scientific reports

Check for updates

OPEN

Gesture recognition for hearing impaired people using an ensemble of deep learning models with improving beluga whale optimization-based hyperparameter tuning

Mohammed Assiri[1,3] & Mahmoud M. Selim[2,3]✉

Sign language (SL) is the linguistics of speech and hearing-impaired individuals. The hand gesture is the primary model employed in SL by speech and hearing-challenged people to talk with themselves and ordinary persons. At present, hand gesture detection plays a vital part, and it is commonly employed in numerous applications worldwide. Hand gesture detection systems can aid in transmission between machines and humans by aiding these sets of people. Machine learning (ML) is a subdivision of artificial intelligence (AI), which concentrates on the growth of a method. The main challenge in hand gesture detection is that machines do not directly understand human language. A standard medium is required to facilitate communication between humans and machines. Hand gesture recognition (GR) serves as this medium, enabling commands for computer interaction that specifically benefit hearing-impaired and elderly individuals. This study proposes a Gesture Recognition for Hearing Impaired People Using an Ensemble of Deep Learning Models with Improving Beluga Whale Optimization (GRHIP-EDLIBWO) model. The main intention of the GRHIP-EDLIBWO model framework for GR is to assist as a valuable tool for developing accessible communication systems for hearing-impaired individuals. To accomplish that, the GRHIP-EDLIBWO method initially performs image preprocessing using a Sobel filter (SF) to enhance edge detection and extract critical gesture features. For the feature extraction process, the squeeze-and-excitation capsule network (SE-CapsNet) effectively captures spatial hierarchies and complex relationships within gesture patterns. In addition, an ensemble of classification processes, such as bidirectional gated recurrent unit (BiGRU), Variational Autoencoder (VAE), and bidirectional long short-term memory (BiLSTM) technique, is employed. Finally, the improved beluga whale optimization (IBWO) method is implemented for the hyperparameter tuning of the three ensemble models. To achieve a robust classification result with the GRHIP-EDLIBWO approach, extensive simulations are conducted on an Indian SL (ISL) dataset. The performance validation of the GRHIP-EDLIBWO approach portrayed a superior accuracy value of 98.72% over existing models.

Keywords Gesture recognition, Sign language, Hearing impaired people, Ensemble deep learning, Image preprocessing

Globally, deaf and dumb people struggle to express their feelings to others. Several tasks are proficient for hearing-impaired and speech-impaired individuals in public places to express themselves to normal individuals[1]. Gestures are created from any physical movement or state. Gestures might be made with your face, hands, shoulders, arms, feet, legs, or a mixture, but the most common hand gestures are possible[2]. GR is an area in computer science that interprets human gestures through mathematical models[3]. A gesture is a

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, P.O. BOX 16273, 3963 Al-Kharj, Saudi Arabia. [2]Department of Mathematics, College of Science and Humanities, Prince Sattam Bin Abdulaziz University, 11942 Al-Kharj, Saudi Arabia. [3]King Salman Center for Disability Research, 11614 Riyadh, Saudi Arabia. ✉email: m.selim@psau.edu.sa

physical movement that permits individuals to interact with information and meaning with each other. Vision-based and data gloves methods are dual alternate methods for human–computer communication. Hand motion classification and detection were considered part of the vision-based method during the research. The logical models generate an adaptable and convenient interface among users and gadgets to utilize hand gestures. Hand gestures are a communication method and the more general GR method[4]. One of the most popular examples of a hand gesture method is SL. It is a linguistic method that utilizes hand motions along with other motions. For instance, most hearing-impaired individuals employ SL worldwide[5]. The three elementary components of SL are finger spelling, non-manual characteristics, and word-level sign vocabulary. SL is one of the most efficient methods to interact with hearing-impaired individuals[6]. SLR is a standard application of hand GR. It is frequently considered only deaf individual depend on SLs for conveying their thoughts. Hand gestures are body language features transmitted over the shape created by the hand, finger position, and centre of the palm. Hand gestures might be categorized into dynamic and static. It indicates the static gesture signifies the shape of the hand, while the dynamic gesture includes the sequence of hand gestures like waving[7]. There's a mixture of hand gestures inside a gesture; for instance, a handshake differs from one individual to another and modifies based on location and period[8]. The major dissimilarity between gesture and posture is that posture aims more at the shape of the hand, while gesture aims at hand movements. The vital methods of hand gesture investigation might be characterized by the camera vision-based and wearable glove-based sensor methods. Hand gestures provide an inspiring area of inquiry that can assist communication and deliver natural means of interaction utilized through various applications. Formerly, hand gesture detection was attained with wearable sensors connected instantly to the hand with gloves. This sensor recognizes physical responses based on finger bending or hand movements[9]. The information was gathered and then processed utilizing a computer related to the gloves with wire. This glove-based sensor method is created portable using a sensor connected to the microcontroller. AI provides reliable and suitable methods in various advanced applications due to employing the learning principle role[10]. Deep learning (DL) and ML employ multi-layers for learning data and give better forecast outcomes.

This study proposes a Gesture Recognition for Hearing Impaired People Using an Ensemble of Deep Learning Models with Improving Beluga Whale Optimization (GRHIP-EDLIBWO) model. The main intention of the GRHIP-EDLIBWO model framework for GR is to assist as a valuable tool for developing accessible communication systems for hearing-impaired individuals. To accomplish that, the GRHIP-EDLIBWO method initially performs image preprocessing using a Sobel filter (SF) to enhance edge detection and extract critical gesture features. For the feature extraction process, the squeeze-and-excitation capsule network (SE-CapsNet) effectively captures spatial hierarchies and complex relationships within gesture patterns. In addition, an ensemble of classification processes such as BiGRU, Variational Autoencoder (VAE), and bidirectional long short-term memory (BiLSTM) technique is employed. Finally, the improved beluga whale optimization (IBWO) method is implemented for the hyperparameter tuning of the three ensemble models. To achieve a robust classification result with the GRHIP-EDLIBWO approach, extensive simulations are conducted on an Indian SL (ISL) dataset. The key contribution of the GRHIP-EDLIBWO approach is listed below.

