Received: July 2, 2021. Revised: August 16, 2021. 168

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hand Gesture Recognition using Auto Encoder with Bi-direction Long Short Term Memory

Md Shoaibuddin Madni[1]* C. Vijaya[2]

[1]*Department of Electronics and Communication Engineering, Navodaya Institute of Technology, Raichur, India*
[2]*Department of Electronics and Communication Engineering, SDM College of Engineering and Technology, Dharwad, India*
* Corresponding author's Email: mdshoaibuddin.madni@navodaya.edu.in

Abstract: Hand Gestures provide a means for a series of interactive communications with human beings who are deaf and dumb. Deep learning based models have been revolutionized for achieving human level performance through computer vision that automates, classifies and detects presence of objects in an image. Various challenges were faced through this process such as variation in the hand size, color, illumination, skin tone, view point, complex natural backgrounds etc, because of the various conditions present on the images. The present research overcomes the problem of high dimensional data and improved the classification accuracy in gesture pattern recognition for the incomplete data. Then, multilevel segmentation method and morphological model are used to segment the gesture regions and furthermore, Auto Encoder (AE) algorithm is applied to reduce the dimension of the data. Additionally, Bi-Long Short Term Memory (LSTM) classifier is applied to classify the alphabets, numbers, and video symbols. In the experimental phase, the improved AE-BiLSTM model achieved effective performance in recognition of hand gesture in terms of accuracy. Compared to the existing CNN and PCA based uni-modal feature-level fusion model, the proposed AE-BiLSTM model showed a maximum value of 99.85 % in terms of accuracy for ISL dataset and 99.75% for NUS dataset.

Keywords: AE-BiLSTM, Deep learning, Dimensionality reduction, Hand gesture, Multilevel segmentation.

1. Introduction

The Sign Language provides a system for performing communications by utilizing visual signs and gestures that are used to interact with Hearing and Speech Impaired (HSI) people. In order to set up an effective communication among them, a knowledge based familiar sign language is essential for them [1]. The main goal is to develop a model that will provide assistive technology and will enable and empower effective communication with the HSI community. As the shape and size of the body differ from person to person, the gesture of two persons or even of the same person who performed it at the very first time may or may not be the same which was an issue in Sign Language Recognition system (SLR) [2]. The present research introduces an effective algorithm for input hand gesture translation using Indian Sign Language (ISL) which will be helpful for translating English text to sign language in speech format. Various existing deep learning models were developed for hand sign gesture recognition such as Convolution Neural Network (CNN), Artificial Neural Network (ANN), and Long-short Term Memory (LSTM) models, which were utilized by using standard datasets such as Indian Sign Language (ISL), and National University of Singapore (NUS) datasets for the gesture recognition but unfortunately due to certain problems faced these models showed lower performances [3-5]. The present research developed an algorithm for detecting and segmenting the hand region from the depth image which is obtained from the Microsoft Kinect sensor. There are various existing models that analyzed more number of features faced challenges as an ineffective classical ML algorithm in case of the high dimensional data

[6]. Thus, it is better to apply pre-processing step for removing an irrelevant feature to overcome the dimensionality problem [7]. The correct selection of features improves learning speed, or simplicity in the model [8]. The main idea is to combine the outputs of several number of single feature selection models to obtain better results [9, 10]. The proposed method provides an accurate solution and the human gestures that are captured by the sensor gives rise to a large amount of data which would contain redundant, incomplete or uncorrelated. The Data dimensionality reduction is performed by the Auto encoder that represent the low dimensional space which improved the classification process.

The structure of the paper is as follows: Section 2 describes about the existing methods utilized for hand gesture recognition, Section 3 explains about the proposed method followed by the explanation of its flow of steps. Section 4 discusses the results obtained for the proposed method and stresses on the importance of the proposed method by comparing it with the existing models. The Section 5 discusses about the conclusion and future work for the research.

2. Literature review

The problems occurred in the existing models for hand gesture detection are as follows:

Gangrade [11] developed ORB (Oriented FAST and Rotated BRIEF) for the ISL dataset. The developed model was fused with various classification techniques to obtain an, optimum result. The developed ORB model worked better for the static hand gestures alone as it failed to handle the signs that were dynamic and continuous in nature. However, the developed model failed to recognize hand gestures properly when overlapping of each hand was occurred.

Gangrade and Bharti [12] developed Convolution Neural networks (CNN) model for hand gesture detection using ISL. The developed algorithm worked well in the cluttered environment such as skin color background and hand overlapped with the face and the feature were invariant to rotation and scaling. The limitation of the developed CNN was that it worked well with static ISL signs only and did not work well for dynamic and continuous signs.

