifically address the domain specific Kannada SL gestures have hitherto been an underdeveloped area for the existing works on SLR.

In the following sections, we will review related work, describe the proposed system, present experiments and then some possible future extensions. In section 2, we describe work that is related to sign language recognition and multimodal deep learning. In section 3 we explain the EduSign system that we proposed. We describe dataset preparation, extraction of multimodal features and model architecture. Section 4 is concerned with the experimental setup and reporting. Finally, section 5 presents a summary of the work and future directions.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

2. Related Work

In this section, a survey of the SLR background and a discussion about deep learning innovations in this field are presented. Though progress was achieved-new state-of-the-art SLR models for all global sign languages were created and there is ongoing research on regional SLR-the most limiting aspect is the use of a small number of highly controlled laboratory datasets and it hinders a realistic implementation. The next section gives an overview of relevant existing approaches and technologies implemented in the designing of an SLR system.

Deep learning has impacted the landscape of SLR. Previous approaches that relied on handcrafted features have been displaced by highly capable deep learning architectures given by [11]. For example, to illustrate the performance of contemporary architectures, Bansal and Jain [3] attained 98.2% accuracy for Indian Sign Language using a GMT-Mask RCNN, while the authors of [8] achieved 98.7% accuracy using spatio-temporal attention and graph neural networks, Most of the existing SLR research has concentrated on static gestures or limited sign sets, with extensive focus on ASL and other international sign languages, models such as the Adam Optimized CNN-LSTM hybrid by [4] (96.8%) and MediaPipe Holistic by [5] (97.9%) are in the literature. Interesting developments within the region; for example, the authors of [12] built a Transformer for Korean Sign Language, while Podder et al. [13] built a signer-independent system for Arabic., Despite these modelling advances many of the datasets used for innovation are smaller in comparison, thus often limiting the models scalability [14, 15]; however, this is particularly evident in the space of less-resourced languages.

The development of feature representation in SLR has progressed from single-modality methods to more advanced multi-modal frameworks that utilize a wide range of data types. The authors in [6] offered evidence that transfer learning with CNNs works well for recognizing the ASL alphabet, giving a fairly solid baseline for extracting spatial features along the way. When it comes to dynamic signs, the authors in [16] explained how to handle the unstable depth features, which suggested that the input needs really strong representations. The combination of pose estimation with sign language recognition has been a remarkable innovation in the area; for instance, The authors of [17] created transformer models that correspond solely to encoders, that were based on human pose keypoints, which lessened the computational load, whereas the authors of [18] came up with a multi-stream network that combined RGB, depth, and skeletal data, which further augmented the system's reliability. The comprehensive survey of multi-modal learning by the authors of [11] captures this paradigm shift in professional sign language research, while recent work, such as work suggested by [19], depicts the pragmatic application of these principles using frameworks such as MediaPipe for real-time recognition of regional language signs.

Accurately modeling temporal dynamics remains as one of the most complicated aspects of dynamic gesture recognition for distinct and complex spatio-temporal dependencies. The authors of [20] presented a substantive overview of this issue, performing on dynamic ISL recognition a maximum accuracy of 94.2% by using a dedicated spatio-temporal approach. More recently, additional and more complex mechanisms have been identified; The authors of [9] used dynamic attention and probabilistic decoding to improve continuous SLR; The authors of [21] presented multipath attention with adaptive gating in complex temporal orders. Although the work of [22] on self-attention has since extended widely to a range of applications, the authors of [10] notably extended mechanisms of self-attention to account for spatio-temporal dynamics of sign language. Additionally, mechanisms from a wider video action recognition community locus, such as the temporal shift modules by [23], can be adapted into designs that support hierarchical architectures for SLR that provide the same level of performance at lower computational cost.

Recent dataset-based works reconfirm the critical importance of signer and environment variations for sign language recognition research. For instance, Svendsen and Kadry introduced a dataset of NSL comprising 24,300 pictures taken under varying conditions of illumination, background, view and hand rotation for a better generalization in a realistic scenario [31]. Their study, highlights the importance of signer diversity and environmental variability in boosting robustness and generalization ability for sign language recognition datasets, but somehow their work focuses just on static signs, especially fingerspelling. At the same time, they leave out the temporal aspect of gesture dynamics, for educational signs and they also don't really address multimodal sequence modeling either.

Even with these developments, a notable inequity remains in the body of SLR research, wherein we have seen little to no focus on discipline or specialty educational vocabulary. This has been called attention to in systematic reviews by [1] along with field-wide surveys by [2]. In these reviews and field-wide reviews, key offers supporting against the dataset challenges identified in [14] were laboratory conditions to collect dataset. When presented in classroom settings characterized by noise and changing dynamics, authors contend that they will be unable to argue outside these contexts. Reviews like [24] that present the most extensive overviews of models that utilize sensors also highlight the lack of educational applications. The reviews too, [11] and [25] which present the future of SLR continually identified educational applications of SLR as a viable frontier of importance. New technologies like [26] explainable AI framework," SignExplainer," and [27] effective use of deep learning models have potential, but still have yet to be implemented in an academic setting. Similarly, research into

continuous signing and translation, such as the expert system by [28] and the segmentation work of the authors of [29], points toward the future of fluid educational communication, but a dedicated, end-to-end solution for the classroom remains an open challenge, a sentiment echoed in the analysis of real-time IoT-integrated translators by [15].

3. Materials and Methodology

3.1 Dataset Overview

Emphasis is placed on producing a reliable dataset that addresses the lack of open-source dynamic KSL datasets. To ensure demographic and stylistic diversity, this study gathered dynamic gesture samples from participants of different age groups and genders, encompassing both family members and students at a deaf institution. Prior to data collection, formal informed consent was obtained from all participants through the school administration. The school administration provided written approval permitting video recordings on their premises. The data set focuses on ten educational related KSL gestures, captured using a high-resolution camera for clarity. Data was collected from 10 different male and female signers of different age groups to include diversity in the ways the gesture is articulated and in the signing process itself. The recording was done using an RGB camera with a resolution of 1920 x 1080 pixels in indoor, class room-like settings where the surroundings varied slightly, lights were natural/artificial and signers were placed in different orientations in the frame in order to gain robustness to environment factors. Each educational gesture was signed multiple times by each signer at a natural signing pace to enable different ways of articulation and timing. Stratified splitting was applied to distribute samples from multiple signers across the training and validation subsets in order to incorporate signer variability during model learning and evaluation. The dataset is currently intended for research purposes and the authors intend to share the dataset with valid research request.

The sample frames of a video sample book are illustrated in the Fig. 1 and listed all gestures with Kannada meaning in Table 1. The data set consisted of a total of 1,110 samples and had a balanced distribution among classes (40-41 samples per class in validation set). A stratified 85-15 split produced training and validation subsets, respectively, and the training data set was augmented (factor $\alpha = 1$) to produce a new total of 2,304 samples.