- The GRHIP-EDLIBWO model utilizes SF to improve edge sharpness and structural clarity in GR images, significantly improving the distinction between gesture boundaries and the background. This refined edge detection assists in extracting more meaningful features during subsequent stages. It also assists in improving the overall accuracy of classification by reducing noise and emphasizing critical spatial details.
- The GRHIP-EDLIBWO method employs SE-CapsNet for capturing spatial hierarchies and intrinsic gesture relationships by incorporating capsule structures with channel attention mechanisms, concentrating on the most important features. This deep representation allows the model to handle discrepancies in gesture shape and orientation effectively, improving the discriminative power of the extracted features and resulting in more accurate classification outputs.
- The GRHIP-EDLIBWO approach utilizes BiGRU to capture long-term dependencies, VAE for robust latent space representation, and BiLSTM for bidirectional context learning, incorporating their strengths to enhance model accuracy. This fusion improves the capability of the model for handling inherent sequential patterns and discrepancies in gesture data, making the classification process more reliable and resilient to diverse complexities in input data.
- The GRHIP-EDLIBWO methodology implements IBWO to efficiently explore the parameter space, resulting in faster convergence and more precise model configurations. This method improves the generalization ability of the technique and improves classification performance. By utilizing the merits of IBWO, the model attains better results while mitigating computational time.
- The novelty of the GRHIP-EDLIBWO technique is in the integration of SE-CapsNet with a diverse ensemble of temporal BiGRU, BiLSTM, and generative VAE classifiers, presenting a multi-perspective learning process specifically for GR. This integration employs spatial hierarchies, sequential modeling, and latent space representation to improve feature extraction and classification accuracy. Also, IBWO-based hyperparameter tuning optimizes model parameters, ensuring effective convergence and boosting performance for precise gesture classification.

Literature of works

Sümbül[11] introduces wearable glove-mounted flex sensors and MEMS-based accelerometer arrays for recognizing hand movements to SL Recognition (SLR). The multi-layer perceptron feed-forward neural network (ML-PFFNN) was selected as the particular ANN model to establish the gestures. To make a complete hand movement database, both accelerometer data and flex sensor were utilized to make Pulse Width Modulation (PWM) values that act as input methods. Hossain et al.[12] aim to improve MediaPipe hand tracking technology and an accurate hand motion recognition method employing CNN. This method also employs MediaPipe's

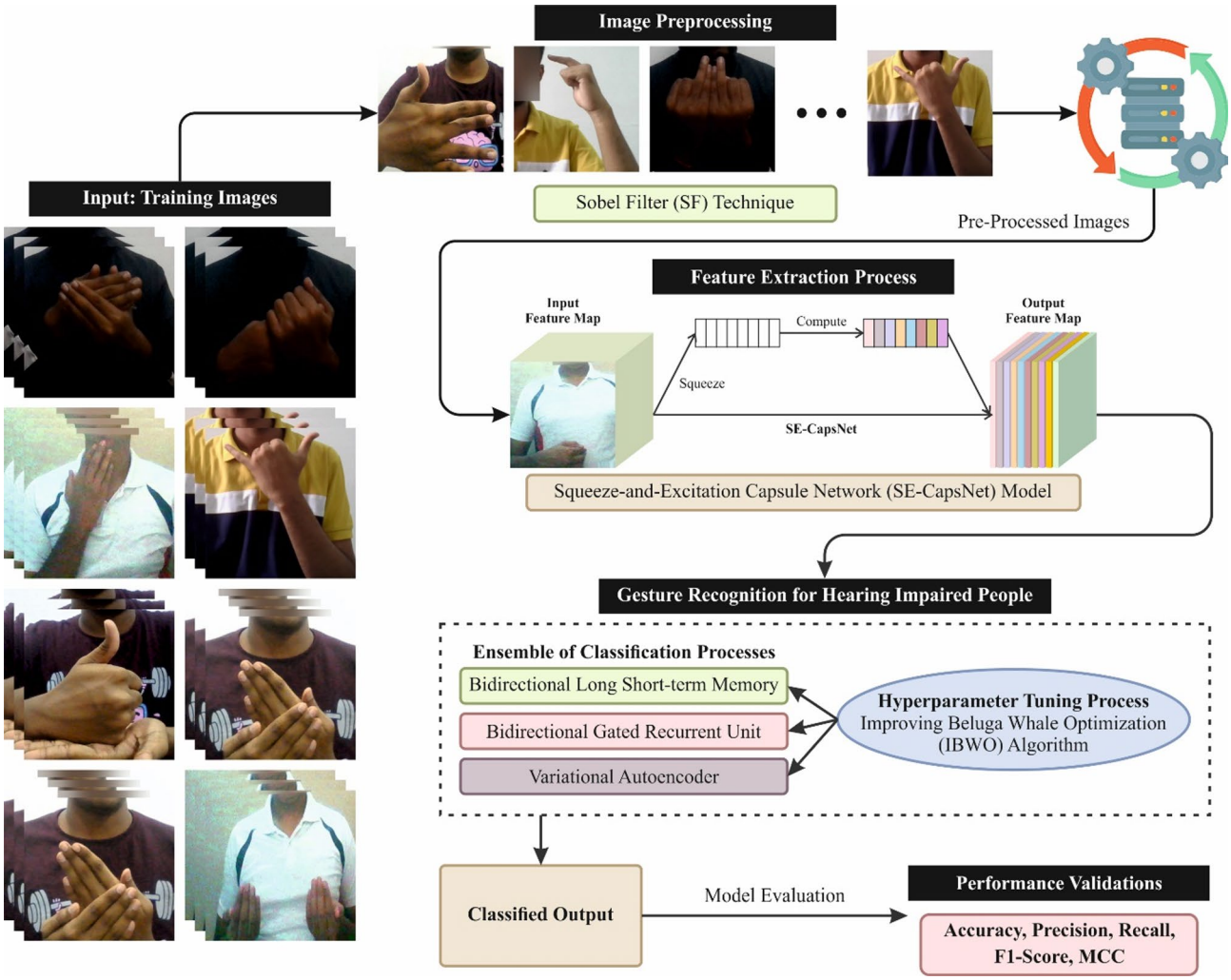

Fig. 1. Overall workflow of GRHIP-EDLIBWO model.

developed hand keypoint technology to key points from images, remove data precisely, orientations, and hand movements. These characteristics lead to a trained CNN method for classification. Vyshnavi et al.[13] projected a Gesture Language Translator method utilizing advanced CNN structures comprising ResNet and VGG-16. The SL Translator method associates the ability of ResNet-50 and VGG-16 methods utilizing ensemble learning. Additionally, this paper integrates data preprocessing methods to improve the consistency and quality of input SL images, guaranteeing optimum CNN method performances. In[14], a CNN model to recognize static alphabet gestures in the American SL (ASL) was developed. This method comprises three stages: A preprocessing stage for removing the ROI, a classification, and a feature extraction stage. Shinde et al.[15] introduce a strong DL-based solution for precise recognition and detection of ISL movements. The presented method incorporates developed methods like LSTM and MediaPipe Holistic for feature extraction systems for SL translation and detection. Feature extraction is implemented by utilizing the MediaPipe Holistic models. An LSTM model and a Recurrent Neural Network (RNN) method were applied through the DL stage using the Keras or TensorFlow structures.