Yong Soon Tan [13] developed CNN with Spatial Pyramid Pooling (SPP) for hand gesture recognition. The model CNN was combined with SPP for hand gesture recognition and it was developed to overcome the conventional pooling problem using multilevel pooling extended the features that were fed for the connected layer. The inputs were varying with respect to the sizes yielded better results for fixed length of features. However, the gray scale images showed complex background and were misclassified because of the complexities in image backgrounds. The nine gray scale images showed complex background were misclassified because of the complexities in color image backgrounds were provided with better clear illustration.

Chandr and Lal [14] developed a CNN model for training the color datasets in order to classify the hand postures using the Bayer images. The developed model showed that Bayer patterned images were used for classification of hand poses. The SqueezeNet was used in the developed method that adopted hand pose recognition and this improved the performance. However, the CNN model used for classification process was sensitive for the blurred images as they were obtained during removal of noises.

Madni and Vijaya [15] developed Hand Gesture Recognition using Semi Vectorial Multilevel Segmentation Method with Improved ReliefF Algorithm. In the experimental phase, the performance of the proposed and improved reliefF-K-nearest neighbour model is analysed in light of Matthew's correlation coefficient, accuracy, sensitivity, specificity, and f-score. For overall 16 distinct alphabets; A, B, D, E, F, G, H, K, P, R, T, U, W, X, Y, Z, the proposed reliefF-K-nearest neighbour model achieves average accuracy. However, the Semi Vectorial Multilevel Segmentation showed dimension problem due to extraction of irrelevant features led to complexity in the system.

3. Proposed methodology

The block diagram of the proposed AE-BiLSTM method is shown in the Fig. 1. The steps involved in the model are as follows, (i) Data Collection, (ii) Pre-processing, (iii) Segmentation, (iv) Dimension reduction, and (v) Classification.

3.1 Image dataset

The present research work is implemented on 3 types of datasets for the evaluation of hand gestures recognition at the stage of classification. The Three datasets are Indian Sign Language dataset, National University of Singapore (NUS), and Mendeley dataset. The description for the 3 datasets are as follows.

3.1.1. ISL dataset

The ISL dataset consists of hand gesture images each of which represents the signs used for numbers and alphabets. The dataset consists of 36 classes each having numbers from 1-10 and alphabets from A to

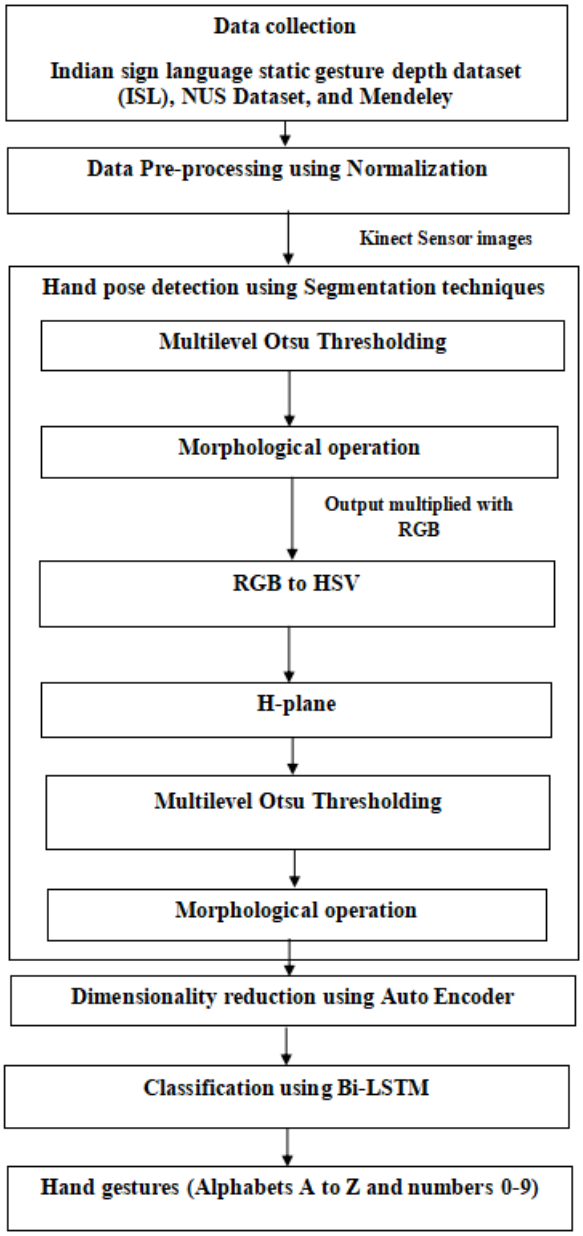

Figure. 1 Flow chart for the proposed method

Z. 12 different people who poses for the signs are captured and each person provides 100 different posed images having changes in angle and position of the hand. The present research considers only depth images for the research [16]. A sample depth image from the ISL dataset is as shown in the Fig. 2.

