Multimedia Tools and Applications FOR APPROVAL
https://doi.org/10.1007/s11042-023-14428-8

An optimized automated recognition of infant sign language using enhanced convolution neural network and deep LSTM

Vamsidhar Enireddy[1] · J. Anitha[2] · N. Mahendra[3] · G. Kishore[4]

Received: 23 March 2022 / Revised: 3 December 2022 / Accepted: 21 January 2023

Abstract
In the world, several sign languages (SL) are used, and BSL (Baby Sign Language) is the process of communication between the parents and baby using gestures. Communication by gestures is a non-verbal process that utilizes motion to pass on realities, expressions and feelings to people. SL is the communication mode in which the information is conveyed via movement of body parts like cheeks, eyebrows and head. Even though many research works based on SL are available, research in BSL remains a challenge. Hence, this paper presents an optimization-based automated recognition of the deep BSL system, which determines the gesture signalled by the kids. Initially, the image frames are extracted from the videos and data augmentation processes are performed. After pre-processing, the features are extracted from the frames using the Enhanced Convolution Neural Network (ECNN). The optimal characteristics are then selected by a new Life Choice Based Optimizer (LCBO). Finally, the classification is carried out by the Deep Long Short-Term Memory (DLSTM) scheme. The implementation is performed on the Python platform, and the performances are evaluated using several performance metrics such as accuracy, precision, kappa, f1-score and recall. The performance of the proposed approach (ECNN-DLSTM) is compared with several deep and machine learning approaches and obtains an accuracy of 99% and a kappa of 96%.

Keywords Baby sign language · Automated recognition · Computer vision · Optimization

✉ Vamsidhar Enireddy
enireddy.vamsidhar@gmail.com

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, (D), Andhra Pradesh, Guntur 522502, India

[2] Department of Computer Science and Engineering, Malla Reddy Engineering College (Autonomous), Hyderabad, Telangana 500100, India

[3] Miracle Educational Society Group of Institutions, Miracle City, Andhra Pradesh 535216, India

[4] Department of CSE, RISE Krishna Sai Prakasam Group of Institutions, Ongole, Andhra Pradesh 523272, India

1 Introduction

In day-to-day life, communication plays a major role in sharing opinions, thoughts, emotions and information. This method utilizes verbal and non-verbal processes; verbal uses spoken words, and non-verbal utilizes posture and gestures of the people as the communication. The other form of communication language is a sign language which is most significant to large groups of individuals in society. It is also used as an effective communication medium with hearing-impaired persons and can greatly influence a child's ability to communicate, such as apraxia, Down syndrome, and autism spectrum disorder [30]. The variability in motion profile, the shape of a hand, and the situation of body parts such as hand and face contributing to each sign are dissimilar in all sign language. Therefore, visual sign language identification in computer vision is a difficult research zone [1]. Sign language recognition decomposed the obstacles for sign language operators in society. Several communication technologies have been developed to support written or spoken language [36].

Additionally, the structural way of hand gestures includes visual signs and motions, which are utilized to help the speech-impaired and deaf community with regular contact [25]. The face, body, head, arm, hand and fingers are different body parts utilized for sign language recognition. Expression/non-manual signals, location, hand shape, palm orientation, and movement are the five main parameters in sign language recognition [27].

In general, sign language is utilized by people with either speech or hearing disabilities to talk with others. Non-verbal contact is the communication of one's feelings and thoughts. The origins of sign language have been focused on several geographical and dialect locations [32]. But, sign language is overly huge and complex for special needs infants or kids whose speech improvement has not been initiated. Therefore, BSL usage proved more beneficial and served the purpose in such cases. The habit of crying for the baby's needs can be gradually replaced by non-verbal communication [9]. During the preverbal stage, the bonding of kids with their parents also improved. Several software packages provide online teaching of Sign language. Thus, needed precise software which understands sign language [5]. Deep learning consists of constructing consistent and reliable models to serve this purpose [6, 26, 34, 37].

Many research efforts in [4, 8, 10, 12, 20, 29, 33, 39] have recently been conducted for sign language recognition with deep learning. Developing a sign language recognition framework for translating sign sentences or words into the voice or text is considered one of the foremost challenges [28]. In addition, sign recognition from video sequences is one of the greatest substantial trials in behaviour understanding and computer vision because it produces the ability to know and identify human gestures for controlling some devices. Normally, gestures play a noticeable role in daily communication and frequently deliver gesticulating person's emotions in expressive body motion [18]. This type of research is motivated by three significant factors with complementary goals. First, the sign language framework is possibly advantageous in supporting communication among members of India's hearing communities and the deaf [24]. Second, advanced image processing is applied to create a spatial data presentation. For an automated system, simple optical image processing produced a vital response [3]. Third, a strong outcome is obtained by developing a framework with machine learning thoughts.

1.1 Motivation

Hand gestures change in orientation of the hand's shape and fingers. Therefore, non-linearity is one of the gesture properties that must be determined. Metadata information of gesture images can be utilized for recognizing the gesture. This process is the combination of two processes.

They are extracting features and classification. Before recognizing the gestures, the image features should be extracted. After that, these features must be classified. Hence, the major challenge is extracting and applying the features for classification.

Deep learning (DL) models are the branch of ML, and it has progressive networks which learn about the input provided during the training phase. Over the last two decades, numerous approaches have been developed to combine various automatic SL analysis methods. Recently, DL approaches have attained better results in various applications especially in SL. Unlike the existing model, DL can determine the hidden information in the original feature, which enhances the model's efficiency. Hence, for automatically recognizing gestures and hand signs of infants, this work uses Enhanced Convolution Neural Network (ECNN) with Deep Long Short-Term Memory (DLSTM) is used. The foremost contribution of this research work is discussed below:

- This paper presents automated recognition of BSL using ECNN with deep LSTM based classification. The BSL is a communication method among the mommies and their kids via gestures, obviously conveying their desires.
- In contrast, the proposed architecture adopts ECNN as a feature extraction module. The dilated based convolution has enhanced the CNN to enhance recognition accuracy.
- Feature selection influences the speed of analysis, and the proposed work utilizes a new metaheuristic algorithm named LCBO to select the optimal features. LCBO is the first developed algorithm for sign language recognition, which has not yet been implemented for feature selection. The LCBO algorithm follows an effective fitness function to select the best features.
- Finally, DLSTM for feature classification is introduced to identify the sign language and to improve the classification accuracy.

This research paper's remaining structure is organized: The following section 1. 2 gives a small review of recent related research works. Next, section 2 gives the proposed methodology, and section 3 gives the results discussion and performance analysis. Finally, the conclusion is given in section 4.

1.2 Recent related works: A review

This section reviews some recent works related to BSL recognition and recognition of sign language.

Sulochana Nadgeri and Arun Kumar [22] presented a baby sign language recognition. This scheme was an image texture-based strategy for classifying and understanding the BSL. Here, feature extraction was activated by the Gray Level Co-Occurrence Matrix operated on the still jpeg images for the sixty-odd baby sign's dataset. The machine learning approaches named random forest (RF), and K-nearest neighbour (KNN) were utilized for the effective classification. The introduced scheme achieved 73% of classification accuracy. The implementation outcomes clearly expressed the emotions and desires shown by the outcomes.

Runpeng Cui et al. [7] presented a continuous recognition of sign language (SL) system using DNN (deep neural networks). The SL sentence's videos are transcribed into ordered gloss label sequences. The main issue in existing literature pieces was the minimum ability to capture the temporal data. DCNN activated the feature selection module with stacked temporal merging layers, and bi-directional recurrent neural networks performed the sequence learning module. A relative improvement of more than 15% was achieved and outperformed the existing approaches.

