BdSL-Net: A Hybrid CNN-LSTM-Attention Framework for Real Time Bangla Sign Language Recognition

Md Faisal Hasan[1], Md. Nazmus Sakib Sheam[1], Uzzwal Kumar Biswas[1], Jarin Tasnim Tonvi[1] *, and Syed Ahsanul Kabir[1]

[1] Department of Computer Science and Engineering, Green University of Bangladesh, Narayanganj 1461, Dhaka, Bangladesh
{faisalhasan.work, 123sheamfeni, uzzwalbiswas70}@gmail.com, {jarin*, kabir}@cse.green.edu.bd

Abstract. Communication barriers between the Deaf and hard-of-hearing (DHH) community and the hearing population in Bangladesh persist due to the lack of automated Bengali Sign Language (BdSL) translation tools. This study proposes BdSL-Net, a real-time BdSL recognition framework based on computer vision and deep learning. A custom dataset of video 2,000 samples covering 40 BdSL signs was developed, from which 1,662 skeletal keypoints were extracted per frame using MediaPipe Holistic. The proposed hybrid neural architecture integrates a 1D Convolutional Neural Network (CNN) for spatial feature extraction, a Long Short-Term Memory (LSTM) for temporal sequence modeling, and an Attention mechanism to highlight the most discriminative motion segments. BdSL-Net achieved 96.08% accuracy and was implemented as a real-time prototype. Explainable AI (XAI) analysis further validated that the model effectively attends to crucial temporal features within gesture sequences. The results demonstrate BdSL-Net's potential as a vision-based assistive technology for bridging communication gaps and enabling future continuous BdSL translation. The findings confirm that the CNN–LSTM–Attention hybrid model offers high-accuracy recognition of BdSL gestures and provides a viable proof of concept for vision-based assistive communication technologies in low-resource linguistic contexts.

Keywords: Bengali Sign Language (BdSL), Real-time Gesture Recognition, Hybrid Deep Learning Model, CNN-LSTM-Attention Network, Explainable Artificial Intelligence (XAI)

1 Introduction

Sign language is an primary method of communicating with the Deaf and hard-of-hearing individuals around the world. It allows personal agency to be expressed, ideas are shared, and everyday encounters between people which lack an auditory component. The lack of an extensive interpretation system of the

* Corresponding author: jarin@cse.green.edu.bd

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_90

Bengali Sign Language (BdSL) in Bangladesh can be seen as the main reason behind continuing communication barriers in the country between the hearing and the DHH populations. Although BdSL has been widely adopted by the Deaf community, the absence of a scalable and real-time translational infrastructure converting the gestural input into written or spoken Bengali highlights a timely and critical concern regarding the need to address the gap in communication through the assistance of the technologies.

The modern advancement in machine learning as well as computer vision, deep learning systems and algorithms have opened up new possibilities in automatic sign-language recognition (SLR) [1] [2]. Traditional machine-learning models traditionally rely on manual feature extraction, which makes them too limited in scalability and predictive accuracy. On the other hand, deep-learning models have the ability to independently learn a spatial and temporal representation of raw visual information [3]. Despite this, the design of a robust BdSL recognition system requires overcoming a range of obstacles, such as large inter-gesture variability, lack of data with annotations, and severe real-time operating requirements.

To address these difficulties, the current paper presents the hybrid neural-network that combines 1D-CNN, LSTM and attention mechanism. The CNN part learns spatial interactions between skeletal keypoints, the LSTM component learns temporal correlations between gestural sequences and the attention component emphasizes selectively frames that are the most discriminative to sign semantics. Using MP Holistic, 1,662 keypoints per frame were extracted, which means that the results coded extensive spatial information about the signer, hands, and facial expressions[4] [5].

The collection of 2,000 samples, labeled with each of 40 BdSL gestures, was gathered to be used both as a training set and as an evaluation set. The suggested CNN-LSTM-Attention model achieved 96.08 percentage accuracy, which proves a good command of accurate gesture classification [6]. Moreover, an Explainable AI (XAI) evaluation was used to show that the network pays the right respect to the most informative and temporal parts of each gesture, which validates its transparency and reliability [7]. There was also a real-time demonstrative prototype that was also designed to show the viability of the model as a practical assistive communication tool.

To summarize, these are the contribution of this research:

- Creation of a customized set of BdSL data on 2,000 annotated samples that includes 40 commonly used signs.
- Amalgamation system of deep-learning model, which combines CNN, LSTM as well as Attention modules in order to extract spatial-temporal features.
- Classification accuracy of high 96.08%, and thus, the robustness and generalization capability of the model.
- Deployment of an actual translation prototype in order to support effective communication support.
- Performing an Explainable AI analysis to clarify how the model makes decisions and to introduce a transparency of method.