Figure. 2 Sample depth image from the ISL dataset

Figure. 3 Sample image from NUS dataset

3.1.2. NUS dataset

The NUS hand posture dataset I contains 10 classes of postures each having 24 sample images that capture various sizes and positions of hand present on the image frames [17]. The sample image from NUS dataset is as shown in the Fig. 3.

3.1.2.1. Dataset I

The NUS hand posture dataset has 10 classes of postures wherein 24 samples are present for each class that usually capture the varied size and position of the hand with the image frame. It includes both greyscale and color images that consist of 160×120 pixels.

3.1.2.2. Dataset II

The image dataset consists of complex natural backgrounds that contain varied hand sizes and shapes. The dataset subjects consist of females and males whose ages range from 22 to 56 years and have 10 hand postures.

Both these datasets are available in both colour as well as grey scale images having 160×120 pixels.

3.1.3. Mendeley dataset

The Mendeley dataset consists of videos that have two sets of data organized, one set consists of captured video sequences and the other consists of cropped video sequences, where the objects are present on the excessive background. The male and female subjects with varied hand sizes and skin tones are included in the dataset which are shown in Fig. 4. The size of the images is uniform, that is, 500×600 pixels. There are mainly eight hand gestures present in the dataset which are ‘accident’, ‘call’, ‘doctor’, ‘help’, ‘hot’, ‘lose’, ‘pain’ and ‘thief’ [16].

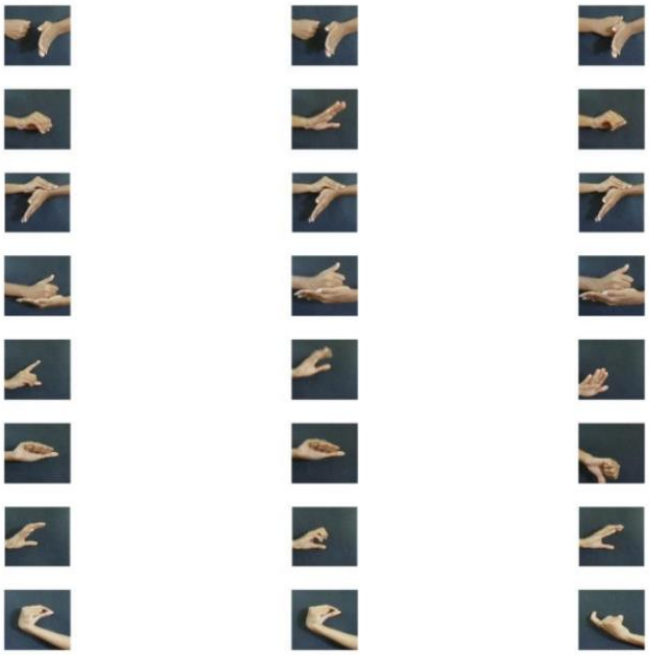

Figure. 4 Sample image from the Mendeley dataset

3.2 Image pre-processing

The image normalization is the process where the range of pixel intensity values are changed. The main purpose is to perform conversion of an input image to the range of values that are normal for sensing. Thus it is known as normalization. The present research performs the function that produces normalization for the input depth image of the channel as it consists of the information about the distance from where the object is viewed from the surface point. The depth image represents pixels of each image in terms of number of bits. Then the values present on the image scale ranging from 0 and 255 are normalized which is shown in the Fig. 5. The digital image undergoes linear normalization by using the below Eq. (1).

$$Output_{channel} = 255 \times \frac{(inputchannel - min)}{max - min} \quad (1)$$

There are several advancements on the sensor technology for recognition of gestures, and in this research study, Kinect sensor is used to detect and enable the hand regions from the images for precise segmentation. Next, a segmentation algorithm helps to detect and segment the hand region using the depth image obtained from the pre-processing which is explained further.

