IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 January 2024, accepted 21 February 2024, date of publication 26 February 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370684

RESEARCH ARTICLE

MediSign: An Attention-Based CNN-BiLSTM Approach of Classifying Word Level Signs for Patient-Doctor Interaction in Hearing Impaired Community

MD. AMIMUL IHSAN[1], ABRAR FAIAZ ERAM[2], LUTFUN NAHAR[1], AND MUHAMMAD ABDUL KADIR[1], (Member, IEEE)

[1]Department of Biomedical Physics and Technology, University of Dhaka, Dhaka 1000, Bangladesh
[2]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

Corresponding author: Muhammad Abdul Kadir (kadir@du.ac.bd)

This work was supported by the International Science Program (ISP) of Uppsala University, Sweden.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Review Committee of the Department of Biomedical Physics and Technology, University of Dhaka.

ABSTRACT Along with day-to-day communication, receiving medical care is quite challenging for the hearing impaired and mute population, especially in developing countries where medical facilities are not as modernized as in the West. A word-level sign language interpretation system that is aimed toward detecting medically relevant signs can allow smooth communication between doctors and hearing impaired patients, ensuring seamless medical care. To that end, a dataset from twenty distinct signers of diverse backgrounds performing 30 frequently used words in patient-doctor interaction was created. The proposed system has been built employing MobileNetV2 in conjunction with an attention-based Bidirectional LSTM network to achieve robust classification, where the validation accuracy and f1- scores were 95.83% and 93%, respectively. Notably, the accuracy of the proposed model surpasses the recent word-level sign language classification method in a medical context by 5%. Furthermore, the comparison of evaluation metrics with contemporary word-level sign language recognition models in American, Arabic, and German Sign Language further affirmed the capability of the proposed architecture.

INDEX TERMS Attention, BiLSTM, MobileNetV2, patient-doctor interaction, sign language.

I. INTRODUCTION

Disabilities have crippled more than 15% population in the world, and a major portion of them have an auditory disability, rendering them unable to interpret sound and vibration. The number is pushing towards 70 million living and breathing human beings who cannot communicate using speech or hear the words persons without disability utter. In most cases, sign language serves as the hearing and speech-impaired population's sole means of communication. However, when people with speech and hearing impairments speak with those who do not understand sign language, this method of communication is ineffective.

Along with day-to-day communication, receiving medical care is quite challenging for the hearing impaired and mute population, especially in developing countries. A capable individual needs to accompany them when they go to medical centers and explain their symptoms to the medical professionals. The hearing impaired person who is facing health challenges may not be able to properly explain his symptoms if the human interpreter is not well-versed in the sign language or is uncooperative. Research has indicated that the presence of inadequate communication between healthcare providers and patients who are hearing impaired or hard of hearing (HOH) remains a significant obstacle in

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

the provision of patient-centered care and the establishment of trust. Despite legal requirements for American Sign Language (ASL) medical interpreters in the United States [1], many clinicians still rely on inadequate communication methods, such as writing notes or assuming patients can lip-read [2].

This can lead to misunderstandings, misdiagnoses, and inadequate treatment, which can have serious consequences for the patient's health and well-being. Automatic sign language interpretation in healthcare settings can help address this issue by facilitating a dependable mode of interaction between healthcare professionals and individuals have hearing impairments. Automatic sign language interpretation can be provided through various technologies such as video remote interpreting (VRI), where an interpreter is connected to the healthcare provider and patient via video conferencing [3], or through the use of artificial intelligence (AI) powered avatars that can interpret sign language in real-time [4]. These technologies can facilitate accurate and efficient communication between healthcare providers and patients, ensuring that patients receive the appropriate care and treatment they need. Additionally, automatic sign language interpretation can also help reduce the stigma associated with being hearing impaired, promoting inclusivity and equality in healthcare access.

Signs used in patient-doctor interactions consist of continuous gestures or finger-spelling in any sign language dialect. Therefore, an interpretation system made for such purposes will focus more on Dynamic Sign Language Recognition (DSLR). The rapid progress of deep learning theory [5] has led to 2 remarkable accomplishments in object detection [6] and gesture recognition [7], as seen by the impressive results obtained using data-driven approaches. The utilization of video sequences in sign language identification allows for the comprehensive utilization of temporal information, which is a vital component in the process of recognizing gestures. This approach differs from the method that relies solely on hand forms and the motion trajectory of hand gestures. Because hands are very small in comparison to the entire environment, backgrounds may overshadow the effective spatial characteristics of gestures. Learning spatiotemporal features concurrently will, therefore, be more crucial for dynamic sign language recognition.

Specific American Sign Language (ASL) or Bangladeshi Sign Language (BSL) datasets for signs related to medical communication were not found during this analysis; therefore, a dataset containing 3596 videos of 30 medical ASL signs used in patient-doctor interaction in different environments, performed by 20 sign language learners of different age, height, skin orientation was created. The size of the dataset pertains to literature based on various sign languages, for instance, [8], [9].

Therefore, the prominent contributions of this work are:

- Formation of a new dataset containing 3596 videos of 30 ASL signs used globally in everyday patient-doctor interactions.
- An attention-boosted deep learning method with a pre-trained CNN (MobileNetV2) model as a feature extractor and a bi-directional LSTM model to detect temporal information and classify the signs.
- A method that is familiar with multiple backgrounds, illumination, skin tone, and gender of subjects and is not affected by it without using segmentation algorithms.

The paper continues as follows. Section II reviews relevant literature, including datasets, models, and results. Section III describes the methodology, covering data preparation, model specifications, experiments, and evaluation metrics. Section IV presents results with comparative analyses and limitations. Finally, Section V concludes the study.

II. RELATED WORKS

Kim et al. [10] conducted an investigation on the matter of finger-spelling identification utilizing hand-shape features through deep neural networks. The investigation included 3684-word instances with accuracies of 92% and 83% for signer-dependent and multi-signer settings, respectively. Literature [11] and [12] also discussed the recognition of finger-spelling in sign language. With a dataset containing 4200 signs in total, Aly et al. [12] produced an average accuracy of 99.5%. Regarding the method based on hand shapes, it was able to convey the meaning of relatively straightforward hand gestures, such as the 31 alphabets and numbers; however, the system's functionality was limited by intricate motion gestures as a result of its inadequate incorporation of hand shape in conjunction with consistent hand movement. In contrast, certain researchers focused solely on examining the motion trajectories of particular hand movements. With long short-term memory, Mohandes et al. [13] only detected hand motions based on hand motion trajectory. The classification of hand motions was accomplished with the aid of sensor technology, including the jump motion controller [14], digital glove data, surface electromyography accelerometer, and gyroscope. In this work [14], using 432 signs, support vector machine (SVM) and deep neural network (DNN) recognition rates were 80.30% and 93.81%, respectively, for 26 letters in the experimental results. The rates for a combination of 26 letters and 10 digits were somewhat lower, at roughly 72.79% and 88.79%, respectively. However, these methods are limited to a specific set of manual motions, such as gesticulation and waving of the hand. The identification of the attributes of hand configurations and the trajectory of movement associated with each hand gesture serves as the foundation for numerous studies. Numerous relevant studies have been carried out with success. For instance, Kumar et al. [15] provided a multimodal framework for the recognition of separate sign language by testing on a dataset of 7500 Indian Sign Language (ISL) movements, encompassing single- and double-handed gestures, reaching overall accuracies of 97.85% and 94.55% for single and double-handed signs, respectively using Kinect [16] sensor devices. The system focuses on distinguishing between one-hand and dual-hand

signs by utilizing a track model. Using RGB-D data and a Sparse Observation description, Wang et al. [17] were able to discern sign language through the analysis of movements of the hands as well as postures. In their study, Savur and Sahin [18] employed Surface Electromyography (sEMG) signals in conjunction with a Support Vector Machine (SVM) classifier for the purpose of recognizing American sign language. The surface electromyography (EMG) signals are acquired by means of external sensors placed on the hand. However, these sensor-based systems have drawbacks in terms of user comfort and practicality.