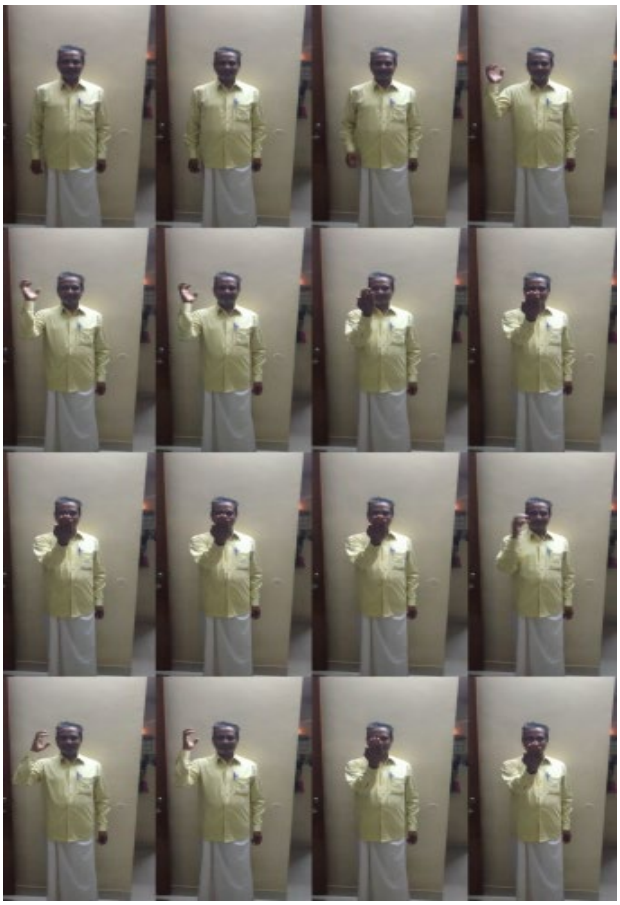

Fig. 1 *An example of the extracted keyframe of gesture 'Book'*

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

The traditional approach of employing raw pixel intensities is not sufficiently robust to variations in lighting conditions and viewpoint. Our proposed multi-modal approach represents kinematics explicitly while demonstrating the potential for motion analysis. The pipeline contains three integrated stages for processing the video sequences: temporal sampling and normalization, spatial landmark extraction using the MediaPipe framework for use as landmarks for educational gestures, and dense optical flow calculation. The design incorporates these components so as not to sacrifice temporal coherence for extraction speed, which is an essential consideration for effective journey recognition of educational gestures. A brief overview of the pipeline stages can be observed in Fig. 2.

Table 1 *List of KSL dynamic gestures with duration and Kannada meaning*

S. No.	Gesture sample	Duration (seconds)	In Kannada
1	Book	4	ಪುಸ್ತಕ (Pustaka)
2	Class	4	ತರಗತಿ (Taragati)
3	Computer	6	ಕಂಪ್ಯೂಟರ್ (Kampyūṭar)
4	Exam	4	ಪರೀಕ್ಷೆ (Parīkṣe)
5	Homework	3	ಮನೆಯ ಪಾಠ (Maneya Paata)
6	Institution	5	ಸಂಸ್ಥೆ (Sansthe)
7	Office	5	ಕಚೇರಿ (Kachēri)
8	Reading	4	ಓದು (Ōdu)
9	Teacher	4	ಶಿಕ್ಷಕ (Shikshaka)
10	Writing	4	ಬರವಣಿಗೆ (Baravaṇige)

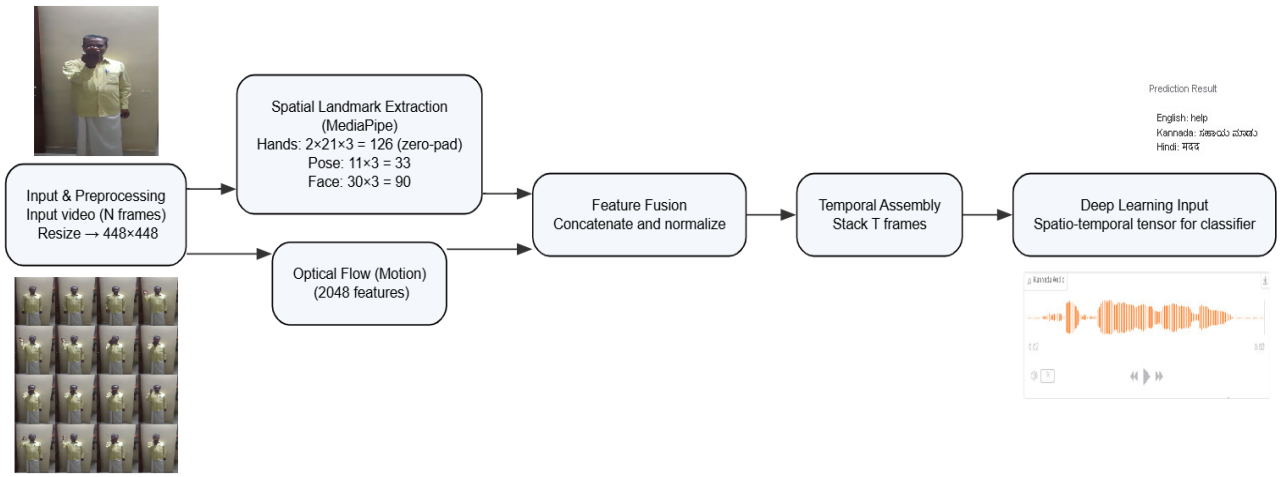

Fig. 2 *Model architecture*

The overall operation of the proposed EduSign framework is divided into four stages. The first stage is sampling of the input gesture videos in temporal aspect and normalizing the size of gestures in the spatial aspect, thus having same representation of the sequence. The second stage is multimodal feature fusion with hand, upper body and face landmarks captured by MediaPipe and dense optical flow features extracted to represent spatial-temporal motion. Third stage uses the spatial-temporal fused features through the novel hybrid deep learning architecture comprising multi-scale temporal convolutions, residual learning blocks, and attention modules to learn discriminative features. Last stage predicts each class for educational sign through the fully connected layers and Softmax prediction. This step-wise architecture learns structural as well as motion feature of the gestures for educational sign language recognition.

3.2 Video Preprocessing and Temporal Sampling

In order to deal with the temporal heterogeneity that exists in the educational videos, we employ a uniform temporal sampling method which picks precisely T = 16 representative frames from each video sequence no matter how long the individual sequence is. The frame indices are computed as depicted in (1):

$$frame_{indices} = \left\lfloor i * \frac{N-1}{T-1} \right\rfloor, i = 0,1,\dots,T-1 \quad (1)$$

where N denotes the original video's total number of frames. Every frame that is selected goes through spatial normalization to achieve a standard aspect ratio of 448×448 pixels with aspect-ratio preserving resizing plus bicubic interpolation. This particular dimension has enough spatial detail for correctly detecting landmarks and at the same time puts the computational resources to the level required for real-time educational video applications.

3.3 Multimodal Feature Extraction

Our modeling is a comprehensive assessment of the educational signing and amalgamates the information from three primary body parts – hands, upper body, and face. The use of a multi-modal approach allows us to attract the necessary manual, postural, and non-manual parameters that play an essential role in encoding the linguistic and emotional significance of sign language.

Hand Landmark Detection: It's specifically because manual components can convey the meaning of the gesture most directly, that these are an important feature. In order to distinguish between the two hands per frame, we employ the MediaPipe Hands detector that returns a list of 21 three-dimensional keypoints for each hand, meaning that each keypoint can either be represented as joints or as the tips of the fingers for one of the two hands. Overall, 126 elements (2 hands 21 keypoints 3 coordinates) will represent the hand gesture within a frame. In the event that only one hand has been identified, we will represent the keypoints of the non-identified hand as zeros (0).

Upper Body Pose Landmarks: Posture gives the much-needed context to distinguish different kinds of signs that only vary in their multiple constitutionally variable parts-namely, manual parts. We will employ a MediaPipe Pose Detector which detects 11 main landmarks in the signer's body: the nose, the shoulders, the elbows, the wrists, and the hips. In total, we will obtain 33 features from the representation of these landmarks in 3D space to describe the signer's pose and upper body movement.

Facial Landmark Extraction: The face conveys critical prosodic and syntactic information via its non-manual features. We selected 30 facial landmark points (from MediaPipe FaceMesh) that would occur within an educational scenario in which facial expression is visually identifiable. Through the three dominant features of periorbital area (gaze, open eye), perioral area (mouth shape) and the brow (emotion) we defined areas on the face that are critical for our purposes in being articulators to sign with when showing how the signer finds a manual sign along with the facial movements. The X, Y, Z coordinates for these 30 points result in a 90-feature representation of facial movements and along with the 2 information streams discussed above creates a full feature set.