In[16], a substantial hand GR method is developed using deep ensemble NN. A pre-trained system was primarily intended to use the VGG-16 structure by a self-attention layer fixed through the VGG-16 structure. This self-attention component allows us to absorb the possibly differentiating image attributes for greater dissimilarity between gesture groups. Afterwards, a weighted ensembling method was presented that utilizes the additional data added to the base model to increase the entire system's performance. In[17], a novel method for real-world hand sign identification with assistance from CNN and OpenCV is developed with DL and a fusion of computer vision (CV). The association of OpenCV and CNN projects a promising avenue for improving communication and accessibility, particularly in surroundings where verbal communication is non-existent or limited. Pre-trained structures like MobileNet and ResNet are associated with the CNN method, which utilizes ensemble learning. Izzalhaqqi[18] projected the enlargement of the hybrid CNN-LSTM method explicitly intended to identify static alphabetical gestures in Indonesian SL (BISINDO). These self-collected imageries endured preprocessing and augmentation phases before the training model. The hyper-parameter and structured pattern of the hybrid method have been adjusted utilizing the Randomized Search CV approach. Ramadan and Abd-Alsabour[19] introduce a real-time GR system for laptop control, replacing the requirement

for a mouse or keyboard. It assists functions such as mouse movement, clicks, scrolling, keyboard shortcuts, file management, media control, volume/brightness adjustment, slideshow navigation, and launching frequently used apps. Rajalakshmi et al.[20] propose a Hybrid Neural Network (HNN) model for recognizing Indian and Russian SL, utilizing 3D Convolution for static gestures and a modified auto-encoder with hybrid attention for dynamic gestures. It also comprises face/hand detection with GradCam and Camshift and a new multi-signer dataset. Rajalakshmi et al.[21] present a hybrid deep neural network (DNN) technique for recognizing Indian and Russian sign gestures, utilizing 3D CNN for spatial features, Bi-LSTM with attention for temporal features, and a hybrid attention module for gesture differentiation.

Shanthi et al.[22] utilize MediaPipe and OpenCV to enable painting through intuitive hand gestures captured by a webcam. Users can draw, erase, and switch tools without conventional input devices, presenting a seamless, immersive digital art experience. Ramkarthik and Benita[23] developed an advanced SL recognition system that integrates CV and DL models to improve communication for individuals with hearing impairments, promoting inclusivity and accessibility. Narayan and Jain[24] present the Image Transformer Model (ITM) for hand GR, utilizing attention mechanisms (AMs) and optimized with Particle Swarm Optimization (PSO) to improve spatial correlation detection, outperforming traditional methods. Hasan and Adnan[25] introduce EMPATH, a computational framework that enhances Bangla SL (BdSL) recognition using Ensemble Learning, MediaPipe Holistic, and an Attention-based Transformer. Marzouk et al.[26] introduce an automated GR using artificial rabbits' optimization with DL (AGR-ARODL) method. The technique also uses median filtering (MF) for image preprocessing, SE-ResNet-50 for feature extraction, and Artificial Rabbit Optimization (ARO) for hyperparameter selection. The Deep Belief Network (DBN) model is employed for hand gesture detection. Tounsi et al.[27] present the Comprehensive Learning Salp Swarm Algorithm with Ensemble DL (CLSSA-EDL) technique utilizing DenseNet201 for feature extraction and hyperparameters optimized by the CLSSA system. It also employs an ensemble model with a Stacked Autoencoder (SAE), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) for detection and classification. Manoharan and Sivagnanam[28] improve hand GR by utilizing VGG16 with transfer learning, improved by an AM model. The technique also includes custom layers like flattening, ReLU activation, and SoftMax, with innovative preprocessing techniques, ensuring robustness in recognizing dynamic gestures.

Despite the advancements in hand GR, existing methods face limitations in handling dynamic gestures, real-time performance, and robustness in diverse environments. Many models rely heavily on pre-trained networks, which may not generalize well to varied SLs or noisy data. The integration of AMs, while beneficial, often increases computational complexity. Furthermore, the reliance on specific hardware or sensors, such as accelerometers or MediaPipe, limits accessibility and scalability. Moreover, datasets are often limited in diversity, resulting in challenges in multi-signer and multi-language recognition. Future research should address these gaps to enhance adaptability, scalability, and real-time performance.

Proposed methodology

In this study, the GRHIP-EDLIBWO approach is proposed. The aim is to assist as a valuable tool for developing accessible communication systems for hearing-impaired individuals. To accomplish that, the proposed GRHIP-EDLIBWO approach involves image preprocessing, feature extraction, an ensemble of the classification process, and hyperparameter tuning. Figure 1 characterizes the complete workflow of the GRHIP-EDLIBWO model.

Image preprocessing

At first, the presented GRHIP-EDLIBWO method initially performs image preprocessing by using an SF model to enhance edge detection and extract critical gesture features[29]. This effectual edge detection technique is commonly used in image processing to emphasize regions of interest. Its advantage is in its simplicity and efficiency, making it computationally less expensive compared to more intrinsic methods. Detecting edges in the image improves crucial features such as boundaries and contours, which is significant for accurate analysis, particularly in tasks like object recognition or segmentation. Unlike more advanced filters, the SF performs well in real-time applications where processing speed is a priority. Additionally, it works well on both small and large-scale images, making it versatile and widely applicable in diverse scenarios. Overall, its performance and computational efficiency balance justify its use in various image preprocessing tasks.

