Sign3DNet: An Enhanced 3D CNN Architecture for Bengali Word-Level Sign Language Recognition

Safi Ullah Chowdhury, Nasima Begum*, Tanjina Helaly, Rashik Rahman
Department of Computer Science and Engineering,
University of Asia Pacific, Dhaka, Bangladesh

***Abstract*—Automated recognition of sign languages has been playing an important role in breaking barriers to communication and inclusion for the deaf and mute community. Several studies have been conducted on Bengali Sign Language (BdSL). However, Bengali Word-Level Sign Language (BdWLSL) remains unexplored due to the lack of large annotated datasets and a stable model. Therefore, in this research, we introduced a large-scale Bengali word-level video dataset and proposed a modified 3D Convolutional Neural Network (CNN) architecture for word-level BdSL recognition, emphasizing its ability to capture the spatial and temporal dynamics from video data. The proposed strategy represents strong performance in Bengali word-level sign language recognition by utilizing the spatio-temporal pattern captured by the modified 3D CNN architecture. The proposed model demonstrates its potential for practical use by successfully learning complex hand movements straight from raw video data. The proposed CNN model is benchmarked against traditional deep learning techniques, Temporal Shift Module (TSM), Long Short-Term Memory (LSTM), and default 3D-CNN, providing a comprehensive comparison of their strengths and limitations. Experiments are conducted using a structured video dataset containing 102 Bengali sign-word classes. To ensure privacy, the volunteers' faces were blurred and only landmark data extracted using MediaPipe, rendered on black backgrounds, were used for training. The experimental result analysis shows that the performance of the proposed 3D-CNN model achieves a satisfactory accuracy of 58.25%, demonstrating its potential for word-level sign language recognition tasks. To our knowledge, this is the very first pilot study for BdWLSL recognition. Hence, we consider the recognition rate 58.25% of the proposed modified 3D-CNN architecture to be satisfactory and a potential scope for future researchers in the same field.**

***Keywords*—Bengali sign word recognition; computer vision; deep learning; convolutional neural network; spatial-temporal dynamics; video data**

I. Introduction

In a world where effective communication is at the heart of human connection, the deaf and mute community faces unique and formidable challenges. The limited accessibility of their distinct sign language has led to profound difficulties in expressing thoughts, feelings, and ideas. Bridging this communication gap is fundamental. Automatic interpretation of sign languages is difficult compared to spoken languages since it consists of signed motions, as well as hand gestures and facial expressions. Thus, although word-level sign language recognition is a well-established field for sentence-level recognition, it presents several difficulties specifically in languages different from ASL (American Sign Language) and BSL (British Sign Language). For example, there is limited awareness of sign languages like Bengali Sign Language (BdSL); the lack of a substantial volume of high-quality annotated datasets and strong models to capture both local and global hand movements contributes to this research.

Previous works ASL [1] have emphasized the requirement for large word-level data sets coupled with enhanced deep learning (DL) models to handle the inherent spatio-temporal characteristics in sign recognition. With advancements in computer vision and deep learning, specially 3D Convolutional Neural Network (CNNs), it is now possible to capture both spatial and temporal dynamics in sign laguage data, which helps to recognize complex gestures across video frames. However, 3D CNNs require extensive datasets and multidimensional data representations, such as 3D joint positions, which are not always possible for sign languages, including BdSL. To address this issue, this research proposes a new architecture for word-level sign language recognition. It is trained and tested on a structured dataset with 102 sign classes, prepared using MediaPipe landmarks to focus only on key gesture components.

Additionally, we compared the performance of three distinct models which are: (1) Modified 3D CNN that directly incorporates entire video frames to efficiently capture both spatial and temporal features; (2) an LSTM model that utilizes sequential data to recognize temporal patterns within sign gestures; and (3) TSM, efficient in capturing the temporal dependencies through frame feature shifting for sign word classification. Each model comes with some strengths and weaknesses when used in the analysis of sign language data outcomes of each model are measured based on the accuracy, precision, recall, F1-score and mAP (Mean Average Precision).

The main contributions of this research are as follows:

- We have developed a large-scale, annotated BdSL sign word-level and sentence-level video dataset for deep learning (DL) applications.
- We developed a modified robust 3D-CNN architecture after adjusting hyperparameters called Sign3DNet to increase the generalization property of the system. To our knowledge, the proposed Sign3DNet is the first pilot architecture for recognizing Bengali Sign Words.
- Finally, we evaluated the performance of the proposed Sign3DNet architecture for our developed dataset, and other existing datasets.
- We also demonstrated the standardization of our developed dataset across three model types (proposed

*Corresponding authors.

modified 3D CNN, LSTM, and TSM), providing an extensive evaluation.

The remainder of this paper is structured as follows: Section II represents the literature review. The dataset description is provided in Section III. The proposed methodology is described in Section IV. Section V demonstrates the analysis of the experimental results. Lastly, Section VI concludes the paper with some future works.

II. Literature Review

Sign language recognition has been vital in the reduction of communicative barriers for the Deaf and hard of hearing people, especially with word and sentence recognition. The recognition of sign language can be widely categorized into two types, namely the static and dynamic gesture recognition. Static gestural phrases are used to depict alphabets and digits, while dynamic gestural phrases can depict words or sentences at one go and hence need sequential recognition of gestures. Sign languages like American Sign Language (ASL) [1] have well-documented datasets, such as Purdue RVL-SLLL ASL Database [2], Boston ASLLVD [3], and RWTH-BOSTON-50 [4], which provide valuable data for various recognition tasks. However, the involved datasets fail to sufficiently cover different signers and instances, which becomes a drawback in terms of large-scale classification for achieving higher vocabulary. Large datasets for other languages like DEVISIGN [5] for Chinese Sign Language have been developed to address similar limitations by including thousands of signs performed by multiple signers, contributing to more robust recognition frameworks.