A multi-modal dynamic recognition of the sign language approach was presented by Yanqiu Liao et al. [19]. This scheme depended on the BILSTM and deep 3-dimensional residual ConvNet. Potential issues in larger video sequences, low recognition accuracy and recognition of complex hand gestures were the drawbacks of the present dynamic sign language identification approaches. The B3D ResNet obtained the spatiotemporal features from the video sequences. After feature analysis, a middle score was recognized for each action in the video sequence. The proposed approach achieved 89.8% and 86.9% recognition accuracy over DEVISIGN_D and SLR_Dataset.

Javed Imran and Balasubramanian Raman [13] presented a scheme without the requirement of hand segmentation for the automatic recognition of sign language. Initially, there are three different motion templates: images of RGB motion, motion history, and dynamic. These three templates fine-tuned three ConvNets trained over the dataset of ImageNet. Besides, fine-tuning prevents learning entire parameters from scratch. Depending on the Kernel-based extreme learning machine (KELM), a fusion technique was introduced to combine the output of three ConvNets. ConvNet-based deep features accompanied by the introduced KELM-based fusion were strong for any kind of human motion recognition. A large relative enhancement of over 15% was achieved and shown by the implementation outcomes.

Walaa Aly et al. [2] introduced a user independent recognition framework. This recognition system was utilized for the American Sign-Language alphabet with depth images. Because of the robustness of background and illumination distinctions, many problems were avoided by exploiting depth information. Rather than the actual hand-crafted feature extraction approaches, CNN architectures based principle component analysis (PCA) feature extraction was applied for the best feature extraction. The linear SVM classifier recognized the obtained features. The introduced scheme outperforms the experimental outcomes compared to the existing recognition accuracy.

Sarfaraz Masood et al. presented the real-time identification of sign language gestures from video sequences [21]. Both the spatial and temporal features were presented in the video sequences. The inception model named deep CNN trained the spatial features. The temporal features were trained by a recurrent neural network (RNN). The introduced scheme for the large set of images obtained 95.2% of accuracy.

Saunders et al. [31] proposed a 3D-multi-channel SL using mixture density and new progressive transformer networks. The transformer network introduced a counter decoding that makes variable-length continuous sequence generation. This work presented a data augmentation procedure for reducing prediction drift, and mixture density was used for producing expressive and realistic sign sequences. This work was evaluated on the Phoenix14k T dataset and obtained better results.

Kowdiki and Khaparde [17] introduced automatic hand gesture identification by hybrid classification models. The input image was pre-processed by converting the image into grayscale, and the image was enhanced by histogram equalization. Then the image was segmented by canny edge detection and active contour processes. Then, the feature was extracted and selected using optimal features. The classifier neural network was used to classify the hand gesture. The optimization technique DH-GWO (Deer Hunting- Grey Wolf Optimization) was used for optimal feature selection and weight updation of the classifier.

Wangchuk et al. [35] introduced the Bhutanese SL digits recognition model using CNN. This work has four phases pre-processing, feature extraction and recognition. The dataset was generated and implemented using several SL models. Initially, the image was pre-processed and augmented. Then, the features were extracted and classified using CNN. The accuracy of training and testing obtained by the CNN model was 99.9% and 97.6%.

Gao et al. [11] introduced Chinese SL recognition using RNN with the transducer. This work's visual hierarchy transcription was designed to capture the video's spatial and temporal features. Then, a lexical prediction network was used to extract efficient contextual information. Finally, RNN with a transducer was used for learning to map among the sequence video features and sentence level labels. This model was evaluated on the CSL dataset and achieved an accuracy of 0.914.

Kamruzzaman [14] introduced SL recognition and the generation of Arabic speech using CNN. The major aim of this research was to introduce a method for persons who have a disability in speech and to reduce the complexities of sign language. This work was also effectively utilized to recognize hand gestures for HCI (Human-computer interaction). But, the model was in the initial phase and required accuracy improvement. Table 1 presents the advantages and challenges of the existing works.

In most existing works, hand-crafted features were used for extracting the features. These manual extraction processes were complex, and every meaningful feature may not be extracted. Even though many research works are presented in SL, the research in BSL is limited. Further, the DL models used in the existing works suffered from the overfitting issue and have high computational complexity. Hence, in order to overcome these challenges, an enhanced CNN with the DLSTM approach is introduced in this paper.

2 Proposed methodology

A system of the BSL recognition system is proposed in this paper. Sign language enables one to talk with naturally emerging hearing babies before they can use spoken communication reliably and easily, known as infant signing. As suggested by the BSL research, it is a promising type of communication that positively affects children's socio-emotional development. The researchers can observe infants more closely by familiarizing communication with a much younger. Teaching infants signs enhanced the emotional and cognitive development found by research on BSL. BSL upsurges the verbal development rate and can improve the child's IQ far from slowing down speech together. Figure 1 demonstrates the system architecture of the BSL recognition model. It follows the flow like the frames extracted from the captured video. Figure 1 is the basic system architecture for sign language identification. The sign language is recognized from the video or images based on feature classification and extraction.

Initially, the BSL video is captured and then extracted the frames from the video, and the frames are transmitted to the pre-processing. The modified CNN is utilized after the pre-processing is done to convert the images into feature vectors, and then the features are transmitted to the feature selection. The high-level features are selected by the new optimization named Life-choice based optimizer. The selected features are sent to the deep LSTM. The existing works utilize the CNN-LSTM [38] in Chinese sign language recognition. The recorded BSL videos derived the labelled images with modified CNN. Before the flattened layer creates the vectorized image representations, the modified CNN model implements the subsampling and convolution layers. The respective vectors are consecutively fed into the deep LSTM for a BSL video. In order to identify the gesture, these vectors are utilized in the deep LSTM. Figure 2 shows the System Architecture of the proposed ECNN-DLSTM based BSL Recognition.

Table 1 Advantages and challenges of the existing works

Authors and Citation	Methods	Advantages	Challenges
Sulochana Nadgeri and Arun Kumar [22]	RF and KNN	This model achieved better accuracy of 73%	Depends on the hand-crafted feature extraction model
Runpeng Cui et al. [7]	DNN	Provided better representations for the gestures	Multiple modalities require high exploration
Yanqiu Liao et al. [19]	BILSTM and deep 3-dimensional residual ConvNet	Improved accuracy and SL recognition	Doesn't recognize the complex SL
Javed Imran and Balasubramanian Raman [13]	ConvNet-KELM	More accurate	In some cases, this model achieved poor results, and the computational time was high
Walaa Aly et al. [2]	PCA	This model doesn't require labelled data and doesn't need additional GPU	Increasing the number of convolutional layers may increase the space and time complexities
Sarfaraz Masood et al. [21]	CNN-RNN	Efficiently extracted the spatial and temporal features	Has overfitting issues
Saunders et al. [31]	Transformer networks	Generated more expressive and realistic sign pose	More number of misclassification occurs
Kowdiki and Khaparde [17]	DH-GWO	Provided useful information with less correlation	Rely on hand-crafted features and suffer from the convergence issue
Wangchuk et al. [35]	CNN	Fast convergence and overfitting were reduced	Has high computational complexity
Gao et al. [11]	RNN	Efficiently learned the alignment among label and video sequences with variable length	The training process was slow
Kamruzzaman [14]	CNN	Presented an effective model for patients with hearing disability	Recognition speed was low

