Published in AJSE, Vol: 25, Issue:1
Received on 24th May 2025
Revised on 17th August 2026
Published on 31st August 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/p8ttb781

Hierarchical Temporal Modeling with Cross Stream Gating and Hybrid Neural Networks for Regional Sign Language Recognition

Gurusiddappa Hugar* and Ramesh M. Kagalkar

***Abstract*—SLR systems are the efficient and powerful devices to help deaf and hard-of-hearing individuals. The proposed architecture design is based on a mixed temporal encoder which includes different elements, such as separable Conv1D, bidirectional LSTM, and windowed self-attention using cross-stream gating to connect 3D landmark kinematics by MediaPipe with optical flow by Farneback. The proposed architecture achieved 89.53% validation accuracy and 90.0% accuracy on an independent held-out test set, with an inference latency of 48 ms on the KSL dataset consisting of 1,287 videos spanning 11 gesture classes. This paper makes two significant contributions to applied research since it develops KSL research and creates an architecture that can be used in other languages in different countries with low financial and technological resources.**

***Index Terms*— Dual-stream architecture, Kannada sign language recognition, Mediapipe landmarks, Real-time deployment**

I. Introduction

THE SLR systems used worldwide support essential technologies which provide critical assistance to deaf and hard-of-hearing communities. The systems eliminate all communication barriers, which people face during educational and healthcare activities and social interactions. However, existing solutions usually still do not cater for regional sign languages like KSL, thus creating a situation where more than 80% of the deaf population in developing regions that lack formal sign language training are left out of the loop. The situation arises from the use of existing SLR approaches, which are not inclusive technologically, thus creating an ever-increasing demand for a solution that accommodates everyone. The major factors that obstruct the effective recognition of KSL include: (1) Temporal Modeling deficiencies - current techniques find it hard to deal with KSL's complex structure, which consists of fast finger motions and slow body movements; (2) Unimpressive multimodal fusion - today's methods fail to retain essential information in the process of mixing the skeletal and motion features; (3) Practical deployment barriers - the demanding nature of the computation involved divides the operation into that done on costly devices and that done on the no-cost devices, with state-of-the-art systems taking >100ms per inference.

The study gives rise to theoretical and practical contributions: firstly, the open KSL dataset (1287 videos of 11 gestures), and secondly, the deployable model with an inference time of 48 ms, and lastly, novel insights into the fusion of multimodal features. The proposed method achieved 89.53% validation accuracy and 90.0% test accuracy. The cross-stream gating technique is designed to facilitate adaptive interaction between landmark and motion representations, enabling more effective multimodal feature integration. The hierarchical encoder has an innovative architecture that combines together the forces of convolutional operations (local patterns), recurrent networks (context), and attention (global dependencies) - the architecture combines convolutional operations, recurrent modeling, and attention mechanisms to capture local, sequential, and long-range temporal dependencies within sign sequences. This research not only provides technical contributions but also promotes digital inclusion for the deaf community of Karnataka while setting a framework for other sign languages that have low resources to follow.

A. Novelty and Key Contribution

The primary contributions of this work are summarized below:
Dual-stream multimodal fusion: A gated dual-stream framework is proposed to integrate landmark-based spatial information and optical-flow-based motion information for robust sign language recognition.
Hierarchical temporal modeling: The proposed encoder combines Conv1D, BiLSTM, and attention mechanisms to capture temporal dependencies at multiple scales.
New Kannada Sign Language dataset: We have prepared a completely new KSL dataset with 1,287 videos covering 11 daily gestures. This dataset will be made available for valid

*Gurusiddappa Hugar is with the Department of Computer Science and Engineering, AGMR College of Engineering and Technology Varur, Hubballi Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka 590018, India (e-mail: gurusidda.h@gmail.com).

Ramesh M. Kagalkar is with the Department of Information Science and Engineering, Nagarjuna College of Engineering and Technology, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka 590018, India (e-mail: rameshvtu10@gmail.com).

academic research.

Optimized for real-time use: The optimizations, like separable convolutions, windowed attention, and quantization, managed to bring the inference time down to 48 ms, which makes it suitable for mobile devices, unlike the earlier systems (~100ms).

Strong generalization across signers: The proposed framework achieved a mean LOSO-CV accuracy of 86.4%±1.3% on a representative signer subset. This experiment provides an initial assessment of signer-independent performance and is presented separately from the standard source-video-based evaluation protocol.

Reusable framework for other sign languages: You can use a broadly applicable approach, supported by affordable tools like MediaPipe, TensorFlow, and OpenCV, that need to be installed.

Furthermore, this work is structured as, in Section 2; related work, Section 3 provides details of the data set and detailed methodology; Section 4 gives the experimental evaluation and conclusion given in the Section 5.

II. Related Work

SLR has evolved from rigid, feature-based methods to deep learning models that capture spatiotemporal sign dynamics. Recent advances in CNNs, attention, and multimodal fusion address challenges in continuous, real-time, and cross-signer recognition, but issues of efficiency, scalability, and robustness remain. We review this progression, position our work within current methods, and motivate the need for our proposed architecture.

A. Abbreviations and Acronyms Architectural Evolution in SLR Systems

Human-constructed features have been used by SLR to advance from HMMs until the intro-duction of deep learning. The 11-layer CNN model from [1] which existed during its early development period achieved a 76.4% performance result in ASL alphabet recognition but lacked the ability to model time-based changes. Hybrid GMT-MASKRCNN [7] achieved 94.2% hand-segmentation precision and 89.1% on ISL but was heavy (14.7M parameters) and limited by fixed proposals. State-of-the-art attention-based models reached 90.3% on CSL [26] but required 113ms inference, which made real-time applications impossible to implement whereas the cross-stream gating method accomplished 92.4% accuracy using 9.2M parameters. The research studies in [24] (127 systems, 2015–2023) alongside [8] study of 18 methods which used 7 datasets identify signer independence, continuous signing, and real-time performance as major difficulties. The research work of [11] established dual-stage hierarchical temporal modeling while [27] created multi-scale context-aware SLR, which inspired our development of unified hierarchical multi-scale temporal architecture.