2 Related Work

Deep learning has redefined the analysis of sequence data, particularly in analyzing sequence data tasks like sign language recognition which involves a profound interpretation of space-based characteristics and dynamism. Sign language is inherently temporal and expressive and motion based such as, position of hands, physical movements, facial expression as well as timing. This complexity requires the neural-network techniques that are capable of dealing with both the patterns in vision and time-limited patterns.

Convolutional Neural Networks (CNNs) are famously known to extract spatial elements in images. As an example, Huang et al. (2017) used CNN on static American Sign Language (ASL) poses. They obtained a precision of 85%, and this demonstrated the usefulness of this method in deciphering hand shapes and formations [8]. Recurrent Neural Networks (RNNs) are popular its because, Long Short-Term Memory (LSTM) networks in order to identify gestures with a temporal dimension. Pigou et al. (2015) used LSTM to stream continuous signs with an accuracy of 88% [9]. LSTMs are good at time-sequences since they remember the past so they are especially applicable to gesture recognition.

Another significant advance was the introduction of the Attention mechanism by Vaswani et al. (2017) [10]. By creating a system for recognizing motion-based words, Tonvi et al. (2024) filled a major research void in dynamic Bangla Sign Language (BdSL). Their conceptual method achieved 80% accuracy for dynamic signs and 99% accuracy for static signs by combining machine learning models with the MediaPipe holistic framework [11].

Attention allows models to prioritize important parts of an input sequence, helping the network focus on the most informative frames. This has been successfully applied to sign language recognition as seen in Camgoz et al. (2018), where integrating Attention mechanisms raised ASL recognition accuracy to 92% [12]. In the context of Bangla Sign Language (BdSL), research is growing but still faces limitations.

M. R. Haque et al. (2023) combined DenseNet201 and ResNet50V and explored an real-time system for recognizing static Bangla Sign Language words, achieving a recognition accuracy of 93% [13]. Tazalli et al. (2022) employed YOLOv5 to recognize 34 BdSL words, achieving 51.44% accuracy a reflection of limitations in dataset size and diversity [14]. On other hand, S. M. M. M. Mahin et al. (2023) focused on detecting 27 phrases, their system achieved an accuracy of 92.07% on LSTM [15].

More recently, extensive datasets such as BdSLW401 (2025), with 401 signs and over 100,000 video clips, have underscored the increasing attention toward BdSL recognition [16]. Hybrid bilingual models like YOLOv11, covering both BdSL and ASL alphabets, show promising directions for multilingual sign language systems [17]. M. Abdul et al. (2025) Using a new 27-action Bangla Sign Language dataset, researchers created a novel method utilizing an LSTM network. With a high accuracy of 95.01%, their model was able to successfully classify movement sequences [18].

A custom CNN model was proposed by Fatema et al. (2024) to recognize 38 different BdSL gestures with an accuracy of 92.33%. The model proved to be more effective than pre-trained architectures such as VGG16 and ResNet50 [19].

Another study Tapu et al. (2024) presented a lightweight CNN with a self-attention mechanism to address the need for computational efficiency. This model demonstrated that low-resource solutions can match with a competitive accuracy of 93.47% [20].

Our approach differs by addresses that gap with a deep learning model that combines CNNs, LSTMs, and Attention layers to process a custom 40-word dynamic BdSL dataset. Trained on 2,000 gesture sequences collected via webcam, the model achieves an accuracy of 96.08%, demonstrating its capability to deliver practical and high-performance real-time recognition.

3 Dataset

To train and test the proposed BdSL recognition system, we created a special video-based dataset whose purpose is to evaluate this research. The data consists of 2,000 video samples, and balanced across 40 different words of BdSL, with a wide range of different expressions, nouns, verbs, and situational phrases [21]. Of these, 34 of them are dynamic signs, the expressions where movement of hands is associated, and six signs are static signs whose postures of the hands remain constant with only minor movement.

To ensure the dataset is balanced with each of the signs, 50 samples have been taken for every word, with randomly recruited participants capturing the signs, under different environmental and lighting conditions to have diversity of the data set and be able to generalize to different situations. The samples were recorded with a conventional RGB web-camera at 30 frames per second (FPS), with a room like environment background and distance to keep the real as possible for everyday use.

