2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM 2025)
27-28 June, 2025, Gazipur-1707, Bangladesh

Multimodal Deep Learning Approach for Bangla Sign Language Recognition: Integrating Spatial and Geometric Features

Sumaiya Rahman
Dept. of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
rsumaiya335@gmail.com

Shyla Afroge
Dept. of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
shyla.ruet@gmail.com

Abduz Zami
Dept. of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
abduz.zami@gmail.com

Mir Md. Jahangir Kabir
Transdisciplinary School
University of Technology Sydney
New South Wales, Australia
mmjahangir.kabir@gmail.com

***Abstract*—In South Asian countries like Bangladesh, the deaf and hard-of-hearing communities predominantly use Bangla Sign Language (BdSL) for communication. However, existing sign language recognition systems often depend primarily on spatial data derived from images or geometric features extracted from hand landmarks, which limits their applicability and effectiveness. This study proposes a unique multimodal deep learning architecture that improves BdSL recognition by merging CNN-based spatial and landmark-based geometric information. The model is trained on the BdSL47 dataset, which contains 37,103 images and uses real-time data augmentation to improve generalizability. The proposed architecture consists of two concurrent streams: a CNN that extracts spatial characteristics from RGB images and a fully connected network that processes 63-dimensional hand position data. These representations are combined and enhanced with fully connected layers, resulting in robust categorization. Experimental evaluations using 10-fold cross-validation show that the proposed approach beats both classic machine learning classifiers and cutting-edge deep learning models, with a remarkable 99.96% accuracy. The inclusion of an NVIDIA Tesla P100 GPU assures computing power, making the model appropriate for real-time applications. The findings set a new standard for BdSL recognition and demonstrate the efficacy of multimodal learning in sign language classification.**

***Index Terms*—Sign Language, Bangla Sign Language, Multimodal, Sign Language Classification**

I. Introduction

Bangla Sign Language (BdSL) serves as a vital communication medium for the deaf and hard-of-hearing community in Bangladesh, facilitating interactions through a series of structured hand gestures and movements. Over 3 million people who are deaf or hard of hearing reside in Bangladesh [1]. For them, sign language serves as a vital means of communication, making it crucial for expressing their social, emotional, and linguistic development [2]. Despite its importance, the development of automated systems for BdSL recognition remains a complex challenge due to the intricate nature of the gestures and the subtle variations in hand movements.

Sign language recognition (SLR) has shifted from sensor-based methods to vision-based approaches, with deep learning techniques, such as CNNs, improving accuracy significantly [3]. Early methods for Bangla Sign Language (BdSL) focused on static gesture recognition, such as recognizing BdSL alphabets [4]. More recent studies have introduced advanced machine learning methods, including Graph Neural Networks, to enhance BdSL recognition performance [5].

Building upon these foundations, this study proposes a unique multimodal approach that synergistically combines spatial and geometric features to enhance the accuracy and robustness of BdSL recognition systems. By integrating spatial information, which captures the visual appearance of gestures, with geometric features that represent the structural configurations of hand movements, the proposed method seeks to overcome the inherent limitations of unimodal systems and to provide an in-depth understanding of BdSL gestures.

The integration of these modalities is anticipated to improve the system's ability to discern subtle differences between similar gestures, thereby advancing the field of sign language recognition and contributing to more inclusive communication technologies for the BdSL community. The contributions of this study are as follows:

- A novel multi-model approach is introduced that effectively integrates spatial and geometric features for the classification of Bengali sign languages.
- The proposed method outperforms existing approaches in this domain, setting a new benchmark for performance in Bangla sign language classification.
- This study proposes a highly efficient, lightweight model that achieves superior results while maintaining com-

putational efficiency, making it suitable for real-time applications.

The organization of the paper is as follows: Section II offers an in-depth review of related work in the field. Section III outlines the methodology employed in this research. The experimental setup and implementation specifics are detailed in Section IV. Section V presents the results and includes a discussion. Finally, the paper concludes with Section VI.