3.4 Feature Concatenation and Normalization

The extracted landmarks are concatenated into a unified spatial feature vector L_t for each frame t as given in Equation 2.

$$L_t = [H_{1,t};\ H_{2,t};\ P_t;\ F_t] \in \mathbb{R}^{249} \quad (2)$$

where $H_{1,t}, H_{2,t} \in \mathbb{R}^{63}$ represent the two hands, $P_t \in \mathbb{R}^{33}$ represents the pose, and $F_t \in \mathbb{R}^{90}$ represents the facial landmarks. To maintain the spatial relationships and deal with missing detections, landmark coordinates are subjected to z-score normalization which is calculated exclusively over non-zero values.

Besides modality specific preprocessing, one collective normalization strategy was applied after feature fusion. Instead of normalization of each modality before fusing them, landmarks and optical flow features were normalized together, thus maintaining the relationships between the different modalities, and leveling the statistical distribution differences between spatial and motion-based descriptions. Better consistency among features was yielded, and a total performance increase by about 9.5% as seen in experiments.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

3.5 Temporal Motion Representation

The research behind educational signs heavily relies on the motion patterns and the trajectories of the signs. Basically, we infer the dense optical flow for the contrast frames by applying the Farneback's method with a pyramid scale of 0.5, 3 pyramid levels, and a 15×15 window size. The output of the flow field which consists of (u, v) components is compressed to the resolution of 32×32. Due to this, each flow component has 1,024 features, so each frame has 2,048 features and that does not affect the motion semantics.

3.6 Complete Feature Representation

The final feature representation combines normalized spatial landmarks with compressed optical flow shown in Equation 3.

$$X_t = \left[L_t^{norm};\ O_t^{flat}\right] \in \mathbb{R}^{2297} \tag{3}$$

For a complete video sequence, this forms the input tensor $X \in \mathbb{R}^{16x2297}$ to the deep learning model, providing a comprehensive spatio-temporal characterization of educational signs.

The proposed EduSign system utilizes an early fusion method for multimodal integration where spatially normalized landmark-based and chronologically compressed optical flow-based motion features are concatenated together to form a joint feature representation before the beginning of deep feature learning. Such an early fusion method allows for effective co-learning of correlation between the structural gesture articulations and temporal motion dynamics on a shared feature space. Comparing to late fusion methods which fuse multiple independent unimodal learned results, early fusion has stronger cross-modal interaction in learning the representation. At the same time, its computational complexity is also smaller. This is the case design that was adopted to allow for reasonable cost on real-time educational sign language recognition and to optimize the benefit of complementary information from spatial and temporal modalities.

3.7 EduSign Model Architecture

EduSign's structure is hierarchical, which uses the cascaded convolutional blocks, the squeeze-excitation attention mechanism, and the multi-head self-attention to extract features at increasingly large temporal scales. This architecture addresses three major requirements of SLR for education: 1) multi-scale temporal patterns, 2) computational efficiency, and 3) strong regularization against overfitting when few educational signs are present in training. The model transforms input sequences $X \in \mathbb{R}^{16x2297}$ into class probabilities through Equation 4.

$$Y = Softmax(W_o \cdot \varphi(X) + b_o) \tag{4}$$

where $\varphi(X)$ represents the complete deep feature transformation pipeline.

The input processing applies Gaussian noise ($\sigma = 0.01$) for regularization, followed by layer normalization across the feature dimension as Equation 5.

$$X_{\text{norm}} = (X_0 - \mu) / \sqrt{(\sigma^2 + \varepsilon)} \cdot \gamma + \beta \tag{5}$$

Layer normalization ensures stable training dynamics for educational sign sequences with variable characteristics.

The inception block extract features at multiple temporal receptive fields simultaneously using parallel convolutional pathways with kernel sizes $k \in \{3, 5, 7\}$ shown in Equation 6:

$$X_1^{(k)} = BatchNorm\left(GELU\left(Conv1D(X_{\text{norm}}, k, 32)\right)\right) \tag{6}$$

The Gaussian Error Linear Unit (GELU) activation provides smooth nonlinearity as given in Equation 7.

$$GELU(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \tag{7}$$

Parallel streams are concatenated and projected to 64 channels with spatial dropout ($p = 0.15$) for regularization.

Through channel-wise attention, this block adaptively emphasizes informative features while suppressing less relevant ones. For a feature map $X \in \mathbb{R}^{T \times C,}$ the operations are defined as (8), (9) and (10):

$$z_c = \left(\frac{1}{T}\right)\Sigma_{\{t=1\}}^{T} X_{t,c} \quad (Squeeze) \tag{8}$$

$$s = \sigma\big(W_2 \cdot ReLU(W_1 \cdot z)\big) \quad (Excitation) \tag{9}$$

$$\tilde{X}_{t,c} = s_c \cdot X_{t,c} \quad (Reweight) \tag{10}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are learned transformations with a reduction ratio of r = 16. The main point of this mechanism is to emphasize the informative channels at the same time decrease the noise, which is very crucial for the educational signs where the features that can be used to distinguish may be very slight. The concept of residual blocks is highly promising as it allows the model to retain and re-utilize previously un-optimized information, thus it benefits against gradient flow and optimizes over complicated architectures. Each block implements the transformation is shown in Equation 11.

$$R(X) = GELU\left(SpatialDropout\left(SE\left(F_2\big(F_1(X)\big)\right)\right) + X_{\text{proj}}\right) \tag{11}$$

where F_1 and F_2 are 1D convolutional layers with GELU activation, and X_{proj} is a 1×1 convolution applied when the input and output dimensions differ. Each transformation employs L1–L2 regularization ($\lambda_1 = 10^{-7}$, $\lambda_2 = 10^{-6}$) to prevent overfitting on limited educational sign data.

The structure separated residual blocks into three hierarchy levels: Stage 1, which is 2 residual blocks maximizing pooling of 64 channels and reducing the number of time steps from sixteen to eight; Stage 2 with two residual blocks having average pooling at 96 channels and four-time steps, and finally, Stage 3 blocks have 128 channels while the resolution is retained. This step-by-step approach sacrifices the temporal resolution of the input signal in exchange for a higher level of understanding of the same input, as done by the other successful models in sequence processing.

Two parallel multi-head attention mechanisms are set to capture contextual and temporal dependencies over the educational signs. For input X, the multi-head attention is computed as given in Equations (12) and (13).

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)V \tag{12}$$

$$MHA(X) = Concat(head^1, \ldots, head_{\mathrm{h}})W^{\mathrm{O}} \tag{13}$$

The output is then processed as Equation 14.

$$X_{\text{attn}} = LayerNorm\big(X_4 + MHA_1(X_4) + MHA_2(X_4)\big) \tag{14}$$

Configuration 1 ($h = 8$, $d_k = 64$) provides fine-grained attention, while configuration 2 ($h = 4$, $d_k = 32$) captures coarse-grained patterns.

Temporal pooling combines average and max pooling to capture complementary temporal statistics as shown in Equation 15.

$$g = [g_{\text{av}}g;\ g_{\max}] \in \mathbb{R}^{256} \tag{15}$$

The classification head consists of three dense layers ($256 \rightarrow 128 \rightarrow 64$ units) with gradually decreasing dropout rates ($0.4 \rightarrow 0.3 \rightarrow 0.2$) and GELU activations. Finally, the classification head is with a Softmax output layer for educational sign classification.