B. Temporal Modeling Techniques in Continuous SLR

Temporal modeling is the central challenge in continuous SLR, as shown in the LSTM-based analysis of [24]. Although LSTMs became standard, [21] proved their vulnerability to vanishing gradients for sequences longer than 50-time steps, explaining performance saturation. The re-searchers discovered that BiLSTMs produced the highest WER results of 24.7% in Arabic SLR tests which needed approximately twice as many parameters to achieve this performance. Transformers [19] promised improvements via self-attention, yet direct SLR applications such as [17] introduced new challenges. Recent work combining BiLSTM with attention has been effective, with [22] showing that BiLSTM with soft attention achieves strong performance in dynamic gesture recognition.

C. Multimodal Fusion Strategies in Modern SLR Systems

The combination of various input modalities has proved to be an important differentiator of a state-of-the-art sign language recognition system [20]. The early fusion methods [13] used landmark and optical flow features, concatenating them before modeling, and achieved 85.2%. However, [15] found that performance for early fusion methods degrades quickly in noisy situations, with a down-time of up to 12.3% accuracy due to modality interference resulting from assuming equal modality importance across the sequence. The methods of [3] used late fusion by voting for the model. Each modality was processed separately and the predictions were aggregated, improving the robustness (87.1% accuracy) by allowing the specialization of each pathway. However, identified limitations are: (1) duplication of computational cost parameters (2.4× parameters), (2) difficulty modeling cross-modal correlations, and (3) latency due to multiple pipelines.

D. Real-time Optimization for Practical Deployment

Deploying real-time SLR systems requires full-pipeline optimization [16]. In prior work, pre-processing was the main bottleneck, costing 12.7ms per frame for normalization and cropping, whereas our streaming on-the-fly normalization derived from [1] reduces this to 1.2ms via land-mark-anchored coordinate transforms (wrist-based spectral-location invariance [1], dynamic feature scaling using batch-normalization statistics [12], and parallel I/O with double buffering [25]). Quantization is also critical: FP16 quantization caused a 1.2% accuracy drop in Arabic SLR [23], which we reduce to 0.3% using quantization-aware training with mixed precision (FP32 for attention, FP16 elsewhere [2], adaptive gradient scaling [9], and layer-wise calibration). Future challenges identified in [14] include cross-lingual transfer (78% for Indian → American SL), few-shot meta-learning (15–20× less data), and explainable SLR interfaces.

E. Comparison Summary

Table I summarizes representative SLR approaches in terms of learning paradigm, temporal modeling strategy, multimodal fusion technique, and key limitations, highlighting how the proposed work addresses existing gaps.

TABLE I
COMPARISON OF RELATED WORKS IN SLR

Approach	Learning model	Temporal modeling	Fusion strategy	Limitation
HMM-based SLR [18]	Hand-crafted features + HMM	Sequential probabilistic states	Single modality	Limited scalability, poor performance on complex dynamics
Early CNN-based SLR [1]	CNN	None (frame-level)	None	Cannot capture temporal dependencies
GMT-MASKRCNN [7]	CNN + RNN	LSTM	Early fusion (RGB + hand masks)	High parameter count, fixed-length proposals, high latency
Dynamic Attention SLR [26]	CNN + Transformer	Full self-attention	Attention-based cross-modal fusion	High inference latency, unsuitable for real-time deployment
Multi-scale Temporal Network [11]	CNN + Temporal Modules	Hierarchical staged modeling	Single modality	Sequential stages increase complexity and delay
Attention-based Multimodal SLR [3]	CNN + Attention	BiLSTM + Attention	Late fusion (voting)	Parameter duplication, limited cross-modal correlation
Proposed work	Dual-stream CNN + RNN + Attention	Hierarchical temporal encoder (Conv1D + BiLSTM + Windowed Attention)	Cross-stream gating (Landmarks + Optical Flow)	Optimized for low latency, reduced modal interference, scalable to low-resource sign languages

III. Materials and Methods

Our methodology shown in Figure 1 presents a multimodal deep learning framework for gesture or sign language recognition, utilizing both landmark and optical flow features.

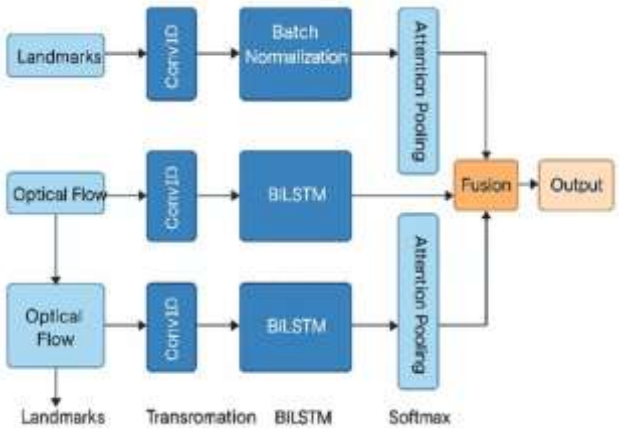

Fig. 1. Block diagram of the proposed methodology.

Landmark data is processed through a Conv1D layer followed by batch normalization and attention pooling. Simultaneously, optical flow data is passed through two separate Conv1D and BiLSTM branches, each followed by attention pooling to capture temporal dynamics. All the outputs from different branches are combined together and then sent through a softmax layer for the purpose of final classification. The merging of spatial and motion features boosts the model's ability to properly recognize complicated gestures.