II. Related Works

The sign language recognition (SLR) research domain is focused on addressing the communication barriers faced by individuals with hearing and speech impairments. By improving computational techniques to recognize sign language, SLR systems have the potential to foster greater inclusivity and accessibility. The latest advances in deep learning have significantly enhanced the performance of Bangla Sign Language (BdSL) systems, improving their accuracy and generalization capabilities across diverse datasets and real-world conditions.

Das et al. [6] introduced a hybrid deep learning model that combines a CNN-based feature extractor and a random forest classifier. Evaluated on the Ishara-Bochon and Ishara-Lipi datasets, the first open-access multipurpose datasets for BdSL. The model received an F1 score of 91.47% for character recognition and 97.37% for digit recognition. Additionally, the study presented a background reduction strategy that dramatically enhanced recognition rates, demonstrating the strength of hybrid models in BdSL recognition.

Hadiuzzaman et al. [7] developed the BAUST Lipi dataset and proposed a hybrid-CNN model that includes several convolutional layers, activation functions, dropout mechanisms, and LSTM layers. The model attained 97.92% accuracy, demonstrating the efficiency of CNN-LSTM integration in recognizing complicated BdSL motions.

Rayeed et al. [8] introduced the BdSL47 dataset, which uses depth-based features to recognize BdSLs. In addition, they developed an Artificial Neural Network (ANN) model with four hidden layers and dropout mechanisms that received an F1 score of 97.84%, highlighting the superiority of depth-based features over standard 2D techniques.

Miah et al. [9] proposed BenSignNet, a CNN-based model with a concatenated segmentation technique that employs YCbCr, HSV, and watershed algorithms. The authors trained the model on numerous datasets, including '38 BdSL,' 'KU-BdSL,' and 'Ishara-Lipi.' The model attained accuracies of 94.00%, 99.60%, and 99.60% across datasets, proving its great generalization capabilities.

Abedin et al. [10] demonstrated the Concatenated BdSL Network, which combines CNN-based image networks and a pose estimation module to extract spatial and keypoint-based geometric characteristics. The model attained 91.51% accuracy, highlighting the potential of integrating CNNs with posture estimates to recognize complicated BdSL symbols.

Hassan et al. [11] created a basic but successful CNN-based BdSL gesture detection system that achieved roughly 92% accuracy with manually collected sign images. This study demonstrated the feasibility of using CNN models to construct low-cost BdSL recognition systems.

Nihal et al. [12] used transfer learning and zero-shot learning (ZSL) to solve the problem of detecting unseen BdSL indications. The study evaluated ZSL on a dataset of 35,149 photos with differences in backdrop, lighting, hand position, and skin tone, demonstrating its ability to recognize unique BdSL indications that were not present during training.

TABLE I
Dataset Statistics of the BdSL47 Dataset

Attribute	Value
Sign type	Static, 1 handed
No of users	10
No of alphabet signs	37
Total no of input image	37103
Image resulation	640x480

III. Methodology

Existing sign language recognition methods primarily rely on either spatial features, such as those extracted with Convolutional Neural Networks (CNNs), or geometric features, such as those derived through posture estimation. However, few research have investigated integrating these two approaches into a single framework. This gap emphasizes the necessity of an improved approach for recognizing Bangla Sign Language (BdSL).

To overcome this challenge, we proposed a multimodal deep learning framework that combines convolutional spatial features with geometric representations derived from landmarks. By combining these alternative modalities, our framework intended to increase the resilience, accuracy, and generalization of BdSL identification systems, particularly when dealing with problems such as differences in signing styles, lighting circumstances, and backdrop surroundings.

A. Dataset Description

For this study, we used the BdSL47 dataset, a comprehensive depth-based dataset for Bangla sign language that contains both alphabet and digit images [8]. We took images from this dataset that correspond to 37 Bangla alphabet signs, for a total of 37,103 input images. All of the images in the dataset were one-handed static signs captured in RGB format at 640x480 pixels. In Table I, we represented the dataset's information in a tabular format.