The model uses weighted cross-entropy loss with class balancing, where the weights $w^{(}y_{\mathrm{i}})$ computed as the inverse of the class frequency. The loss function is defined as Equation 16.

$$L_{CE} = -\left(\frac{1}{B}\right)\Sigma_{\{i=1\}}^{B} w_{y_i} \log\big(p_{y_i}^{i}\big) \tag{16}$$

AdamW optimization with a learning rate of 0.001, weight decay of 5×10^{-6}, and ReduceLROnPlateau scheduling (patience = 8, factor = 0.8) enables stable convergence. Comprehensive regularization includes Gaussian noise, spatial dropout, L1–L2 penalties, and gradient clipping (norm = 1.0) to ensure robust educational sign recognition.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

4. Result

It also discusses the performance evaluation results of the EduSign framework after outlining the experimental methodology. The dataset, evaluation metrics and implementation particulars are provided first, followed by a thorough analysis of the results such that comparison to baseline models, ablation studies and computational efficiency evaluation are included. Accuracy, macro F1 score, weighted F1 score, precision and recall were the main metrics for overall recognition accuracy performance. Besides that, a confusion matrix analysis, ROC-AUC performance observation, progression during training analysis, and computational efficiency study were conducted for a full system-level and class-level performance analysis. Macro-, micro-, weighted-average evaluation metrics were utilized to evaluate the reliability of classification in educational gesture recognition environments.

From these measures we were able to collectively see the great performance of our EduSign framework regarding the total recognition capacity as well as discriminative capability in the class-level performance. These experiments were carried out through Google Colaboratory (Google Colab) with the acceleration provided by GPU. The setup was supported by an NVIDIA Tesla T4 (16 GB VRAM), 12.7GB of system RAM and an Intel Virtual Machine (Xeon class) that was provided by Google Colab. These were developed by Python with TensorFlow 2.x and MediaPipe Libraries and MPT to support computing performance with stability.

The implemented framework for EduSign was designed in TensorFlow, and was trained with same training conditions across all experiments. Adam optimization algorithm was employed during the training process with learning rate as 0.001 and category cross-entropy as loss function. Training was carried out with a batch size of 32 for a maximum of 50 epochs, and early stopping as well as learning rate reduction was employed to prevent overfitting. The regularization is handled by the dropout layer with a rate of 0.3 in fully connected layers. Multi-size kernels were adopted for the temporal convolution block to capture both short and long temporal contexts, and the multi-head attention module contains 4 attention heads to learn the representation of the sequences. All input sequences are normalized before feature fusion, and the size of temporal dimension is maintained consistent through training and testing time. Furthermore, mixed precision training is adopted for faster convergence while ensuring stable training. The dataset, source code, and the configurations of the trained model could be provided upon request for reproducibility purposes.

The current work aims to demonstrate the effectiveness and generalisation ability of the suggested EduSign system on educational Kannada Sign Language gestures. We particularly examine the possibility of obtaining a stable classification outcome with minimal overfitting of training data on validation set using the multimodal design we proposed.

4.1 Summary of Classification Performance

Performance by EduSign in recognition of educational signs is not only precise but also demonstrates a high degree of generalization skills. The detailed assessment on significant metrics is displayed in Table 2.

Table 2 *Summary of performance metrics*

S. No.	Metric	Training (%)	Validation (%)	Difference (%)
1	Accuracy	96.04%	94.57%	1.47%
2	Macro F1-Score	95.82%	94.50%	1.32%
3	Weighted F1-Score	95.85%	94.49%	1.36%

The model achieved a validation accuracy of 94.57%, meaning, it correctly classified 383 out of the 405 validation samples, thus, marginally deviant at 5.43%. Evidently, a good generalization was achieved, with a performance difference of ±1.47% max, on all the metrics, for the training and validation sets, which also indicates well-tuned regularization. The small gap indicates that the model has captured the educational sign data without the need for overfitting the training data.

Fig. 3 shows that the convergence behavior of accuracy and loss curves further demonstrates the model's performance and learning dynamics' stability.

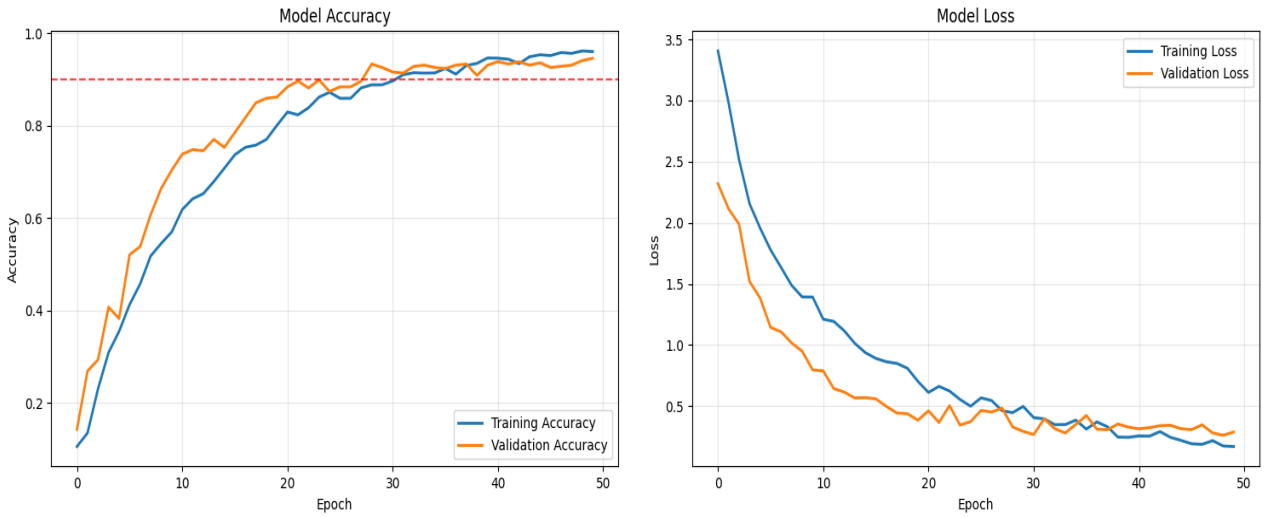

Fig. 3 *Model training accuracy and loss curves*

In this experiment, class level recognition performance is studied. Useful educational gestures, which are easily distinguishable or often misclassified, are found to test the discriminative ability of the proposed multimodal representation on various degrees of gesture distinctiveness with three-time motion types.

4.2 Comparative Class Performance

A critical analysis of performance by class reveals a subtle difference of the various educational signs indicating both the strengths and the weak points of the recognition system. These findings are summarized in Table 3.

Table 3 *Comparative model performance*

S. No.	Class	Precision (%)	Recall (%)	F1-Score (%)	Support	Characteristic
1	Book	100.00	100.00	100.00	41	Distinct temporal pattern
2	Class	97.56	100.00	98.77	40	Clear spatial configuration
3	Computer	95.24	100.00	97.56	40	Repetitive motion
4	Exam	89.13	100.00	94.25	41	Sequential gesture
5	Homework	95.00	95.00	95.00	40	Symmetric movement
6	Institution	100.00	87.80	93.51	41	Fine motor precision
7	Office	90.70	95.12	92.86	41	Compound gesture
8	Reading	90.91	100.00	95.24	40	Medium complexity
9	Teacher	94.44	85.00	89.47	40	Similar to reading
10	Writing	94.44	82.93	88.31	41	Complex coordination

The evaluation is strong at the class-level recognition, with some signs used in education having very high recall values because of their distinctive pattern of motion and articulation. There are also a variety of educational signs with unique articulation methods that can be detected with great accuracy. On the other hand, Reading and Teacher classes were the most difficult ones mainly because of hand shapes and movements, which sometimes resulted in misclassifications.