An RCNN for continuous sign language recognition that is entirely reliant on video sequence, which achieved a Word Error Rate of 38.7%, was created by Cui et al. [19]. The method was evaluated on RWTH-PHOENIX-Weather multi-signer 2014, which is a publicly available benchmark dataset for continuous sign language recognition containing 6841 instances/sequences. A framework for a Hierarchical Attention Network (HAN) incorporating a latent space was put forth by Huang et al. [20] to facilitate the development of global and regional video representations of attributes. Two open-source continuous SLR datasets are used in these experiments, one for CSL and the other is the German sign language dataset RWTH-PHOENIX-Weather (Koller, Forster, and Ney 2015), each consisting of 25,000 and 6841 instances, respectively. The proposed model has an accuracy of 82.7%. Motion information was included in static photographs by Köpüklü et al. [21]. Tested using the Jester, ChaLearn LAP IsoGD, and NVIDIA Dynamic Hand Gesture Datasets, the model produces competitive classification accuracies of 96.28%, 57.4%, and state-of-the-art 84.7%, for a total of 148,092 gesture videos. Nevertheless, the use of this technology proved challenging on a universal scale due to the inability to convert certain hand motion information into a static image. In the study conducted by the authors in [22], they presented a novel spotting-recognition architecture designed specifically for the purpose of extensive, persistent gesture recognition. For a total of 22,535 video samples, the architecture yielded a Jaccard Index of 0.6103. The sequence-to-sequence method was utilized by Camgoz et al. [23] to learn the sign language recognition issues, producing a Word Error Rate of 43.1% in a database of 69,832 signs. The process of categorizing continuous sign language clips is enhanced by employing elements such as movement and shapes, which enable a more nuanced and precise classification. Kishore et al. [24] were able to understand continuous sign language. The dataset comprised 500 signs, and the model reached an average matching score of 92.5%. An artificial neural network classification method was employed to perform continuous sign language identification on a dataset of 180 selfie images, resulting in an average Word Matching Score of almost 90% by Gondu et al. [25]. Deep learning technology is predominantly employed in the implementation of dynamic sign language recognition methods, such as the methods described in [26] and [27] to employ convolutional neural networks (CNNs) for the purpose of retrieving selective features from hand gestures in a dataset of 1.4 million images [26] which resulted in top-1 and top-5 error rates of 37.5% and 17.0%, respectively. Donahue et al. [27] introduced their model through the use of the TACoS multilevel dataset, which has 44,762 videos. In [28], to learn video sequences using recurrent neural networks (RNNs), the model employs decoder LSTMs and encoder LSTMs for input sequence mapping on the UCF-101 and HMDB-51 datasets which respectively contain 13,320 and 5100 videos. For learning spatiotemporal sequential features integrating CNNs and RNNs, [29] makes use of the KTH dataset and obtains an accuracy of 94.39%. The KTH dataset is divided into two parts: KTH-1 containing 599 long sequences, where each contains multiple iterations of the same action, and KTH-2 containing 2391 sequences, where each sequence contains a single action. Furthermore, Yue-Hei Ng et al. [30] introduced multiple models throughout two distinct databases: UCF-101, containing 13,320 videos, and Sports 1 Million datasets consisting of 1.2 million YouTube sports videos. The highest recorded accuracy of 88.6% was achieved using the LSTM with 30 Frame Unroll (Optical Flow + Image Frames) method in the UCF-101 dataset. The aforementioned methods exhibited superior performance compared to the method that relied on hand shapes and motion trajectories in the context of dynamic sign language recognition utilizing a video sequence. Hu et al. [31] introduce a correlation module (CorrNet) to compute correlation maps between the current frame and adjacent frames, identifying trajectories of all spatial patches. The architecture is tested on four separate datasets: PHOENIX14 (6841 videos), PHOENIX14-T (8247 videos), CSL-Daily (20,654 videos), and CSL (25,000 videos). A novel attention-based approach called SLGTformer for Sign Language Recognition was proposed by Song and Xiang [32], which utilizes decoupled graph and temporal self-attention to learn the spatiotemporal dynamics of skeleton key points. It achieves Top-1 and Top-5 recognition rates of 47.42% and 79.58% on the World-Level American Sign Language (WLASL) dataset, which contains 21,083 samples. Furthermore, Li et al. [33] proposed a novel pose-based temporal graph convolution network (Pose-TGCN) that models spatial and temporal dependencies in human pose trajectories simultaneously. The approach is tested using the WLASL dataset and reaches 62.63% at top-10 accuracy. Using the PHOENIX14 and PHOENIX14-T datasets, Hao et al. [34] introduced a method called Self-Mutual Knowledge Distillation (SMKD) to enhance the discriminative power of both the visual and contextual modules in Continuous Sign Language Recognition (CSLR) by focusing on short-term and long-term information simultaneously. The visual and contextual modules share the weights of their corresponding classifiers and are trained with the Connectionist Temporal Classification (CTC) loss. A system for word-level sign language recognition using the Transformer model, with a focus on low

computational cost and potential usage on hand-held devices, was presented by Bohacek and Hruz [35]. With 63.18% and 43.78% recognition rates on the WLASL100 (2038 videos) and WLASL300 (5117 videos) datasets, respectively, and 100% test recognition accuracy on LSA64 (3200 videos), it achieves state-of-the-art performance on all the datasets. Kun Xia et al. [36] presented a MobileNet-YOLOv3- based model with a hand-held device for 12 medical signs through the use of a dataset comprising 4000 samples and attained an identification accuracy of 90.77%. Da Silva et al. [37] applied I3D and LSTM to train a dataset with 5000 videos for recognition of Brazilian sign language (Libras). the model was applied to datasets comprising Brazilian Sign Language (Libras) and Argentinian Sign Language (LSA), achieving high accuracy rates of 99.80% and 100%, respectively. A model consisting of a CNN and LSTM integrated model referred to as the Long-term Recurrent Convolutional Network (LRCN) model was applied by Das et al. [38] for Indian Sign Language recognition in a medical context from a dataset of 288 videos, yielding an accuracy of 67.53%. The overall related works are summarized in Table 1.

III. METHODOLOGY

The study's primary focus is the analysis of sign language words within patient-doctor interactions, achieved through the collection and processing of video data. A dataset comprising 30 distinct signs performed by 20 subjects was gathered, reflecting medical terminology used in healthcare settings. The videos are then preprocessed to contain a fixed number of frames in optimum dimensions, which are later used to train and evaluate the proposed deep-learning model.