B. Data Pre-processing

Data pre-processing is crucial for ensuring high-quality and reliable input data for model training. Here, we used two types of datasets: Image Data (RGB Hand Sign Images) extracted from the BdSL47 dataset and Geometric Data (Hand Landmarks) which were a set of three dimensional coordinates representing hand pose estimation generated by MediaPipe [13]. Fig. 1 shows how MediaPipe detects hands and their landmark keypoints from some sample Bangla sign images. The data was preprocessed through the following steps:

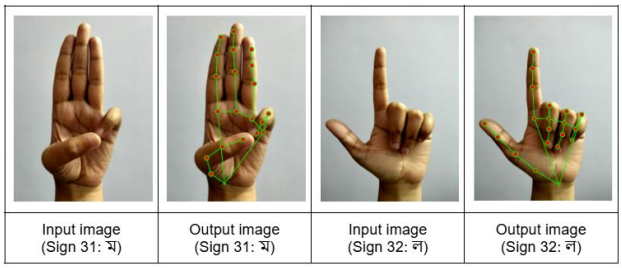


Fig. 1. Sample Bangla Sign Language images with corresponding hand keypoints generated using MediaPipe.

The RGB Hand Sign Images were resized to 64x64 pixels for consistency among samples. Then, each image was normalized to the [0, 1] range by dividing pixel values by 255.

The corresponding landmark features were then standardized using z-score normalization to provide interoperability with deep learning models. One-hot encoding was also applied to transform categorical class labels into numerical representations.

C. Real-time Data Augmentation

In this study, real-time data augmentation was employed to increase model generalization and reduce overfitting. To enhance gesture recognition, random rotation within the range (-12° to +12°) and horizontal flipping were applied. To provide variations in gesture size, images were downsized at random by a factor of 0.2. Shifts were also used to simulate natural hand movements across the X and Y axes. This technique increases the diversity of training samples without requiring additional data collection.

D. Multimodal CNN Architecture

Our model employs a multimodal deep learning strategy to recognize Bangla Sign Language by leveraging both visual spatial features and geometric hand-shape structures. Fig. 2 shows the architectural summary of the multimodal CNN model. The architecture consists of two concurrent streams: a Convolutional Neural Network (CNN) for extracting the spatial characteristics from RGB images and a fully connected network for processing 63-dimensional landmark-based hand pose features. The architecture comprises four key stages:

1) Spatial Feature Extraction: The spatial feature extraction stream processed 64×64×3 RGB pictures using a deep CNN with three convolutional blocks including 64, 128, and 256 filters, respectively. The model employs ELU activation, He Normal weight initialization, Batch Normalization for stability, Max pooling (2x2) for downsampling, and a 0.3 dropout rate to prevent overfitting. A Global Average Pooling (GAP) layer was applied to reduce the spatial dimensions of the feature map F while preserving essential spatial characteristics. To achieve a compact spatial feature representation, a fully linked layer with 128 neurons and ReLU activation was used:

$$f_{\text{spatial}} = \text{ReLU}(W_{\text{fc1}} \cdot \text{GAP}(F) + b_{\text{fc1}}) \tag{1}$$

Here, W_{fc1} and b_{fc1} denote the weight matrix and bias vector of the fully connected layer, respectively.

2) Geometric Feature Extraction: The geometric feature extraction pipeline handled 63-dimensional landmark-based hand pose features based on key points on the signer's hand. The extracted features were first reshaped using a Flatten layer, then two fully connected (dense) layers with 256 and 128 neurons, respectively. The ReLU activation function was used to induce non-linearity, which improves the model's ability to record complicated patterns. Batch normalization was used to stabilize training and a dropout rate of 0.3 is used to reduce overfitting, resulting in robust feature learning. The transformation of the landmark-based features is expressed as follows:

$$f_{\text{geo}}^{(1)} = \text{ReLU}(W_1 \cdot x_{\text{landmarks}} + b_1) \tag{2}$$

$$f_{\text{geo}}^{(2)} = \text{ReLU}(W_2 \cdot f_{\text{geo}}^{(1)} + b_2) \tag{3}$$

Here, $x_{\text{landmarks}}$ denotes the 63-dimensional vector of landmark coordinates. W_1, W_2 are the weight matrices of the fully connected layers, and b_1, b_2 are the corresponding biases.