Afterwards the video samples were processed frame-by-frame using MediaPipe Holistic to extract anatomical landmarks, which were flattened into a feature vector of 1,662 numerical values per frame. This vector represents the coordinate data (x, y, z, and visibility) for facial landmarks, hand joints, and full-body skeletal coordinates, As shown in, Table 1. These structured features were utilized to model the spatio-temporal dynamics of each gesture; consequently, they replaced raw sample pixels and greatly decreased computational complexity without losing vital movement information.

Dataset was then split into 80-to-20 ratio for training (80%) and testing (20%) subsets with samples per class being evenly distributed across the both subsets. Such a setup allowed the model to acquire knowledge on inter as well as intra-class differences[22].

Table 1: Breakdown of the 1,662-keypoints, Feature Vector extracted per frame using MediaPipe Holistic.

Component	Landmarks	Attributes per Landmark	Total Values
Face	468	3 (x, y, z)	$468 \times 3 = 1,404$
Pose	33	4 (x, y, z, vis)	$33 \times 4 = 132$
Left Hand	21	3 (x, y, z)	$21 \times 3 = 63$
Right Hand	21	3 (x, y, z)	$21 \times 3 = 63$
Total	**543**	**-**	**1,662**

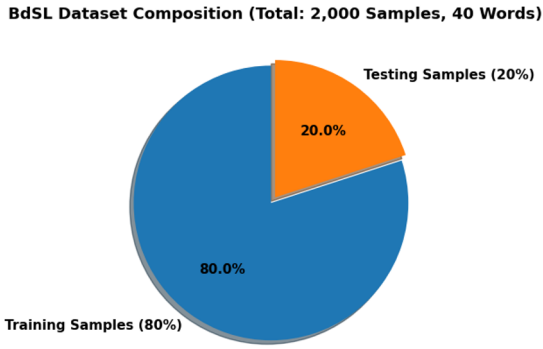

Fig. 1: Data split visualized of the custom Bangla Sign Language (BdSL) dataset

The dataset, As layed in Table 2 also offers a strong basis to further research in Bengali Sign Language recognition since it is one of the first systematically curated BdSL datasets that balanced dynamic and static gestures and thus enabled the use of deep learning-based time series-modeling.

Table 2: Overview of the Bengali Sign Language (BdSL) Dataset

Category	Count / Description	Remarks
Total Words	34 Dynamic, 6 Static	40
Dynamic Words	Involving temporal motion	34
Static Words	Represented by fixed posture	6
Samples per Word	Multiple individuals	50
Total Samples	40×50	2,000
Training Samples	80% of total data	1,600
Testing Samples	20% of total data	400
Frame Rate	Standard RGB camerae	30 FPS
Keypoints	MediaPipe Holistic Extraction	1,662

4 Methodology

Our methodology is structured as a multi-stage pipeline designed to deliver both high predictive accuracy and model interpretability. As shown in Fig: 2, the framework begins with dataset creation, followed by systematic preprocessing and augmentation to ensure robust learning. The system adopts a custom CNN+LSTM+Attention model then fine-tuned for multi-class classification across 40 bangla words. Finally, Using the gradient of the predicted class score with respect to the input features was calculated. The absolute value of these gradients creates a saliency map, where higher values indicate features that were more influential in the model's prediction.

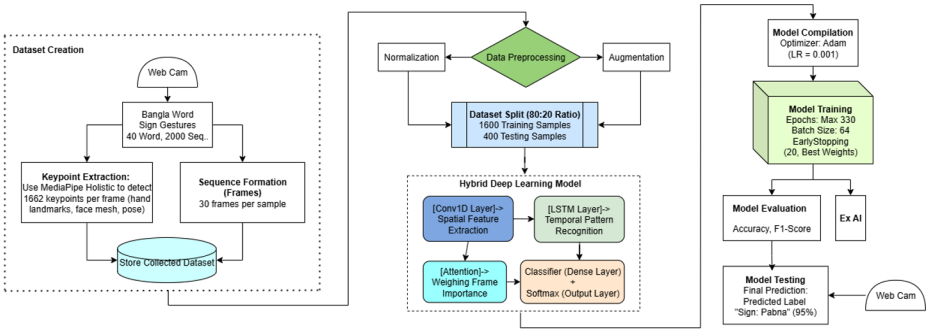

Fig. 2: Workflow of CNN+LSTM+Attention. Steps include dataset creation, preprocessing and augmentation, Bangla word classification, and Ex-AI visualization for interpretability.

4.1 Data Preprocessing

To improve the model with respect to performance and strength, we have used two major preprocessing methods. To start with, we did the normalization process, where we converted all the keypoint coordinates in such a way that it is relative to the left shoulder of the signer, As illustrated in Fig: 3, This transformation makes the analysis scale-independent and position-independent of the subject in the frame.