3.3 Image segmentation

The depth images obtained undergo segmentation in order to binarize the image on the basis of intensities of the pixel. The segmentation of

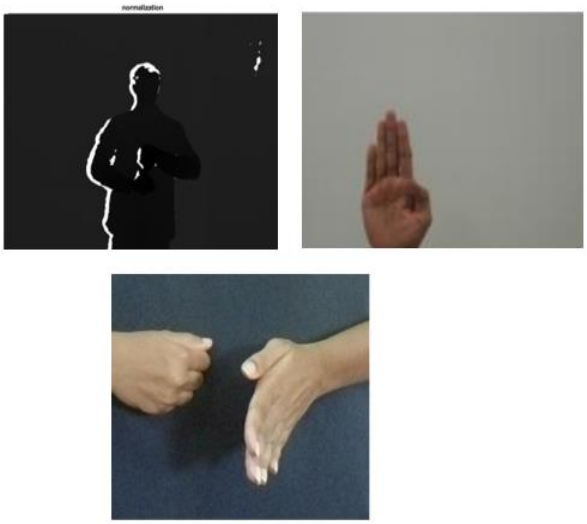

Figure. 5 Pre-processed images obtained from ISL, NUS, and Mendeley dataset

important regions is done by using thresholding technique, as the residue portion still needs to be removed accurately. In order to perform such a function, Multilevel-Otsu threshold algorithm performs pixel separation of an input image into different classes that separates gray levels based on the intensity values. Several thresholds are calculated using multi-Otsu that determines classes of the desired number. The main reason for using Multilevel Otsu thresholding and Morphological operator techniques is to eliminate unwanted regions even after masking using Morphology operation. Morphological operations are performed for processing functions based on the shapes and sizes.

The weighted within the class probabilities are calculated using the below Eq. (2)

$$q_1(t) = \sum_{i=1}^{t} P(i), \quad q_2(t) = \sum_{i=t+1}^{I} P(i), \quad q_n(t) = \sum_{i=I+t+1}^{n} P(i) \tag{2}$$

The threshold value ranges from 1 to t

$q_{1\,..,n}$ is weighted class within pixel probabilities P of the foreground and background.

The class means are given in Eq. (3)

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)}, \mu_2(t) = \sum_{i=t=1}^{I} \frac{iP(i)}{q_2(t)}, \ldots\ldots, \mu_n(t) = \sum_{i=n}^{t} \frac{iP(i)}{q_n(t)} \tag{3}$$

Where, μ_1 and μ_2 are the average gray level values

An input image that is having a structuring element is used for performing morphological operations and obtains an output image without losing its properties. The morphological operation is performed for each pixel of an input image that corresponds to the neighborhood pixels. The shape and size of an image is chosen based on the neighborhood pixels and morphological operation is performed for constructing specific shapes for an input image. The specific pixels were not clear as the intensities of the darker pixels were difficult to distinguish. Therefore, the obtained image pixels are multiplied with the RGB to obtain the images having RGB pixels. Each pixel element of an RGB image should be multiplied with the corresponded co-ordinate element of the other image which are multiplying an RGB image with the constant color planes. The RGB image pixels are converted into HSV (Hue Saturation Value) scale that provides a numerical readout of an image corresponding to the color names. The main purpose of using HSV model instead of RGB image is because, the RGB images define the color with respect to the primary color combinations. As the identification of color plays an important role, the HSV model is preferred more than the RGB model. In case of RGB image, it is difficult for separating the information of luminance from that of the color. Thus, H planes are known as geometrical half planes that are defined based on the chromatic color present in the linear color space. The H distortion needs to be eliminated and corresponding to the input, the pixels should be enhanced. The outcome proved that the effective approach improves the results for the hue distortion because the two color enhancement improved the image enhancement. The H-plane is obtained and the it underg morpholo thresholding process and Multi level Otsu thresholding approach is as shown in Fig. 6.