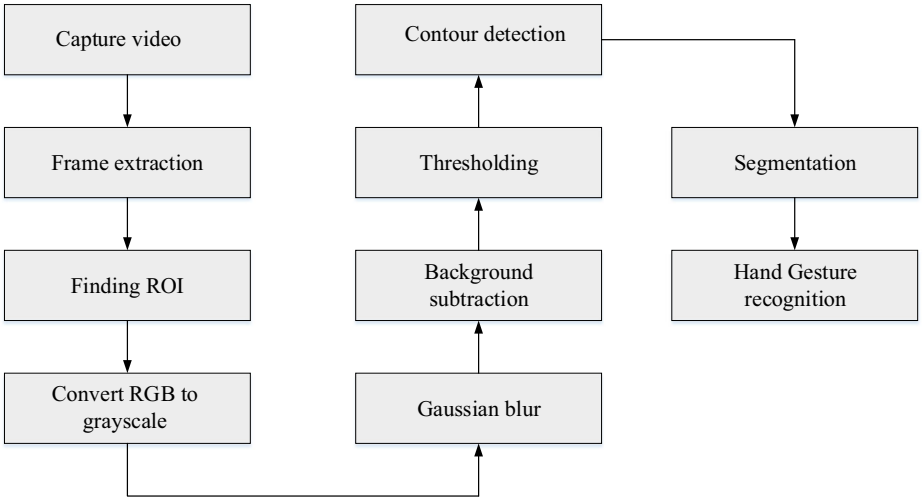

Fig. 1 Sign language Recognition model's System Structural design

2.1 Pre-processing of data

The video is changed into frames once the video is taken by the camera device and pre-processed. The images with hand gestures, faces, and various circumstances are changed into the format needed for the system during the pre-processing phase. Every image is pre-processed by appealing to the region in bounding box coordinates.

In a real-time situation, to identify the hand region, the video recorded by the baby monitor to identify the gesture made by an infant is pre-processed. During pre-processing, the frames are recovered from each video, the coordinates of the region of interest (ROI) are determined, and the image in that region is extracted. The RGB image is changed into a Gray scale image for easy and further processing. The Gaussian Blur filtering is applied with zero standard deviation and 7 by 7 kernel to apply image smoothing.

$$G(a,b) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \tag{1}$$

where σ is the standard deviation of Gaussian distribution, a and b are the distance from the origin in the horizontal and vertical axes.

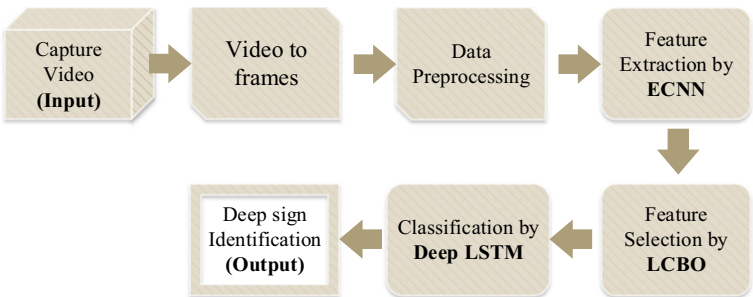

Fig. 2 System Architecture of proposed ECNN-DLSTM based BSL Recognition

The background subtraction approach is the computer vision approach that stores only the ROI in the dataset for further processing. This approach segments the hand (foreground) from the background. Initially, when the application runs, it obtains background and develops the background model. Then, when the hand is placed in ROI, it utilizes the background model for segmenting the foreground from the background. The background subtraction algorithm determines the absolute difference between the current and the background frame. The running average scheme finds a new gesture made in the video sequence.

$$dst(a,b) = (1-\beta).dst(a,b) + \beta.src(a,b) \tag{2}$$

Where *dst* is the destination image, *src* is the source image, β represents the weight of the input image; $\beta = 0.5$ this term decides the updating speed. Simple thresholding is activated with a threshold limit of 25 and a maximum value of 255.

$$dist\,(a,b) = \begin{cases} 1\,when\,src\,(a,b) > threshold \\ 0\,when\,src\,(a,b) \le threshold \end{cases} \tag{3}$$

The maximum contour region is made, and the threshold image achieves the contours. The segmented image recognizes the gesture made by the baby.

2.2 Augmentation of data

image augmentation is performed at the time of the training phase. In order to enhance the accuracy, the model should get trained on more images than the actual images. In addition, it ensures the model is not overfitted with a unique type of image.

2.3 Enhanced CNN based feature extraction

This section introduces an enhanced convolution network design to recognize sign language. The actual CNN is enhanced by the dilated-based convolution named ECNN, which enhances recognition accuracy. In recognizing the sign language framework, extracting good features is the most significant step. Several approaches are introduced to implement the recognition of the sign language system, and most schemes are based on hand-crafted features. In computer vision problems, the features learned by CNN have gained much attention due to the deep learning algorithm's success in being applied to the sign language recognition framework. The CNN models have overfitting issues and high computational complexity. Hence, the accuracy may reduce, and the results may degrade when the size of the network increase; In this scheme, some hidden and deep features are extracted to extract the meaningful features; hence, the classification efficiency is enhanced.

The new dilated convolution based network presents a fully connected layer and dilated convolution layer (Dilat-Net). This network extracts features from the pre-processed image. Since the video sequence has both spatial and temporal features, it is used for extracting the spatial-temporal features. Dilat-Net with the receptive field is trained for learning features of every sign, and this network extracts the features using CNN's pooling and convolution layer. The initial layers are used to extract the low-level features, and the final layers are used to extract high-level features. Hence, efficient and fast recognition has been achieved by expanding the receptive convolution kernel field. The Dilat-Net has several merits for sign language recognition compared with other CNN. Initially, the actual convolution layer exchanged by the Dilat convolution layer in Dilat-Net minimizes the computation cost. The

weight in the original convolution layer matched to 1 × 1 size area in the actual image, whereas in the Dilat-Net, it matched to 2 × 2 area of size. It shows that the Dilat-Net has a superior interesting field than the normal convolution, even with similar weight constraints. In addition, the Dilat convolution has fewer network weight parameters if the dilated and actual convolution networks have similar receptive fields.

Secondly, the Dilat-Net removes the pooling layer since each pixel's data in the low-resolution image has a key role in the identification process. It ensures that the image data is not lost during the recognition procedure and produces enhanced recognition accuracy. While growing the receptive arena, the pooling layer loses the image data. But, the Dilat-Net expands the receptive field without losing image data. Relative to the pooling layer, the Dilat convolution with 3 × 3 receptive fields has no loss of image data. The Dilat-Net has a very modest network structure relating to other CNN that would cut down the network training's computation cost. Figure 3 demonstrates Dilat-Net's structure introduced for sign language recognition.

The input layer is the first layer of the network. The CNN framework produces 128 size vector representations. The whole input image has a size of 256 × 256 × 1. A dilated convolution layer 1 is the second layer and produces 16 feature maps using a 3 × 3 filter and the dilation rate 1. Dilated convolution layer 2 again produces 16 feature maps using a 3 × 3 filter and the dilation rate 2. The third layer is the fully connected layer, which more accurately recognizes the features and creates the feature data obtained from the dilat convolution layer. ReLu, as the activation function adopted in the Dilat-Nets, has a minimum cost and less computation time than the sigmoid function.