3) Feature Fusion: The spatial and geometric feature vectors were integrated to form a unified multimodal representation, which was further refined using fully connected layers. A dense layer of 256 neurons activated with ReLU was followed by Batch Normalization and Dropout (0.3), which improves regularization. To improve stability and maximize feature learning, another thick layer of 128 neurons with ReLU activation was added, along with Batch Normalization and Dropout (0.3). The fusion of the spatial and geometric feature representations is formulated as follows:

$$f_{\text{fused}} = f_{\text{spatial}} \oplus f_{\text{geometric}} \tag{4}$$

$$f_{\text{fused}} = \text{ReLU}(W_3 \cdot f_{\text{fused}} + b_3) \tag{5}$$

$$f_{\text{fused}} = \text{ReLU}(W_4 \cdot f_{\text{fused}} + b_4) \tag{6}$$

Here, $\oplus$ denotes the concatenation operation. W_3, W_4 are the weight matrices, and b_3, b_4 are the corresponding bias vectors of the fully connected layers.

4) Classification Output: Finally, the fused feature vector is passed through a softmax activation layer to predict one of the 37 Bangla sign language classes:

$$y = \text{softmax}(W_{\text{out}} \cdot f_{\text{fused}} + b_{\text{out}}) \tag{7}$$

Here, W_{out} and b_{out} denote the weight matrix and bias vector of the output layer, respectively.

IV. Experimental Setup and Implementation

In our study, all experiments were performed on the Kaggle platform using the NVIDIA Tesla P100 GPU to accelerate training and enhance deep learning performance. The studies used strategies such as hyperparameter tuning and cross-validation to improve model performance, provide reliable generalization, and reduce overfitting. To preserve class balance, the dataset was randomly shuffled and divided into 80% for training and 20% for testing.

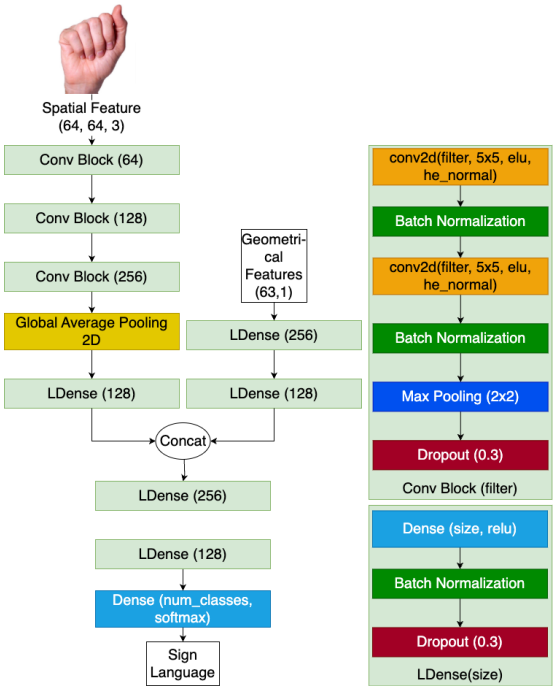


Fig. 2. The Multimodal CNN Architecture

A. Hyper-parameter Tuning

The model's performance was optimized through hyperparameter tuning, which is critical to improving accuracy. Adam was used as the optimizer with a learning rate of 0.001, and the accuracy of the validation was monitored throughout the training. To avoid overfitting, early stoppage was applied with a patience of 11 epochs, and the learning rate was reduced by a factor of 0.5 after 7 epochs without improvement. These techniques ensure that the model converges effectively without overfitting. As the loss function, categorical cross entropy was employed.