A. DATASET OVERVIEW

The selection of signs used in the collection was carefully vetted to encompass a wide array of medical terminology and phrases that are essential inside healthcare environments. These 30 distinct signs encapsulate a diverse spectrum of medical concepts, from common ailments like "Cold" and "Headache" to critical terms such as "Emergency" and "Hospital." In order to comprehensively depict the intricate and multifaceted nature of sign language communication, the dataset incorporates contributions from a diverse group of 20 individuals. The deliberate inclusion of individuals from many genders, age groups, and skin tones in sign language communication aims to accurately reflect the wide range of variances observed in real-world contexts, and it also allows the model to learn the physiological differences of different signers and be robust to such changes. It is worth noting that the videos were recorded using personal smartphones and tripods, ensuring that the data collection process was practical and accessible while still upholding the necessary technical standards. Furthermore, videos were captured at different times of the day to account for potential variations in lighting conditions, thereby enhancing the dataset's robustness. Table 2 and 3 display the 10th frame of the signs under investigation in this work.

B. PREPROCESSING

Thirty frames were extracted from each video, ensuring that this subset effectively captured the essence of sign language expression. This is because the dataset comprises sign words that are gesture-based, not finger-spelled. As a result, even with the first 30 frames, the uniqueness of a given gesture in the dataset can be encapsulated. These frames were uniformly resized to a standardized dimension of 100×100 pixels, striking a balance between retaining essential information and managing computational resources.

The rationale behind this approach was twofold. Firstly, it ensured that we preserved the critical temporal aspects of sign language communication by maintaining a fixed sequence of frames from each video. Secondly, the resizing to a consistent dimension not only reduced the computational burden but also ensured uniformity across the dataset. This uniformity was crucial for training and deploying deep learning models, as it allowed us to focus on the essence of the signs without being overwhelmed by extraneous visual data.

C. PROPOSED MODEL

Fig.1 contains the end-to-end approach from frame extraction to classification, where a fixed number of frames is taken as the input. CNN backbone for this operation has been chosen to be MobileNetV2, and to capture the temporal pattern, Bidirectional LSTM layers backed by attention mechanism were used. Then, fully connected dense layers were used to classify the videos.

D. FEATURE EXTRACTION

During this phase, transfer learning with a pre-trained CNN, which is MobileNetV2 [36] in this case, is used to extract features from the video frames. MobileNetV2 can achieve similar accuracy with significantly lower resource demand and fewer number of layers (53). A study conducted by Podder et al. [39] reported superior accuracy results for MobileNetV2 compared to other pre-trained CNNs in deep learning applications. The extracted frames from the preprocessing pipeline are traversed through 53 layers of MobileNetV2 and fed into a BiLSTM system with an attention mechanism to extract temporal information.

E. ATTENTION-BOOSTED BIDIRECTIONAL LSTM

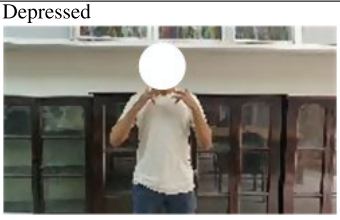
Conventional LSTMs, while adept at retaining past information, lack the ability to comprehend all the necessary context of time-based data, like videos, which necessitates the consideration of both future and previous information. This study adopted an LSTM configuration that is bidirectional to mitigate this limitation, enabling the extraction of information according to context from videos in both temporal directions.

The illustrated BiLSTM architecture in Figure 2 consists of two distinct LSTMs, a forward and a backward unit. These independently traverse the input sequence in opposite directions, each maintaining its own cell along with a hidden

TABLE 1. Synopsis of literature review.

Authors	Language	Dataset	Model	Results
Kim et al. [10]	American Sign Language (ASL)	Locally made dataset with 3684-word instances	Tandem Hidden Markov Model (HMM) with DNN classifiers	Accuracy 92% (signer-dependent setting) and 83% (multi-signer settings)
Kang et al. [11]	American Sign Language (ASL)	31,000 depth maps using a depth sensor, Creative Senz3D camera	CNN	Accuracy 99.99% (observed signers) and 83.58% to 85.49% (new signers)
Aly et al. [12]	Arabic Sign Language (ArSL)	Dataset using Softkinect sensor 4200 signs	PCANet	Accuracy 99.5%
Chong et al. [14]	American Sign Language (ASL)	432 signs using Leap Motion Controller (LMC)	SVM and DNN	Accuracy 72.79% (SVM) and 88.79% (DNN)
Kumar et al. [15]	Indian Sign Language (ISL)	7500 signs using Microsoft Kinect and LMC	Hidden Markov Model (HMM) and Bidirectional Long Short-Term Memory Neural Network (BLSTM-NN)	Accuracy 97.85% (single handed signs) and 94.55% (double handed signs)
Wang et al. [17]	Chinese Sign Language	Dataset A with 1850 signs and Dataset B with 3000 signs using Kinect sensor	Sparse Observation (SO) based model	Top 1, Top 5, and Top 10 values of 0.744, 0.891, and 0.926 respectively
Savur et al. [18]	American Sign Language (ASL)	sEMG signal collected using eight-channel Bio Radio 150 CleveMed device (2080 samples)	SVM	Accuracy 91% (offline system) and 82.3% (real-time system)
Cui et al. [19]	German Sign Language	RWTH-PHOENIX-Weather multi-signer 2014 containing 6841 instances	Recurrent Convolutional Neural Networks (RCNN)	Word Error Rate 38.7%
Huang et al. [20]	Chinese and German Sign Language	Locally made (25,000 signs) and RWTH-PHOENIX-Weather (6841 signs)	Hierarchical Attention Network with Latent Space (LS-HAN)	Accuracy 82.7%
Kopuklu et al. [21]	Distinct Gestures	Jester, ChaLearn LAP IsoGD, and NVIDIA Dynamic Hand Gesture Datasets (148,092 gesture videos)	CNN with Motion Fused Frames (MFF)	Accuracies 96.28%, 57.4%, and 84.7% respectively
Liu et al. [22]	Chinese Sign Language	ChaLearn LAP ConGD Dataset (22,535 video samples)	Two streams Faster R-CNN, C3D model, and a linear SVM	Jaccard Index 0.6103
Camgoz et al. [23]	Danish, New Zealand and German Sign Language	One-Million Hands dataset (69,832 signs)	SubUNets	Word Error Rate 43.1%
Kishore et al. [24]	American Sign Language (ASL)	Locally made dataset with 500 signs	Fuzzy Inference Engine (FIS)	Average matching score 92.5%
Gondu et al. [25]	Indian Sign Language (ISL)	180 selfie images	ANN	Word Matching Score nearly 90%
Krizhevsky et al. [26]	Distinctive Objects	ImageNet (1.4 million images)	CNN	Top 1 error rate 37.5%, and Top 5, error rate 17%
Donahue et al. [27]	Distinctive Human Actions	TACoS multilevel dataset (44,762 videos)	Recurrent Convolutional Neural Networks (RCNN) with LSTM	BLEU-4 score 28.8%
Srivastava et al. [28]	Distinctive Human Actions	UCF-101 dataset (13,320 videos) and HMDB-51 dataset (5100 videos)	RNN with LSTM	Accuracy 75.8% (UCF-101 dataset) and 44.1% (HMDB-51 dataset)
Baccouche et al. [29]	Distinctive Human Actions	KTH datasets: KTH-1 (599 sequences) and KTH-2 (2391 sequences)	3D-CNN and RNN	Accuracy 94.39%
Yue-Hei Ng et al. [30]	Distinctive Human Actions	UCF-101 dataset (13,320 videos) and Sports 1 million dataset (1.2 million videos)	CNN and LSTM	Highest Accuracy 88.6% (using LSTM)
Hu et al. [31]	German and Chinese Sign Language	PHOENIX14 (6841 videos), PHOENIX14-T (8247 videos), CSL-Daily (20,654 videos), and CSL (25,000 videos)	Correlation Network (CorrNet)	Word Error Rate (WER): 19.4% (PHOENIX14), 20.5% (PHOENIX14-T), 30.1% (CSL-Daily), and 0.8% (CSL)
Song et al. [32]	American Sign Language (ASL)	WLASL2000 (21,083 samples)	SLGTformer	Top-1 and Top-5 recognition rates 47.42% and 79.58%
Li et al. [33]	American Sign Language (ASL)	WLASL2000 (21,083 samples)	Holistic visual appearance-based approach, 2D human pose-based approach, and a novel pose-based temporal graph convolution network (Pose-TGCN)	Top-10 Accuracy 62.63%
Hao et al. [34]	German Sign Language	PHOENIX14 (6841 sentences), and PHOENIX14-T (8247 sentences)	Self-Mutual Knowledge Distillation (SMKD)	Word Error Rate (WER) 21% (), and 22.45% ().
Bohacek et al. [35]	American and Argentinian Sign Language	WLASL100 (2038 videos – ASL). WLASL300 (5117 videos – ASL), and LSA64 dataset (3200 videos – Argentinian SL)	Sign Pose-based Transformer	Recognition Rates: 63.18% (WLASL100), 43.78% (WLASL300), and 100% (LSA64)
Xia et al. [36]	Chinese Sign Language	Locally made dataset with 4000 samples	MobileNet-YOLOv3	Accuracy 90.77%
da Silva et al. [37]	Brazilian Sign Language (Libras) and Argentinian Sign Language (LSA)	Locally made dataset with 5000 videos	I3D and LSTM	Accuracy 99.80% (Libras), and 100% (LSA)
Das et al. [38]	Indian Sign Language (ISL)	Locally made dataset with 288 videos	Long-term Recurrent Convolutional Network (LRCN)	Accuracy 67.53%