Rectified linear unit (RLU) It is followed by an activation function and the convolution layers application in CNNs. When the input is 0, the output value has zero; otherwise, the output is similar to the input value but is like a linear function. Eq. 4 represents the mathematical form;

$$f(a) = max(a, 0) \tag{4}$$

Softmax layer In the case of multi-class generalization, the Softmax function is deliberated as the logistic sigmoid function. Softmax is the implemented activation function, represented as:

$$P\left(\frac{C_r}{u}\right) = \frac{e^{a_r}}{\sum_j^k e^{a_j}} \tag{5}$$

The satisfied conditions are $0 \le P\left(\frac{C_r}{u}\right) \le 1$ and $\sum_{j=1}^{k} P\left(\frac{C_j}{u}\right) = 1$. Where, $P(C_j)$ indicates the probability of class prior and $P\left(\frac{u}{C_j}\right)$ denotes the conditional probability given class$r$; $a_r = \ln\left(P\left(\frac{u}{C_j}\right) P(C_r)\right)$.

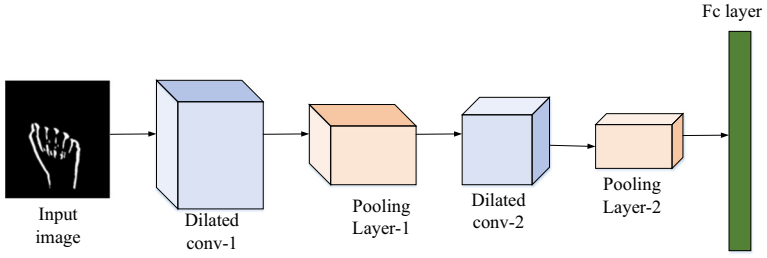

Fig. 3 Architecture of a Dilated-CNN (ECNN)

2.4 New LCBO based feature selection

Feature selection is the approach utilized to obtain the maximum classification performance, which is determining the fewer features between the redundant feature spaces. In this work, the LCBO [15] method is employed in the feature selection process to choose the ideal feature subset with a minimum fitness function. In order to choose the optimal feature combination, the proposed LCBO algorithm is utilized; while reducing the number of selected features, the classification accuracy is maximized. The LCBO algorithm imitates the life cycle of a human being, which is also inspired by an existing but recent algorithm named Jaya that uses selective influence.

2.4.1 Learning from the mutual greatest group

The optimal feature learned by the LCBO algorithms for an assumed population, represented as X with arranged fitness values:

$$\chi_a^{'} = \frac{\sum_{h=1}^{N} (r(h) * \chi_h)}{N} \tag{6}$$

The parameters in the algorithm are represented as N. The parameter χ_a indicates the a^{th} search agent in the process and $\chi_a^{'}$ indicates that χ_a is updated if it has the optimal fitness.

2.4.2 Fitness evaluation

The multi-objective optimization issue is feature selection, which considers the maximum classification accuracy and the minimum number of feature selections. The highest classification accuracy and a minimum number of features are considered the best solution for the feature selection issue. The proposed fitness function assessed every solution that depends on the KNN classifier for evaluating its classification accuracy and the number of features chosen.

$$fitness = \alpha \gamma_R(E) + \beta \frac{|S|}{|W|} \tag{7}$$

Where, $\alpha$and β are the parameters, represented as $\alpha = [0, 1]$ and $\beta = (1 - \alpha)$, respectively. $\gamma_R(E)$ denoted the KNN classifier's error rate, $|S|$ represented the selected subset of features and $|W|$denoted the entire features from the feature extraction.

2.4.3 Knowing very next best

One must be clever to realize the present location instead of entirely focusing on massive targets. The current position to a better position is very important; hence, the present target is prioritized.

$$\eta_1 = 1 - \frac{(presentchances - 1)}{(totalchances - 1)} \tag{8}$$

$$\eta_2 = 1 - \eta_1 \tag{9}$$

$$bestDiff = \eta_1 * C_1 * (\chi_1 - \chi_a) \tag{10}$$

$$betterDiff = \eta_2 * C_1 * (\chi_{a-1} - \chi_a) \tag{11}$$

$$\chi_a^{'} = \chi_a + rand() * betterDiff + rand() * bestDiff \tag{12}$$

Where, η_1 and η_2 are linearly varied from 0 to 1 and 1 to 0, individually. C_1 denoted the constant value (C_1 = 2.35), *rand*() is the random number and χ_{a-1} denoted the search agent's position. χ_1 denoted the best position of the search agent. χ_a's position will only be updated to $\chi_a^{'}$ if $\chi_a^{'}$ has better fitness than χ_a.

2.4.4 Reviewing mistakes

The method defined by Eq. 13 is the Avi escape scheme and has been utilized as a generalized scheme to improve the exploration of an algorithm.

$$\chi_a^{'} = \chi_{\max} - (\chi_a - \chi_{\min}) * rand() \tag{13}$$

Where, $\chi_{\min}$ and $\chi_{\max}$ are the lower and upper bound values, correspondingly.

The introduced LCBO is initiated with the number of iterations, upper and lower bounds and size of the population. Initially, the populated is created, and equivalent fitness is calculated by eq. (7). Until the total iterations are exhausted or the target fitness has been obtained, the feature's positions and fitness are updated iteratively.

Pseudocode of the LCBO

Input: Extracted features
Output: Optimal features
Initialize the population $\chi_a, (a = 1, 2, ... population)$ //features
Initialize $presentchance, C_1$
Evaluate N
Evaluate the fitness value using equation (7) and sort the population
While $(presentchance < totalchances)$
For every feature **do**
  $Z = r(\ )$
  if $(Z > 0.875)$
    update present feature by equation (6)
  else if $(Z < 0.70)$
    update η_1, η_2 by equations (8) and (9)
    update present feature by equations (10, 11 and 12)
  else
    update present feature by equation (13)
  end if
end for
  $\chi = sort(\chi)$
  $presentchance = presentchance + 1$
end while
return χ_1

The optimal feature set is selected, and the selected features are forwarded to the deep LSTM for classification.

2.5 Deep LSTM based classification

An enhanced model of the standard LSTM is called deep LSTM with several hidden LSTM layers. The additional hidden layers learn the process of the prior layers. An alternative outcome that requires minimum neurons and faster training can be achieved by increasing the network's depth. Figure 4 demonstrates the architecture of deep LSTM. The LSTM network has the vanishing gradient problem, and it is an extended model of the RNN. The information is stored in the memory blocks, and the data is accessed for a long time.

The expressions for input gate, forget gate and output gate of the LSTM at time t can be described as

$$i_t = \sigma_i(W_{xi}x_i + W_{hi}h_{t-1} + W_{ci}c_{t-1}) \tag{14}$$

$$o_t = \sigma_o(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1}) \tag{15}$$

$$f_t = \sigma_f\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1}\right) \tag{16}$$

$$c_t = f_t c_{t-1} + i_t\sigma_c(W_{xc}x_t + W_{hc}h_{t-1}) \tag{17}$$

$$h_t = o_t\sigma_t(c_t) \tag{18}$$

Where, Windicates the weight allocated to the link between two network units; σindicates the non-linearity function;$x_t$denotes the input vector; the forget gate, output gate, hidden state vector, and

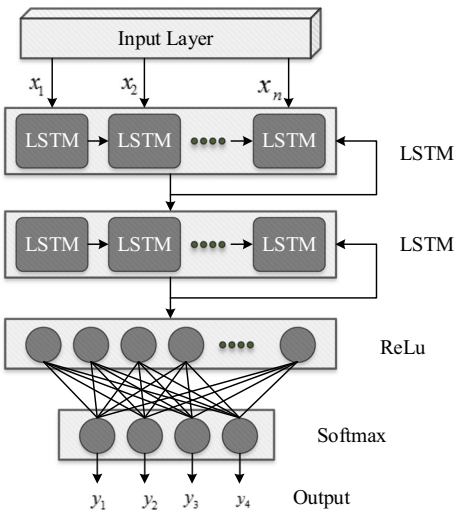

Fig. 4 Architecture of Deep LSTM Networks

input gate are represented as f_t, o_t, h_t *and* i_t, respectively. The softmax function estimates the distribution by imposing the likelihood loss as the cost function. It is expressed as:

$$y = soft\max(WT + b) \tag{19}$$

where $T$is a feature vector and $b$is the bias value. In the training process, the total loss of the model at every iteration is computed using cross-entropy and L_2regularization. It is expressed as:

$$Loss = -\frac{1}{p}\sum_{j=1}^{p} t_j \log\left(\rho(y)\right) + \alpha\|\theta^2\| \tag{20}$$

Where t is the true label probability, $\rho(y)$ is the probability of every class with the softmax, p is the number of targets and α is the L_2 regularization parameter.