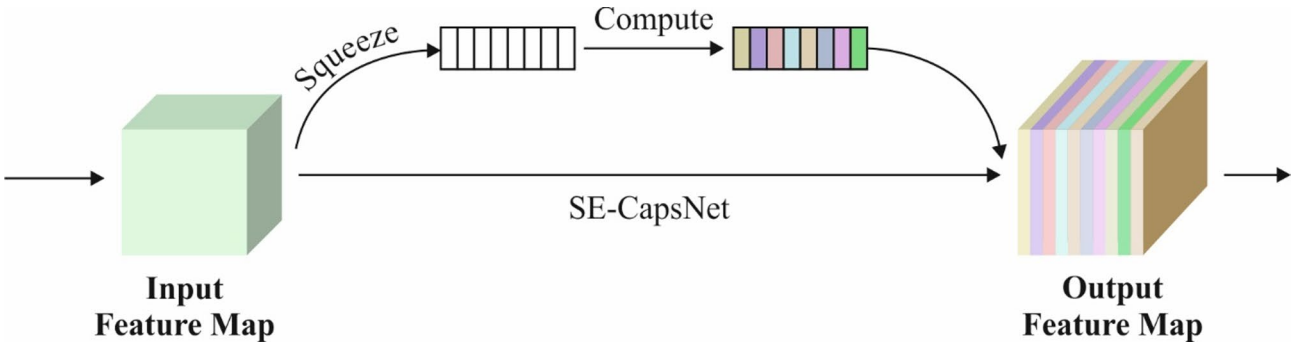

Fig. 2. SE-CapsNet architecture.

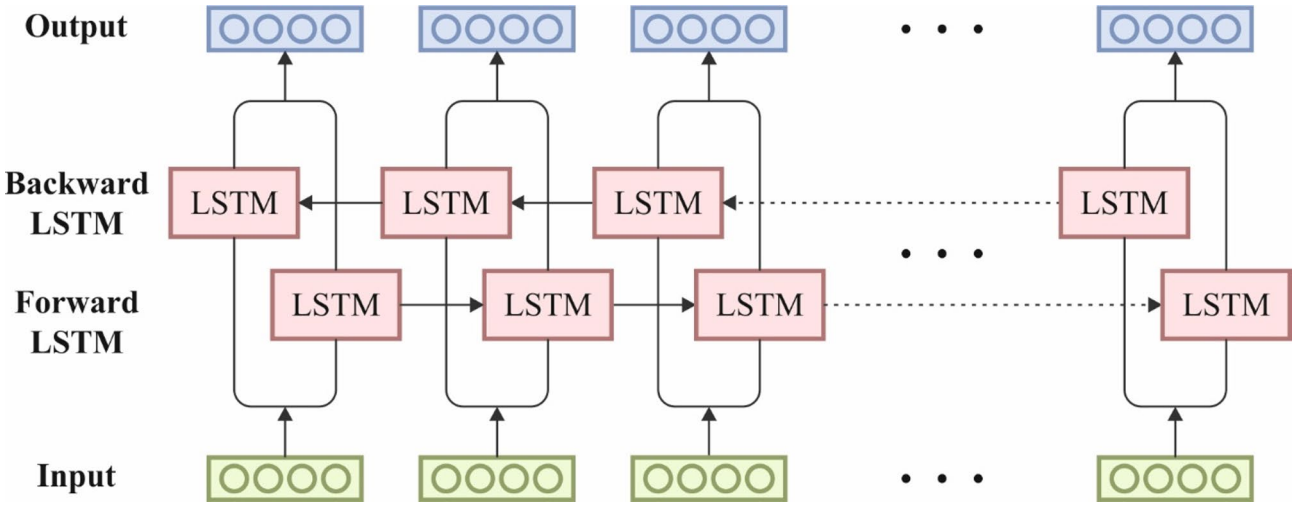

Fig. 3. Structure of BiLSTM model.

The SF is a general edge-recognition method employed in image preprocessing for gesture detection. The framework of GR for hearing-challenged individuals helps to highlight the contours and boundaries of hand gestures, making it simpler to differentiate between diverse movements and shapes. The SF functions by calculating the gradient of the image strength at every pixel, highlighting regions with significant variations in intensity. This boosts the related feature in the image, which is vital for precisely classifying gestures. Using SF, the method can attain enhanced feature extraction and segmentation, leading to better recognition performance.

SE-CapsNet-based feature extraction

For feature extraction, the SE-CapsNet is utilized to effectively capture spatial hierarchies and complex relationships within gesture patterns[30]. This robust model combines the advantages of CapsNet and Squeeze-and-Excitation (SE) blocks. CapsNet presents an enhanced handling of spatial hierarchies and rotations in image data, making it particularly effectual for object recognition and segmentation tasks. The addition of SE blocks improves the AM of the model, enabling it to concentrate on more crucial features, thereby enhancing overall performance in tasks such as classification. Compared to conventional CNNs, SE-CapsNet outperforms in handling complex patterns with fewer parameters, making it less prone to overfitting. However, one trade-off is that CapsNet models generally exhibit higher computational complexity and longer training times due to their more complex architecture. Despite this, SE-CapsNet presents a good balance of feature extraction capability and efficiency, making it appropriate for applications needing accurate and robust performance, specifically in scenarios where capturing spatial relationships is significant. Figure 2 illustrates the structure of the SE-CapsNet model.

The DL used in the associated area of malware recognition has achieved brilliant study performances; particularly, the capsule system estimates the relationship amongst attributes, and this model provides advantages once used in more minor instances. This work associations channel AMs, named the SE blocks, utilizing the capsule system to produce the SE-CapsNet method, mainly formed from the succeeding four layers.

(1) Convolutional Layer

The initial layer is the basic convolutional layer tailored for removing local attributes. It employs 3x3 convolution kernels with step dimensions of 1 in association with the activation function of $ReLU$.