B. Cross Validation

In this experiment, 10-fold cross-validation with 50 epochs at each fold was used to evaluate the model's generalizability. The dataset was divided into 10 folds, with each fold being used once for training and once for validation.

V. Results & Discussions

Utilizing the BdSL47 dataset, extensive evaluations were conducted to assess the effectiveness of the proposed multimodal deep learning approach. We then compared our findings to classic machine learning classifiers and cutting-edge models.

A. Performance of our approach

Our proposed multimodal deep learning model achieved an outstanding accuracy of 99.96% on the BdSL-47 dataset, significantly surpassing previous methods. Fig. 3 presents some Bangla sign images along with their matching real and predicted labels, allowing for a visual representation of the model's classification results.

Fig. 3. Bangla sign images with actual and predicted labels

The model demonstrates enhanced resilience to changes in occlusion, lighting, and hand orientation by incorporating structural and visual information. The results underscore the effectiveness of multimodal learning in Bangla Sign Language recognition and its potential for real-time, assistive communication technologies.

B. Comparison with State-of-the-Art Models

To evaluate the effectiveness of our proposed Bangla Sign Language (BdSL) recognition model, we conducted a comparative analysis with existing state-of-the-art approaches that have been applied to the BdSL47 dataset. This involved assessing performance across several traditional and ensemble machine learning classifiers to verify the efficacy of the proposed model.

Table II summarizes the reported metrics—accuracy, precision, recall, and F1-score—for classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest Classifier (RFC), Histogram-based Gradient Boosting Classifier (HistGBC), and LightGBM. According to their results, LightGBM exhibited the best accuracy of 98.84%, which was followed by RFC (98.53%) and HistGBC (98.81%). In contrast, our proposed CNN-based multimodal model achieved an accuracy of 99.96%, significantly outperforming the previously reported models across all metrics.

This performance gain highlights the strength of our architecture, which effectively combines spatial features extracted by convolutional layers with geometric landmark-based inputs. By achieving a higher degree of accuracy while using the same benchmark dataset, our approach demonstrates a clear advancement in Bangla Sign Language recognition and holds strong potential for deployment in real-world assistive communication systems.

Fig. 4 shows our model's confusion matrix, which illustrates its performance across different classes. The matrix demonstrates that the model achieves nearly flawless classification, with negligible misclassifications across all categories. In addition, Fig. 5 illustrates the training and validation loss

TABLE II
PERFORMANCE COMPARISON OF TRADITIONAL AND ENSEMBLE CLASSIFIERS ON THE BDSL47 SIGN ALPHABET DATASET (37 LABELS)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
KNN	98.49	98.50	98.49	98.49
SVM	96.55	96.58	96.55	96.55
LR	86.31	86.45	86.31	86.30
RFC	98.53	98.53	98.53	98.54
HistGBC	98.81	98.82	98.81	98.81
Light-GBM	98.84	98.85	98.84	98.84
Proposed model	**99.96**	99.96	99.96	99.96