Initially, the weights are selected by the uniform random numbers in the LSTM network. In the hidden layer, the function of activation *tanh* is applied. To overcome the over fitting issue, a dropout with a rate of 0.5 is employed. To update the weights, backpropagation is implemented. Overfitting is one of the most significant issues in DNN, whereas dropout is one of the latest significant methods to avoid overfitting. During the training process, the units are randomly dropped out at the specified rates at the dropout layer. This approach minimizes overfitting by preventing the units from being too compliant with each other. Because of this, the dropout was placed in two different layers in the designed networks. Finally, the classified results were received for recognition.

3 Implementation results and discussion

This section explains the discussion of the implementation results and the performance analysis of the proposed approach with the existing scheme. The implementation is activated on the Python 3.6 (TensorFlow) platform. The performances are compared with several existing algorithms such as Naïve Bayes, Random Forest, KNN, SVM, VGG16, BILSTM and SVM with RBF kernel. Each machine learning algorithm is implemented for our proposed dataset and obtains the performance outcomes. A brief explanation of existing algorithms is given in below:

a) ***Naïve Bayes:*** *One of the best common probabilistic approaches to predict. To train the model, parameters such as variance, mean and a limited dataset are enough for this approach. It also provides better outcomes and is easy to interpret* [18].
b) ***Random Forest:*** *It is successfully introduced for regression and classification schemes* [24]. *This ensemble model predicts the results by evaluating the average of several independent base models' predictions* [3]. *In this model, no pruning is utilized, and each tree is established to its conceivable extent. The random forest produces better performance and less generalization error rate than the tree classifiers.*
c) ***KNN:*** *It is one of the easiest techniques and algorithms in machine learning* [7, 22]. *An item is labelled by the algorithm using the majority of votes from the neighbors. When* $k = 1$ *then assigning an entity to its nearby neighbour, take the most care in assigning a K value.*
d) ***SVM:*** *It is mainly utilized to classify binary placed in a hyperplane that splits the data into two classes* [2]. *The kernel feature is utilized in this approach, which transmits the*

data into a D dimensional space and is separable; so, the data is not linearly separable because SVM turns the information. Hence, computational issues are caused by the data transmitted into such space and over-fitting issues. The simple dot product is utilized in SVM, and there is no requirement to handle the data openly; the simple dot product removes the concerns.

e) ***VGG16:*** *From the University of Oxford, K. Simonyan and A. Zisserman developed the VGG16* [13]. *The VGG16 was arranged and employed NVIDIA Titan Black GPUs for a considerable length of time. The initial convolution layer has a fixed size of 224 × 224 RGB images* [16]. *VGG16 is a large network, and slow to train are the main disadvantages.*

f) ***SVM and RBF kernel:*** *The RBF kernel is the foremost popular kernel among all the kernels in SVM. Whole tweet scores are in parametric form, which means different penalties of existing SVM-RBF at different runtimes.*

3.1 Dataset description: Personalized BSL dataset

The data set is created by capturing the videos of people playing the BSL. Each individual's video is recorded for 20 frames and then pre-processed with the respective frames and protected with the respective sign name as the folder name into the folder. A total of 20 subjects are taken to create the dataset. The dataset is generated for 6 baby signs representing Play, More, Water, Yes, Hungry, and Mom. Kids and babies with special needs are required these baby signs and are the most basic requirements. So, the most basic dataset is created with the essential signs. Finally, 100 images of each sign were presented in the dataset and split into 70% for training, validation and 30% for testing.

Figure 5 demonstrates the common sign samples utilized for the sign language of infants (0 to 1 year). Figure 6 demonstrates the basic sign language images for our proposed scheme and customized images. RGB images of one signer, which indicates a certain sign using a hand, are recorded using the digital Nikon D3400 camera. Each sign is carried out by ten time by one signer by changing the orientation and lighting conditions.

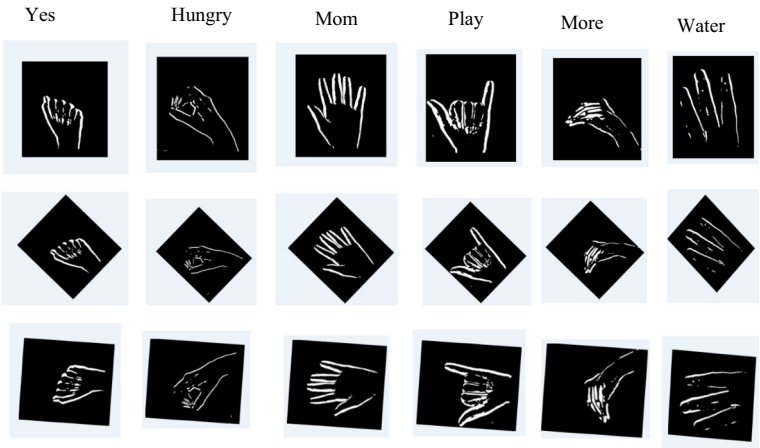

Fig. 5 Basic sign sample images utilized in the BSL

Fig. 6 Sample Images of Baby Sign Language from the Dataset

3.2 Performance metrics

The following are the performance metrics used to evaluate the performance of the proposed and existing systems.

Accuracy A evaluation of the correctness of the detection is known as accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

Precision The ratio of projected positives to real positives is precision.

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

Recall Total positives recognized accurately are known as recall.

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

F1-score The vocal means of precision and recall is the F1-score.

$$F{-}measure = \frac{2TP}{2TP + FP + FN} \tag{24}$$

Kappa The kappa performance metric compares the accuracy of observed and expected classifiers.

3.3 Performance results analysis

This section compares the performance outcomes obtained by the proposed scheme. The proposed approach is evaluated using performance metrics like accuracy, kappa, recall, f1-score, and precision. Besides, the performances of the proposed approach are compared with several machine learning algorithms named Naïve Bayes, Random Forest, KNN, SVM, VGG16, BILSTM and SVM with RBF kernel.

Figure 7 gives a few other sample outcomes generated by the proposed scheme. Most signs are predicted correctly, whereas few are misclassified. In Fig. 7a, 94% of the images are classified as hungry, 100% of the images are classified as a mom, 96% of the images are classified as yes, 100% of the images are classified as water, 98% of the images are classified as more and 99% of the images are classified as play. In Fig. 7b, 90% of the images are classified as hungry, 95% of the images are classified as a mom, 93% of the images are classified as yes, 94% of the images are classified as water, 94% of the images are classified as more and 93% of the images are classified as play.