Existing sign language recognition approaches are typically centered around three phases, such as feature extraction, temporal modeling, and classification. Classical approaches used some hand-made features including motion/sensor-based features such as SIFT [6], HOG [7] and/or adopting frequency domain [8] to represent the hand poses and then using HMMs [9] or DTW [10] to model temporal components. In recent years, CNNs and 3D CNNs have been used widely to solve spatial and temporal data at the same time. However, there is still an issue associated with the application of the mentioned methods – the limited size and quality of the databases. Thus, reliable assessment and, particularly, transfer of models to new sign languages continues to be problematic, particularly for low-resource sign languages without large annotated datasets.

For Bengali Sign Language (BdSL) research is still in its infancy and mainly involves only static gesture recognition. First data sets are given with limited alphabets and digits of Bangla such as Ishara-Lipi [11] and Ishara-Bochon [12]. Following investigations used deep learning-based image recognition approaches like YOLO, CNN, and AlexNet [13], [14] on static BdSL data to categorize isolated words. Relatively little has been done in the way of developing dynamic word-level BdSL (WL-BdSL) datasets and recognition. Previous studies of dynamic BdSL are limited to a small set of static signs and do not reflect continuous word-level recognition issues. In ASL, the current state of static BdSL recognition with large vocabularies has recently used MediaPipe keypoint extraction, but dynamic video recognition for wider vocabularies has not been studied thoroughly.

In the case of Bengali Sign Language (BdSL), studies are still very limited, and early studies' main concern is on recognition of static gestures only. There are few datasets for letter-level and static word recognition in Bangla such as BornoNet [15], BDSL49 [16], Ishara-Lipi [11], and BenSignNet [17]. More specifically, it is recognized that Ishara-Lipi is a benchmark dataset for sign language that is described by the use of modern methods involving two-handed signs. This means that while BornoNet is particularly focused on the letter-level classification the dataset BDSL49 offers a large amount of Bangla characters. BenSignNet proposed a new CNN model that is trained on three different sets of BdSL, such as BdSL, Ishara-Lipi, and KU-BDSL [18] with differing performances.

However, efforts toward dynamic, word-level BdSL (WL-BdSL) recognition remain limited. Most of the work done in this area pertains to very small, arty static sign sets that do not suffice for the sort of word-level continuous reading tasks with larger, open vocabularies. However, despite recent attempts to adapt MediaPipe for static BdSL keypoint extraction, there is a lack of study regarding its application for dynamic video recognition. This gap points to an urgent need to work on the advances in the comprehensive ample of datasets and the state of the art of the recognition models addressing the issues of temporal and spatial dynamics in the continuous BdSL. To overcome these imperatives, we have developed a novel Bengali Sign Language dataset for dynamic word-level recognition, consisting of 102 emergent Bengali Sign words and including more than 30600 samples captured from various viewing angles.

III. Dataset Description

Bengali language has about one million words, and mute and deaf people use those sign words for their daily communication. Therefore, a video dataset is essential for Bengali Word Level Sign Language (BdWLSL) recognition. However, to our knowledge, there are no such video datasets for Bengali Sign Word except [19]. Thus, in this research, we developed a structured BdWLSL video dataset for our recognition task. The data set is used to support advanced deep learning applications for continuous sign language recognition. This dataset also solves the problem of limited large-scale BdWLSL dataset by providing a wide variety of video material to capture the details of BdWLSL gestures with 102-word classes. Each label in Table I represents a Bengali word. The data set includes various video samples that capture variability in signers, lighting, background, and angles as shown in Fig. 1. All videos used in training were collected with proper consent. For privacy, we applied automated face blurring to each frame. We then extracted hand and facial landmarks using MediaPipe. These landmarks were projected over a black background, and only these versions were used for model training. This ensured that the model learned from gestures, not identity or background noise.

Our developed dataset has different variations. For example, videos are captured by different signers at different angles and various lighting conditions. After recording all the videos, we used MediaPipe to generate landmarks for the hand and face regions. At the same time, we removed the raw versions of the videos and used a black background with the landmarks to construct the data. Fig. 2 shows the approaches which