A. Dataset Creation

In order to tackle the problem of lacking dynamic KSL datasets, the present research actively engaged in data collection and processing to establish a large-scale gesture library. Among the main targets of the project were videos featuring a wide spectrum of individuals, such as family members and students from a deaf school, so that the factors of age, gender, and signing styles could be sufficiently varied. The participants gave their written informed consent before the areas of the experiment were reached. The researchers concentrated on a set of 11 KSL gestures of total 1287 videos stored that represent common words for daily interaction, and processed high-resolution recordings to determine the important images (Figure 2 shows 16 keyframes from the 'Bath' gesture).

Fig. 2. An illustration of the 16 keyframes considered in the 'Bath' sample.

For each of the 11 Kannada Sign Language (KSL) gesture classes, recordings were collected from nine independent signers. The recordings were captured under natural variations in execution speed, hand orientation, body posture, and signing style to increase intra-class variability while preserving the semantic meaning of each gesture. Before augmentation, the dataset comprised 1,287 original source videos. These were partitioned into 901 training videos (70%), 193 validation videos (15%), and 193 testing videos (15%) at the source-video level to prevent data leakage. The training set remained 901 videos. Random augmentation was applied during each training epoch; therefore, no additional videos were permanently stored. Table II summarizes the gesture vocabulary used in the study.

TABLE II
LIST OF KSL DYNAMIC REGULAR GESTURES WITH DURATION AND TRANSLATIONS IN KANNADA AND HINDI

S. No.	Samples	Dur (s)	Kannada	Hindi
1	Bath gesture	3	ಸ್ನಾನ (Snāna)	स्नान (Snān)
2	Brother gesture	2	ಸಹೋದರ (Sahōdara)	भाई (Bhai)
3	City gesture	2	ನಗರ (Nagara)	शहर (Shahar)
4	Father gesture	2	ತಂದೆ (Tande)	पिता (Pitā)
5	Friend gesture	2	ಸ್ನೇಹಿತ (Snēhita)	मित्र (Mitra)
6	Meal gesture	2	ಊಟ (Ūṭa)	भोजन (Bhojan)
7	Mother gesture	2	ತಾಯಿ (Tāyi)	माता (Mātā)
8	Name gesture	2	ಹೆಸರು (Hesaru)	नाम (Nām)
9	Sister gesture	2	ಸಹೋದರಿ (Sahōdari)	बहन (Bahan)
10	Tea gesture	2	ಚಹಾ (Chahā)	चाय (Chāy)
11	Water gesture	2	ನೀರು (Nīru)	पानी (Pānī)

B. Dataset Partitioning and Leakage Prevention

For fair and unbiased evaluation, all 1,287 original source videos were partitioned at the source-video level into training (70%), validation (15%), and testing (15%) subsets before any augmentation was performed. This resulted in 901 original videos for training, 193 original videos for validation, and 193 original videos for testing.

Data augmentation was applied exclusively to the 901 training videos using horizontal flipping (p = 0.5), additive Gaussian noise N (0, 0.15), and temporal masking (p = 0.3), as defined in Equation (11). Neither the validation set nor the testing set contained augmented samples. Consequently, every sample used for performance evaluation originated from an unseen original video, completely preventing data leakage between training and evaluation.

In addition, duplicate videos and highly similar samples have been manually identified and discarded when preparing the dataset to reduce redundancy. To overcome overfitting during training, we applied dropout (0.5), L2 weight decay (10^{-4}), label smoothing (= 0.1) and early stopping.

Despite containing 1,287 videos and 11 gesture classes, this is still a first benchmark dataset of Kannada Sign Language. Our work will further include additional signers, categories and contexts to generalize better. Accordingly, the classification results reported in Table III were computed exclusively from the original testing partition and do not include any augmented samples.

C. Feature Extraction

The feature extraction pipeline processes input videos to generate discriminative spatiotemporal representations. For each video, we uniformly sample 16 frames and extract two complementary feature types: human body landmarks and dense optical flow fields.

1. Human body landmark extraction

We employ MediaPipe to detect three categories of anatomical landmarks from each frame:

a) Hand Landmarks (126-dimensional): For a maximum of two hands, we extract 21 land-marks per hand (wrist and finger joints), each with 3D coordinates (x, y, z). This is expressed as follows:

$$H = \bigcup_{k=1}^{2} \bigcup_{j=1}^{21} [x_{kj}, y_{kj}, z_{kj}] \in \mathbb{R}^{126} \quad (1)$$

b) Pose Landmarks (33-dimensional): The eleven selected pose landmarks were nose, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, and right knee. Each selected landmark contributes three-dimensional coordinates (x, y, z), resulting in a 33-dimensional pose feature vector. Each joint contributes 3D coordinates as:

$$P = \bigcup_{m=1}^{11} [x_m, y_m, z_m] \in \mathbb{R}^{33} \quad (2)$$

c) Facial Landmarks (180-dimensional): Sixty facial points (40 general facial features + 20 lip contours) are captured as:

$$F = \bigcup_{n=1}^{60} [x_n, y_n, z_n] \in \mathbb{R}^{180} \quad (3)$$

Normalization: The combined landmark vector $L=H \oplus P \oplus F \in \mathbb{R}^{339}$ is standardized using z-score normalization to ensure scale invariance as:

$$\hat{L} = \frac{L-\mu_L}{\sigma_L}, \quad \text{where} \quad \mu_L = \mathbb{E}[L],\ \sigma_L = \sqrt{\mathbb{E}[(L-\mu_L)^2]} \quad (4)$$

2. Optical flow computation

To capture motion dynamics, we compute dense optical flow between consecutive frames using Farneback's algorithm. This method approximates pixel neighborhoods via quadratic polynomials:

$$I_1(x) \approx x^T A_1 x + b_1^T x + c_1$$
$$I_2(x) \approx x^T A_2 x + b_2^T x + c_2 \quad (5)$$