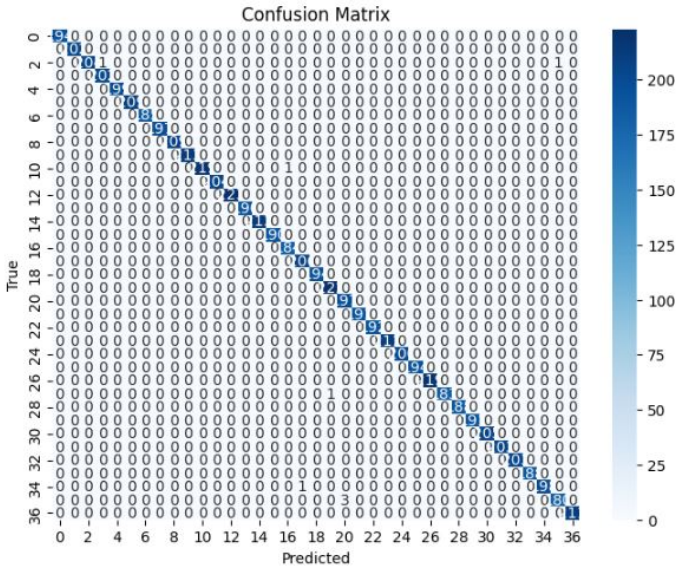
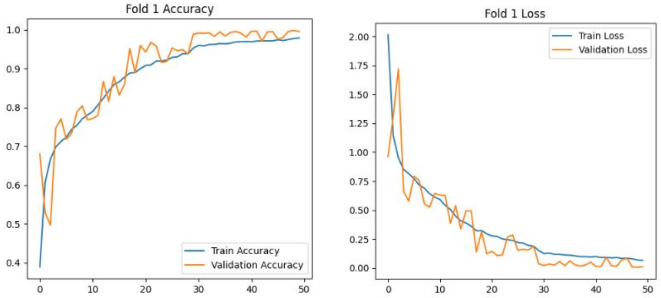


Fig. 4. Confusion Matrix of the Proposed Model

and accuracy curves for our model. The graph illustrates the model's performance during training and validation for Fold 1 of the 10-fold cross-validation, illustrating its convergence and generalizability.



(a) Accuracy Curve (b) Loss Curve

Fig. 5. Training and validation loss and accuracy curves.

Table III displays the classification report, which exhibits the performance of the proposed model for recognizing Bangla Sign Language. The model achieves near-perfect precision, recall, and F1 scores across all classes, with an overall accuracy of 99.96%. This indicates the efficiency of integrating spatial and geometric cues for accurate recognition and classification of Bangla Sign Language gestures.

TABLE III
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	1.0000	1.0000	1.0000	194
1	1.0000	1.0000	1.0000	202
2	1.0000	0.9903	0.9951	206
3	0.9951	1.0000	0.9975	202
4	1.0000	1.0000	1.0000	199
5	1.0000	1.0000	1.0000	209
6	1.0000	1.0000	1.0000	185
7	1.0000	1.0000	1.0000	197
8	1.0000	1.0000	1.0000	209
9	1.0000	1.0000	1.0000	213
10	1.0000	1.0000	1.0000	220
11	1.0000	1.0000	1.0000	204
12	1.0000	1.0000	1.0000	223
13	1.0000	1.0000	1.0000	190
14	1.0000	1.0000	1.0000	213
15	1.0000	1.0000	1.0000	190
16	1.0000	1.0000	1.0000	184
17	0.9952	1.0000	0.9976	206
18	1.0000	1.0000	1.0000	194
19	1.0000	1.0000	1.0000	222
20	1.0000	1.0000	1.0000	197
21	1.0000	1.0000	1.0000	191
22	1.0000	1.0000	1.0000	192
23	1.0000	1.0000	1.0000	211
24	1.0000	1.0000	1.0000	200
25	1.0000	1.0000	1.0000	194
26	1.0000	1.0000	1.0000	219
27	1.0000	1.0000	1.0000	184
28	1.0000	1.0000	1.0000	188
29	1.0000	1.0000	1.0000	191
30	1.0000	1.0000	1.0000	205
31	1.0000	1.0000	1.0000	203
32	1.0000	1.0000	1.0000	203
33	1.0000	1.0000	1.0000	188
34	1.0000	0.9950	0.9975	199
35	0.9946	1.0000	0.9973	183
36	1.0000	1.0000	1.0000	211
Accuracy			0.9996	7421
Macro Avg	0.9996	0.9996	0.9996	7421
Weighted Avg	0.9996	0.9996	0.9996	7421

By fusing CNN-extracted spatial features with landmark-derived geometric data, the proposed framework significantly improves the robustness and precision of Bangla Sign Language (BdSL) recognition. Effective preprocessing, real-time data augmentation, and feature fusion all improve generalization, while dropout and batch normalization reduce overfitting.