TABLE. I. BENGALI WORD-LEVEL SIGN LANGUAGE (BDWLSL) CLASSES

Label	Word	Label	Word	Label	Word	Label	Word
0	পরিবার (Family)	26	কফি (Coffee)	52	টেলিভিশন (Television)	78	পেশা (Occupation)
1	সম্পর্ক (Relation)	27	রুটি (Bread)	53	ক্রিকেট (Cricket)	79	গৃহিণী (Housewife)
2	বাবা (Father)	28	সবজি (Vegetable)	54	ফুটবল (Football)	80	স্যার (Sir)
3	মা (Mother)	29	ডাল (Dal)	55	হাঁটা (Walking)	81	ম্যাডাম (Madam)
4	ভাই (Brother)	30	মিষ্টি (Sweet)	56	কথা বলা (Talking)	82	ডাক্তার (Doctor)
5	বোন (Sister)	31	পানি (Water)	57	ঘুমানো (Sleeping)	83	শ্রমিক (Labour)
6	স্বামী (Husband)	32	দুধ (Milk)	58	আসা (Coming)	84	উকিল (Lawyer)
7	স্ত্রী (Wife)	33	আম (Mango)	59	যাওয়া (Going)	85	পুলিশ (Police)
8	চাচা (Paternal Uncle)	34	কাঁঠাল (Jackfruit)	60	দেখা (Looking)	86	সাংবাদিক (Journalist)
9	মামা (Maternal Uncle)	35	আপেল (Apple)	61	বিভাগ (Division)	87	প্রকৌশলী (Engineer)
10	খালা (Maternal Aunt)	36	কমলা (Orange)	62	ঢাকা (Dhaka)	88	বাংলা (Bangla)
11	ফুফু (Paternal Aunt)	37	কলা (Banana)	63	চট্টগ্রাম (Chattogram)	89	ইশারা (Sign)
12	বন্ধু (Friend)	38	আঙ্গুর (Grape)	64	সিলেট (Sylhet)	90	ভাষা (Language)
13	কালো (Black)	39	পেপে (Papaya)	65	খুলনা (Khulna)	91	আমি (Me)
14	সাদা (White)	40	স্কুল (School)	66	রজশাহী (Rajshahi)	92	তুমি/আপনি (You)
15	লাল (Red)	41	কলেজ (College)	67	রংপুর (Rangpur)	93	আমার (My)
16	নীল (Blue)	42	বিশ্ববিদ্যালয় (University)	68	ময়মনসিংহ (Maymenshing)	94	আমাদের (Our)
17	হলুদ (Yellow)	43	অফিস (Office)	69	দিন (Day)	95	হাজার (Thousand)
18	সবুজ (Green)	44	কাজ (Work)	70	শনিবার (Saturday)	96	লাখ (Million)
19	খাদ্য (Food)	45	কম্পিউটার (Computer)	71	রবিবার (Sunday)	97	ছেলে (Boy)
20	ফলমূল (Fruits)	46	মোবাইল (Mobile)	72	সোমবার (Monday)	98	মেয়ে (Girl)
21	ভাত (Rice)	47	ইন্টারনেট (Internet)	73	মঙ্গলবার (Tuesday)	99	রোজা (Fasting)
22	মাছ (Fish)	48	বই (Book)	74	বুধবার (Wednesday)	100	ঈদ (Eid)
23	মাংস (Meat)	49	কলম (Pen)	75	বৃহস্পতিবার (Thursday)	101	কোরবানি ঈদ (Eid-Al-Adha)
24	মুরগি (Chicken)	50	বাতি (Light)	76	শুক্রবার (Friday)		
25	চা (Tea)	51	পাখা (Fan)	77	পেশা (Occupation)		

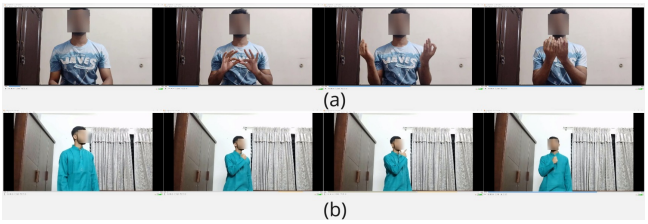


Fig. 1. Sample of Bengali word-level video data (a: পরিবার (Family); b: বাবা (Father)).

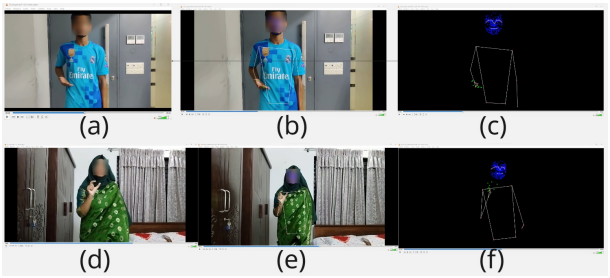


Fig. 2. Sample word data directory (a-c: বাংলা (Bangla), d-f: ভাষা (Language)).

is used to understand the condition of the dataset in various scenarios. To maintain volunteer privacy, all facial features were blurred during preprocessing using automated detection methods. For training, only the background-removed landmark videos were used. The dataset includes signers from diverse backgrounds to improve inclusivity and generalization performance. Table II summarizes the attributes of our developed dataset, providing a detailed breakdown of its structure and contents. In total, the dataset expands to 91,800 samples.

As illustrated in Fig. 3, each frame used for training was processed by first applying face blurring for privacy, followed by landmark extraction using MediaPipe. The final training samples consisted of black-background videos containing only the key landmark features of the sign gestures. Each word-level sign sample is labeled with the appropriate Bengali words. In each class, there are approximately 300 videos and they are divided into three parts: the first part is for training, the second one is for validation and the last part is for testing. The train, validation, and testing datasets contain 75%, 15%, and 15% of the total dataset, respectively. Table III shows the data distribution between the train , validataion and test set of the BdWLSL dataset. As shown in the Table III , there are around 210 video samples in the train set and 30 samples in the validation and test set of each class. All video frames were resized to 64x64 using OpenCV's bilinear interpolation method. This dimension was chosen based on a balance between visual fidelity of hand gestures and computational

TABLE. II. ATTRIBUTES FOR PROPOSED DATASET

Attribute	Value
No. of cameras	20+
Camera position	Multiple Viewing Angle (front, left, right)
No. of signers	39
No. of sign words	102
No. of word videos/class	300
Types of video	3 (Raw, Landmark with Background, Landmark without Background)
Total No. of Videos (Word)	91,800
File Type	mp4
Length/video length (sec)	1-2

efficiency. Larger resolutions led to increased training cost with negligible improvement in performance.