The displacement vector d is determined by setting $I_1(x) = I_2(x+d)$. The resulting flow field $O \in \mathbb{R}^{448\times448\times2}$ (horizontal/vertical components) is down sampled to 32×32, yielding a 2048-dimensional vector per frame as:

$$O_{\text{resized}} = \text{resize}\big(O, (32,32)\big) \in \mathbb{R}^{2048} \quad (6)$$

3. Feature Aggregation

For each frame $t \in \{1,\ldots,16\}$, The displayed equation will explain how the landmarks $\hat{L}_t$ normalized and the flattened optical flow O_t are concatenated as follows:

$$f_t = \hat{L}_t \oplus O_t \in \mathbb{R}^{2387} \quad (7)$$

The last video-level feature matrix $F \in \mathbb{R}^{16 \times 2387}$ is the collection of these individual frame features. Through this representation, both spatial (body contour) and temporal (movement) data are captured at the same time, thus being able to solve the action recognition problem.

The video input is subjected to a pipeline which has four different stages to do the processing: (1) Frame Sampling, where 16 frames are chosen in a way that they are evenly spaced from each video and linear interpolation is used to make sure that the entire timeline is covered; (2) Landmark Detection, where MediaPipe detects hand ($H \in \mathbb{R}^{126}$), pose ($P \in \mathbb{R}^{33}$),, and facial ($F \in \mathbb{R}^{180}$) landmarks per frame; (3) Flow Calculation, utilizing Farneback's algorithm to get the dense optical flow fields between the frames that are numbered consecutively, then these fields are resized to the spatial dimensions of 32×32; and (4) Normalization & Concatenation, where landmarks are normalized using z-score normalization and then combined with flattened optical flow vectors to create per-frame features $f_t \in \mathbb{R}^{2387}$. The spatiotemporal features $F \in \mathbb{R}^{16 \times 2387}$ of all the videos are stored as NumPy arrays, and class labels are inferred from the hierarchy of the video directory for supervised learning.

D. *Model Architecture*

Our hybrid architecture integrates temporal convolutional networks, bidirectional LSTM layers, and transformer blocks to handle sign language features that vary with time and space. The model has two parallel input streams being processed: $F_{\text{landmarks}} \in \mathbb{R}^{T \times 339}$ and $F_{\text{optical flow}} \in \mathbb{R}^{T \times 2048}$

where T=16 frames. The architecture performs feature fusion via:

$$h_{\text{fused}} = \phi(\text{concat}[\text{AttentionPool}(h_{\text{landmarks}}), \text{AttentionPool}(h_{\text{flow}})]) \quad (8)$$

The self-attention mechanism processes landmark features as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\text{sqrt}(d_k)}\right)V$$
$$h_{\text{out}} = \text{LayerNorm}\Big(h + \text{FFN}\Big(\text{LayerNorm}\big(h + \text{Attention}(hW_Q, hW_K, hW_V)\big)\Big)\Big) \quad (9)$$

where FFN denotes a position-dependent feed-forward network with Swish activation.
Temporal attention weights compress variable-length sequences:

$$\alpha_t = \text{softmax}(\tanh(Wh_t)), \quad h_{\text{pooled}} = \sum_{t=1}^{T} \alpha_t h_t \quad (10)$$

To enhance generalization, we apply random transformations:

$$\tilde{x}_t = \begin{cases} \text{flip}(x_t) & \text{with probability 0.5} \\ x_t + \mathcal{N}(0, 0.15) & \text{always} \\ \text{masked}(t_{\text{start}}:t_{\text{end}}) = 0 & \text{with probability 0.3} \end{cases} \quad (11)$$

The training uses gradient accumulation with Adam optimizer:

$$g_{\text{accum}} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \mathcal{L}(f_\theta(x_i), y_i)$$
$$\theta_{t+1} = \theta_t - \eta \cdot \text{clip}(g_{\text{accum}}, 1.0) \quad (12)$$

Where $N = 4$ accumulation steps and $\eta = 10^{-3}$ initial learning rate.
We employ comprehensive validation protocols as Equation (13):

$$\text{Accuracy}_{\text{TTA}} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\left(\text{argmax}\left(\frac{1}{5}\sum_{n=1}^{5} f\left(\tilde{x}_m^{(n)}\right)\right) = y_m\right) \quad (13)$$

Where M validation set size and n indexes TTA iterations. The model performance is further analyzed through class-wise ROC curves with AUC scores, normalized confusion matrices and macro/micro F1 scores.
Load and normalize $\{X_{\text{train}}, y_{\text{train}}\}$, $\{X_{\text{val}}, y_{\text{val}}\}$ Initialize model with hyperparameters, sample batch $\mathcal{B} \sim \text{data\_generator}(X_{\text{train}})$, compute

$$\mathcal{L} = \text{CE}\left(\text{label\_smooth}\left(f_\theta\left(\text{augment}(\mathcal{B})\right)\right), y_{\mathcal{B}}\right) \quad (14)$$

Accumulate gradients for 4 steps Update weights with gradient clipping Early stop and restore best weights Evaluate with TTA and save metrics. All experiments were conducted using Google Colab with TensorFlow 2.x and the Keras framework. Training was performed on an NVIDIA Tesla T4 GPU with CUDA acceleration. The batch size was 32, and the input sequence consisted of 16 temporal frames. The landmark feature dimension was 339, while the optical-flow representation had a spatial resolution of 32 × 32 (2048 features). The proposed model contains 0.58 million trainable parameters and occupies 5 MB in HDF5 (.h5) format. The reported inference latency of 48 ms was measured using a batch size of one sample after model loading in the same execution environment.