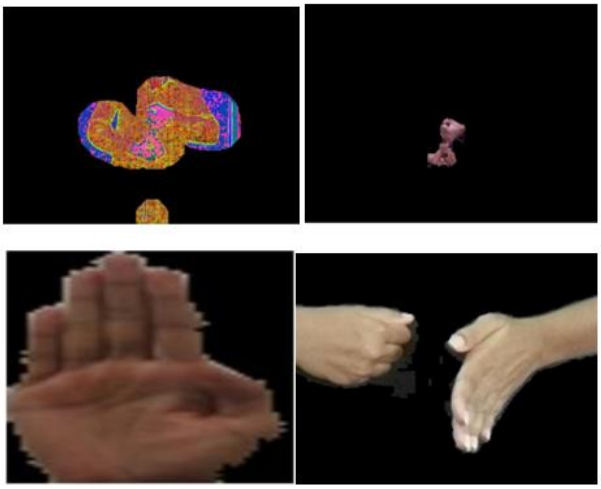

Figure. 6 Images obtained after segmentation using Multilevel Otsu thresholding and Morphological operator techniques

3.4 Proposed dimension reduction using auto-encoder with Bi-LSTM

The dimensionality reduction is performed in order to reduce the feature space for obtaining the statistical and stable model that avoids the dimensionality problem. The present research work uses auto encoder for dimensionality reduction. AE, which is an Artificial Neural Network, compresses the data into a data with low dimension reconstructs the input data as shown in the Fig. 7. The AE identifies the data for representing lower dimension and focuses on important features in order to remove the noise and obtain data redundancy The AE has an encoder and a decoder. The encoder encodes the high dimensional data into the lower dimension data. The decoder will take the low dimension data and again it will reconstruct the high dimensional data which minimizes the reconstruction loss. The reconstruction of the original data through the reduced feature space is obtained. The lower level features are used for feeding the deep learning model and thus executes as a hierarchy at lower levels. Using the back propagation algorithm, the training of network in supervised mode is performed to adjust the weights.

The mapped input data from the encoder represents the hidden data and the reconstruction of input data is performed from the hidden representation, where $\{x_n\}_{n=1}^{N}$ represents the hidden encoder vector is going to be calculated as x_n and $\hat{x}_n$which is known as the decoder vector from the output layer. The process of encoding is done by using the Eq. (4)

$$h_n = f(W_1 x_n + b_1) \quad (4)$$

Where f is known as the encoding function.
W_1 is known as the encoder's weight matrix
b_1 is known as the bias vector which is calculated by using Eq. (5)

$$\hat{x}_n = g(W_2 h_n + b_2) \quad (5)$$

Where g is known as the decoding function
W_2 is known as the decoder weight matrix
b_2 is known as the bias vector.
Therefore the reconstruction error ϕ is minimized which sets the auto encoder and the values optimized by using the following Eq. (6)

$$\phi = \underset{\theta,\theta'}{arg\,min} \frac{1}{n} \sum_{i=1}^{n} L(x^i, \hat{x}^i) \quad (6)$$

L Represents a loss function which is defined as shown in Eq. (7)

$$L(x, \hat{x}) = \|x - \hat{x}\|^2 \quad (7)$$

arg min is the argument of minimum for the instance which attains smallest value for the loss function. Stacked Auto Encoders combined with dropout model were applied for training the weight matrix from the frequency spectra that belongs to the vibration of signals.

3.5 Hand gesture classification using Bi-LSTM

The optimal pixel's values are selected from the previous step and in the next step, classification of images based on the gesture identification is performed. In this research work, a multi- objective machine learning classification algorithm is used to detect and classify the hand gestures into 26 English Alphabets and 0-9 numerical digits. Bi-LSTMs are the extension of LSTM that improve the performance based on the classification problems occurred in the sequence. This will provide an additional context about the network and the results generates faster and learns the problem. The bidirectional sequence was provided initially and was justified in the domain for speech recognition as the utterance is as a whole is used for interpreted being rather than performing linear interpretation. The Bi-LSTM will calculate the sequence from the opposite direction up to the forward hidden sequence and then again to the backward hidden sequence. The encoded vector forms the concatenation up to the final forward and backward outputs. It calculates the input sequences $x = (x_1, x_2, \ldots, x_n)$ from an opposite direction towards the forward hidden sequence $\vec{h}_t = (\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_n)$ which is shown in Eq (8 and 9). The backward sequence $\overleftarrow{h}_{t=}(\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_n)$. An encoded vector y_t is formed by the process of concatenation of the output obtained from the forward and the backward outputs $y_t = [\vec{h}_t, \overleftarrow{h}_t]$ is as shown in the Eq. (10).