Table 2 gives the implementation results on the BSL dataset. The transfer learning model VGG16 achieves 94% accuracy. The accuracy has increased to 98% with the implementation of the proposed model. Compared with existing machine learning models, the proposed scheme achieved 97% precision, 99% recall, 98% F1-score, 96% kappa and 98% testing accuracy. Further, in this comparison, The SVM and Naïve Bayes classifier have a high computational time of 1.17 and 5.43 s, respectively. Moreover, the proposed ECNN-DLSTM achieves the lowest computational time of 0.75 s compared to other approaches. It shows that the proposed method has less computational complexity.

Table 3 represents the performance of the ECNN-LSTM model with and without the augmentation process. It is shown that the precision and recall are only 96.2% and 92% before the augmentation, respectively. After augmentation, the proposed ECNN-LSTM obtains better precision and recall values of 97% and 99%. Hence, it is proved that when compared to without augmentation, with augmentation achieves better outcomes. Table 4

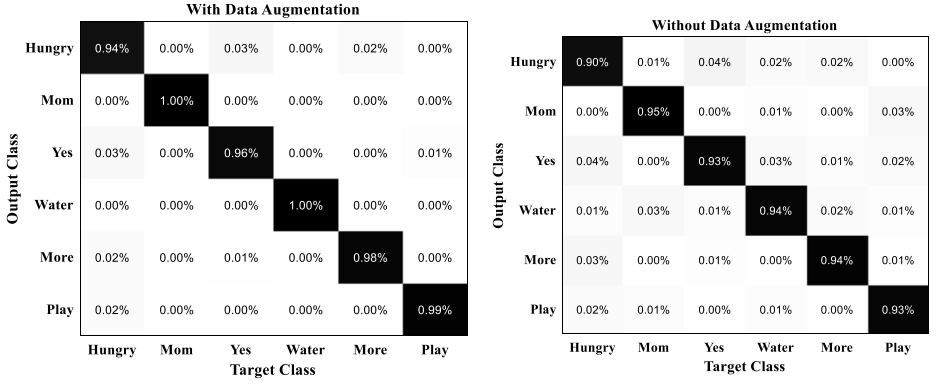

Fig. 7 Confusion matrix of ECNN-DLSTM of (**a**) with augmentation (**b**) without augmentation

Table 2 Implementation Calculation over the BSL Dataset

Architecture Models	Train Accuracy	Test Accuracy	Kappa	F1-Score	Precision	Recall	Time (s)
Proposed Approach (ECNN-LSTM)	99	98	96	98	97	99	0.75
BILSTM	98	96.7	94	97	96	98.3	3.18
VGG16	96	94	88	94	92	96	1.89
SVM+RBF kernel	92	89	83	89	84	95	7.32
SVM	90	87	80	87	82	93	5.43
KNN	84	76	70	76	71	82	0.98
Random Forest	90	85	73	85	80	91	1.26
Naïve Bayes	82	75	62	75	71	79	1.17

presents the performance of ECNN-LSTM with and without pre-processing. The metrics like train accuracy, kappa and F1-score are compared in this comparison. When compared to without pre-processing, the pre-processing with augmentation obtained better performance. Hence, it is proved that pre-processing is essential for BSL.

Figure 8 represents the accuracy and loss curves of the proposed ECNN-DLSTM model. Here, the epoch value is varied for 150 iterations, and the performances are evaluated. In Fig. 8a, when the value of the epoch is increased, the accuracy is also increased. The training accuracy for data augmentation is approximately 100% after the 25th iteration. Similarly, the loss is approximately 0.1 for the training and validation losses. However, in Fig. 8b, the validation accuracy (78%) and loss (0.24) are less than with data augmentation outcomes. Hence, it is proved that the proposed ECNN-DLSTM achieved better results when the data augmentation was applied to the dataset.

Fig. 9 demonstrates the recall performance comparison of the proposed approach using existing algorithms such as Naïve Bayes, KNN, Random Forest, SVM, VGG16, BILSTM and SVM with RBF kernel. The proposed scheme achieved a maximum of 99% recall performance than the others. The detection accuracy of the proposed approach is high because the optimal features are utilized for classification. Based on the selected features, an optimal classification is done using DLSTM, improving detection accuracy. The existing VGG16 and BILSTM models achieved 96% and 98.3% recall. Compared with VGG16 and BILSTM, the proposed method achieves a 3% improvement in accuracy metric.

Table 3 Performance of ECNN-LSTM with and without augmentation

Architecture Models	Train Accuracy	Test Accuracy	Kappa	F1-Score	Precision	Recall
ECNN-LSTM (with augmentation)	99	98	96	98	97	99
ECNN-LSTM (without augmentation)	92	95	95	97	96.2	92

Table 4 Performance of ECNN-LSTM with and without pre-processing

Methods	Train Accuracy	Kappa	F1-Score
Pre-processing+augmentation	99	96	98
Without pre-processing+augmentation	97.2	97.5	96

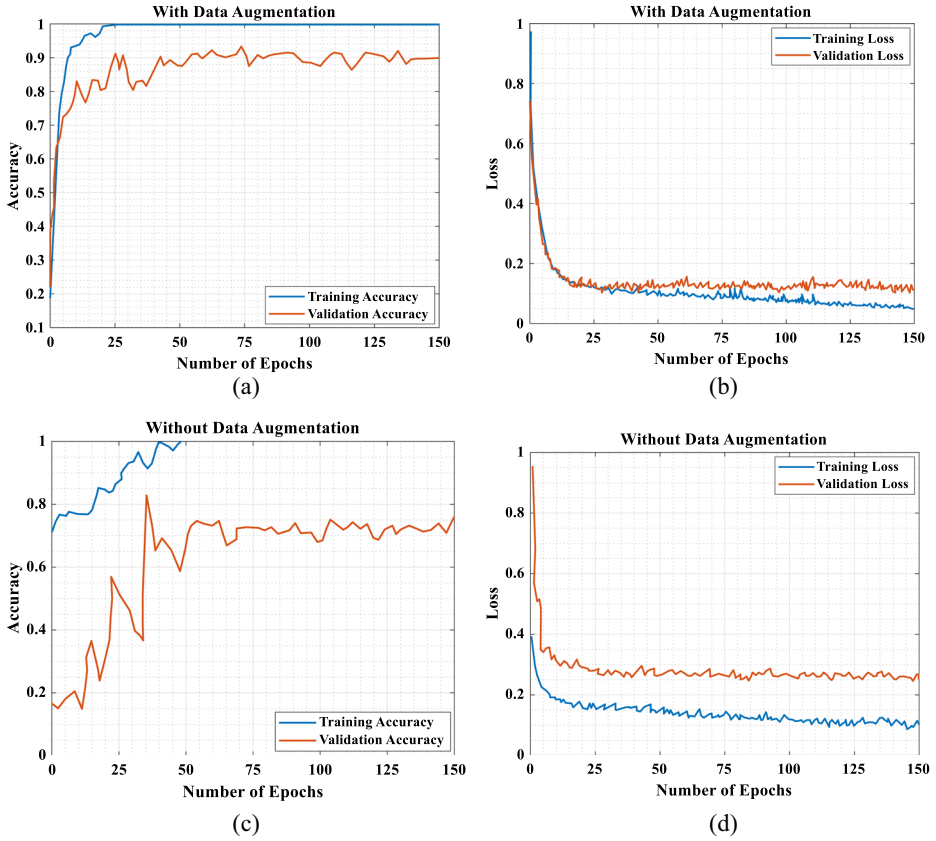

Fig. 8 Accuracy and loss curves of (**a**) with augmentation (**b**) without augmentation