(2) SE Layer

The SE blocks are easier to apply, may enhance the method's feature extraction capability, and are helpful in classification, which comprises dual processes: excitation and squeeze. The objective squeeze process aims to gain the global characteristics of the provided channels. uc characterizes the Cth feature mapping that is output through the convolution layers. Over global average pooling, they may gain channel-to-channel information zc. The Excitation process is the learned channel weights method. At the same time, σ characterizes the activation function of the sigmoid, δ signifies the function of $ReLU$, and W_1 and W_2 represent dimensionality-increasing and -reducing activities, correspondingly. A non-linear network interaction is rapidly learned during excitation, acquiring the learned channel weights. Lastly, the new feature mapping improves the learned channel weighting over the scaling process to gain attention for feature mapping as the SE block's output. The computation equations are demonstrated in (1)-(3):

$$zc = F_{sq}(uc) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} uc(i,j) \tag{1}$$

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(Wz)) \tag{2}$$

$$\widetilde{x_C} = F_{scale}(u_C, s_C) = s_C \cdot u_C \tag{3}$$

(3) Primarycaps Layer

Afterwards, the block of SE captures all feature maps through its equivalent attention weighting as the PrimaryCap's input. This layer is distinct from the standard convolutional layer. Based on its description, these layers and capsules may be obtained and may further be named vectors that can store more statistics.

(4) Digitcaps Layer

It is applied for storing non- and Ponzi capsules. The vector characterizes the last output. The capsule system applies a function of squashing. However, keeping the vector direction, the output vector length is used as the likelihood of the existence of an object. The computation equations amongst capsules i and j are exposed in (4)-(6):

$$\widehat{u}_{j|i} = W_{ij} u_i \tag{4}$$

$$s_j = \sum_i c_{ij} \widehat{u}_{j|i} \tag{5}$$

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{6}$$

Here W_{ij} characterizes the weighted matrix, demonstrating the association amongst capsules i and j, and $\widehat{u}$ symbolizes the prediction, in which the ith lower-level capsule establishes the jth higher-level capsule. c_{ij} denotes the coupling coefficient gained over dynamic routing. The output v_j is estimated using the outcome of the last squashing function.

Ensemble of classification process

In addition, an ensemble of classification processes such as BiLSTM**,** BiGRU, and VAE techniques are employed, which presents significant advantages in classification tasks, especially in sequential or time-series data. Bidirectional Long Short-Term Memory (BiLSTM) and BiGRU are designed to capture both past and future context in a sequence, enhancing the model's ability to understand temporal dependencies more effectively than unidirectional models. These techniques are valuable when dealing with complex data where the context is not strictly linear. VAE, conversely, is a generative model that can learn efficient data representations and handle uncertainty, making it helpful in enhancing model robustness and generalization. Compared to conventional models like CNNs, these techniques are more adept at handling sequential data and missing information, resulting in an improved performance in classification tasks with variable input lengths and patterns. However, VAE models are more computationally intensive due to the requirement for training both an encoder and a decoder. Still, the benefit is seen in more flexible and robust features for classification.

BiLSTM classifier

An LSTM method characterizes variation of the RNNs, showing robust handling and analytic abilities regarding time-series data[31]. The LSTM element comprises three gates- the forget, the input, and the output gates—which handle information unidirectionally from left to right. This framework restricts the component's data-capturing capability. Researchers have improved the LSTM and presented Bi-LSTM techniques to gain complete knowledge. A Bi-LSTM is made of numerous LSTM elements. The underlying computation principle transmits valuable data for following calculations by forgetting and recalling novel data within the state of the cell while removing irrelevant data. The particular computation procedure is as demonstrated: Initially, receive the input of the present moments and the hidden layer (HL) of the preceding point, and pass over the forgetting gate over the sigmoid functions:

$$\begin{cases} f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_t\right) \\ \sigma(x) = \frac{1}{1+e^{-x}} \end{cases} \tag{7}$$

In this case, x_t, h_{t-1}, w_f, and b_t characterize the present input, the HL at the preceding point, and the weighted and biased matrix of the forgetting gates. According to the data from x_i and h_{t-1}, the input gate particularly collects the objective data into the cell state C_t at the upgraded time t:

$$\begin{cases} q_t = \sigma\left(w_q\left[h_{t-1}, x_t\right] + b_q\right) \\ a_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \\ C_t = f_t C_{t-1} + q_t a_t \\ \tanh(x) = \frac{e^x - e - x}{e^x + e - x} \end{cases} \tag{8}$$

In such cases, a_t and q_t characterize the output correspondingly according to the input node and gate. In the same way, w_c and w_q indicate the weighted matrices of input nodes and gates individually. In contrast, b_q and b_c correspondingly indicate the biased matrices of input nodes and gates. Lastly, C_t and C_{t-1} specify the states of the cell at times t and $t-1$, correspondingly. They gain the output gate 0_t according to the data from h_{t-1} and x_t, and more control over how much data should be applied as the present output state or HL derived from the output gate o_t and the present cell state C_t:

$$\begin{cases} o_t = \sigma\left(w_0\left[h_{t-1'} x_t\right] + b_0\right) \\ h_t = o_t \tanh\left(C_{t-i}\right) \end{cases} \tag{9}$$

Phrases	Class Labels	Images for Phrases
Attendance	L-1	40
Book	L-2	40
Careful	L-3	40
Congratulations	L-4	40
File	L-5	40
Good Morning	L-6	40
Happy Birthday	L-7	40
How are you	L-8	40
Hungry	L-9	40
I Need Help	L-10	40
Keepsmile	L-11	40
Mistake	L-12	40
Opinion	L-13	40
Practice	L-14	40
Team	L-15	40
Thirsty	L-16	40
Together	L-17	40
Understand	L-18	40
Wait	L-19	40
Write	L-20	40
Total number of images		800

Table 1. Details of the dataset.

while o_t, w_0, and b_o represent the output, weighted matrix, and biased matrix of the output gates, and h_t signifies the HL of the present point. Finally, bi-directional HL is integrated, and the conditions of the forward and backward HLs are joined by a weighted coefficient as shown, so offering the complete assessment of the data:

$$\begin{cases} h_{t1} = LSTM\left(x_t, h_{t-1}\right) \\ h_{t2} = LSTM\left(x_t, h_{t+1}\right) \\ y_t = w_{h1}h_{t1} + w_{h2}h_{t2} + b_y \end{cases} \tag{10}$$

Here, w_{h1} and w_{h2} represent the related loop-weighted matrix of the backward and forward LSTM layers; correspondingly, y_t signifies the last HL. Figure 3 illustrates the architecture of the BiLSTM technique.