Second, we used data augmentation, synthetic expansion of the dataset by adding Gaussian noise and producing horizontally-mirrored Keypoints, As illustrated in Fig: 4. This enhancement plan enables better generalization to unknown data and prevents the likelihood of overfitting.

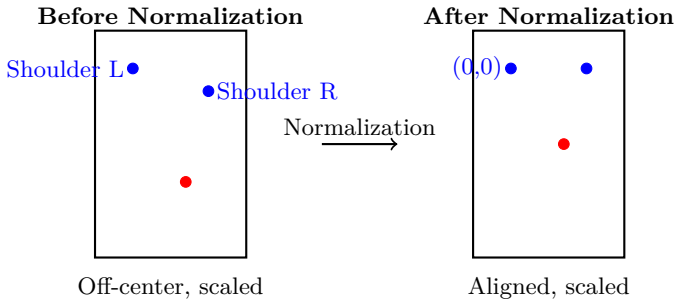

Fig. 3: Effect of keypoint normalization: coordinates are transformed relative to the left shoulder, focusing on gesture form and movement.

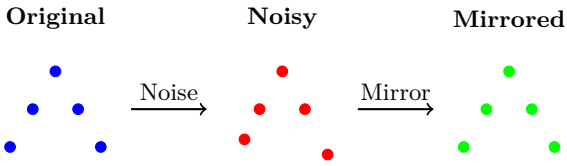

Fig. 4: Illustration of data augmentation: Gaussian noise perturbs keypoints slightly, and horizontal mirroring flips the gesture to create new training samples.

4.2 Model Architecture

The proposed model takes a hybrid neural network design to suit both spatial and temporal complexities involved in the gestures of the Bangla Sign Language (BdSL). The architecture uses 1D-CNN, LSTM as well as attention mechanism in order to make the sign recognition robust, As shown in Fig: 5. This hybrid structure allows the network to firstly capture low-level spatial information of skeletal keypoints within each frame, then model the temporal dynamics over the entire sequence of frames and finally focus attention on the most discriminative frames in order to be classified [23].

By definition, assume that $X \in \mathbb{R}^{T \times F}$ has to be an input sequence of gestures, and T denotes the total number of temporal frames, F denotes total count of features (1,662 keypoints) that are extracted by MP Holistic.

The 1D CNN layer processes the input sequence to obtain spatial representations on the keypoints at every time instance. As such, a series of feature maps represented by C are produced.

$$C = \text{CNN}_{1D}(X) \quad \text{where } C \in \mathbb{R}^{T' \times D} \tag{1}$$

Here, T' is the possibly down-sampled time length and D denotes the number of total dimensions from the extracted features of a CNN.

This leads to the sequence of higher level features, which are then fed to a LSTM layer, which is what learns the temporal correlations between the various frames. The LSTM gives a series of hidden states, which is represented by $H = (h_1, h_2, \ldots, h_{T'})$.

$$H = \text{LSTM}(C) \quad \text{where } H \in \mathbb{R}^{T' \times H_{dim}} \tag{2}$$

Where H_{dim} is dimension for LSTM hidden states.

After this is done, an attention mechanism is implemented into the LSTM network's hidden states. This process gives the model the ability to impose dynamic importance weights which are denoted by ,α_t, on each temporal step and therefore focuses on the most significant elements of the gesture sequence. The resulting context vector, c, is the weighted combination of a combination of the hidden states.

$$\alpha_t = \frac{\exp(\text{score}(h_t))}{\sum_{j=1}^{T'} \exp(\text{score}(h_j))} \tag{3}$$

$$c = \sum_{t=1}^{T'} \alpha_t h_t \tag{4}$$

This is then followed by the propagation of the context vector c, a representation of the totality of the gesture sequence - through a fully connected (FC) layer with its parameters represented by a bias vector $b \in \mathbb{R}^K$, and a weight matrix $W \in \mathbb{R}^{K \times H_{dim}}$. The cardinality of BdSL sign classes is represented by a parameter that is abbreviated as $K = 40$. The final logits, o, is obtained in this operation.

$$o = Wc + b \tag{5}$$

The likelihoods of the various classes given as $\hat{y}$ are estimated by means of the softmax function applied to the logits:

$$\hat{y}_k = p(y = k|X) = \frac{\exp(o_k)}{\sum_{j=1}^{K} \exp(o_j)}, \quad k = 1, \ldots, K \tag{6}$$

To train the network, we optimize the model parameters by decreasing the as much as loss of categorical cross-entropy, $\mathcal{L}$, compared to the anticipated probability $\hat{y}$ and the one-hot encoded true labels y.