TABLE 2. 10th frame extracted from each signs for Subject 7.

Anxiety 	Asthma 	Bandage 
Blood 	Blood Pressure 	Broke 
Burn 	Cold 	Constipate 
Cut 	Depressed 	Diarrhea 
Disease 	Doctor 	Emergency 
Fever 	Headache 	Hospital 
Infection 	Itch 	Nauseous 

state. The forward LSTM's hidden state (h_t^f) ingests input vectors (x_t) in chronological order ($t = 1, 2, \ldots, T$), accumulating information from the past. On the contrary, the backward LSTM's hidden state (h_t^b) encounters input vectors in reverse order ($t = T, T-1, \ldots, 1$), allowing it to incorporate future context. The ultimate result (y_t) of

TABLE 3. 10th frame extracted from each signs for subject 7 (contd).

Pain	Patient	Redness
Sore Throat	Sprain	Swallow
Treatment	Urgent Care (UC)	Vomit

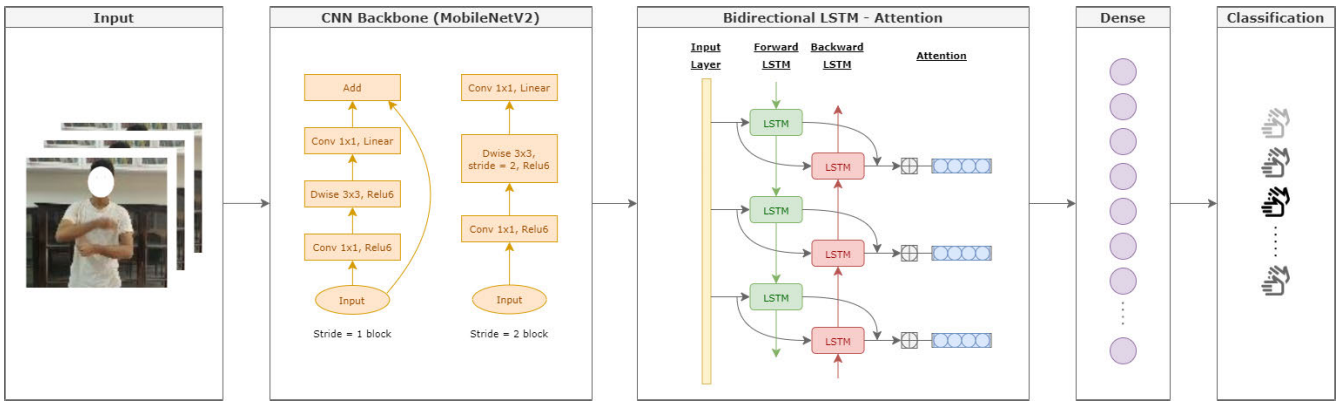


FIGURE 1. Proposed medical sign classification model.

the BiLSTM is formed by encapsulating the information captured by h_t^b and h_t^b, as defined by the provided equations.

$$h_t^f = \tanh\left(W_{xh}^f x_t + W_{hh}^f h_{t-1}^f + b_h^f\right) \quad (1)$$

$$h_t^b = \tanh\left(W_{xh}^b x_t + W_{hh}^b h_{t+1}^b + b_h^b\right) \quad (2)$$

$$y_t = W_{hy}^f h_t^f + W_{hh}^b h_t^b + b_y \quad (3)$$

Building upon the equations presented above, where h_{t-1} and h_{t+1} represent the preceding and subsequent hidden states, and W and b represent weight and bias vectors, respectively. The proposed model integrates an attention layer with the BiLSTM configuration, as suggested by [40]. This system allows the network to focus in a selective manner on crucial points by giving them greater weight, thereby enhancing its capability to encapsulate pertinent information. Essentially, the model based on attention aims to break down complicated problems into simpler, sequentially processed focus areas. For long input sequences, attention provides a weighting scheme to extract important features and not get bogged down in noise. Furthermore, attention regularization has been shown to improve gradient flow during training, allowing the BiLSTM models to be trained better for some problems. At time t, considering the $i-th$ BiLSTM, the final hidden state, $h_{i(t)}$, is calculated this way:

$$h_{i(t)} = \left[h_t^f,\ h_t^b\right] \quad (4)$$

Therefore, at time t, the calculation of attention modality is as follows:

$$e_{i(t)} = \tanh\left(W_a h_{i(t)} + b_a\right) \quad (5)$$

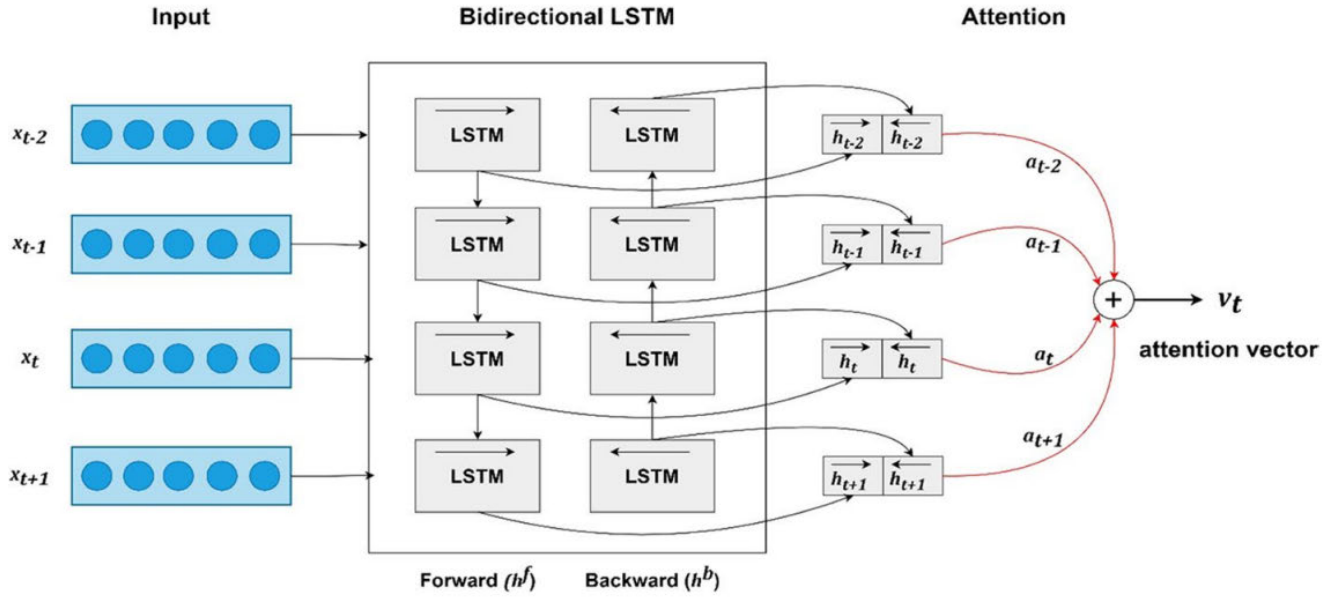


FIGURE 2. The architecture of a bidirectional LSTM model with an integrated attention mechanism [40].

$$a_{i(t)} = \frac{\exp\left(e_{i(t)}\right)}{\sum_{j=1}^{T} \exp\left(e_{j(t)}\right)} \tag{6}$$

$$v_t = \sum_{j=1}^{T} a_{i(t)} h_{i(t)} \tag{7}$$

The relevance of $h_{i(t)}$ is evaluated by passing it through a fully-connected layer, generating an attention energy score $e_{i(t)}$. Subsequently, a softmax layer transforms $e_{i(t)}$ into a probability distribution $a_{i(t)}$, representing the relative importance of each BiLSTM output vector. Ultimately, the attention vector (v_t) is formed by combining the $i-th$ BiLSTM output vectors at time t, weighted by their corresponding attention probabilities $(a_{i(t)})$.

F. FULLY CONNECTED DENSE LAYER

At the end of our model, we add a fully connected network to act as a classifier. This layer generates a probability distribution across thirty categories by utilizing the softmax activation (for example, fever, patient, cold, etc.). This probability distribution tells us how likely the input belongs to each category. The model uses an activation function called 'softmax,' represented by σ, which is expressed as:

$$\sigma\left(z_i\right) = \frac{\exp\left(z_i\right)}{\sum_{c=1}^{N} \exp\left(z_c\right)} \tag{8}$$

Equation 8 employs the element-wise exponential function of the input vector (z_i) represented by $\exp\left(z_i\right)$. This indicates that each element of z_i is raised to the power of e. Notably, N denotes the number of potential classes (30 in this context), signifying that the calculation applies to each component of the vector. Likewise, $\exp\left(z_c\right)$ denotes the element-wise exponentiation of the output vector (z_c).

G. OVERVIEW OF EXPERIMENT

Fusing MobileNetV2 architecture with attention-integrated Bidirectional LSTM is a result of various experimental assessments. Initially, it was observed that the effectiveness of two concurrent layers of BiLSTMs surpassed that of a single one. Subsequent evaluations involved exploring different quantities of hidden units, ranging from 64 to 512. An equivalent number of hidden units were employed in both layers to maintain consistency. The optimization process of the loss function was done using the Adam method [41], while additional hyperparameters such as the batch size, dropout, and learning rates were determined by systemic experimentation using different values. The specific values of hyperparameters are detailed in Table 4 and the overall algorithm has been summarized in Fig. 3.

TABLE 4. List of tuned hyperparameters of the proposed attention mechanism-based CNN-BiLSTM mode.

Parameters	Value
Input frames	30
Input frame dimensions	100×100×3
Number of BiLSTM layers (Attention-based)	2
BiLSTM (number of hidden units)	256
Dropout rate	0.25
Optimizer	Adam
Learning rate	0.001
Loss function	Categorical Cross-entropy
Training epochs	100
Batch Size	16

H. HARDWARE AND SOFTWARE CONFIGURATION

Tensorflow [42] and Keras [43], which are open-source libraries, were used to conduct this experiment. The entire

Input:
Total dataset $D = \{(A_k, Y_k),\ k = 1, 2, 3, \ldots, N\}$, where A_k = input video clip assigned with label Y_k.
Training dataset $T = \{T_i, \forall\, T_i = (a_p, y_p) \in D\}$, where $i = 1, 2, 3, \ldots, n_T$ and $p = 1, 2, 3, \ldots, N_T$.
Validation dataset $V = \{V_j, \forall\, V_j = (a_q, y_q) \in D\}$, where $j = 1, 2, 3, \ldots, n_V$ and $q = 1, 2, 3, \ldots, N_V$.
Here, $T \cap V = \emptyset$
Number of epochs = ep_num
Feature set of total dataset = f_D
Feature set of training dataset = f_T $[f_T \subset f_D]$
Feature set of validation dataset = f_V $[f_V \subset f_D]$

Output:
Trained model $model_{epoch}$, Accuracy $accuracy_{epoch}$, and F1-score $f1\_score_{epoch}$

for all $A_k \in D$ $(k = 1, 2, 3, \ldots, N)$
  Extract frames $(f_1, f_2, \ldots, f_{30})$ *of video clip* A_k
  Resize frames to 100×100 *pixels*
end for

for $ep = 0, ep\_num$ **do**
  for all $A_k \in D$ $(k = 1, 2, 3, \ldots, N)$
    Extract video frame features f_{C_k} *using MobileNetV2*
    Append the feature set $\boldsymbol{f_D}$ *with video frame features* f_{A_k}
  end for
  Train CNN − BiLSTM_with_attention model for T : $model_{ep}.fit(f_T, y_T)$
  Validate the model for epoch : $result_{epoch} \leftarrow model_{ep}.evaluate(f_V, y_V)$
end for
Generate confusion matrix : $conf\_matrix_{ep} \leftarrow Confusion\_Matrix(result_{ep}.predicted, result_{ep}.actual)$
Calculate $precision_{ep}, recall_{epoch}, accuracy_{ep}$ *and* $f1\_score_{ep}$ *using* $conf\_matrix_{ep}$
return $model_{ep}, accuracy_{ep}, f1\_score_{ep}$

FIGURE 3. Proposed algorithm.

process took place in a local setup on Visual Studio Code with the help of 64 GB RAM, Intel Core i5-11400 processor, and NVIDIA RTX 3050 GPU. Furthermore, the dataset was split into 80% for training and 20% for validation ratio.