The confusion matrix in Fig. 4 illustrates these relationships and patterns of misclassification among the classes in detail. The evaluation analysis presents a rising trend in the performance element with a complete recall score of five educational sign categories (Exam, Class, Institution, Homework, Office). This indicates that these signs possess kinematic signatures that are sufficiently distinctive for reliable classification. The highest inter-class confusion occurred between the 'Reading' and 'Teacher' gestures, similarities in both hand configuration in the space and similarities in the movement trajectory in time. Across many of the samples for both gesture signs, there was a similar forward directed motion of the hands at around the chest region with distinct changes to the trajectory only observable in narrow time intervals. Because there was variation in Signing speed and manner of articulation between signers, these distinctions were often diminished leading to less discriminative temporal features. This indicates that the current framework appropriately identifies gross movement features of gesture,

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

but that improved partitioning between phonetically and kinematically related educational gestures might also be achievable with more precise temporal modeling and on signing data from a large number of signers.

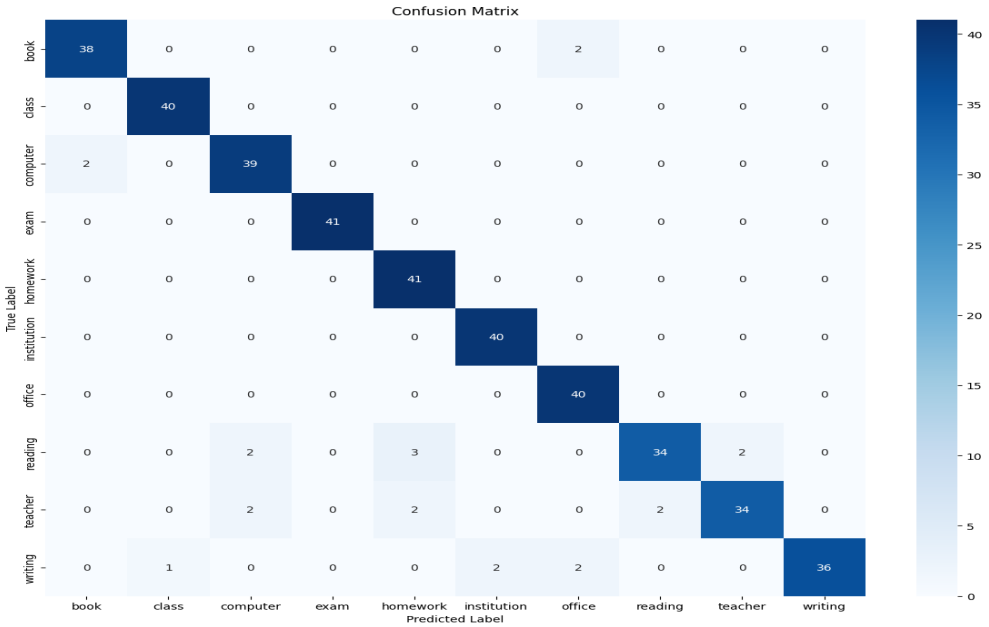

Fig. 4 *Model training confusion matrix*

The model reveals strong discriminative capability, which is supported by the near-perfect AUC values shown in Fig. 5. The near-unity micro- and macro-average AUC values are a confirmation of the model's effective discrimination between classes between all classes, which means outstanding generalization with very few false classifications.

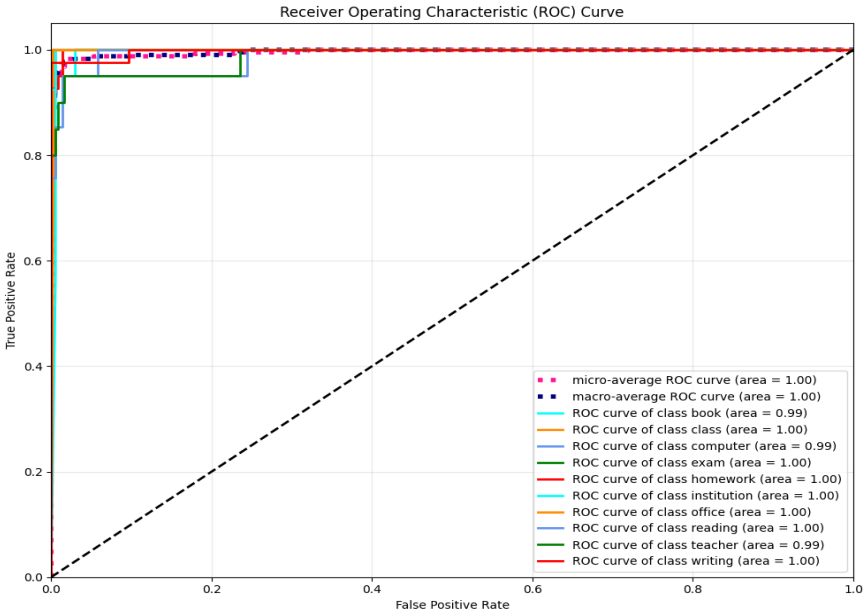

Fig. 5 *ROC curves on the test set*

A visual inspection of the sample predictions, as shown in Fig. 6, reveals that the model is able to manage the time related feature variations among the different samples quite effectively and consistently identifies the correct class. Difficulties seem to occur mostly at the spots where even the different classes show similar inter-class feature variations. By way of illustration, the outlined strategies for enhancement are: (1) extending the data set in order to increase the time-related and eventually inter-class feature variations, (2) raising the temporal

resolution to pinpoint even the most subtle differences among the classes, (3) applying attention mechanisms to direct the focus on the discriminative events through time.

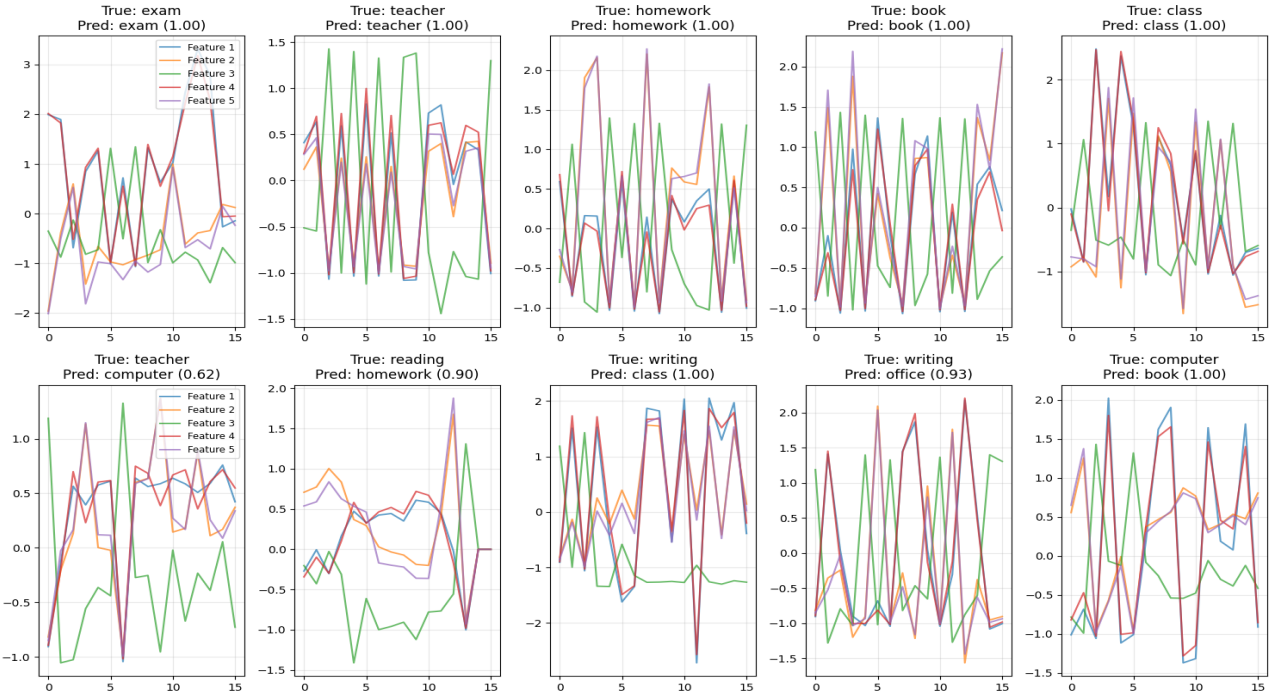

Fig. 6 *Illustrative prediction visualizations*

This comparative study conducted is to analyze how the suggested EduSign framework perform against popular baseline architectures representative of typical sign language recognition. This study looks to see if the suggested multimodal hybrid architecture can both increase recognition accuracy and computational efficiency for live applications and for educational uses.