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log(\hat{y}_{i,k}) \tag{7}$$

and N denotes cardinality of sample set on a batch. Adam optimizer with a learning rate of 1×10^{-3} was used to train.

To capture temporal dependencies, the model runs input feature sequences through stacked LSTM units and an embedding layer. Then, to improve context

understanding, attention modules give the most instructive time steps larger weights. A final dense-softmax layer generates class probabilities for sequence classification, while dropout layers in between stages enhance generalization.

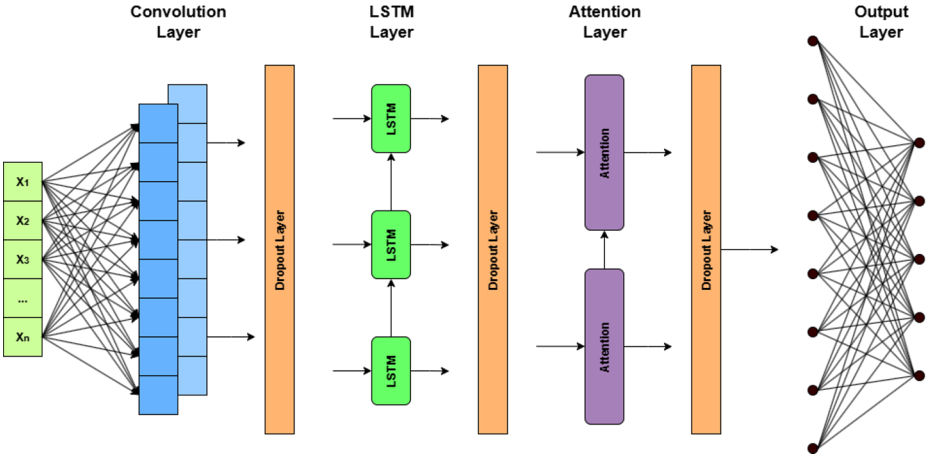

Fig. 5: Detailed schematic of proposed CNN-LSTM-Attention hybrid model flow: CNN extracts spatial features, LSTM captures the sequence, and the Attention layer focuses important frames.

4.3 Explainability with Saliency Map

We analyzed the learned attention weights (α_t) during inference in order to quantitatively evaluate the attention mechanism's contribution. Through the dynamic identification of the most discriminative frames within an input sequence, this procedure offers empirical proof of the model's capacity to carry out temporal feature selection.

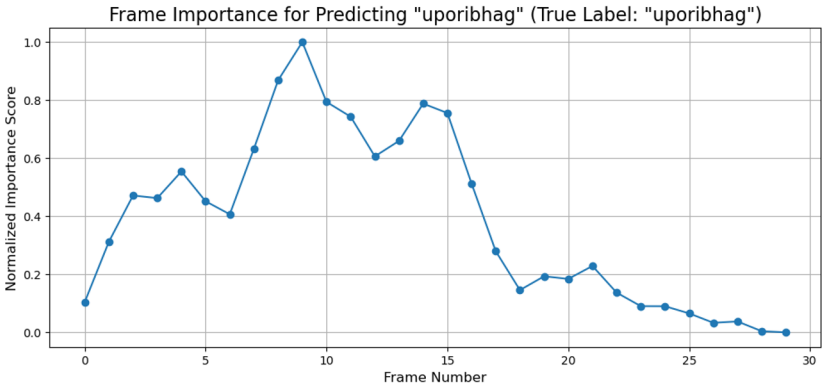

Fig. 6: Frame Importance Visualization.

As shown in Fig: 6, the distribution of normalized attention weights for a correctly classified instance of the sign uporibhag across all temporal steps ($t \in [0, 29]$). The sequence of LSTM hidden states $(h_1, h_2, \ldots, h_n)$ is passed through a feed-forward network and then a Softmax function calculates the attention weights. The weighted sum of these hidden states is then used to compute the final context vector c:

$$c = \sum_{t=1}^{n} \alpha_t h_t \tag{8}$$

A very localized and sparse weight distribution can be seen in the plot. Important technical observations consist of:

Saliency Localization: With a clear global maximum at frame $t = 9$, the model assigns significant weights to a particular temporal window. This suggests that the main contributors to the final context vector are the hidden state h_9 and its close temporal neighbors. From the standpoint of machine learning, the model has discovered that this particular subsequence encodes the most prominent, class-discriminative characteristics for "uporibhag".