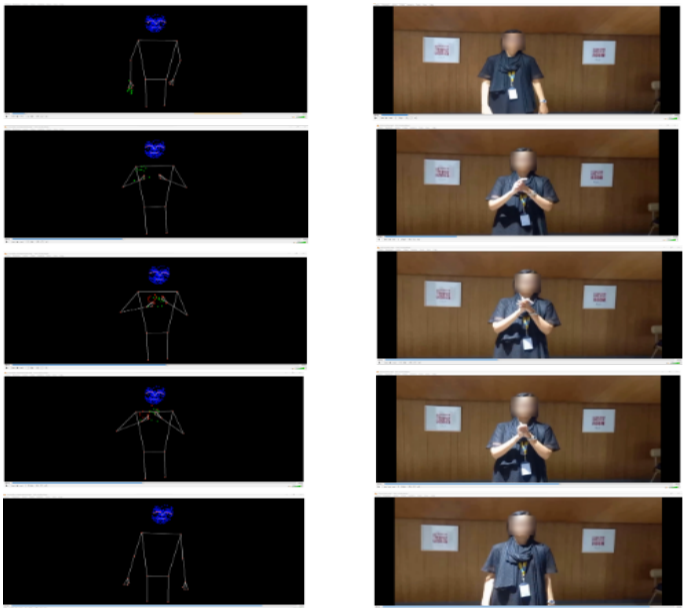


Fig. 3. Sample of the training data showing (left) the blurred original frame and (right) the extracted landmark frame on a black background for the sign "বন্ধু" (Friend).

TABLE. III. CLASS WISE DATA DISTRIBUTION

Label	Train	Test	Validation
0	210	30	30
1	210	30	30
2	210	30	30
3	210	30	30
...	...	...	...
101	210	30	30

IV. METHODOLOGY

This section describes the proposed methodology for word-level video sign language recognition. In this work, we introduce a novel CNN model specifically designed for word-level video sign language recognition, which is described in detail in Section IV(A). Fig. 4 shows the workflow of the proposed methodology.

To establish a baseline, we trained a few deep learning models on our developed dataset, including Standard 3D CNN, TSM, and LSTM. Among them, the LSTM model performed very well, achieving a moderate accuracy. However, the LSTM model is primarily suited for temporal data, whereas video data inherently possesses both spatio-temporal features. The TSM model is concentrated on the temporal dependencies through the shifting between feature maps in frames, losing its spatial information in video frames. On the other hand, the standard 3D CNN is designed to handle both spatio-temporal patterns directly. Therefore, to maximize performance for this application, we concentrated on modifying the 3D CNN architecture.

A. Proposed Modified 3D CNN Architecture

The proposed 3D CNN model is optimized for spatio-temporal feature extraction while balancing computational efficiency. The architecture employs the Depthwise Convolution and BatchNormalization techniques, along with the ReLU activation function, at multiple locations, as illustrated in Fig. 5. The model is fed with input videos, each one with 20 frames at 64×64 px grayscale. Convolution is done through the three Conv3D layers initially, extracting spatial and temporal patterns over spatial dimensions (x, y) and temporal (t) using the 3D kernels. Mathematically, the 3D convolution operation at the position (x, y, t) is defined as follows:

$$F(x,y,t) = \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} W(i,j,k)\cdot I(x+i,y+j,t+k) \quad (1)$$

Where $W(i, j, k)$ represents the 3D convolution kernel weights, $I(x + i, y + j, t + k)$ denotes the input video frame at spatial position (x, y) where t, is time and K is the kernel size. To stabilize training and accelerate convergence, batch normalization is applied after each convolutional layer, normalizing activations by Eq. (2)-(3) as follows:

$$\hat{x} = \frac{x - \mu}{\sigma} \quad (2)$$

$$y = \gamma\hat{x} + \beta \quad (3)$$

Here, the learnable parameters γ, β, μ and σ represent the mean and standard deviation of the batch. The MaxPooling3D layers are used to downsample the overall dimension of the model for efficient computation. MaxPooling3D was selected for its capability to retain the most prominent features while reducing computational load. Compared to Average pooling and L2 pooling, MaxPooling showed higher validation accuracy in early experiments on this dataset. The last one is a Global Average Pooling (GAP) layer to reduce input spatial dimensions by averaging feature maps to generate a fixed-length feature vector. The feature relationships within this vector are then captured by fully connected layers with ReLU activation. Finally, classification is performed using a

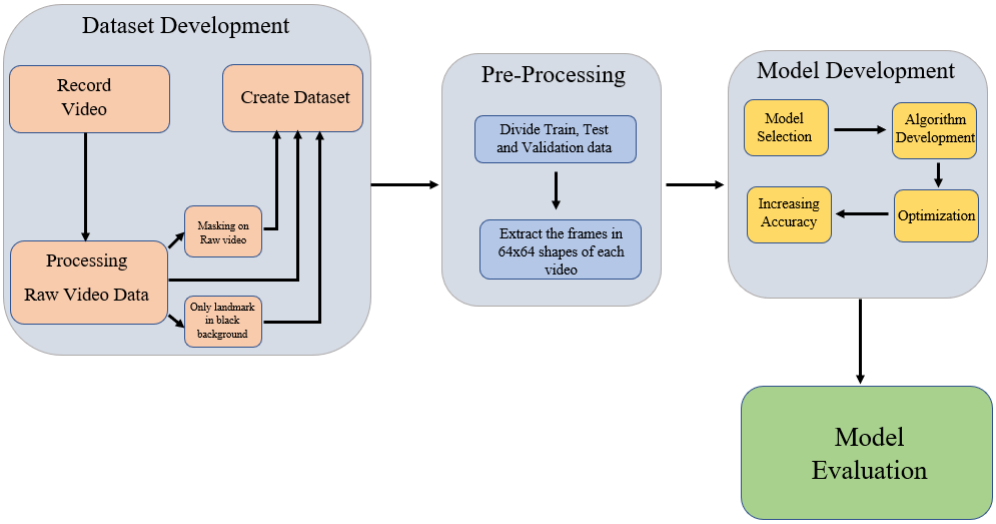


Fig. 4. Overall flow-diagram of proposed methodology.