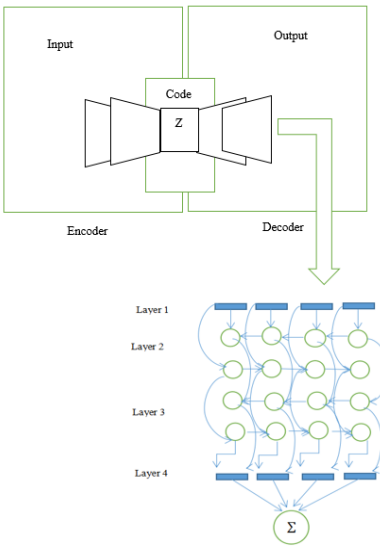

Figure. 7 Proposed AE-BiLSTM

$$\vec{h}_t = \sigma(W_{\vec{h}x} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1}) \quad (8)$$

$$\overleftarrow{h}_t = \sigma(W_{\vec{h}x} x_t + W_{\vec{h}\vec{h}} \overleftarrow{h}_{t-1}) \quad (9)$$

$$y_t = W_{y\vec{h}} \vec{h}_t + W_{y\vec{h}} \overleftarrow{h}_t \quad (10)$$

Where y_t represents the outputs obtained for $(y_1 \ldots, y_n \ldots\ldots, y_t)$
σ is the function of backward and forward process.

The obtained sequence will be ranging between the 0 and 1 values and this sequence is considered as an input. The output values are assigned as 0 and once the cumulative sum for all the input values are obtained in sequence, then threshold is generated among the output values ranging from 0 to 1.

4. Results and discussion

In this research article, the proposed AE-BiLSTM model is simulated using MATLAB (2019a) environment with windows 10 operating system, with the configurations as 8GB RAM, Intel i5 processor, and 4 TB hard disk. In this section, the performance of the proposed AE-BiLSTM model is evaluated.

4.1 Performance measures

In this scenario, Indian sign language database is used for validating the effectiveness of the proposed and existing models using accuracy, precision, recall, Fowlkes-Mallows Index (FMI), and Critical Success Index (CSI) as the parameters. The mathematical expressions are represented in Eqs. (11) to (15).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (13)$$

Fowlkes-Mallows Index (FMI)

$$= \sqrt{\frac{TP}{TP+FN} \times \frac{TP}{TP+FP}} \quad (14)$$

Critical Success Index (CSI)

$$= \frac{TP}{TP+FN+FP} \times 100 \quad (15)$$

Where, TP is denoted as true positive, TN is stated as true negative, FP is stated as false positive, and FN is represented as false negative.

4.2 Quantitative analysis

The Fig. 8 shows the results obtained for the proposed AE-Bi-LSTM model as well as the existing models such as Probabilistic Neural Network (PNN), Decision Tree, ANN, and AdaBooster models, and these classifiers were evaluated by individually combining them with various dimensionality reduction algorithms. The algorithms used for dimension reduction were, Independent Component Analysis (ICA), Principal Component Analysis (PCA), R-PCA, and finally the AE. The PNN classifier obtained low performances even after being combined with all these algorithms.

Similarly, the DT used was unstable, as whenever there was small change in the data it leads to large variations in the structure of optimal decision tree. The model was often inaccurate and many more predictors that were used performed better with respect to the similar data. Thus, the DT also showed less performances for all the dimension reduction techniques. The ANN was not able to generate the previous state values and thus optimum results were not produced. The Ada Boost classifier had few disadvantages when large datasets such as NUS, ISL datasets it showed vulnerable with respect to the uniform noise which lead to over-fitting. Whereas, the Bi-directional LSTM performed well for all of the dimension reduction algorithms as it enabled additional training from left then to right and from right to left improved training performances as shown in Fig. 9.

The additional parameters involved was beneficial for tuning the parameters and thus the Bi-LSTM offered better predictions when compared to other classifiers.

4.3 Comparative analysis

Table 1 shows the comparative analysis of the existing and proposed models in terms of accuracy. In [11] the Oriented FAST and algorithms were used and the results obtained were with respect to the ISL dataset. The features utilized showed high dimension