Temporal Feature Attenuation: The weights given to frames in the terminal phase ($t > 22$) and initial phase ($t = 0$ to $t = 2$) are greatly reduced, almost to zero. This behavior is important because it shows that the model has learned to be invariant to non-discriminative information. The preparatory and retractional phases of the gesture, which have lower signal content than the core sign execution, are usually represented by these attenuated frames. The model successfully filters temporal noise by down-weighting these segments.

Inductive Bias for Robustness: This empirical finding confirms that the attention mechanism is a potent inductive bias. It compels the model to build a representation that is concentrated on a small number of significant events rather than being uniformly dependent on every frame. Because the classifier's decision is largely based on the high-importance, information-rich segments of

the gesture, this improves the model's resilience to temporal variations, such as changes in signing speed or slight hesitations.

5 Result

A held-out test set of 400 Bangla Sign Language (BdSL) sequences was used to assess the suggested hybrid model. Overall accuracy, precision (per-class), recall, F1-score, confusion matrix analysis [24]. These metrics are crucial in order to understanding the models performance, decision-making ability and robustness of signs.

5.1 Quantitative Analysis

Through an overall classification accuracy of **96.08%**, the model exhibits strong qualitative results. The result confirms that hybrid CNN-LSTM-Attention architecture and the chosen feature extraction pipeline have been successful at capturing all the complicated dynamic words or spatio-temporal patterns. Table 3 shows an summary of proposed model's performance based on macro-averaged metrics.

Table 3: Evaluation Results of the Proposed Model using Standard Classification Metrics on test set.

Metric	**Score (%)**
Accuracy	96.08
Precision (Macro)	94.25
Recall (Macro)	92.25
F1-Score (Macro)	96.00

For further analysis of the efficacy of the model on a per-class basis, we generate a confusion matrix, which can be seen in Fig: 7. With a high true positive rate across most of the 40 sign classes and few off-diagonal entries (misclassifications), the matrix shows a strong diagonal concentration.

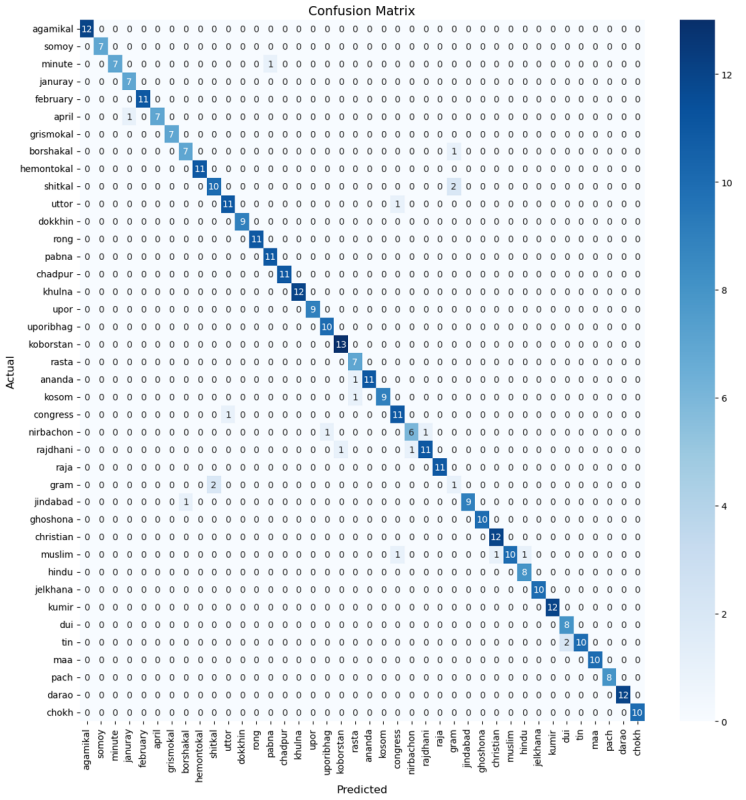

Fig. 7: Confusion Matrix

5.2 Generalization to Unknown Users

The capacity of an model to generalize to unknown users is a crucial test of robustness. We tested the trained model using a user whose data was not included in the training set after deploying it in a real-time system (Fig: 8). Strong signer-independent performance was demonstrated by the system, which was able to identify signs with high confidence. The use of MediaPipe skeletal keypoints, which abstract away signer-specific visual features like appearance and attire, and data augmentation, which exposed the model to a greater range of gestural variations during training, are primarily responsible for this generalization.