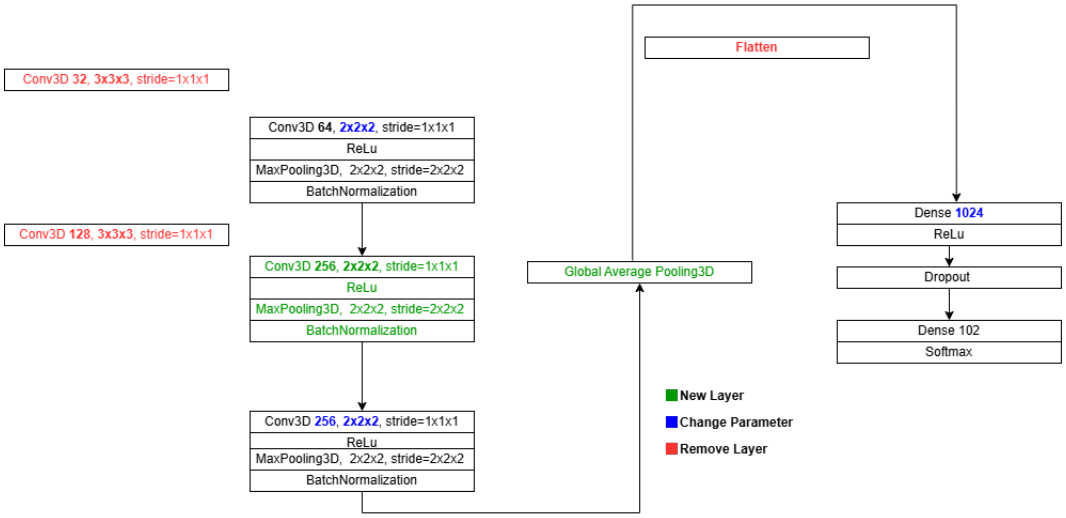


Fig. 5. Architectural modification of 3D CNN.

dense layer with a softmax activation function, converting the network output into class probabilities:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{4}$$

where $P(y_i)$ is the probability of class i, z_i is the output score for class i, and N represents the number of classes. Softmax activation function was used in the final layer to output a probability distribution over the 102 sign classes. This allows for intuitive interpretation of prediction confidence and is suitable for multi-class classification tasks. This allows for intuitive interpretation of prediction confidence and is suitable for multi-class classification tasks.

The network is trained using the categorical cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

Specifically, if y_i denotes the ground truth label (and 1 if the class is the correct class, and 0 otherwise) and $\hat{y}_i$ is the predicted probability. Model parameters are updated using the Adam optimizer with adaptive learning rates for improved convergence. The layers and their parameters are described in Table IV.

A 13-layer modified 3D CNN architecture is proposed where the computational cost is optimized for Bengali Word-level Sign Language (BdWSL) recognition using computational efficiency and strong recognition capabilities together. The proposed model greatly improves the classification performance by integrating Conv3D, batch normalization, pooling,

TABLE. IV. MODIFIED 3D CNN MODEL LAYERS AND THEIR PARAMETERS

Level	Name of the Layer	Parameters	Number of Layers
1	Input3D	-	1
2	Conv3D	Kernel size: 3x3x3 Filters: 64, 256, 256 Activation: Leaky ReLU	3
3	MaxPooling3D	Pool size: 2x2x2	3
4	BatchNormalization	-	3
5	GlobalAveragePooling3D	-	1
6	Dense	Units: 1024, Number of Classes Activation: ReLU, Softmax	2
Total Layers			**13**

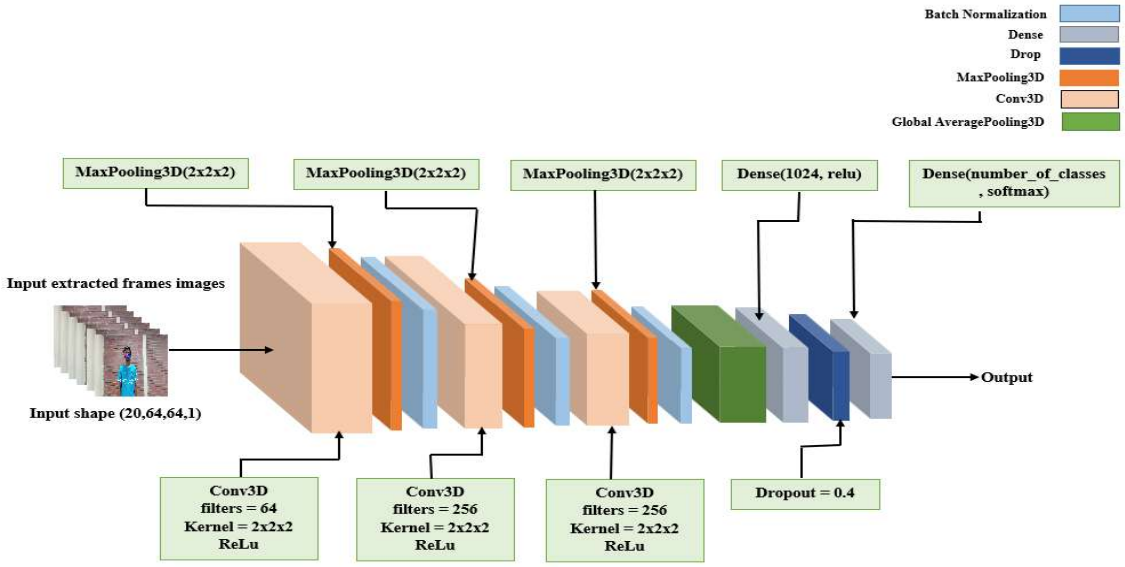


Fig. 6. Modified 3D CNN architecture.

and GAP, which allows the model to capture and process spatio-temporal features. An overview of the proposed 3D CNN architecture is illustrated in Fig. 6.

B. Experimental Setup

The 3D CNN model is deployed with modifications to effectively capture the temporal and spatial dynamics and details of the sign gestures. The model is optimized using the Adam optimizer. In 3D CNN, a sequential Conv3D architecture is used during the study, and involves down-sampling layers, batch normalization, and global pooling for feature extraction for 3D CNN. The model is trained using the Adam optimizer with a learning rate of 0.002. The loss function categorical cross-entropy is used, which is suitable for multiclass classification tasks.