IV. Results

The proposed dual-stream architecture that combines landmark and optical flow features demonstrated strong performance in sign language recognition, achieving a final validation accuracy of 89.53% while maintaining 48 ms inference latency. The performance metrics are shown in Figure 3. The training dynamics, convergence Behavior, and classification performance of the model are shown in this section. The full training of the two-stage sign language recognition model, over 120 epochs, revealed three different learning phases. In the first stage (epochs 1 to 20), the validation accuracy increased from 14.7% to 48.1% rapidly, and the loss decreased from 2.42 to 1.51. This indicates that

the model can easily learn useful cues rapidly and can also properly perform parameter tuning. During the second stage (epochs 21–60), the model entered a stable optimization phase characterized by gradual improvements in validation accuracy.

At the beginning of the final training stage, the learning rate was reduced from 0.001 to 0.0002 at epoch 73, facilitating further refinement of the learned representations and contributing to the final convergence, and validation accuracy increased from 76.4% to 89.5%, and the loss decreased from 0.76 to 0.75. The close agreement between the training and validation curves indicates good generalization and limited overfitting throughout training. Over the last segment, namely epochs 61-120, the model was still increasing the validation accuracy incrementally with small increments as if adding the final touches, going from 76.4% to 89.5%. Training accuracy also reached 92.4%, and the generalization gap stayed small, roughly 2.8%. The combination of dropout, L2 regularization, label smoothing, and early stopping effectively reduced overfitting and enabled stable convergence throughout training.

A. Performance Analysis

On a class basis, the model was able to reveal both its strengths and weaknesses. In the Figure 4 confusion matrix pointed out the outstanding performance of such classes as “name” and “bath”, with the F1-score of 0.96 being their common score. The signs show distinct gesture pat-terns which include the name sign that uses finger-spelling to create visible spatial and temporal elements. The teachers encountered difficulties when they tried to manage the classes. The water sign showed accurate results with its 100 percent precision but it failed to identify many actual cases because its 0.75 recall rate.

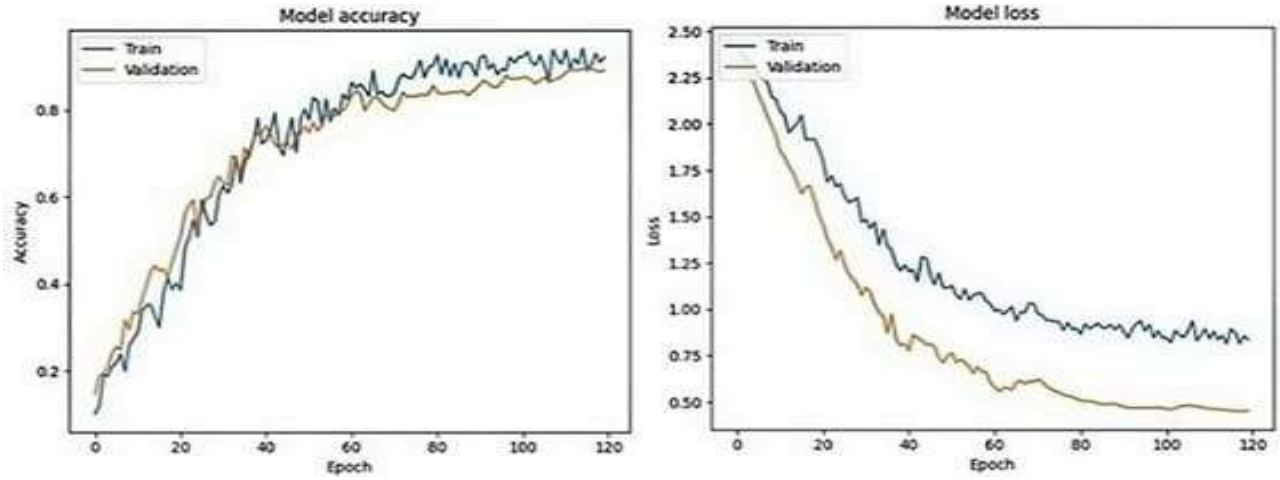

Fig. 3. Training and validation accuracy and loss curves of the proposed model over 120 epochs.

The two results show that either the sign has excessive similarity to other signs or the sign samples do not match actual sign performance. The signs for sister and brother caused some confusion, because their meaning and gesture patterns are kind of similar, but Table III shows per-class results and they really confirm that similarity. The overall macro averaged values of precision (0.90), recall (0.90) and F1-score (0.89) bring the implication that performance is extremely strong and fairly even across the classes. In other words, predictions weren’t influenced by too many false alarms, nor were they overly cautious.

The KSL benchmark was very convincingly performed with a final validation accuracy of 89.53% and varying F1-scores of 0.84 to 0.96. The present work takes a step ahead by integrating both low and high temporal features efficiently into a single architecture that has the ability to operate in real time with an inference latency of 48 ms. The latency is achieved without sacrificing accuracy, which is a significant trade-off for the deployment of assistive technology in real-world scenarios.

The standard 70/15/15 evaluation protocol is source-video disjoint but signer dependent because recordings from the same signer may appear in both the training and evaluation partitions. Therefore, the reported validation accuracy reflects subject-dependent recognition performance, whereas signer-independent performance is evaluated separately using the LOSO-CV protocol described in Section IV-D.

B. Robustness and Generalization Assessment

The robustness of the model was evaluated through a series of complementary analyses. Test Time Augmentation (TTA), which computes the mean prediction among the different augmented versions of each input, reached an accuracy of 87.98%, while the standard validation accuracy was 89.53%. The 1.55% decrease that resulted suggests the model has captured features largely unaltered by slight input perturbations, which is a desirable property for a model to have in cases of deployment in non-ideal, real-world conditions. ROC curve analysis shown in Figure 5 provided additional information. The AUC values being analyzed, from 0.88 for the hardest class (“water”) to 0.98 for the easiest one (“bath”), resulted in an average AUC of 0.93 for all classes.