BiGRU classifier

The GRU is a variation of the RNN intended to tackle the vanishing gradient problem, which conventional RNNs face when handling longer sequences[32]. The features of GRU are a reset gate and an update gate to adjust the information flow. The GRU method facilitates the network architecture by integrating the forget gate and input gate established in the LSTM structure into a particular update gate. The specific equation is as shown:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \tag{11}$$

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right) \tag{12}$$

$$\widetilde{h}_t = \tanh\left(r_t \otimes U_h h_{t-1} + W_h x_t\right) \tag{13}$$

$$h_t = \left(1 - z_t\right) \otimes \widetilde{h}_t + z_t \otimes h_{t-1} \tag{14}$$

In this case, z_t and r_t characterize the reset gate and update gate conditions, correspondingly, at tth time. The $tanh$ function signifies a non-linear activation function. h_{t-1} represents the HL at the preceding time step, whereas h_t characterizes the output of the HL. The representation $\widetilde{h}$ denotes the candidate HL, and x_t signifies the input data. W and U represent weighted matrices, σ characterizes the function of the sigmoid and $\otimes$ symbolizes the Hadamard product.

The GRU technique handles information only in the forward direction, which can manage essential details associated with the backward sequence data. A BiGRU has been presented to improve precision and gather information before and after variations. This method associations the HL from the forward or backward GRU directions, permitting bidirectional data processing in sequential order. This model allows the technique to gain more complete time series data, forecast, and enhance complete model performance. The particular equation is as shown:

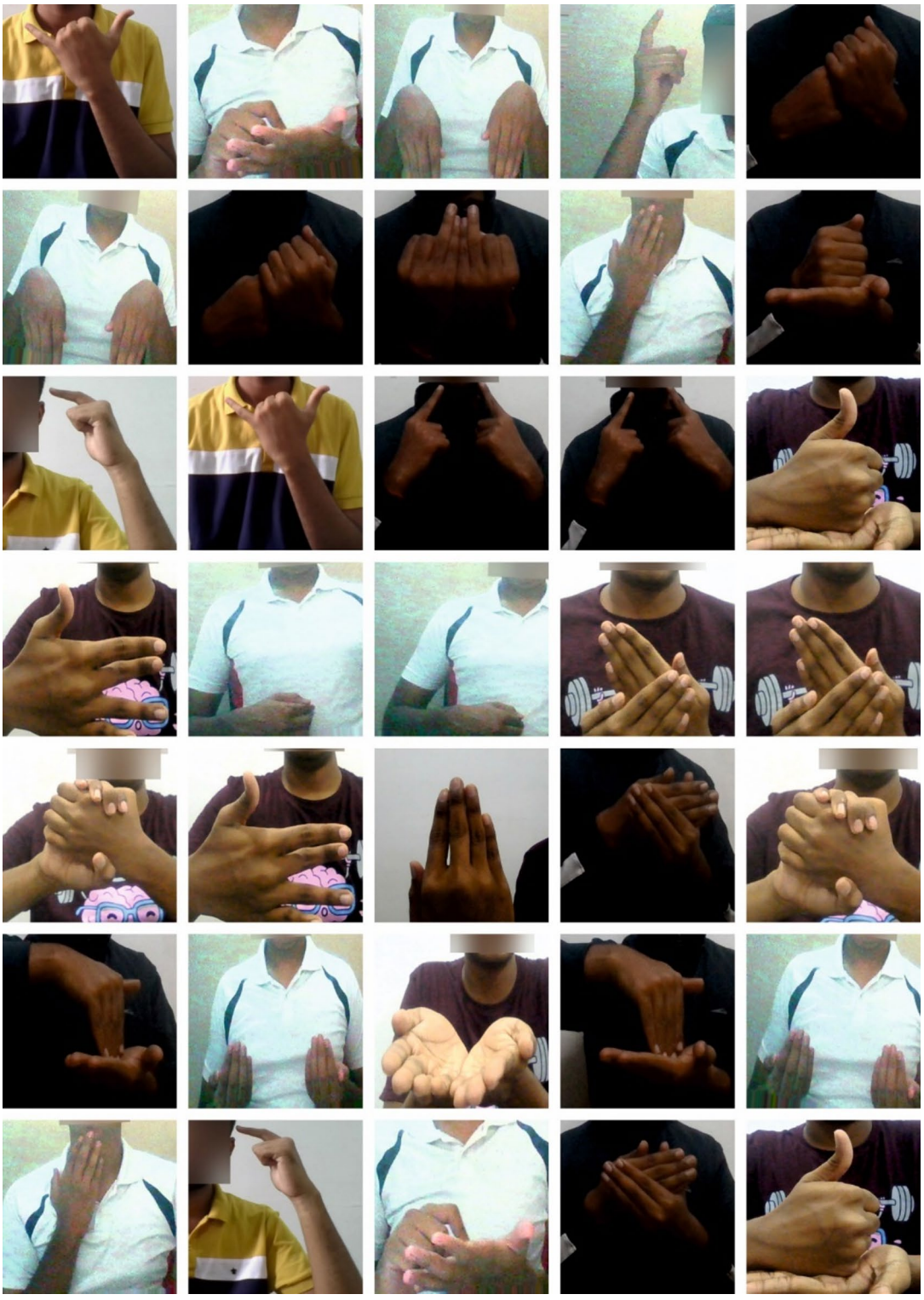

Fig. 4. Sample images.

$$\begin{cases} \overrightarrow{\mathrm{h_t}} = GRU\left(x_t, \overrightarrow{\mathrm{h_{t-1}}}\right) \\ \overleftarrow{h_t} = GRU\left(x_t, \overleftarrow{h_{t-1}}\right) \\ h_t = \alpha_t \overrightarrow{h_t} + \beta_t \overleftarrow{h_t} + b_t \end{cases} \tag{15}$$