For consistency across the model, the video frames are normalized with an equivalent diagonal bounding box of 64 pixels in standardized subsets of frames that focus on the signer's areas of interest, where applicable. All frames for a video are prepared by applying spatial-temporal augmentation. The complete setup of the experimental environment is summarized in Table V.

TABLE. V. EXPERIMENTAL ENVIRONMENT SETUP

Component	Specification
Operating System	Windows 11
GPU	NVIDIA RTX 4090 (24GB VRAM)
CPU	Intel Core i9-12900K
RAM	32GB DDR4
Framework	TensorFlow 2.10
CUDA Version	11.6
cuDNN Version	8.4
Programming Language	Python 3.9
Libraries	NumPy 1.23.5, OpenCV 4.6, Matplotlib 3.5

Algorithm 1 Sign-3D: An Optimized 3D Convolutional Network for Word-Level Sign Recognition

Input: Video sequence V
Output: Predicted class label C
Function Preprocess_Video(V):
    Resize each frame in V to 64×64
    Convert frames to grayscale
    Normalize pixel values
    return Processed video V'
$V' \leftarrow$ Preprocess_Video(V)
Function Extract_Features(V'):
    Apply 3D Convolution to extract spatio-temporal features
    Apply Batch Normalization for stabilization
    Apply Max Pooling for dimensionality reduction
    return Feature map F
$F \leftarrow$ Extract_Features(V')
Function Classify(F):
    Apply Global Average Pooling to reduce feature dimensions
    Flatten the output
    Pass through Fully Connected Layers with ReLU activation
    Apply Softmax activation for final classification
    return Predicted class C
$C \leftarrow$ Classify(F)
Return Predicted Class C

V. Result and Discussion

In this section, the experimental results and model evaluations are quantitatively analyzed. We have compared the proposed modified 3D CNN model with various other models previously used for sign language recognition.

A. Evaluation Metrics

We evaluated the performance of the models from the point of view of their ability to recognize signs concerning situational change in the sign language at the word level, where adjusting the position or the size of a hand may results in misclassification of a sign. Evaluation metrics are determined using the following Eq. (6)-(11):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F1\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{11}$$

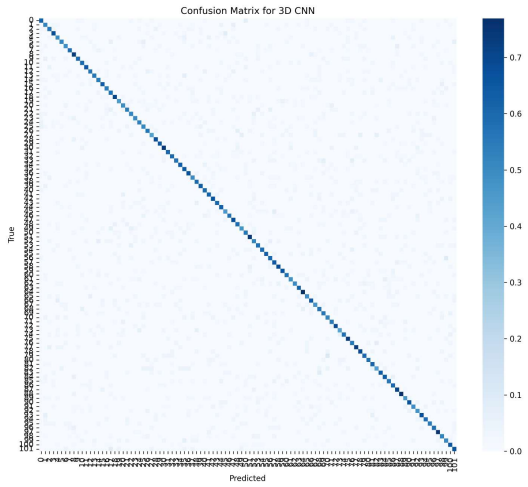
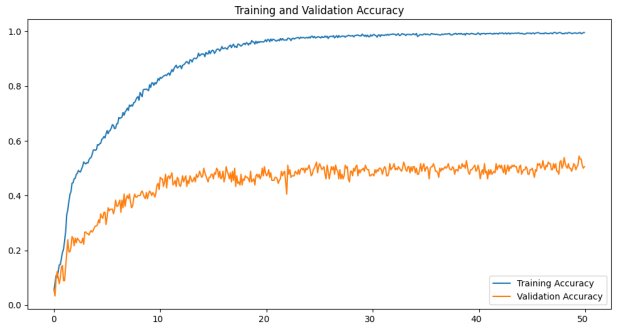
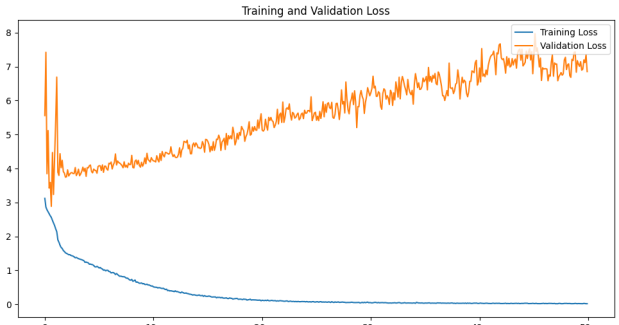


Fig. 7. Confusion matrix of modified 3D CNN model.