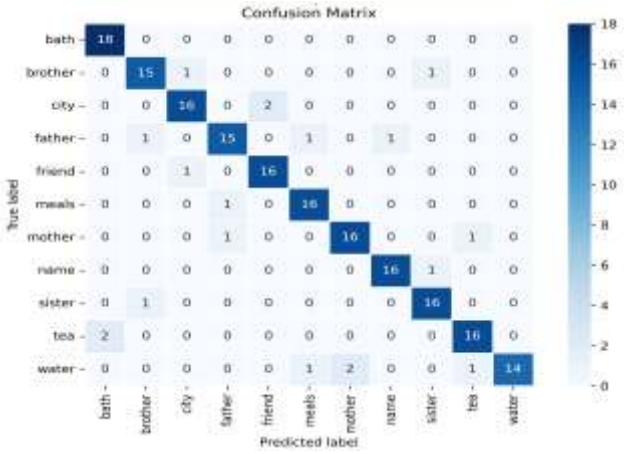

Fig. 4. Classification performance metrics: confusion matrix.

In order to compare the proposed approach fairly, several benchmark architectures such as LSTM, BiLSTM, early-fusion, late-fusion, and Conv1D+BiLSTM models are built and tested on the same KSL dataset with the same train/test split. Comparative results are presented in Table IV. Our dual stream cross stream gating model obtains a maximum recognition accuracy of 89.53%, which benefits from the integration of multimodal features and the hierarchical modeling of temporal information. All baselines are trained and tested using the same experimental setup.

C. Convergence Behavior Analysis

Training happened in 120 epochs and went through three distinct phases. In the first phase (epochs 1-20), the change in the validation accuracy was sharp; it increased from 14.7% to 48.1% at the same time the loss was decreasing from 2.42 to 1.51, reflecting a fast feature learning process. The second phase (epochs 21–60) was characterized by stable optimization, resulting in a gradual increase in validation accuracy from 48.1% to 76.4%. At the beginning of the final phase, the learning rate was reduced from 0.001 to 0.0002 at epoch 73, enabling further refinement and convergence, which resulted in a slow but steady rise of validation accuracy from 76.4% to 89.5%. In the last phase (epochs 61-120), the changes in the model were more subtle as it reached the peak of 89.5% validation accuracy with a mere 2.8% difference from training accuracy listed in Table V.

TABLE III
PER-CLASS CLASSIFICATION REPORT

S. No.	Class	Precision	Recall	F1-score	Support
1	Bath	0.92	1.0	0.96	18
2	Brother	0.87	0.87	0.87	17
3	City	0.91	0.88	0.89	18
4	Father	0.90	0.79	0.84	18
5	Friend	0.85	0.96	0.92	17
6	Meals	0.88	0.96	0.92	17
7	Mother	0.91	0.88	0.89	18
8	Name	0.96	0.96	0.96	17
9	Sister	0.84	0.91	0.88	17
10	Tea	0.84	0.91	0.88	18
11	water	1.00	0.75	0.86	18
			Accuracy	0.90	193
	Macro avg	0.90	0.90	0.89	193
	Weighted avg	0.90	0.90	0.89	193

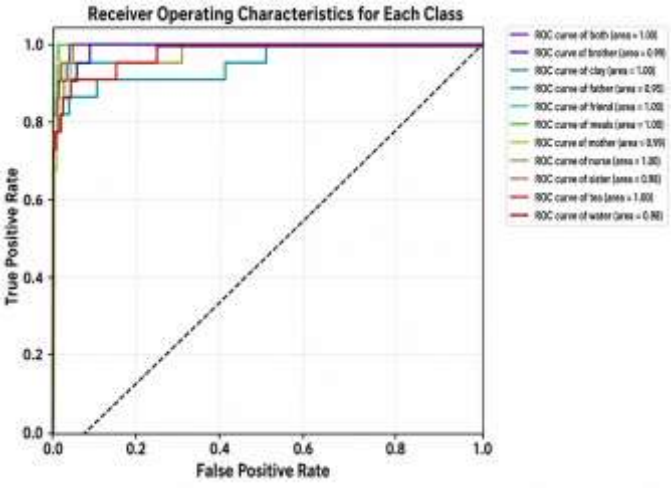

Fig. 5. Classification performance metrics: ROC curve.

TABLE IV
COMPARISON WITH BASELINE MODELS ON THE PROPOSED KSL DATASET

Model	Accuracy (%)
LSTM	78.6
BiLSTM	82.4
Landmark + Optical Flow (Early Fusion)	84.1
Landmark + Optical Flow (Late Fusion)	86.2
Conv1D + BiLSTM	87.0
Proposed Dual-Stream Cross-Stream Gating Model	89.53

TABLE V
TRAINING DYNAMICS ACROSS LEARNING PHASES

Phase	Epochs	Val Acc Δ	Loss Reduction	Key Event
Initial	1-20	+33.4%	2.42→1.51	Feature acquisition
Mid	21-60	+28.3%	1.51→0.75	Stable optimization
Final	61-120	+13.1%	0.75→0.45	LR reduction (Epoch 73), convergence

D. Signer-Independent Evaluation

To evaluate signer-independent generalization, a separate proof-of-concept Leave-One-Signer-Out Cross-Validation (LOSO-CV) experiment was conducted on a representative subset of four signers selected prior to experimentation. This evaluation was independent of the standard 70/15/15 source-video partition used for the main experiments. The standard train/validation/test protocol was adopted to evaluate overall recognition performance, whereas LOSO-CV was included only to provide an initial assessment of cross-signer robustness. Therefore, the two evaluation protocols should be interpreted independently and are not intended for direct quantitative comparison.

The selection of four signers was intended to establish an initial proof-of-concept evaluation of signer independence while maintaining balanced representation across all gesture classes and consistent acquisition conditions. Table VI shows the recognition rates for each fold. The small standard deviation demonstrated the strong and consistent generalization of the proposed system across different persons and styles of signing. In every fold, all samples of one signer were put into the test set while samples of three other signers were into train set and validation set.