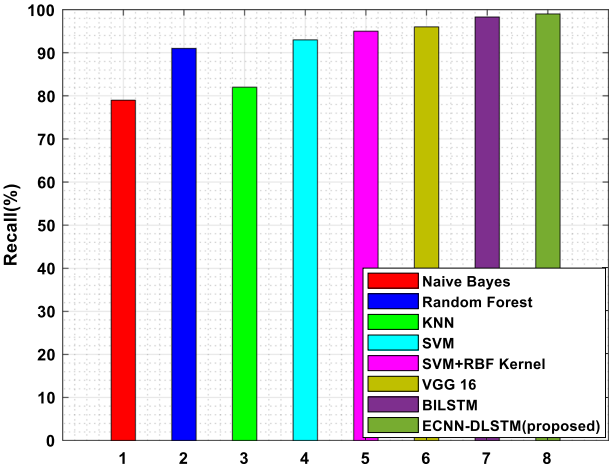

Fig. 9 Performance of Recall

Figure 10 illustrates the F1-score performance comparison of the proposed approach with existing schemes. The proposed (ECNN-DLSTM) scheme achieved a 98% f1-score value, and the existing algorithms such as Naïve Bayes, Random Forest, KNN, SVM, VGG16, BILSTM and SVM with RBF kernel achieved 75%, 85%, 76%, 87%, 94%, 97% and 89%, respectively. The F1-score measure is the vocal means of recall and precision. The proposed scheme's precision and recall value are more than other approaches. Compared to VGG16 and BILSTM, the introduced scheme achieved a 4% more f1-score performance.

The Kappa performance comparison of the proposed approach with existing algorithms is illustrated in Fig. 11. It is the relation between the observed and expected accuracy. The training and testing accuracy of the proposed scheme is high compared to other existing approaches named Naïve Bayes, Random Forest, KNN, SVM, VGG16, BILSTM and SVM with RBF kernel. The proposed strategy achieved a maximum of 98% kappa performance than the others. The existing machine learning algorithms obtained 62%, 73%, 70%, 80%, 88, 94% and 83% kappa performance values.

Figure 12 illustrates the comparative precision and accuracy performance analysis with classifier models. ECNN-DLSTM achieved a 99% maximum accuracy with 97% precision to all other classification models implemented. Thus, the proposed scheme is better for implementing infant sign language identification. Table 5 compares ECNN-DLSTM with recent research works on CNN and MobileNet models. The performance measures like accuracy, F1-score, precision and recall are compared. The comparison proves that the proposed ECNN-DLSTM achieves better accuracy due to the better optimal selection by LCBO. Hence, the experimental results prove that the proposed ECNN-DLSTM model can efficiently utilize BSL recognition.

3.4 Discussion

The simple and humble set of signs used by infants and babies to convey their desires and emotions before communicating orally is known as Baby Sign Language (BSL). Besides, it helps infants with speech weakening concerns who cannot study the adult's huge and complex sign language. It is very hard to know the signs made by babies who practice BSL for non-sign

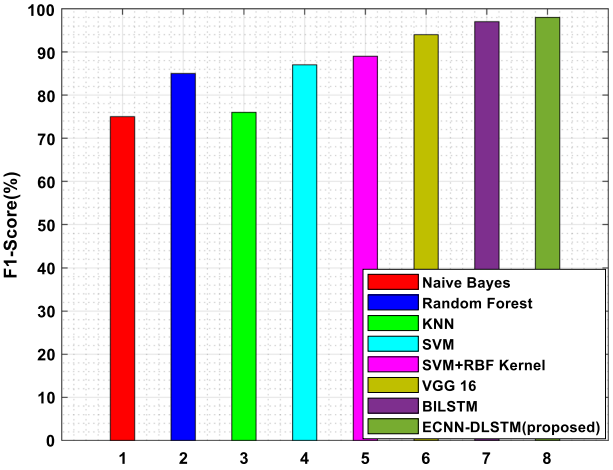

Fig. 10 Performance of F1-score

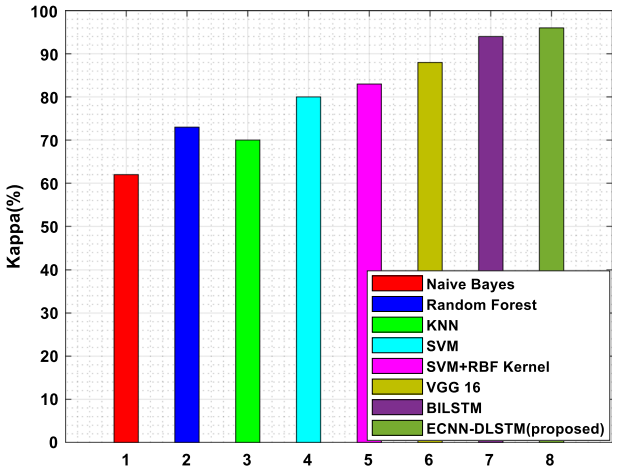

Fig. 11 Performance of Kappa

language users. This work undergoes pre-processing, feature extraction, feature selection, and classification stages. The pre-processing processes are carried out, and the efficient features are extracted by the DL model ECNN. Then the metaheuristic optimization LCBO is used for selecting the optimal features. Finally, classification plays a major role in SL since it takes the extracted features as the input and recognizes the exact gesture. The machine learning models have less generalization ability and are trapped by local optima. The deep learning-based model LSTM is used to recognize baby sign language to overcome these issues.

Since there are no public datasets available for research work to be carried out, the own dataset is generated using random signs. The performances like accuracy, kappa, recall, f1-score, and precision are evaluated for the methods like Naïve Bayes, Random Forest, KNN, SVM, VGG16, BILSTM and SVM with RBF kernel. Initially, the performances without and with augmentation results are carried out. The accuracy and loss curve of the proposed model

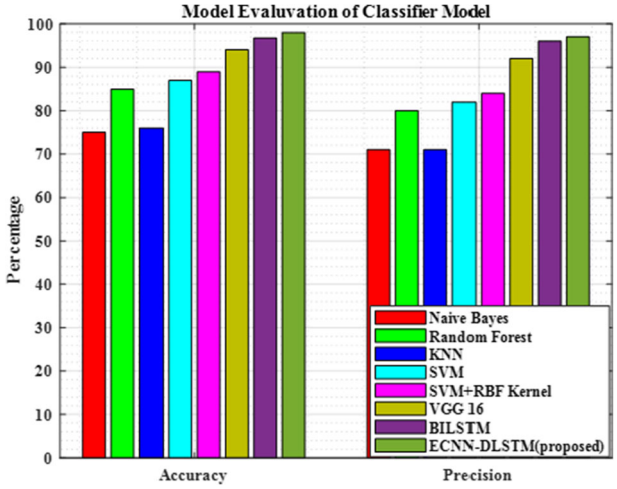

Fig. 12 Precision and Accuracy comparative analysis of classifier models

Table 5 Comparison of ECNN-DLSTM with recent research works

Methods	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Proposed (ECNN-DLSTM)	99	98	97	99
CNN [35]	–	98	98	98
Mobilenet [23]	85.7	85.08	88.2	–

is compared; in this comparison with augmentation, results are better than without the augmentation process.