I. EVALUATION METRICS

In the evaluation of deep learning models, metrics like Precision, Recall, and F1 Score are essential for assessing performance, particularly in classification tasks. These metrics rely on four fundamental elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which are defined as follows:

True Positive (TP): A true positive occurs when the model correctly predicts a positive instance as positive. In other words, it is a correct prediction of a positive outcome.

True Negative (TN): A true negative occurs when the model correctly predicts a negative instance as negative. In other words, it is a correct prediction of a negative outcome.

False Positive (FP): A false positive occurs when the model incorrectly predicts a negative instance as positive. In other words, it is an incorrect prediction of a positive outcome.

False Negative (FN): A false negative occurs when the model incorrectly predicts a positive instance as negative. In other words, it is an incorrect prediction of a negative outcome.

Precision: Precision is a metric that quantifies the accuracy of positive predictions made by the model. It is calculated as the ratio of True Positives to the sum of True Positives and False Positives, expressed as

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

Precision is a metric that evaluates the model's capacity to decrease the occurrence of false positive estimations.

Recall: The metric of recall, alternatively referred to as Sensitivity or True Positive Rate, quantifies the model's capacity to accurately detect all positive cases. The calculation involves determining the proportion of True Positives in relation to the combined total of True Positives and False Negatives, expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Recall assesses the model's capacity to minimize false negatives.

F1 Score: The F1 Score is a statistic that integrates Precision and Recall, offering a harmonious mean between the two measures to achieve balance. It is calculated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

The F1 Score is often employed when there is a need to consider both false positives and false negatives simultaneously.

J. INFORMED CONSENT OF HUMAN SUBJECTS

In adherence to ethical standards, informed consent was obtained from all participants in this study through comprehensive consent forms, ensuring a transparent and voluntary commitment to their involvement in the research.

IV. RESULTS AND ANALYSIS

A. RESULTS

After starting with a baseline CNN-LSTM model with good validation accuracy (85.39%), the focus was given to finding suitable pre-trained CNN architecture for feature extraction. With insight from literature and experience from experimentation with a number of pre-trained models, MobileNetV2 with BiLSTM layers with 64 units each provided an approximately 7% increase in accuracy (92.87%) from the baseline model. It was also observed that the validation accuracy increased with the number of units used in LSTM layers, as with 128 units, the reported accuracy reached 93.18%. Furthermore, by introducing an attention mechanism with BiLSTM [40] and increasing the number of units to 256, the model became capable of being 95.83% accurate while working with unseen data (validation set). This gradual improvement in performance is described in Table 5, and a visual representation is shown in Figure 4.

Furthermore, Table 6 provides a deeper look at the detailed performance evaluation for 30 signs for attention based MobileNetV2-BiLSTM model.

B. PERFORMANCE COMPARISON WITH SEGMENTATION ALGORITHMS

Mediapipe library [44] has recently gained popularity due to being a lightweight and fast framework. Google launched a Kaggle competition called "Google - Isolated Sign Language Recognition" on March 8, 2023 [45]. The goal of the competition was to develop an automatic sign language

TABLE 5. Gradual improvement in evaluation metrics in the experiemental models.

Model	Validation Accuracy	Precision	Recall	F1 Score
CNN BiLSTM Baseline (LSTM Unit=64)	85.39%	82%	82%	81%
MobileNetV2 BiLSTM (LSTM Unit=64)	92.87%	91%	91%	91%
MobileNetV2 BiLSTM with Attention (LSTM Unit=128)	93.18%	92%	92%	92%
MobileNetV2 BiLSTM with Attention (LSTM Unit=256)	95.83%	93%	93%	93%

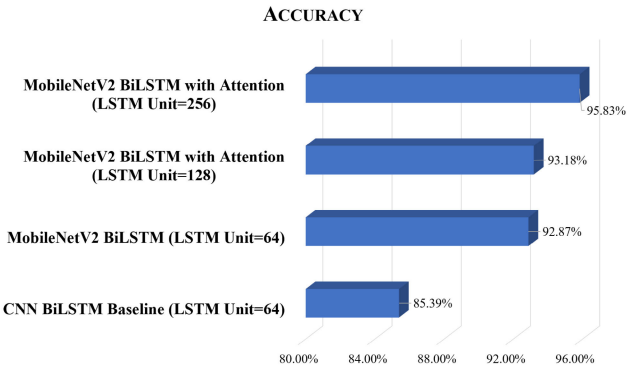


FIGURE 4. Gradual performance improvement in experimental models.

TABLE 6. Performance evaluation for 30 medical signs for the proposed model.

Sign	Precision	Recall	F1-Score
Anxiety	1.00	0.96	0.98
Asthma	1.00	1.00	1.00
Bandage	0.93	0.97	0.95
Blood	1.00	1.00	1.00
Blood Pressure	0.88	0.79	0.83
Broke	0.87	0.90	0.89
Burn	0.85	0.94	0.89
Cold	1.00	1.00	1.00
Constipated	0.95	0.88	0.91
Cut	0.83	0.95	0.89
Depressed	0.87	0.95	0.91
Diarrhea	1.00	0.85	0.92
Disease	0.87	0.95	0.91
Doctor	1.00	0.92	0.96
Emergency	0.97	1.00	0.98
Fever	0.90	1.00	0.95
Headache	0.95	0.88	0.91
Hospital	0.90	1.00	0.95
Infection	1.00	0.86	0.92
Itch	0.86	0.90	0.88
Nauseous	0.95	0.95	0.95
Pain	0.87	0.91	0.89
Patient	1.00	1.00	1.00
Redness	0.91	0.95	0.93
Sore Throat	0.90	0.86	0.88
Sprain	0.88	0.94	0.91
Swallow	0.96	0.93	0.95
Treatment	0.95	0.88	0.91
Urgent Care	1.00	0.96	0.98
Vomit	0.91	0.95	0.93

recognition (ASLR) system that can accurately recognize American Sign Language (ASL) signs from videos. During experimentation, Mediapipe was used to extract landmarks of key points in the face, hands, and pose, where a ResNet50-LSTM model was used to classify the signs with a validation accuracy of 85.43%, which is slightly better than the baseline CNN-BiLSTM model (85.39%). However, after taking a deeper look, it became evident that segmenting using Mediapipe provides good performance in classifying signs with distinct gestures; nevertheless, when it comes to similar gestures, for example, "Patient" and 'Hospital," as shown in Figure 3, accuracy drops to 75% for both, which is a result of segmentation error [46], [47], [48].

On the contrary, attention-based MobileNetV2-BiLSTM can classify "Hospital" at 0.95 F1-Score and "Patient" with 1.0 F1-Score, mitigating the effects of gesture similarity.

FIGURE 5. Gesture similarity of hospital and patient.

C. COMPARATIVE ANALYSIS WITH CONTEMPORARY METHODS

A sign language dataset specifically catering to words that are used in patient-doctor interactions was not available at the time of this work. Although Xia et al. [36] created a private dataset for this purpose, it only consists of 15 words. Das et al. [38] created another one for 6 words with only 288 videos. As a result, for comparison, contemporary word-level sign language classification methods with datasets with similar proportions were used in Table 7 to assess the performance of the proposed model.