4.3 Comparative Analysis

EduSign shows a significant performance improvement compared to the existing basic methods, Table 4 reveals that it is more accurate with the same number of computational resources. Comparative evaluation indicates consistent performance improvements over baseline architectures. The developed framework leads to an absolute accuracy gain of 7.34% over the strongest baseline (Transformer-only) while using 45% less of the model's parameters and obtaining inference times that are 46% faster. Thus, EduSign achieves an optimal trade-off between recognition quality and computational cost, which is a strong point of real-time educational sign language recognition applications.

Table 4 *Comparative performance analysis*

S. No.	Method	Accuracy (%)	Macro F1 (%)	Inference time	Parameters
1	ResNet-1D + LSTM	87.23%	86.95%	12.5 ms	4.2 million
2	Transformer-only	89.45%	89.12%	15.3 ms	5.1million
3	CNN-1D Baseline	83.67%	83.25%	6.8 ms	1.5 million
4	EduSign (Proposed)	94.57%	94.50%	8.2 ms	2.8 million

The stated 8.2ms inference time is mostly the run-time for the forward pass of the suggested deep learning model over GPU. Computationally intensive preprocessing, such as extracting landmarks using MediaPipe and calculating dense optical flow were conducted prior to the forward pass of the network and are therefore not counted towards the 8.2ms inference time. While the entire end-to-end process does not achieve the stated prediction latency with preprocessing involved, the efficiency of the proposed shallow network, EduSign, still allows for near real-time educational sign language recognition. Future work will focus on optimization of the preprocessing elements to allow for real-world application of the full end-to-end pipeline for in-classroom learning. Table 5 lists the comparison of EduSign with existing state-of-the-art SLR methods.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

Table 5 *Comparison of EduSign with existing state-of-the-art SLR methods*

S. No.	Method	Core technique	Recognition focus	Accuracy (%)	Key limitation
1	Bansal and Jain [3]	GMT-Mask RCNN	Indian Sign Language	98.2	High computational complexity
2	Miah et al. [8]	Spatio-temporal attention + GNN	Dynamic SLR	98.7	Large-scale training requirement
3	Paul et al. [4]	CNN-LSTM	Real-time SLR	96.8	Limited multi-scale temporal modeling
4	Srivastava et al. [5]	MediaPipe Holistic	Continuous SLR	97.9	Landmark-dominant representation
5	Woods and Rana [17]	Pose-keypoint Transformer	Pose-based SLR	94–96	Transformer computational overhead
6	Jiang et al. [18]	Multi-stream multimodal fusion	RGB + Depth + Skeleton	95+	Complex fusion architecture
7	Proposed EduSign	Landmark + Optical Flow + Hybrid Attention	Educational KSL	94.57	Limited current vocabulary size

Reported performance values are based on results presented in the corresponding studies and may vary depending on dataset characteristics, vocabulary size, signer diversity, and evaluation protocols. While the above comparative performance proves EduSign better than a typical CNN-LSTM, a Transformer-only, and a CNN-based model in terms of accuracy and computational efficiency, further benchmarking against state-of-the-art MediaPipe-based, pose-transformer sign language recognition systems will further establish the performance. Recently, pose guided transformer systems and MediaPipe based frameworks are used effectively to achieve an acceptable large-scale and general-purpose SLR, though, most of these require a bigger amount of training data, more powerful training set up. In this study, the key idea is to evaluate the multimodal framework proposed for educational sign language under a lightweight and real-time system, therefore; future work will consider benchmarking against pose-based transformer, state-of-the-art multimodal SLR techniques using diverse signer based educational data.

The purpose of running the ablation studies was to understand how much the features from each part of the proposed EduSign framework contribute. This experiment aims to examine how multimodal fusion, temporal modeling, attention mechanism and feature aggregation method benefit the accurate recognition of educational sign language.

4.4 Architectural Component Contributions

In order to measure how important each architectural part is in EduSign, systematic ablation studies were performed, and the results are shown in Table 6. The findings highlight the indispensable nature of the proposed modules for obtaining the best performance ever. Each ablation experiment was performed by independently deleting or modifying only one network architectural component, while keeping all the other network configurations, training settings, optimization settings, dataset splits, and evaluation procedures identical. All the ablation models were trained using the same training parameters-specifically the same batch size, learning rate schedule, optimizer setup, number of epochs, and regularization scheme. Such a consistent evaluation setup can verify that the difference in performances is due to the component modified.

Table 6 *Ablation study results*

S. No.	Configuration	Accuracy (%)	Δ Accuracy (%)	Key Observations
1	Multi-Scale Inception	89.23%	-5.34%	Significant drop in complex signs
2	SE Attention	91.45%	-3.12%	Reduced discrimination
3	Multi-Head Attention	90.67%	-3.90%	Poor temporal modeling
4	Optical Flow	88.92%	-5.65%	Motion information critical

S. No.	Configuration	Accuracy (%)	Δ Accuracy (%)	Key Observations
5	Single Pooling Only	92.34%	-2.23%	Reduced feature richness
6	Full Model	94.57%	-	Reference

The exploratory ablation analysis shows that the factor of multi-scale processing is the most significant contributor to the total accuracy, where the removal of multi-scale processing results in a 5.34% accuracy loss. This reflects the need for multi-scale processing to capture the various temporal dynamics and spatial scales that are characteristic of educational signs. The optical flow model input is also very important with a 5.65% drop in accuracy when not taken into account, which implies that motion information needs to be indirectly accounted for to distinguish signs. The total attention mechanism contributes +7.02% accuracy overall, which includes both squeeze-and-excitation and multi-head attention, when compared to a baseline without attention modules. This demonstrates a clear gain of attention units and is consistent with the importance of considering channel-wise feature recalibration and modeling temporal dependencies in educational sign language recognition.

4.5 Feature Analysis and Training Dynamics

Comprehensive analysis of feature contributions and training behavior provides insights into the model's learning process and the relative importance of different input modalities as shown in Table 7.

Table 7 *Feature contribution analysis*

S. No.	Feature Component	Performance (%)	Contribution (%)	Key Characteristics
1	Hand Landmarks	87.42	42.3	Primary source for sign articulation
2	Pose Landmarks	72.15	28.7	Critical for body context and large movements
3	Facial Landmarks	64.83	19.5	Important for expressive elements
4	Optical Flow Only	88.92	-	Motion information critical (− 5.65% when removed)
5	Combined Normalization	94.57	+9.5	Integrated processing vs. separate

According to the feature analysis, 42.3% discrimination power belongs to the hand landmarks as the primary source for articulation of signs. The pose landmarks provide significant contextual clues about larger scale movements (28.7%), while the facial landmarks provide representation of expressiveness and grammatical structure (19.5%). The optical flow features are thus critical supporting components to temporal discrimination since eliminating them led to a 5.65% accuracy loss. This joint normalization strategy, which was applied once the landmark and optical flow features were fused in a common feature space, achieved a 9.5% accuracy improvement over using normalization on the feature streams individually. Hence, the entire workflow's joint normalisation strategy (applied once the landmark and optical flow features were fused in a common feature space) implemented yielded an accuracy improvement of 9.5% over applying normalisation on the feature streams individually.