4 Conclusion

This research presents an optimized automatic recognition of six basic signs commonly used by infants. In contrast with existing research, this paper presented a deep learning based classification, and the enhanced CNN extracts the features. The proposed scheme introduced a best feature selection module and is activated by a novel LCBO. Numerous machine learning systems like Random Forest, SVM, Naïve Bayes, VGG16, SVM (RBF) and KNN are implemented with the proposed scheme and compared with the results of existing models, the proposed strategy achieved maximum outcomes. A dynamic solution of the proposed ECNN-DLSTM approach is presented, and the combination produces better accuracy due to enhanced feature extraction and recognition. Besides, the optimal features for the best classification are selected by the novel LCBO. The proposed scheme achieved 99% accuracy compared to other classification methods, as shown by the results. Finally, the proposed scheme is the best option for creating a real-time application to screen the babies and communicate better with them. In future, large datasets will be used to improve the accuracy and quality of the model. Further, a new optimization technique will be integrated with the LSTM model to optimize the network's weight, achieving better performance.

Funding No funding is provided for the preparation of the manuscript.

Data availability Data sharing does not apply to this article

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

1. Albanie S, Varol G, Momeni L, Afouras T, Chung JS, Fox N and Zisserman A (2020) BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. arXiv preprint arXiv:2007.12131
2. Aly W, Aly S, Almotairi S (2019) User-independent American sign-language alphabet recognition based on depth image and PCANet features. IEEE Access 7:123138–123150
3. Arora M, Mehta P, Mittal D, Bajaj P (2020) Word-level sign language gesture prediction under different conditions. In: international conference on innovative computing and communications. Springer, Singapore: pp. 427-435

4. Asadi-Aghbolaghi M, Clapés A, Bellantonio M, Jair Escalante H, Ponce-López V, Baró X, Guyon I, Kasaei S, Escalera S (2017) Deep learning for action and gesture recognition in image sequences: a survey. Gesture Recog:539–578
5. Bragg D, Koller O, Bellard M, Berke L, Boudreault P, Braffort A, Caselli N, Huenerfauth M, Kacorri H, Verhoef T and Vogler C (2019) Sign language recognition, generation, and translation: An Interdisciplinary Perspective Computers and Accessibility 16-31
6. Cai W, Liu D, Ning X, Wang C, Xie G (2021) Voxel-based three-view hybrid parallel network for 3D object classification. Displays 69:102076
7. Cui R, Liu H, Zhang C (2019) A deep neural framework for continuous sign language recognition by iterative training. IEEE Transac Multimed 21(7):1880–1891
8. Deng X, Yang S, Zhang Y, Tan P, Chang L, Wang H (2017) Hand3D: Hand Pose Estimation using 3D Neural Network. arXiv:1704.02224
9. Farooq U, Rahim MSM, Sabir N, Hussain A, Abid A (2021) Advances in machine translation for sign language: approaches, limitations, and challenges. Neural Comput Applic 33(21):14357–14399
10. Ferreira P, Cardoso J, Rebelo A (2019) On the role of multi-modal learning in the recognition of sign language. Multimed Tools Appl 78:10035–10056
11. Gao L, Li H, Liu Z, Liu Z, Wan L, Feng W (2021) RNN-transducer based Chinese sign language recognition. Neurocomputing 434:45–54
12. Guo H, Wang G, Chen X (2017) Towards good practices for deep 3D hand pose estimation. arXiv: 1707.07248
13. Imran J, Raman B (2020) Deep motion templates and extreme learning machine for sign language recognition. Vis Comput 36(6):1233–1246
14. Kamruzzaman MM (2020) Arabic sign language recognition and generating Arabic speech using convolutional neural network. Wirel Commun Mob Comput 2020:1–9
15. Khatri A, Gaba A, Rana KPS and Kumar V (2019) A novel life choice-based optimizer. Soft computing 1-21
16. Koller O, Zargaran S, Ney H, Bowden R (2018) Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. Int J Comput Vis 126(12):1311–1325
17. Kowdiki M, Khaparde A (2021) Automatic hand gesture recognition using hybrid meta-heuristic-based feature selection and classification with dynamic time warping. Comput Sci Rev 39:100320
18. Li D, Rodriguez C, Yu X, Li H (2020) Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. Appl Comput Vis:1459–1469
19. Liao Y, Xiong P, Min W, Min W, Lu J (2019) Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. IEEE Access 7:38044–38054
20. Lim K, Tan A, Lee C, Tan S (2019) Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. Multimed Tools Appl 78:19917–19944
21. Masood S, Srivastava A, Thuwal HC and Ahmad M (2018) Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In intelligent engineering informatics. Springer, Singapore 623-632
22. Nadgeri S, Kumar A (2019, July) An image texture based approach in understanding and classifying baby sign language. In 2019 2nd international conference on intelligent computing, instrumentation and control technologies (ICICICT). IEEE 1:854–858
23. Nadgeri S, Kumar D (2020) An analytical study of signs used in baby sign language using Mobilenet framework. In proceedings of the international conference on recent advances in computational techniques (IC-RACT)
24. Naranjo-Zeledón L, Peral J, Ferrández A, Chacón-Rivas M (2019) A systematic mapping of translation-enabling technologies for sign languages. Electronics 8(9):1047
25. Neiva DH, Zanchettin C (2018) Gesture recognition: a review focusing on sign language in a mobile context. Expert Syst Appl 103:159–183
26. Ning X, Gong K, Li W, Zhang L (2021) JWSAA: joint weak saliency and attention aware for person re-identification. Neurocomputing 453:801–811
27. Prietch SS, Pineda IO, Paim PDS, Calleros JMG, García JG, Resmin R (2019) Discussion on image processing for sign language recognition: an overview of the problem complexity. Res Develop Technol: 112–127
28. Rao GA, Kishore PVV (2018) Selfie video based continuous Indian sign language recognition system. Ain Shams Eng J 9(4):1929–1939
29. Rastgoo R, Kiani K, Escalera S (2018) Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. Entropy
30. Rastgoo R, Kiani K, Escalera S (2020) Sign language recognition: a deep survey. Expert systems with Applications113794

31. Saunders B, Camgoz NC, Bowden R (2021) Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. Int J Comput Vis 129(7):2113–2135
32. Sullivan AL, Thayer AJ, Farnsworth EM, Susman-Stillman A (2019) Effects of child care subsidy on school readiness of young children with or at-risk for special needs. Early Child Res Q 47:496–506
33. Wadhawan A, Kumar P (2020) Deep learning-based sign language recognition system for static signs. Neural computing and applications 1–12. https://doi.org/10.1007/s00521-019-04691-y
34. Wang C, Wang X, Zhang J, Zhang L, Bai X, Ning X, Zhou J, Hancock E (2022) Uncertainty estimation for stereo matching based on evidential deep learning. Pattern Recogn 124:108498
35. Wangchuk K, Riyamongkol P, Waranusast R (2021) Real-time bhutanese sign language digits recognition system using convolutional neural network. ICT Express 7(2):215–220
36. Wei S, Chen X, Yang X, Cao S, Zhang X (2016) A component-based vocabulary-extensible sign language gesture recognition framework. Sensors 16(4):556
37. Wu F, Jing XY, Dong X, Hu R, Yue D, Wang L, Ji YM, Wang R, Chen G (2018) Intraspectrum discrimination and interspectrum correlation analysis deep network for multispectral face recognition. IEEE Transac Cyber 50(3):1009–1022
38. Yang S, Zhu Q (2017) Continuous Chinese sign language recognition with CNN-LSTM", Proc. SPIE 10420, Digital Image Processing (ICDIP 2017), 104200F. https://doi.org/10.1117/12.2281671
39. Zheng L, Liang B, Jiang A (2017) Recent Advances of Deep Learning for Sign Language Recognition. 2017 International conference on digital image computing: techniques and applications (DICTA), Sydney, NSW, Australia