5.3 Comparison with Existing Work

When compared to previous research Table 4, our suggested system shows notable improvements in scope, methodology, and performance. Although Haque et al.(2023)'s work offers a basis for static word recognition with a 93% accuracy rate, it is only applicable to 10 static words. Our system not only outperforms this with 99% accuracy on static words, but it also tackles the more

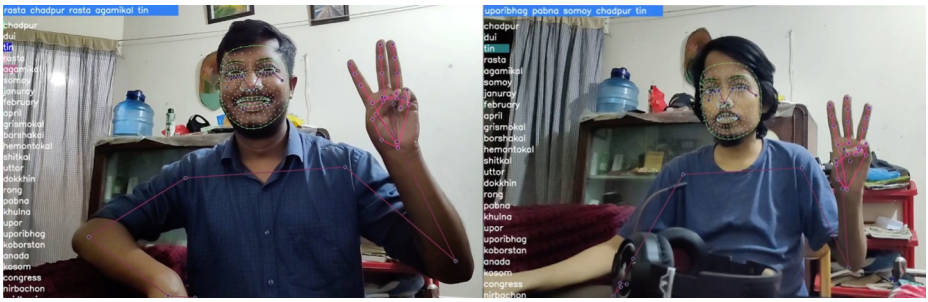

Fig. 8: BdSL signs detection using unknown user

Table 4: Comparative Analysis of Recent BdSL Recognition Studies

Study	**Methodology**	**Dataset Size**	**Accuracy (%)**
Uddin et al.(2023)[25]	BlazePose+ LSTM	Small (sentences)	87.14
Haque et al.(2023)[13]	DenseNet201+ ResNet50-V2	10 words	93.00
Mahin et al.(2023)[15]	MediaPipe+ LSTM	27 phrases	92.07
M. Abdul et al.(2025)[18]	LSTM	27 actions (24.3k images)	95.01
Tapu et al.(2024)[20]	LW-CNN (Attention)	38 gestures	93.47
Our Research	**CNN-LSTM-Attention**	**40 words (2,000 seq.)**	**96.08**

difficult problem of dynamic, motion-based signs, which their research did not address. Additionally, compared to their deep CNN-based models (DenseNet201 and ResNet50V2), our use of the MediaPipe framework provides a more computationally efficient method.

In a comparable way, Mahin et al. used an LSTM model to achieve 92.07% accuracy while focusing on phrases rather than individual words, even though they included both static and dynamic signs. With 80% accuracy, our research sets a strong performance benchmark and focuses on a known gap in BdSL research: the recognition of isolated dynamic words. This distinction is crucial because correctly interpreting longer sign phrases requires first recognizing individual dynamic words. Our work offers a more detailed analysis and a focused solution for this difficult area by creating a new, categorized dataset specifically for this purpose.

6 Conclusion and Future Work

The experimental results validate the effectiveness of the proposed BdSL-Net for isolated Bengali Sign Language (BdSL) recognition. The integration of attention mechanisms, which dynamically emphasize the most discriminative frames, and LSTM layers, which capture temporal dependencies, collectively contribute to the model's high classification accuracy of 96.08%. Qualitative analyses further reinforce these findings and it confirms that the proposed model's ability to generalize across different signers. The system's successful real-time implementation demonstrates its practical applicability and robustness, establishing it as a promising proof-of-concept for vision-based assistive communication technologies and a step toward developing user-independent, continuous BdSL translation systems.

While the proposed framework demonstrates strong performance, certain limitations warrant further investigation. Dataset used in this research was created under controlled environmental conditions such as room and sitting down infront of computer, which may constrain the model's robustness in more diverse real-world settings. Additionally, the current system is designed for recognizing isolated signs rather than continuous sequences. In future the focal point of this research will be allocating enough time and resources to expanding the existing dataset with handful of people, including broader bengali vocabulary. Further efforts will also aim to extend the model toward continuous sign language recognition, incorporating advanced temporal modeling and context-aware mechanisms to enable seamless real-time translation of natural sign communication.

References

1. Cooper, Helen & Holt, Brian & Bowden, Richard. (2011). Sign Language Recognition. doi:10.1007/978-0-85729-997-0_27
2. Rastgoo R, Kiani K, Escalera S. Sign Language Recognition - A Deep Survey. Expert Syst. Appl. 2021;164:113794. 113794. doi:10.1016/J.ESWA.2020.113794
3. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Electron Markets 31, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2
4. Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8
5. Greff, Klaus Srivastava, Rupesh Koutník, Jan Steunebrink, Bas Schmidhuber, Jürgen. (2015). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems. 28. 10.1109/TNNLS.2016.2582924
6. Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," Neurocomputing, vol. 452, pp. 48–62, 2021, doi:10.1016/j.neucom.2021.03.091
7. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Tang, J., Kan, MY., Zhao, D., Li, S., Zan, H. (eds) Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science(), vol 11839. Springer, Cham. https://doi.org/10.1007/978-3-030-32236-6_51