This also exhibits a clear convergence pattern into three stages as the training progresses. Throughout the beginning of this learning (epoch 1-15) the model converges very fast, as it learns to accurately identify specific patterns. The second stage (epochs 16-35) is the one during which the model's recognition becomes much more stable, and the temporal dimension is the one that gets the main attention. The third stage (epochs 36-45) is the one where the model is very slowly learning, starting from epoch 47, which allows the model to get close to 96.04% training accuracy by the end. The final stage focused on refining subtle discriminative gesture features and refining the barely visible but very important aspects of the patterns. The stability and thus the corresponding quality of the model at each epoch is the point that should not be overlooked, as a clear and strong model that can merge the educational sign recognition is the one that has been presented in Table 8.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 75-92

Table 8 *Training run analysis*

S. No.	Training Phase	Start Accuracy (%)	End Accuracy (%)	Learning Characteristics
1	Phase 1 (Epochs 1–15)	12.30	74.80	Rapid learning of basic sign patterns
2	Phase 2 (Epochs 16–35)	74.80	91.20	Steady improvement in temporal modeling
3	Phase 3 (Epochs 36–45)	91.20	96.04	Fine-tuning of subtle features
4	LR Reduction (Epoch 47)	95.60	96.04	0.0008→0.00064 enabled final convergence

4.6 Domain-Specific Performance Analysis

EduSign shows excellent performance within its specific domain because it solves classroom-specific challenges that other large-vocabulary SLR systems fail to recognize together with its Arabic and Korean language-based systems [3, 8, 4] and [13] and [12] implementations. The multi-scale temporal modeling module (kernels 3, 5, 7) captures both fine-grained finger articulations and broader educational gestures, with a 5.34% performance drop when excluded as shown in Table 9.

Table 9 *Summary of contributions and limitations*

S. No.	Dimension	Experimental Evidence	Literature Context	Limitation
1	Multi-Scale Temporal Modeling	5.34% drop when inception module removed	Addresses domain-specific gesture dynamics overlooked in large-vocabulary systems [3, 8, 4]	Limited 10-class vocabulary
2	Hybrid Feature Representation	Improved trajectory-based differentiation	Bridges structural modeling [17] and holistic motion approaches [21, 23]	Cross-signer generalization not explicitly evaluated
3	Balanced Regularization	1.47% training–validation gap	Mitigates small-dataset overfitting issues [14, 24]	Requires broader validation
4	Computational Efficiency	Real-time inference (8.2 ms)	Avoids complex fusion [18] and sensor-based systems [24]; aligns with real-time challenges identified in surveys [2, 11]	Evaluated only on isolated signs
5	Positioning & Accessibility	Domain-focused design for classroom use	Extends regional and community real-time efforts [20, 19, 25, 30]	In-situ classroom testing pending

The hybrid landmark–optical flow representation combines two opposing modeling methods explicit structural modeling [17] and holistic motion representations [21, 23]. The system outperforms single-modality approaches while satisfying real-time operational constraints and specialized SLR solutions which survey results identified [2, 11]. The implementation of balanced regularization decreases overfitting problems while demonstrating a 1.47% training–validation gap which solves small-data set generalization difficulties described in [14, 24]. In contrast to both complicated multi-stream fusion systems [18], along with sensor-based systems [24], EduSign is capable of lightweight real-time inference (8.2 ms). In addition to establishing community-based real-time efforts [20, 19, and 25 - 30] into a practical standard for Classroom Ready Education-based SLR, there is limited evaluation that consists of only the 10 classes of vocabulary sign language and isolated sign language with no explicit cross sign language signer validation.

The system as presented can yield good recognition performance on the tests done in a stratified training-validation procedure, as mentioned above, but not a signer-independent cross-validation is performed in this paper. Given that all the gestures were created by a number of individuals, and they are equally partitioned between train and test sets, the result is basically a measurement of the gesture recognition performance given variability in the signer. Signer-independent cross-validation is a very important problem for the signer-independent speech-understanding system as one can study how many signers there are and how their particular

writing styles, movement styles, and gesture features affect recognition, etc. Thus, leave-one-signer-out and signer-independent tests will be performed on larger sign datasets in the following study.

Although the suggested EduSign framework performs efficiently to recognize signs related to classroom setting using Kannada language, since the dataset comprises only 10 signs it is not feasible to be deployed in large scale. In that way the proposed work should be seen as a proof-of-concept framework specific to classroom setting demonstrating the efficacy of multimodal spatial-temporal modeling in classroom context for sign recognition. A direction for future work would involve expanding the number of signs for larger vocabulary in the realm of instructions within classroom and regular classroom behavior as well as continuous signing at sentence level. It is believed that by collecting diverse signer large-scale datasets of larger vocabulary, a more adaptable framework can be built.

4.7 Limitations of the Study

We present the EduSign framework which has shown good results. But there are limitations in the presented work: firstly, the number of educational Kannada sign language words are just 10 which are aimed at classroom-based scenario, and thus cannot achieve good coverage and suitability for real life education. Secondly, although the signer data has been included in the training and validation data creation, there is no signer-independent cross validation conducted in a strict way, so we cannot measure its robustness against unknown signers. Thirdly, isolated sign words have been used, and the continuous sentence level educational sign language recognition is still a significant real-world research task. And the inference time presented is mainly the network prediction time, not including the time for MediaPipe landmark extraction preprocessing and dense optical flow computation. The experiments are also carried out in an idealized classroom like setting. In a real-world classroom, an experiment needs to be conducted.

5. Conclusion

In this paper, we proposed an educational KSL recognition based on a multimodal deep learning approach named EduSign for classroom-style communication. The method leverages a landmark-based spatial representation and dense optical flow-based temporal motion model for structural and dynamic features extraction. We proposed a lightweight hybrid model combined with multi-scale temporal convolutions, squeeze-and-excitation attention, hierarchical residual learning, and multi-head temporal attention for real-time educational sign recognition. With the proposed educational KSL dataset, our model achieves acceptable performance: validation accuracy of 94.57%, generalization gap of 1.47%, as well as acceptable performance under comparison and ablation experiments, validating multimodal spatial-temporal features fusion for domain specific educational sign language recognition.

From a practical perspective, EduSign can be used as an efficient computation system of near real-time assistive learning systems. The small size model with multi-modality can be applied in intelligent classroom support, assistance of handicapped in learning, intelligent teaching systems, and communication with deaf and hard of hearing students based on AI. The proposed model provides a potential replacement for complicated and computationally intensive transformers-based methods in educational context considering the efficiency of execution.

Future work will focus on three key areas: Expansion of the education sign language vocabulary to provide adequate signing for larger classroom-focused gesture sets, and the development of an end-to-end system that is capable of continuous, sentence level sign recognition. In the future, we will consider using both signer independent evaluation and an adversarial approach for domain adaptation to achieve a more robust sign language recognition system. We will continue exploring the optimization of the end-to-end pipeline including speeding up the pre-processing. Furthermore, we aim to evaluate the real-world classroom setup at the end-to-end level and determine the robustness at varying illumination, backgrounds, and noise. Specifically, adversarial training will be investigated through domain adaptation techniques that encourage signer-invariant feature learning by minimizing signer-specific representations while preserving gesture-discriminative information.