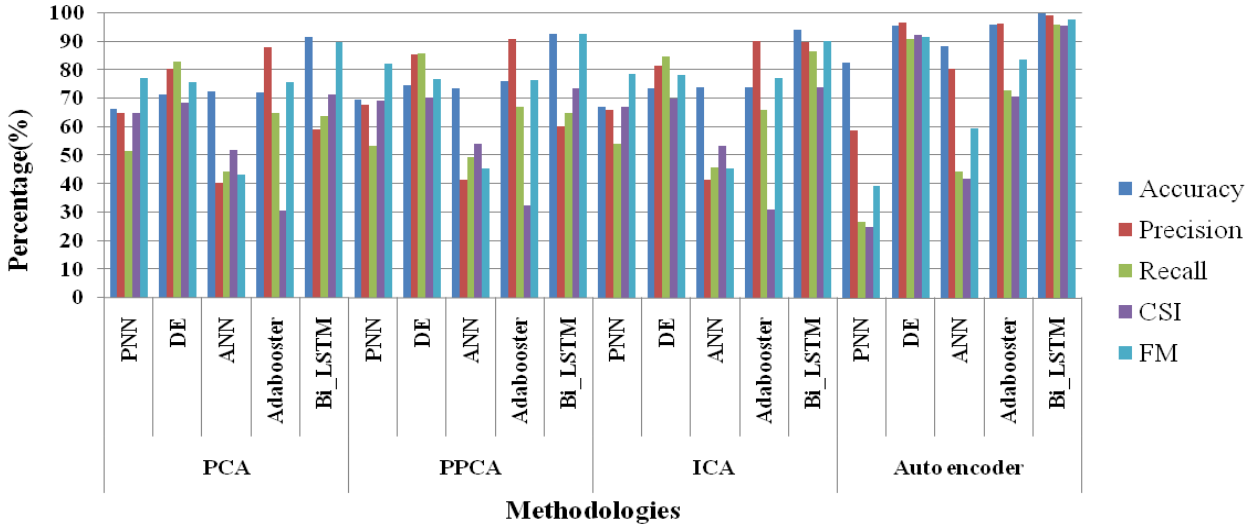

Figure. 8 Results for ISL dataset (numbers) for classifiers with dimensional reduction algorithms

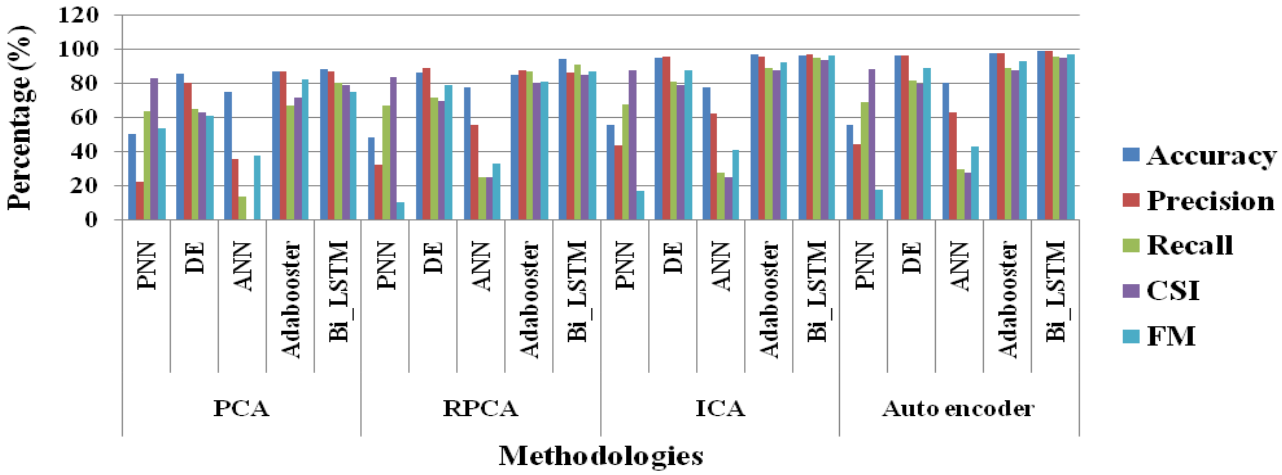

Figure. 9 Results for Mendeley video dataset for classifiers with dimensional reduction algorithms

Table 1. The comparative analysis for the proposed AE-BiLSTM with the existing algorithms

Authors	Methodology	Dataset	Accuracy (%)
Gangrade [11]	Oriented FAST and Rotated BRIEF	ISL	93.26
Jayesh and Bharti [12]	CNN		99
Tan [13]	CNN-SPP	NUS	98.40
Chandra and Lall [14]	CNN		99.67
Madni [15]	Improved reliefF K-nearest neighbour	ISL	98.95
Proposed Method	AE-Bi-LSTM	ISL	99.85
		NUS	99.75

data complexity and thus the accuracy was lowered up-to 93.26 %. Similarly, [12, 14] CNN model was utilized for better automated classification but it failed to analyse for different dataset and obtained accuracy of 99%. The CNN-SPP model [13] was utilized for hand gesture recognition but however it failed to improve the classification accuracy due to the high dimensionality of the data hence it obtained accuracy of 98.40%. Whereas, the proposed AE-Bi-LSTM outperformed with better results when compared to the existing algorithms as the AE is used for dimension reduction, thus, it showed improvement of accuracy for both the datasets; 99.85 % in ISL and 99.75 % in NUS.