It is evident that the proposed attention-based MobileNetV2-BiLSTM method outperforms the Hierarchical Attention Network [20] and CNN-Attention network [49] with similar dataset sizes. Moreover, the dataset for this experiment may be smaller than that of Xia et al. [36], who made use of MobileNet with YOLOv3, but it has 30 words, which is 15 words higher. Moreover, the proposed architecture has a validation accuracy that is 5% higher. Podder et al. [39] used Mediapipe and aimed toward making a signer-independent method with MobileNetV2-LSTM-SelfMLP, where SelfMLP was introduced to

TABLE 7. Juxtaposing proposed model with contemporary work.

Author Reference	Year of Study	Technique	Dataset Language	Dataset Size	Validation Accuracy
Huang et al. [20]	2018	Latent Space Hierarchical Attention Network	German (General)	6841	82.70%
Parelli et al [49]	2020	CNN-Attention	English (General)	3553	91.38%
Xia et al. [36]	2022	MobileNet-YOLOv3	English (Medical)	4000	90.77%
Podder et al [39]	2023	MobileNetV2-LSTM-SelfMLP	Arabic (General)	6667	87.69%
Proposed Model	**2023**	**MobileNetV2 BiLSTM with Attention**	**English (Medical)**	**3593**	**95.83%**

decrease overfitting; nevertheless, using MobilenetV2 with attention-fused Bidirection LSTM provided better performance without using Mediapipe, proving the superiority of the proposed model over a wide range of methodologies.

D. CURRENT LIMITATIONS

While promising, the current results are limited by the small dataset size of only 30 medical signs. Future work should construct a larger and more comprehensive dataset covering over 100 words and sentences reflecting real patient-doctor interactions. This will enable robust validation across a diverse semantic range. Additionally, this study extracted just the first 30 frames due to hardware constraints. We are actively optimizing our software and upgrading our equipment to process full, unbounded sequences for complete analysis. The larger dataset will also facilitate leveraging recent vision transformer networks, which can potentially boost classification accuracy but require sufficient training data.

V. CONCLUSION

In this paper, a word-level sign language detection method was presented, which is an attention-based MobileNetV2-BiLSTM model for a dataset that is related to sign words frequently used in patient-doctor interactions. Although the proposed model outperforms many contemporary sign language recognition mechanisms proposed in recent literature with similar dataset sizes, there are still limitations that must be overcome. Currently, the research team is actively working with the hearing impaired population to understand the challenges during patient-doctor interactions and create a larger dataset with words and sentences that will be used to further develop the model and deploy it within a web app or a mobile application so that it can become mainstream. Furthermore, with a large dataset, state-of-the-art vision transformers can come into play in this context, which was outside the scope of this paper due to having a dataset that is not suitable for implementing this technique. As the model becomes more efficient, along with a diversified dataset for medical communication, we will move one step further to make the world more comfortable for the hard-of-hearing community.

AUTHORS CONTRIBUTION

- Md. Amimul Ihsan: writing: original draft, review & editing, investigation, project management, data curation, methodology, software, validation.
- Abrar Faiaz Eram: writing: review & editing, visualization, data curation
- Lutfun Nahar: writing: review & editing, visualization, data curation
- Muhammad Abdul Kadir: conceptualization, writing: review & editing, project administration, supervision

ACKNOWLEDGMENT

The authors would like to acknowledge the University of Dhaka, Bangladesh, for providing research facilities and covering publication fees. They also acknowledge Dr. Muhammad Enamul Hoque Chowdhury, Qatar University, for his guidance on sign language recognition techniques.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] R. E. Hommes, A. I. Borash, K. Hartwig, and D. DeGracia, "American sign language interpreters perceptions of barriers to healthcare communication in deaf and hard of hearing patients," *J. Community Health*, vol. 43, no. 5, pp. 956–961, Oct. 2018.

[2] G. K. Shuler, L. A. Mistler, K. Torrey, and R. Depukat, "Bridging communication gaps with the deaf," *Nursing*, vol. 43, no. 11, pp. 24–30, Nov. 2013, doi: 10.1097/01.NURSE.0000435197.65529.cd.

[3] P. Kushalnagar, R. Paludneviciene, and R. Kushalnagar, "Video remote interpreting technology in health care: Cross-sectional study of deaf patients' experiences," *JMIR Rehabil. Assistive Technol.*, vol. 6, no. 1, Mar. 2019, Art. no. e13233, doi: 10.2196/13233.

[4] O. B. Hoque, M. I. Jubair, Md. S. Islam, A.-F. Akash, and A. S. Paulson, "Real time Bangladeshi sign language detection using faster R-CNN," in *Proc. Int. Conf. Innov. Eng. Technol. (ICIET)*, Dec. 2018, pp. 1–6, doi: 10.1109/CIET.2018.8660780.

[5] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2017, pp. 476–483, doi: 10.1109/FG.2017.150.

[6] J. Han, X. Ji, X. Hu, D. Zhu, K. Li, X. Jiang, G. Cui, L. Guo, and T. Liu, "Representing and retrieving video shots in human-centric brain imaging space," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2723–2736, Jul. 2013, doi: 10.1109/TIP.2013.2256919.

[7] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[8] H. M. Cooper, E. J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2205–2231, Jul. 2012.

[9] E. J. Ong, O. Koller, N. Pugeault, and R. Bowden, "Sign spotting using hierarchical sequential patterns temporal intervals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1923–1930.

[10] T. Kim, J. Keane, W. Wang, H. Tang, J. Riggle, G. Shakhnarovich, D. Brentari, and K. Livescu, "Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation," *Comput. Speech Lang.*, vol. 46, pp. 209–232, Nov. 2017, doi: 10.1016/j.csl.2017.05.009.

[11] B. Kang, S. Tripathi, and T. Q. Nguyen, "Real-time sign language fingerspelling recognition using convolutional neural networks from depth map," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 136–140, doi: 10.1109/ACPR.2015.7486481.

[12] S. Aly, B. Osman, W. Aly, and M. Saber, "Arabic sign language fingerspelling recognition from depth and intensity images," in *Proc. 12th Int. Comput. Eng. Conf. (ICENCO)*, Dec. 2016, pp. 99–104, doi: 10.1109/ICENCO.2016.7856452.

[13] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to Arabic sign language recognition," *IEEE Trans. Hum.-Mach. Syst.*, vol. 44, no. 4, pp. 551–557, Aug. 2014, doi: 10.1109/THMS.2014.2318280.

[14] T.-W. Chong and B.-G. Lee, "American sign language recognition using leap motion controller with machine learning approach," *Sensors*, vol. 18, no. 10, p. 3554, Oct. 2018, doi: 10.3390/S18103554.

[15] P. Kumar, H. Gauba, P. Pratim Roy, and D. Prosad Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, Oct. 2017, doi: 10.1016/J.NEUCOM.2016.08.132.

[16] K. Lai, J. Konrad, and P. Ishwar, "A gesture-driven computer interface using Kinect," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Apr. 2012, pp. 185–188, doi: 10.1109/SSIAI.2012.6202484.

[17] H. Wang, X. Chai, and X. Chen, "Sparse observation (SO) alignment for sign language recognition," *Neurocomputing*, vol. 175, pp. 674–685, Jan. 2016, doi: 10.1016/J.NEUCOM.2015.10.112.