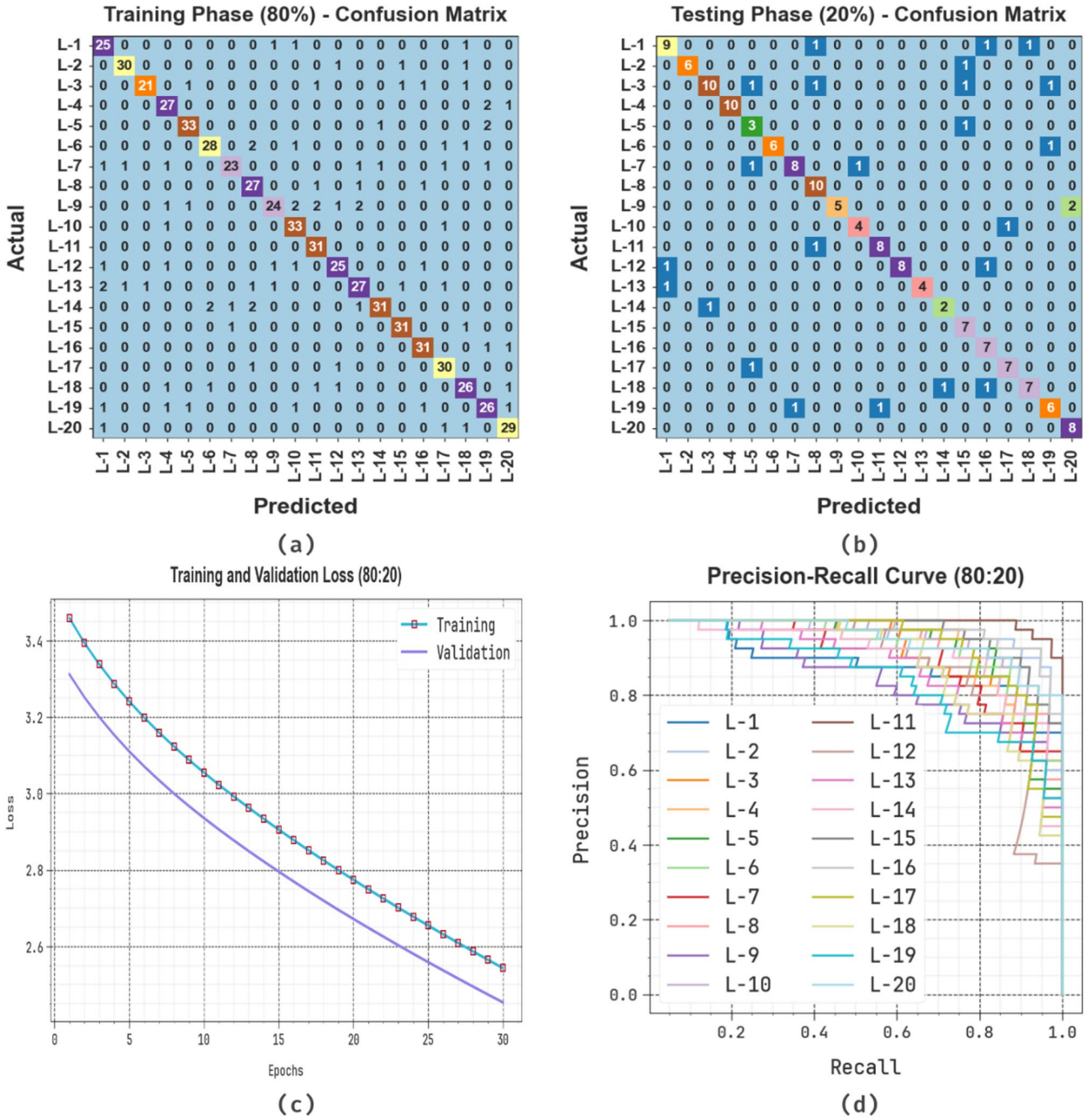

Fig. 5. 80%TRPH and 20%TSPH of (**a**,**b**) of confusion matrix, (**c**,**d**) PR and ROC curves.

while $GRU\ (\cdot)$ characterizes the GRU, $\overrightarrow{\mathrm{h_t}}$ and $\overleftarrow{h_t}$ signify the forward and backward HL, respectively; α_t and β_t represent weights related to the forward and backward HLs, correspondingly; and b_t characterizes the bias.

VAE classifier

A latent representation, signified as $z \in y \subset \mathbb{R}^k$, searches for defining a data sample $x \in X \subset \mathbb{R}^n$ Using lack of information, in other words "n, while maintaining related characteristics like similarities amongst samples of data in X [33]. The theoretic possibility of the efficient latent representation was discovered utilizing the samples and studied using models, namely VAEs, principal component analysis (PCS), and distributed stochastic neighbourhood (t-SNE).

Auto-encoders (AE) contain an encoder. E_ϕ: $X \to y$ and a decoder D_θ: $y \to X$, either neural networks parameterization by weightings φ and θ, correspondingly. The encoding mapping high-dimension data $X \subset \mathbb{R}^n$ to a space of latent vector $y \subset \mathbb{R}^k$. Conversely, the decoding intends to rebuild the novel signal from these compressed models. The main aim is to reduce an error of reconstruction $\mathcal{L}$: $X \times X \to \mathbb{R}$ by repetitively altering the weighting θ and φ. Formally, the optimizer balanced readers.

Class labels	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$	MCC
TRPH (80%)					
L-1	98.59	80.65	89.29	84.75	84.13
L-2	99.22	93.75	90.91	92.31	91.91
L-3	99.06	95.45	80.77	87.50	87.34
L-4	98.59	81.82	90.00	85.71	85.08
L-5	99.06	91.67	91.67	91.67	91.17
L-6	98.75	90.32	84.85	87.50	86.89
L-7	98.59	92.00	76.67	83.64	83.28
L-8	98.59	81.82	90.00	85.71	85.08
L-9	98.12	88.89	72.73	80.00	79.46
L-10	98.75	84.62	94.29	89.19	88.67
L-11	99.22	86.11	100.00	92.54	92.41
L-12	98.59	86.21	83.33	84.75	84.02
L-13	97.97	84.38	77.14	80.60	79.62
L-14	98.75	93.94	83.78	88.57	88.07
L-15	99.22	91.18	93.94	92.54	92.14
L-16	99.06	88.57	93.94	91.18	90.73
L-17	98.91	85.71	93.75	89.55	89.08
L-18	98.28	81.25	83.87	82.54	81.65
L-19	98.12	81.25	81.25	81.25	80.26
L-20	98.91	87.88	90.62	89.23	88.67
Average	98.72	87.37	87.14	87.04	86.48
TSPH (20%)					
L-1	96.88	81.82	75.00	78.26	76.66
L-2	99.38	100.00	85.71	92.31	92.28
L-3	96.88	90.91	71.43	80.00	79.00
L-4	100.00	100.00	100.00	100.00	100.00
L-5	97.50	50.00	75.00	60.00	60.05
L-6	99.38	100.00	85.71	92.31	92.28
L-7	98.12	88.89	80.00	84.21	83.35
L-8	98.12	76.92	100.00	86.96	86.82
L-9	98.75	100.00	71.43	83.33	83.97
L-10	98.75	80.00	80.00	80.00	79.35
L-11	98.75	88.89	88.89	88.89	88.23
L-12	98.75	100.00	80.00	88.89	88.85
L-13	99.38	100.00	80.00	88.89	89.16
L-14	98.75	66.67	66.67	66.67	66.03
L-15	98.12	70.00	100.00	82.35	82.84
L-16	98.12	70.00	100.00	82.35	82.84
L-17	98.75	87.50	87.50	87.50	86.84
L-18	98.12	87.50	77.78	82.35	81.52
L-19	97.50	75.00	75.00	75.00	73.68
L-20	98.75	80.00	100.00	88.89	88.85
Average	98.44	84.70	84.01	83.46	83.13