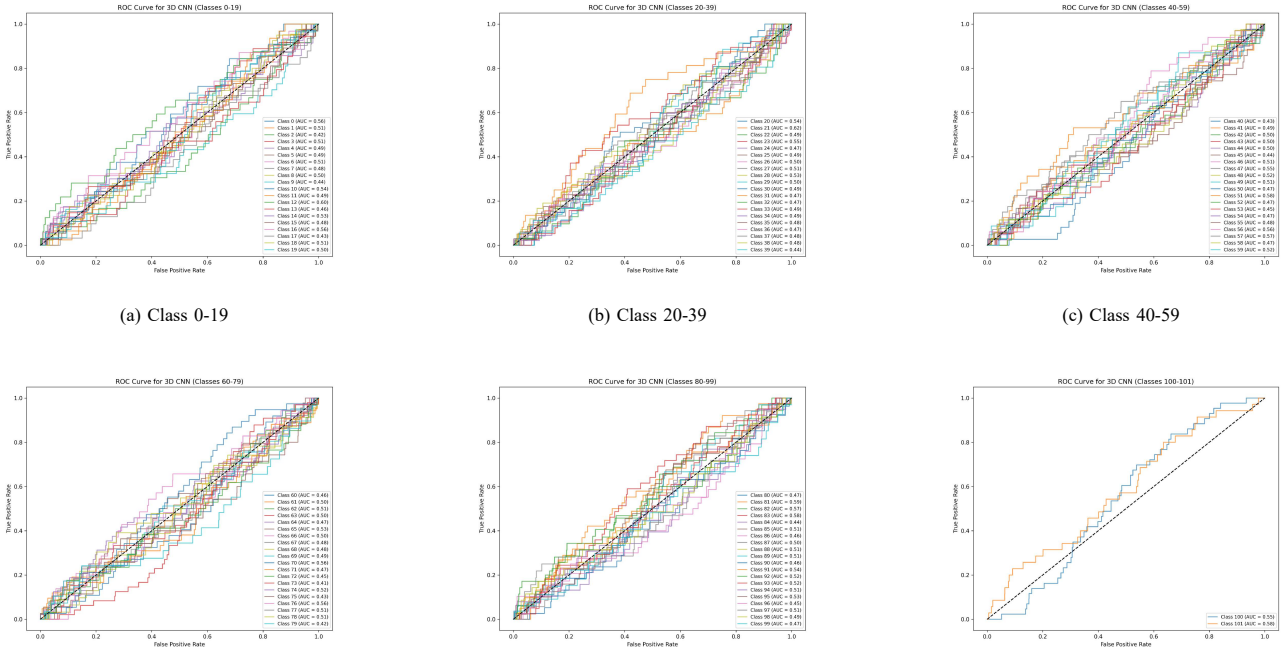
(a) Training and validation accuracy.



(b) Training and validation loss.

Fig. 8. Modified 3D CNN model accuracy and loss graph.

Furthermore, based on the classification of subtypes, the ROC-AUC curve for each class is obtained to assess the performance of the proposed modified 3D CNN model. Using the criterion for the 3D CNN model, the confusion matrix



(a) Class 0-19 (b) Class 20-39 (c) Class 40-59

(d) Class 60-79 (e) Class 80-99 (f) Class 100-101

Fig. 9. ROC and AUC curve of modified 3D CNN model.

TABLE. VI. Performance Comparison of Proposed 3D CNN Model with State-of-art Models

Model	Accuracy	Precision	Recall	F1 Score	Specificity	mAP	Batch Size
LSTM	54.83%	55.53%	54.83%	54.82%	60.12%	43.57%	16
3D CNN (Default)	52.83%	53.72%	52.83%	52.83%	58.37%	45.17%	16
TSM	51.75%	52.60%	51.75%	51.74%	57.92%	44.38%	16
Modified 3D CNN (Proposed)	**58.25%**	**59.17%**	**58.25%**	**58.25%**	**62.50%**	**41.32%**	16

TABLE. VII. Performance Comparison of Proposed 3D CNN Model with Different Datasets

Dataset	No. of Classes	Accuracy	Precision	Recall	F1 Score	Specificity	mAP	Batch Size
BdSLW60 [19]	60	44.62	45.50	44.62	44.62	50.00	55.38	16
Our Dataset	**102**	**58.25**	**59.17**	**58.25**	**58.25**	**62.50**	**41.32**	**16**

is analyzed to better understand the nature of errors and the performance across different classes.

B. Performance Evaluation of Proposed 3D CNN Model

The performance of different baseline models on the video dataset is shown in Table VI. We compared the performance of the proposed modified 3D CNN along with the default 3D CNN, LSTM, and TSM architectures. The results demonstrate varying capabilities of the models in handling the spatio-temporal complexities of sign language recognition.

Among them, the Modified 3D CNN achieved the highest performance with an accuracy of 58.25%, precision of 59.17%, recall of 58.25%, specificity of 62.50%, and an F1-score of 58.25%, demonstrating its capability to differentiate between sign classes effectively. However, its mAP (41.32%) indicates scope for improvement in class-wise performance. Fig. 8a and Fig. 8b shows the training and validation accuracy curve and loss curve, respectively. The optimization is effective, as the training curve stabilizes after a few epochs. A low gap between the training and validation curve indicates reduced overfitting (better generalization capability) of our model. The LSTM model, designed for capturing temporal dependencies, attained an accuracy of 54.83% and F1-score of 54.82%, performing well in modeling sequential data, being limited by processing feature vectors instead of raw frames. Its specificity (60.12%) is slightly lower than the proposed modified 3D CNN, but it has a higher mAP (43.57%), showing a more balanced class-wise performance. The TSM model, despite being structured for temporal feature shifting, performed the lowest, with 51.75% accuracy, 51.74% F1-score, specificity of 57.92%, and mAP of 44.38%, suggesting its limitations in extracting effective spatial features. The default 3D CNN, while leveraging spatio-temporal learning, achieved an accuracy of

52.83% and F1-score of 52.83%, falling behind the modified version due to the lack of optimizations.

With the complexity of multi-class classification across 102 distinct sign classes, high inter-signer variability, and the presence of subtle gesture overlaps between similar signs, the proposed model's accuracy of 58.25% may seem low; however, it accurately captures the complexity of the dataset. The proposed 3D CNN method outperforms the existing baselines, such as LSTM, TSM, and default 3D CNN, thus the 58.25% accuracy marks a significant relative improvement. As the first pilot initiative for Bengali Word-Level Sign Language (BdWLSL) recognition, this result establishes an impactful baseline for future research and development in the field.

Fig. 9 represents the ROC and the AUC graph. Here, the ROC threshold value is plotted for each class, and then the AUC curve for each class is determined. For all the classes, AUC lies between 0.42 and 0.62, making it a very generalized model that can perform well for large-scale sign language recognition. We used the BdSLW60 [19] dataset to verify our proposed model's performance in order to further assess its efficiency. The comparison is displayed in Table VII. An optimized deep learning architecture combined with a well-structured dataset produces noticeably superior performance, as shown in Table VII. This highlights how crucial it is to have a solid, stable model and a well-organized dataset to prevent low performance.