[18] C. Savur and F. Sahin, "Real-time American sign language recognition system using surface EMG signal," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 497–502, doi: 10.1109/ICMLA.2015.212.

[19] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1610–1618, doi: 10.1109/CVPR.2017.175.

[20] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, Apr. 2018, pp. 1–8, doi: 10.1609/AAAI.V32I1.11903.

[21] O. Kopuklu, N. Köse, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*. UT, USA: IEEE, Jun. 2018, pp. 2184–21848, doi: 10.1109/CVPRW.2018.00284.

[22] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*. Italy: IEEE, Oct. 2017, pp. 3056–3064, doi: 10.1109/ICCVW.2017.361.

[23] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3075–3084, doi: 10.1109/ICCV.2017.332.

[24] P. V. V. Kishore, D. Anil Kumar, E. Goutham, and M. Manikanta, "Continuous sign language recognition from tracking and shape features using fuzzy inference engine," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 1–6, doi: 10.1109/WISPNET.2016.7566526.

[25] A. Gondu, P. V. V. Kishore, A. Sastry, D. Anil Kumar, and K. Eepuri, "Selfie continuous sign language recognition with neural network classifier," in *Proc. 2nd Int. Conf. Micro-Electron., Electromagn. Telecommun.*, Mar. 2017, pp. 1–10, doi: 10.1007/978-981-10-4280-5.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[27] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017, doi: 10.1109/TPAMI.2016.2599174.

[28] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 843–852.

[29] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*, A. A. Salah and B. Lepri, Eds. Cham, Switzerland: Springer, 2011, pp. 29–39.

[30] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[31] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. BC, Canada: IEEE, Jun. 2023, pp. 2529–2539, doi: 10.1109/CVPR52729.2023.00249.

[32] N. Song and Y. Xiang, "SLGTformer: An attention-based approach to sign language recognition," 2022, *arXiv:2212.10746*.

[33] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. CO, USA: IEEE, Mar. 2020, pp. 1448–1458, doi: 10.1109/WACV45572.2020.9093512.

[34] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. QC, Canada: IEEE, Oct. 2021, pp. 11283–11292, doi: 10.1109/ICCV48922.2021.01111.

[35] M. Bohácek and M. Hruz, "Sign pose-based transformer for word-level sign language recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*. HI, USA: IEEE, Jan. 2022, pp. 182–191, doi: 10.1109/WACVW54805.2022.00024.

[36] K. Xia, W. Lu, H. Fan, and Q. Zhao, "A sign language recognition system applied to deaf-mute medical consultation," *Sensors*, vol. 22, no. 23, p. 9107, Nov. 2022, doi: 10.3390/S22239107.

[37] D. R. B. da Silva, T. M. U. de Araújo, T. G. do Rêgo, M. A. C. Brandão, and L. M. G. Gonçalves, "A multiple stream architecture for the recognition of signs in Brazilian sign language in the context of health," *Multimedia Tools Appl.*, vol. 83, no. 7, pp. 19767–19785, Jul. 2023.

[38] H. V. Das, K. Mohan, L. Paul, S. Kumaresan, and C. S. Nair, "Transforming consulting atmosphere with Indian sign language translation," *Multimedia Tools Appl.*, vol. 83, no. 5, pp. 13543–13555, Jul. 2023.

[39] K. K. Podder, M. Ezeddin, M. E. H. Chowdhury, M. S. I. Sumon, A. M. Tahir, M. A. Ayari, P. Dutta, A. Khandakar, Z. B. Mahbub, and M. A. Kadir, "Signer-independent Arabic sign language recognition system using deep learning model," *Sensors*, vol. 23, no. 16, p. 7156, Aug. 2023.

[40] K. Yousaf and T. Nawaz, "An attention mechanism-based CNN-BiLSTM classification model for detection of inappropriate content in cartoon videos," *Multimedia Tools Appl.*, pp. 1–24, Sep. 2023, doi: 10.1007/s11042-023-16727-6.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[42] N. Ketkar, "Introduction to keras," in *Deep Learning with Python*. Cham, Switzerland: Springer, 2017, pp. 97–111, doi: 10.1007/978-1-4842-2766-4_7.

[43] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, and J. Dean, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Des. Implement.*, 2016, pp. 265–283.

[44] J.-W. Kim, J.-Y. Choi, E.-J. Ha, and J.-H. Choi, "Human pose estimation using mediapipe pose and optimization method based on a humanoid model," *Appl. Sci.*, vol. 13, no. 4, p. 2700, Feb. 2023.

[45] (2023). *Google-American Sign Language Fingerspelling Recognition*. Accessed: Sep. 11, 2023. [Online]. Available: https://kaggle.com/competitions/asl-fingerspelling

[46] J. L. Lauermann, A. K. Woetzel, M. Treder, M. Alnawaiseh, C. R. Clemens, N. Eter, and F. Alten, "Prevalences of segmentation errors and motion artifacts in OCT-angiography differ among retinal diseases," *Graefe's Arch. Clin. Experim. Ophthalmol.*, vol. 256, no. 10, pp. 1807–1816, Oct. 2018.

[47] G. Gill and R. R. Beichel, "An approach for reducing the error rate in automated lung segmentation," *Comput. Biol. Med.*, vol. 76, pp. 143–153, Sep. 2016.

[48] W. Jia, L. Yang, Z. Jia, W. Zhao, Y. Zhou, and Q. Song, "TIVE: A toolbox for identifying video instance segmentation errors," *Neurocomputing*, vol. 545, Aug. 2023, Art. no. 126321.

[49] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," in *Proc. Comput. Vision–ECCV 2020 Workshops*. Springer, 2020, pp. 249–263.

MD. AMIMUL IHSAN received the B.Sc. degree in electrical and electronic engineering from the Islamic University of Technology and the M.S. degree in biomedical physics and technology majoring in biomedical engineering from the University of Dhaka, in 2023. His research interests include biomedical signal and image processing, AI and ML for image and video analysis, and biomedical instrumentation. He was a recipient of the prestigious International Science Program (ISP) Fellowship from Uppsala University.

ABRAR FAIAZ ERAM is currently pursuing the B.Sc. degree in electrical and electronic engineering with Bangladesh University of Engineering and Technology. His research interests include machine learning, dc–dc converters, quantum computing, topological insulators, and biomedical instrumentation.

LUTFUN NAHAR received the B.Sc. degree in electrical and electronic engineering from Noakhali Science and Technology University with the University Grants Commission Scholarship and the M.S. degree in biomedical physics and technology from the University of Dhaka, with a focus on using deep learning techniques for biomedical image processing. She was a recipient of the National Science and Technology Fellowship (NST) from the Ministry of Science and Technology, Bangladesh.

MUHAMMAD ABDUL KADIR (Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in physics from the University of Dhaka, Bangladesh, and the Ph.D. degree in biomedical physics through a joint program with the University of Dhaka and the University of Warwick. As a Commonwealth Scholar, he was a Visiting Postgraduate Researcher with the University of Warwick, U.K. He is currently a Professor with the Department of Biomedical Physics and Technology, University of Dhaka. His research interests include biomedical instrumentation and the medical applications of electrical impedance techniques. He possesses practical expertise in designing and developing multi-frequency electrical bioimpedance measurement systems. Moreover, his interests extend to biomedical signal and image analysis, with a particular emphasis on employing machine learning techniques for biomedical applications.