8. Jie Huang, Wengang Zhou, Houqiang Li and Weiping Li, "Sign Language Recognition using 3D convolutional neural networks," 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, 2015, pp. 1-6, doi:10.1109/ICME.2015.7177428.
9. Pigou, L., Dieleman, S., Kindermans, PJ., Schrauwen, B. (2015). Sign Language Recognition Using Convolutional Neural Networks. In: Agapito, L., Bronstein, M., Rother, C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science(), vol 8925. Springer, Cham. https://doi.org/10.1007/978-3-319-16178-5_40
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017. doi:10.48550/arXiv.1706.03762
11. Tonvi, Jarin Nijhum, S. Rahman, Towfiq. (2024). Enhancing Real-Time Converter for Bangla Sign Language with the Integration of Dynamic Words. 687-692. 10.1109/ICCIT64611.2024.11022002.
12. Camgoz, Necati Koller, Oscar Hadfield, Simon Bowden, Richard. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. 10.48550/arXiv.2003.13830.
13. Haque A, Pulok RA, Rahman MM, Akter S, Khan N, Haque S. Recognition of Bangladeshi Sign Language (BdSL) Words using Deep Convolutional Neural Networks (DCNNs). Emerg Sci J [Internet]. 2023 Dec. 1 [cited 2025 Nov. 20];7(6):2183-201. Available from: https://www.ijournalse.org/index.php/ESJ/article/view/2062
14. T. Tazalli et al., "Computer Vision-Based Bengali Sign Language To Text Generation," 2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS), Genova, Italy, 2022, pp. 1-6, doi:10.1109/IPAS55744.2022.10052928.
15. S. M. M. M. Mahin, M. R. Islam and S. M. M. Ahsan, "Phrase Level Bangla Sign Language Recognition using Keypoints from Hand Gesture Video," 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 2023, pp. 1-6, doi:10.1109/NCIM59001.2023.10212460.
16. Rubaiyeat, Husne Youssouf, Njayou Hasan, Md Kamrul Mahmud, Hasan. (2025). BdSLW401: Transformer-Based Word-Level Bangla Sign Language Recognition Using Relative Quantization Encoding (RQE). 10.48550/arXiv.2503.02360.
17. Navin, N., Farid, F.A., Rakin, R.Z., Tanzim, S.S., Rahman, M., Rahman, S., Uddin, J. Karim, H.A. (2025). Bilingual sign language recognition: A YOLOv11-based model for Bangla and English alphabets. Journal of Imaging, 11, 134. https://doi.org/10.3390/jimaging11050134
18. Masud, M.A., Das, A., Huq, S. Mondal, S.S. (2025). Bangla sign language detection in real-time by using action recognition and LSTM deep learning model. In: Proceedings of the 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), Rangpur, Bangladesh, pp. 1–6. https://doi.org/10.1109/QPAIN66474.2025.11171951
19. K. F. Tanni, S. Islam, Z. Sultana, T. Alam and M. Mahmood, "DeepBdSL: A Comprehensive Assessment of Deep Learning Architectures for Multiclass Bengali Sign Language Gesture Recognition," 2024 27th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2024, pp. 2219-2224, doi:10.1109/ICCIT64611.2024.11022054.
20. T. K. Tapu, F. Faiaz and A. R. Sikder, "Lightweight Convolutional Neural Network with Self-Attention Mechanism for Bangla Sign Language Recognition," 2025

International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2025, pp. 1-6, doi:10.1109/ECCE64574.2025.11013937.
21. Bangladesh Sign Language Committee, Bengali Sign Language Dictionary, National Centre for Special Education, Ministry of Social Welfare, Dhaka, Bangladesh, 1994. [Online]. Available: https://books.google.com.bd/books?id=oa_0HAAACAAJ
22. Nguyen, Quang Ly, Hai-Bang Ho, Lanh Al-Ansari, Nadhir Lê, Hiệp Van Quan, Tran Prakash, Indra Pham, Binh. (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering. 2021. doi:10.1155/2021/4832864.
23. Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. Suganthan, P.N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151. https://doi.org/10.1016/j.engappai.2022.105151
24. Yacouby, Reda Axman, Dustin. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. 79-91. 10.18653/v1/2020.eval4nlp-1.9.
25. S. K. Akash, D. Chakraborty, M. M. Kaushik, B. S. Babu and M. S. R. Zishan, "Action Recognition Based Real-time Bangla Sign Language Detection and Sentence Formation," 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2023, pp. 311-315, doi:10.1109/ICREST57604.2023.10070072.