C. Limitations and Challenges

Fig. 7 shows the confusion matrix, where it is clearly visible that the model struggled with differentiating signs across the full 102-class. However, the results also highlight the challenges of scaling the model to handle a larger number of classes effectively. Future research should focus on optimizing the model architecture to enhance scalability while preserving a high recognition rate.

VI. Conclusion and Future Work

To enhance the communication of the disabled community, a method for the recognition of Bengali word-level sign language (BdWLSL) is proposed. This paper represents a benchmark study on a proposed modified 3D CNN model using a newly developed large-scale video dataset of 102 BdSL words. The use of black-background landmark data enabled the model to focus effectively on gesture regions. The proposed method is the pilot research for BdWLSL recognition. The experimental result shows that the proposed 3D CNN model achieved the highest recognition rate of 58.25%, indicating better results than LSTM (54.83%) and TSM (51.75%). Among the compared models, the LSTM model shows effectiveness in capturing the temporal dependencies of the sign words. However, the lack of spatial feature extraction limited its overall performance. Though the TSM model is designed for efficient temporal feature shifting, its inability to fully model spatio-temporal relationships affected its effectiveness in word-level sign language recognition. These findings highlighted the distinct strengths and weaknesses of each approach. The experimental analysis demonstrates that it is crucial to utilize advanced architectures like 3D CNN for Bengali Word-level Sign Language recognition as it can catch both the spatio-temporal features at the same time. The 58.25% accuracy of the proposed model is admissible considering the complexity of multi-class recognition with 102 categories, signer variability, and subtle gesture overlaps. This performance sets a valuable benchmark for Bengali word-level sign recognition tasks. Future work can focus on enabling models to generalize across real-world scenarios. We aim to integrate our model into mobile platforms and educational applications to support the deaf and mute community. Real-time sign recognition tools could serve as effective communication aids. Future enhancements include expanding the dataset, incorporating transformer-based or hybrid architectures for improved temporal reasoning, and leveraging multimodal approaches such as combining video with skeletal data or attention mechanisms. These strategies could address challenges related to large class sizes and further improve scalability and recognition performance.

VII. Acknowledgment

We give special thanks to the University of Asia Pacific for supporting this research and thanks to the Institute of Energy, Environment, Research and Development (IEERD), University of Asia Pacific, for funding this research project.

References

[1] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[2] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 2002, pp. 167–172.

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The american sign language lexicon video dataset," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–8.

[4] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of tangent distance and an image distortion model for appearance-based sign language recognition," in *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27*. Springer, 2005, pp. 401–408.

[5] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena, "Devisign: dataset and evaluation for 3d sign language recognition," *Technical report, Beijing, Tech. Rep*, 2015.

[6] M. Oja, S. Kaski, and T. Kohonen, "Bibliography of self-organizing map (som) papers: 1998–2001 addendum," *Neural computing surveys*, vol. 3, no. 1, pp. 1–156, 2003.

[7] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205–2231, 2012.

[8] M. Al-Rousan, K. Assaleh, and A. Tala'a, "Video-based signer-independent arabic sign language recognition using hidden markov models," *Applied Soft Computing*, vol. 9, no. 3, pp. 990–999, 2009.

[9] T. Starner and M. Group, "Visual recognition of american sign language using hidden markov models," 05 1995.

[10] J. Lichtenauer, E. Hendriks, and M. Reinders, "Sign language recognition by combining statistical dtw and independent classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, pp. 2040–6, 12 2008.

[11] M. S. Islam, S. S. S. Mousumi, N. A. Jessan, A. S. A. Rabby, and S. A. Hossain, "Ishara-lipi: The first complete multipurposeopen access dataset of isolated characters for bangla sign language," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–4.

[12] M. S. Islam, S. Sultana Sharmin, N. Jessan, A. S. A. Rabby, S. Abujar, and S. Hossain, *Ishara-Bochon: The First Multipurpose Open Access Dataset for Bangla Sign Language Isolated Digits*, 07 2019, pp. 420–428.

[13] N. Begum, R. Rahman, N. Jahan, S. S. Khan, T. Helaly, A. Haque, and N. Khatun, "Borno-net: a real-time bengali sign-character detection and sentence generation system using quantized yolov4-tiny and lstms," *Applied Sciences*, vol. 13, no. 9, p. 5219, 2023.

[14] N. Begum, S. S. Khan, R. Rahman, A. Haque, N. Khatun, N. Jahan, and T. Helaly, "Qmx-bdsl49: An efficient recognition approach for bengali sign language with quantize modified xception," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.

[15] A. S. A. Rabby, S. Haque, S. Islam, S. Abujar, and S. A. Hossain, "Bornonet: Bangla handwritten characters recognition using convolutional neural network," *Procedia computer science*, vol. 143, pp. 528–535, 2018.

[16] S. S. Khan, A. Haque, N. Khatun, N. Begum, N. Jahan, and T. Helaly, "An evaluation of bdsl 49 dataset using transfer learning techniques: A review," in *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022*. Springer, 2023, pp. 437–447.

[17] 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.

[18] A. A. J. Jim, I. Rafi, M. Z. Akon, U. Biswas, and A.-A. Nahid, "Ku-bdsl: An open dataset for bengali sign language recognition," *Data in Brief*, vol. 51, p. 109797, 2023.

[19] H. A. Rubaiyeat, H. Mahmud, A. Habib, and M. K. Hasan, "Bd-slw60: A word-level bangla sign language dataset," *arXiv preprint arXiv:2402.08635*, 2024.