Furthermore, hyperparameter optimization, 10-fold cross-validation, and an improved training technique provide stability and efficacy. The experimental results on the BdSL47 dataset show that it outperforms existing models, particularly when dealing with differences in signing styles and surroundings. The implementation on an NVIDIA Tesla P100 GPU increases computational efficiency, making it appropriate for real-time applications.

VI. CONCLUSIONS

This study presents a multimodal deep learning architecture that substantially enhances Bangla Sign Language (BdSL)

recognition by integrating spatial information from CNNs with landmark-driven geometric features. The integration of these two modalities allows the model to attain great accuracy and robustness, addressing issues such as differences in signing styles, lighting situations, and backgrounds. Our model improves existing approaches by including rigorous data preparation, real-time augmentation, and feature fusion, resulting in greater generalization and computing efficiency.

The experimental findings on the BdSL47 dataset show that the proposed method performs exceptionally well, with a classification accuracy of 99.96%, exceeding traditional machine learning models and earlier deep learning techniques. The use of hyperparameter modification, 10-fold cross-validation, and dropout regularization enhances the model's stability and effectiveness. Given its high accuracy and efficiency, the proposed framework holds significant promise for real-time applications in assistive communication technology, allowing designers to create more inclusive solutions for the BdSL community. Future work will focus on extending the model to recognize dynamic gestures and enabling real-time deployment to enhance accessibility.

REFERENCES

[1] K. H. Tarafder, N. Akhtar, M. M. Zaman, M. A. Rasel, M. R. Bhuiyan, and P. G. Datta, "Disabling hearing impairment in the bangladeshi population," *The Journal of Laryngology 38; Otology*, vol. 129, no. 2, p. 126–135, 2015.

[2] B. Sarkar, K. Datta, C. Datta, D. Sarkar, S. J. Dutta, I. D. Roy, A. Paul, J. U. Molla, and A. Paul, "A translator for bangla text to sign language," in *2009 Annual IEEE India Conference*. IEEE, 2009, pp. 1–4.

[3] D. Koller, "A survey of sign language recognition techniques," *arXiv preprint arXiv:2008.09918*, 2020.

[4] S. Hossain, D. Sarma, T. Mittra, M. N. Alam, I. Saha, and F. T. Johora, "Bengali hand sign gestures recognition using convolutional neural network," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, pp. 636–641.

[5] H. S. Shahgir, K. S. Sayeed, M. T. Tahmid, T. A. Zaman, and M. Z. U. Alam, "Connecting the dots: Leveraging spatio-temporal graph neural networks for accurate bangla sign language recognition," *arXiv preprint arXiv:2401.12210*, 2024.

[6] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Systems with Applications*, vol. 213, p. 118914, 2023.

[7] M. Hadiuzzaman, M. S. Ali, T. Sultana, A. R. Shafi, A. S. M. Miah, and J. Shin, "Baust lipi: A bdsl dataset with deep learning based bangla sign language recognition," *arXiv preprint arXiv:2408.10518*, 2024.

[8] S. Rayeed, S. T. Tuba, H. Mahmud, M. H. U. M. Md, S. H. M. Md, and K. H. Md, "Bdsl47: A complete depth-based bangla sign alphabet and digit dataset," *Data in Brief*, vol. 51, p. 109799, 2023.

[9] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "Bensignnet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network," *Applied Sciences*, vol. 12, no. 8, p. 3933, 2022.

[10] T. Abedin, K. Prottoy, A. Moshruba, and S. Hakim, "Bangla sign language recognition using concatenated bdsl network. arxiv 2021," *arXiv preprint arXiv:2107.11818*.

[11] N. Hassan, "Bangla sign language gesture recognition system: Using cnn model," *ScienceOpen Preprints*, 2022.

[12] R. A. Nihal, S. Rahman, N. M. Broti, and S. A. Deowan, "Bangla sign alphabet recognition with zero-shot and transfer learning," *Pattern Recognition Letters*, vol. 150, pp. 84–93, 2021.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.