Acknowledgement

The authors acknowledge the support provided by Nagarjuna College of Engineering and Technology, Bengaluru. The authors also thank the participating institution and contributors for facilitating data collection.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The author Gurusiddappa Hugar confirms responsibility for the following: ***study conception and design, data collection, analysis and interpretation of results, and manuscript preparation;*** *Ramesh M. Kagalkar* ***contributed to conceptualization, methodology, and project administration, supervision, manuscript writing, review, and editing****.*

References

[1] Renjith, S., & Manazhy, R. (2024). Sign language recognition and translation: A systematic review. *Multimedia Tools and Applications, 83*, 77077–77127. https://doi.org/10.1007/s11042-024-18583-4

[2] Li, D., Rodriguez, C., Yu, X., & Li, H. (2022). Deep learning for sign language recognition: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Advance online publication. https://doi.org/10.1109/TPAMI.2022.3187421

[3] Bansal, N., & Jain, A. (2025). Word recognition from Indian Sign Language using GMT-MaskRCNN. *Multimedia Tools and Applications, 84*, 2565–2597. https://doi.org/10.1007/s11042-024-20384-8

[4] Paul, S., Abul, M., Walid, M. A. A., Rahman, T., Islam, S., Ahmed, K., & Khan, F. (2024). Adam-based CNNLSTM for real-time sign language recognition. *Bulletin of Electrical Engineering and Informatics, 13*(1), 499–509. https://doi.org/10.11591/eei.v13i1.6059

[5] Srivastava, S., Singh, S., & Pooja. (2024). Continuous sign language recognition using MediaPipe Holistic. *Wireless Personal Communications, 137*, 1455–1468. https://doi.org/10.1007/s11277-024-11356-0

[6] Alsharif, B., Altaher, A. S., Altaher, A., Ilyas, M., & Alalwany, E. (2023). Deep learning for American Sign Language alphabet recognition. *Sensors, 23*(18), Article 7970. https://doi.org/10.3390/s23187970

[7] Gan, S., Yin, Y., Jiang, Z., Xie, L., & Lu, S. (2023). Edge-optimized SLR. In Proceedings of the ACM Multimedia 2023 (pp. 4502–4512). *Association for Computing Machinery*. https://doi.org/10.1145/3581783.3611820

[8] Miah, A. S. M., Hasan, M. A. M., Okuyama, Y., & Toda, T. (2024). Spatio-temporal attention for sign language recognition. *Pattern Analysis and Applications, 27*(2), Article 37. https://doi.org/10.1007/s10044-024-01229-4

[9] Xiong, S., Zou, C., Yun, J., Li, H., & Zhang, W. (2025). Continuous sign language recognition with dynamic attention. *Signal, Image and Video Processing, 19*(2), Article 141. https://doi.org/10.1007/s11760-024-03718-9

[10] Wang, H., & Wang, L. (2023). Modeling temporal dynamics with self-attention for sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. Advance online publication. https://doi.org/10.1109/TCSVT.2023.3268084

[11] Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications, 164*, Article 113794. https://doi.org/10.1016/j.eswa.2020.113794

[12] Shin, J., Miah, A. S. M., Hasan, M. A. M., Okuyama, Y., & Toda, T. (2023). Korean sign language recognition using Transformers. *Applied Sciences, 13*(5), Article 3029. https://doi.org/10.3390/app13053029

[13] Podder, K. K., Ezeddin, M., Chowdhury, M. E. H., Khandakar, A., & Aly, S. A. (2023). Signer-independent Arabic sign language recognition using deep learning. *Sensors, 23*(16), Article 7156. https://doi.org/10.3390/s23167156

[14] De Sisto, M., Vandeghinste, V., Gomez, S. E., Coster, M. D., Shterionov, D., & Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. In Proceedings of the 13th *International Conference on Language Resources and Evaluation (pp. 2478–2487)*. European Language Resources Association.

[15] Papatsimouli, M., Sarigiannidis, P., & Fragulis, G. (2023). Real-time sign language translators: IoT integration. *Technologies, 11(*4), Article 83. https://doi.org/10.3390/technologies11040083

[16] Abdullahi, S., & Chamnongthai, K. (2023). IDF-sign: Addressing inconsistent depth features for dynamic sign word recognition. *IEEE Access, 11*, 88511–88526. https://doi.org/10.1109/ACCESS.2023.3305255

[17] Woods, L. T., & Rana, Z. A. (2023). Pose-keypoint Transformers for sign language recognition. *Mathematics, 11*(9), Article 2129. https://doi.org/10.3390/math11092129

[18] Jiang, S., Sun, B., Wang, L., Liu, Y., & Wu, X. (2023). A multi-stream sign language recognition framework with skeleton-aware training. *IEEE Transactions on Multimedia*. Advance online publication. https://doi.org/10.1109/TMM.2023.3274890

[19] Bora, J., Dehingia, S., Boruah, A., Chetia, A., & Gogoi, D. (2023). Real-time Assamese sign language recognition using MediaPipe. *Procedia Computer Science, 218*, 1384–1393. https://doi.org/10.1016/j.procs.2023.01.117

[20] Sharma, S., & Singh, S. (2023). A spatio-temporal framework for Indian sign language recognition. *Wireless Personal Communications, 132*, 2527–2541. https://doi.org/10.1007/s11277-023-10730-8

[21] Zhang, H., Hu, Z., Yu, D., Wang, L., & Chen, Y. (2024). Multipath attention for video action recognition. *Neural Processing Letters, 56*(3), Article 124. https://doi.org/10.1007/s11063-024-11591-3

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30. Curran Associates. https://doi.org/10.48550/arXiv.1706.03762

[23] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 183–192)*. IEEE. https://doi.org/10.1109/CVPR42600.2020.00026

[24] Taneja, M., Singla, N., Goyal, N., & Jinda, R. (2023). A comprehensive review of sensor-based sign language recognition models. Journal of Xi'an Shiyou University, *Natural Science Edition, 19*(5), 292–302.

[25] Kumar, S., Rani, R., & Chaudhari, U. (2024). Real-time sign language recognition for the disabled community. *MethodsX, 13*, Article 102901. https://doi.org/10.1016/j.mex.2024.102901

[26] Kothadiya, D., Bhatt, C., Rehman, A., Alamri, F., & Saba, T. (2023). SignExplainer: An explainable AI framework for SLR. *IEEE Access, 11*, 47410–47419. https://doi.org/10.1109/ACCESS.2023.3274851

[27] Alshewimy, M. (2023). Efficient deep learning for sign language recognition. *Intelligent Systems with Applications, 20*, Article 200284. https://doi.org/10.1016/j.iswa.2023.200284

[28] Sreemathy, R., Turuk, M., Chaudhary, S., Mishra, A., & Patra, P. K. (2023). Continuous word-level sign language recognition using expert systems. *International Journal of Cognitive Computing in Engineering, 4*, 170–178. https://doi.org/10.1016/j.ijcce.2023.04.002

[29] Jiang, T., Camgoz, N., & Bowden, R. (2024). A comparative study of pose-based sign language recognition systems. *Computer Vision and Image Understanding, 231,* Article 103712. https://doi.org/10.1016/j.cviu.2023.103712

[30] Lakshmi, K., Samiya, M. S., Kumar, M. R., & Singuluri, P. K. (2023). Real-time hand gesture recognition for deaf communication. *International Journal of Intelligent Systems and Applications in Engineering, 11*(6s), 23–37. https://ijisae.org/index.php/IJISAE/article/view/2825

[31] Svendsen, B., & Kadry, S. (2024). A dataset for recognition of Norwegian Sign Language. International *Journal of Mathematics, Statistics, and Computer Science*, 2. https://doi.org/10.59543/ijmscs.v2i.8049