5. Conclusion

In this research, AE-BiLSTM model is proposed to enhance the performance of hand gesture recognition. Initially, the normalization method is undertaken to enhance the visual level of the images, which are collected from the Indian sign language database. Then, multilevel segmentation method and morphological model is used to segment the gesture regions and further AE algorithm is applied to reduce the dimension of the data that completely resolves the problem of dimensionality. Additionally, Bi-LSTM classifier is applied to classify the 16 gesture symbols. In the experimental phase, the improved AE-BiLSTM model achieved effective performance in hand gesture recognition in terms of accuracy. Compared to the existing CNN model and PCA based uni modal feature-level fusion model, the proposed improved AE-BiLSTM model showed maximum of 2 to 4 % improvement in average accuracy. In future work, a new optimization-based clustering algorithm can be included in the proposed model to further improve the performance of hand gesture recognition.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] H. Tang, H. Liu, W. Xiao, and N. Sebe, “Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion”, *Neurocomputing*, Vol. 331, pp. 424-433, 2019.

[2] A. Sharma, A. Mittal, S. Singh, and V. Awatramani, “Hand Gesture Recognition using

Image Processing and Feature Extraction Techniques", *Procedia Computer Science*, Vol. 173, pp. 181-190, 2020.

[3] S. Ameur and A. B. Khalifa, "A novel Hybrid Bidirectional Unidirectional LSTM Network for Dynamic Hand Gesture Recognition with Leap Motion", *Entertainment Computing*, p. 100373, 2020.

[4] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition", *Pattern Recognition*, Vol. 76, pp. 80-94, 2018.

[5] Y. Yao and Y. Fu, "Contour model-based hand-gesture recognition using the Kinect sensor", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 11, pp. 1935-1944, 2014.

[6] G. A. Rao and P. V. V. Kishore, "Selfie video based continuous Indian sign language recognition system", *Ain Shams Engineering Journal*, Vol. 9, No. 4, pp. 1929-1939, 2018.

[7] J. Jeong and Y. Jang, "Max-min hand cropping method for robust hand region extraction in the image-based hand gesture recognition", *Soft Computing*, Vol. 19, No. 4, pp. 815-818, 2015.

[8] P. Bao, A. I. Maqueda, C. R. D. Blanco, and N. García, "Tiny hand gesture recognition without localization via a deep convolutional network", *IEEE Transactions on Consumer Electronics*, Vol. 63, No. 3, pp. 251-257, 2017.

[9] M. F. Wahid, R. Tafreshi, M. A. Sowaidi, and R. Langari, "Subject-independent hand gesture recognition using normalization and machine learning algorithms", *Journal of Computational Science*, Vol. 27, pp. 69-76, 2018.

[10] K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition", *Neural Computing and Applications*, Vol. 28, No. 12, pp. 3941-3951, 2017.

[11] J. Gangrade, J. Bharti, and A. Mulye, "Recognition of Indian sign language using ORB with bag of visual words by Kinect sensor", *IETE Journal of Research*, pp. 1-15, 2020.

[12] J. Gangrade and J. Bharti, "Vision-based Hand Gesture Recognition for Indian Sign Language Using Convolution Neural Network", *IETE Journal of Research*, pp. 1-10, 2020.

[13] Y. S. Tan, K. M. Lim, C. Tee, C. P. Lee, and C. Y. Low, "Convolutional neural network with spatial pyramid pooling for hand gesture recognition", *Neural Computing and Applications*, Vol. 33, No. 10, pp. 5339-5351, 2021.

[14] M. Chandra and B. Lall, "A Novel Method for CNN Training Using Existing Color Datasets for Classifying Hand Postures in Bayer Images", *SN Computer Scienc*e, Vol. 2, No. 2, pp. 1-10, 2021.

[15] M. S. Madni and C. Vijaya, "Hand Gesture Recognition Using Semi Vectorial Multilevel Segmentation Method with Improved ReliefF Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 3, pp. 447-457, 2021.

[16] R. Elakkiya and B. Natarajan, "ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition", *Mendeley Data*, 2021.

[17] P. K. Pisharady, P. Vadakkepat, and L. A. Poh, "Hand posture and face recognition using fuzzy-rough approach", In: *Proc. of Computational Intelligence in Multi-Feature Visual Pattern Recognition*, pp. 63-80, 2014.