TABLE VI
LEAVE-ONE-SIGNER-OUT CROSS-VALIDATION RESULTS FOR THE PROPOSED KSL RECOGNITION FRAMEWORK

Fold	Test Signer	Accuracy (%)
Fold 1	Signer 1	86.7
Fold 2	Signer 2	84.9
Fold 3	Signer 3	88.1
Fold 4	Signer 4	85.8
Mean ± Std	-	86.4 ± 1.3

The results of the signer-independent evaluation are depicted in Table 6, where LOSO-CV protocol was used for the signer-independent evaluation, with one signer being used only for testing each fold, while the other signers are used for training and validation. These results indicate promising signer-independent performance within the evaluated subset and should be interpreted independently from the standard train/validation/test evaluation.

V. Conclusion

Our dual-stream architecture has already performed the hierarchical temporal Modeling at its best and recognized KSL by eliminating the fundamental problems in this area. Recognition of the system involved 89.53% validation accuracy and 48 ms latency, which were significantly improvements compared to previous approaches. The proposed framework's effectiveness and robustness have been clearly demonstrated through comparative assessment against baseline architectures and signer-independent verification. The impact of this research is immediate and extensive. The announcement of the KSL dataset will mark a major turning point for the researchers working with regional sign languages as they will be equipped with an essential resource. Besides, our deployment optimizations will make technology applicable to a wide range of real-world applications, best educational aids and communications assistive devices. The architecture's modularity allows it to be integrated into existing assistive technologies, potentially leading to a substantial change in the accessibility situation of the deaf community in Karnataka.

Future research can focus on three primary directions. The first one is to augment the data set with more gestures and regional differences which will be the reason for the model's power. The second one is to develop adaptive learning methods that could even go beyond the current limitations of few-shot signs learning. The third one is to apply, among others, explainable AI techniques which would be a plus to the model's transparency with users thus building trust.

Author Contributions

Gurusiddappa Hugar: Conceptualization, Methodology, Software, Resources, Investigation, Writing -- original draft.
Ramesh M. Kagalkar: Formal analysis, Investigation, Data curation, Visualization, Validation, Supervision, Project administration, Funding acquisition, Writing – review.
All authors have read and agreed to the published version of the manuscript.

References

[1] B. Alsharif, Y. Ouakrim, N. D. Alotaibi, and A. S. Mahmoud, "Deep learning technology to recognize American sign language alphabet," Sensors, vol. 23, no. 18, p. 7970, Sep. 2023, 10.3390/s23187970.

[2] M. Alshewimy, "Efficient deep learning models based on tension techniques for sign language recognition," Intell. Syst. Appl., vol. 20, p. 200284, Dec. 2023, 10.1016/j.iswa.2023.200284.

[3] S. U. Amin, E. Kavak, B. Kim, J. Kang, and S. Park, "Deep learning based active learning technique for data annotation and improve the overall performance of classification models," Expert Syst. Appl., vol. 228, p. 120391, Oct. 2023, 10.1016/j.eswa.2023.120391.

[4] S. U. Amin, E. Kavak, J. Kang, and S. Park, "An efficient attention-based strategy for anomaly detection in surveillance video," Comput. Syst. Sci. Eng., vol. 46, no. 3, pp. 3939–3958, 2024, 10.32604/csse.2024.049058.

[5] S. U. Amin, B. Kim, Y. Jung, S. Seo, and S. Park, "Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3D convolutions and long short-term memory modules," Adv. Intell. Syst., vol. 7, no. 2, p. 2400065, Feb. 2025, 10.1002/aisy.202400065.

[6] M. Aziz and A. Othman, "Evolution and trends in sign language avatar systems: Unveiling a 40-year journey via systematic review," Multimodal Technol. Interact., vol. 7, no. 10, p. 97, Oct. 2023, 10.3390/mti7100097.

[7] N. Bansal and A. Jain, "Word recognition from Indian sign language in videos using dual feature descriptor and GMT-MASKRCNN recognition technique," Multimedia Tools Appl., vol. 84, no. 3, pp. 2565–2597, Jan. 2025, 10.1007/s11042-024-19317-2.

[8] I. Garibay, "A survey on sign language literature," Mach. Learn. Appl., vol. 14, p. 100504, Dec. 2023, 10.1016/j.mlwa.2023.100504.

[9] B. Hdioud and M. Tirari, "A deep learning-based approach for recognition of Arabic sign language letters," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 4, pp. 452–459, Apr. 2023, 10.14569/IJACSA.2023.0140453.

[10] X. Hong-qin and Z. Yuan-yuan, "Advanced gesture recognition method based on fractional Fourier transform and relevance vector machine for smart home appliances," Comput. Animation Virtual Worlds, vol. 36, no. 1, p. e70011, Jan. 2025, 10.1002/cav.70011.

[11] Z. Huang, W. Xue, Y. Zhou, L. Hu, and X. Lu, "Dual-stage temporal perception network for continuous sign language recognition," Vis. Comput., vol. 41, no. 6, pp. 1971–1986, Apr. 2025, 10.1007/s00371-024-03414-2.

[12] A. Ji, L. Cao, and S. Wang, "Dataglove for sign language recognition of people with hearing and speech impairment via wearable inertial sensors," Sensors, vol. 23, no. 15, p. 6693, Jul. 2023, 10.3390/s23156693.
[13] A. S. M. Miah, M. A. M. Hasan, J. Shin, Y. Okuyama, and Y. Tomioka, "Spatial-temporal attention with graph and general neural network-based sign language recognition," Pattern Anal. Appl., vol. 27, no. 1, p. 37, Mar. 2024, 10.1007/s10044-024-01240-9.
[14] S. Mohsin, M. Mohsin, and M. Alshara, "American sign language recognition based on transfer learning algorithms," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 5s, pp. 390–399, 2023, 10.18201/ijisae.2023.350.
[15] F. M. Najib, "A multi-lingual sign language recognition system using machine learning," Multimedia Tools Appl., vol. 83, no. 5, pp. 1345–1362, Jan. 2024, 10.1007/s11042-023-15457-z.
[16] I. Olmos-Pineda, "Word level sign language recognition via handcrafted features," IEEE Latin Amer. Trans., vol. 21, no. 7, pp. 839–848, Jul. 2023, 10.1109/TLA.2023.10279947.
[17] M. Papatsimouli, P. Sarigiannidis, and G. F. Fragulis, "A survey of advancements in real-time sign language translators: Integration with IoT technology," Technologies, vol. 11, no. 4, p. 83, Aug. 2023, 10.3390/technologies11040083.
[18] S. Paul, M. M. Rahaman, M. A. Hossain, M. R. Islam, and I. H. Sarker, "An Adam-based CNN and LSTM approach for sign language recognition in real time for deaf people," Bull. Electr. Eng. Inform., vol. 13, no. 1, pp. 499–509, Feb. 2024, 10.11591/eei.v13i1.6211.
[19] K. K. Podder, M. E. H. Chowdhury, A. M. Tahir, Z. B. Mahbub, A. Khandakar, and M. S. Hossain, "Signer-independent Arabic sign language recognition system using deep learning model," Sensors, vol. 23, no. 16, p. 7156, Aug. 2023, 10.3390/s23167156.
[20] S. Renjith and R. Manazhy, "Sign language: A systematic review on classification and recognition," Multimedia Tools Appl., vol. 83, no. 24, pp. 77077–77127, Aug. 2024, 10.1007/s11042-024-19563-4.
[21] J. Shin, A. S. M. Miah, M. A. M. Hasan, Y. Tomioka, and Y. Okuyama, "Korean sign language recognition using transformer-based deep neural network," Appl. Sci., vol. 13, no. 5, p. 3029, Mar. 2023, 10.3390/app13053029.
[22] R. P. Singh and L. D. Singh, "Dyhand: Dynamic hand gesture recognition using BiLSTM and soft attention methods," Vis. Comput., vol. 41, no. 1, pp. 41–51, Jan. 2025, 10.1007/s00371-024-03556-3.
[23] R. Sreemathy, M. Turuk, S. Nema, P. Nilesh, and D. Singh, "Continuous word level sign language recognition using an expert system based on machine learning," Int. J. Cogn. Comput. Eng., vol. 4, pp. 170–178, Jun. 2023, 10.1016/j.ijcce.2023.05.001.
[24] A. Wali, A. Zafar, S. D. Ali, K. Javed, U. Tariq, and I. H. Lee, "Recent progress in sign language recognition: A review," Mach. Vis. Appl., vol. 35, no. 1, p. 127, Jan. 2024, 10.1007/s00138-024-01628-x.
[25] L. T. Woods and Z. A. Rana, "Modeling sign language with encoder-only transformers and human pose estimation keypoint data," Mathematics, vol. 11, no. 9, p. 2129, May 2023, 10.3390/math11092129.
[26] S. Xiong, C. Zou, J. Yun, L. Li, and X. Lu, "Continuous sign language recognition enhanced by dynamic attention and maximum backtracking probability decoding," Signal Image Video Process., vol. 19, no. 1, pp. 141–150, Jan. 2025, 10.1007/s11760-024-03531-8.
[27] S. Xue, L. Gao, L. Wan, and W. Feng, "Multi-scale context-aware network for continuous sign language recognition," Virtual Real. Intell. Hardw., vol. 6, no. 4, pp. 323–337, Aug. 2024, 10.1016/j.vrih.2023.09.001.
[28] J. Yao, J. Chen, L. Niu, and B. Sheng, "Scene-aware human pose generation using Transformer," in Proc. 31st ACM Int. Conf. Multimedia, Ottawa, ON, Canada, 2023, pp. 2847–2855, 10.1145/3581783.3611768.
[29] H. Zhang, Y. Zuo, T. Xu, F. Han, and L. Zuo, "Multipath attention and adaptive gating network for video action recognition," Neural Process. Lett., vol. 56, no. 1, p. 124, Feb. 2024, 10.1007/s11063-024-11570-8.

Gurusiddappa Hugar is currently pursuing his Ph.D. Degree in the area of Computer Science and Engineering at Visvesvaraya Technological University (VTU). He has over thirteen years of broad experience in the fields of teaching, academic administration, and scientific research. Presently, he is working as an Assistant Professor in the Department of Computer Science and Engineering at AGMR College of Engineering and Technology, Varur, Hubballi. He has 11 design patent grants in his credit. Published more than 20 research papers in reputed international journals, peer-reviewed IEEE conference proceedings, and indexed book chapters. His works have been documented in few other book chapters for developing the real time hazard detection and speech recognition framework.

Ramesh M. Kagalkar received the Ph.D. degree in Computer and Information Science from VTU in 2019.He has nearly two decades of experience in teaching, academic administration, and scientific research. He is currently working as a Professor in the Department of Information Science and Engineering at Nagarjuna College of Engineering and Technology, Bengaluru, India. Previously, he has served in key institutional leadership roles, including Dean of Research and Development (R&D). His core research interests include Human-Computer Interaction (HCI), Computer Vision, Digital Image Processing, and the Internet of Things (IoT), with a primary focus on developing automated sign language recognition systems and smart assistive technologies. Dr. Kagalkar has published over 45 research articles in reputed international journals and peer-reviewed IEEE conference proceedings. He has authored four academic textbooks, secured more than six Indian patents, and managed research projects funded by bodies such as TEQIP. He serves as an active reviewer for several international journals and holds life memberships in prominent professional engineering societies.