Table 2. GR of GRHIP-EDLIBWO methodology under 80%TRPH and 20%TSPH.

$$\min_{\theta,\phi} \sum\nolimits_{x \in X} \mathcal{L}\left(x, \left(E_\phi \circ D_\theta\right)(x)\right) \tag{16}$$

The function of the objective in the equation enables the encoder of the k important latent features of the signal, or else, the decoding absences the ability for higher-quality reconstruction. VAE offers a probability-based viewpoint and presents standardization on the latent representation to tackle the completeness and continuity of the embedded area. A distribution parametrization by θ allocates a probability to every data sample $x \in X$ signified as $p_\theta(x)$. Demonstrating the relations between the latent representation and the data sample contains marginalize over $z \in y$.

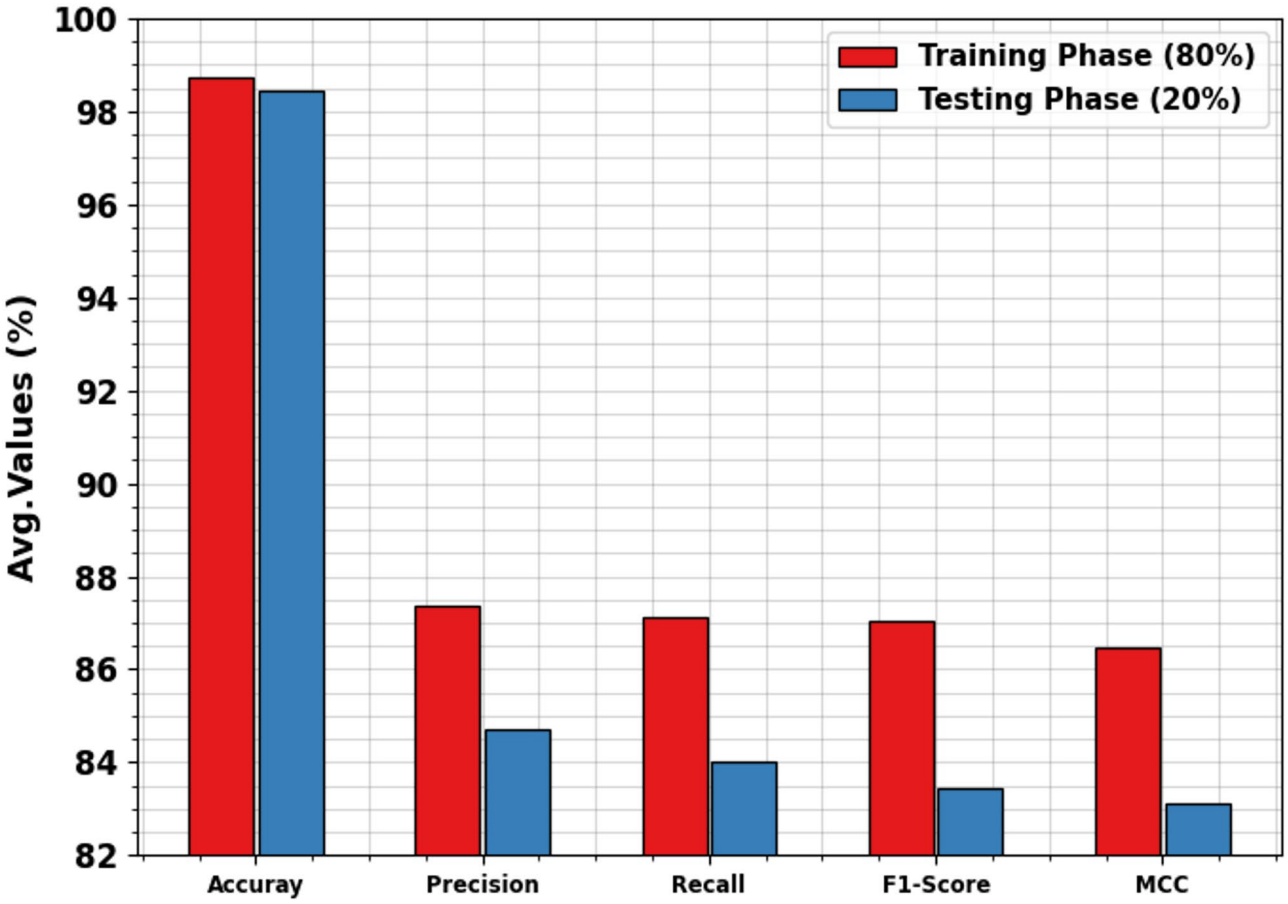

Fig. 6. Average of GRHIP-EDLIBWO technique under 80%TRPH and 20%TSPH.

$$\begin{aligned} p_\theta(x) &= \int_z p_\theta(x,z)\,dz = \int_z p_\theta(x|z)\,p_\theta(z)\,dz \\ &= \int_z p_\theta(z|x)\,p_\theta(x)\,dz \end{aligned} \tag{17}$$

The $p_\theta(x|z), p_\theta(z|x)$, and $p_\theta(z)$ in Eq. (17) are usually mentioned as probability, posterior, and previous. The target is double and contains the discovery of the parameterization, which maximizes $p_\theta(x)$ for the phenomenon of detected samples of data $x \in X \subset \mathbb{R}^n$. Conversely, the posterior $p_\theta(z|x)$ computation is essential to investigate the latent representation. It is challenging to calculate because of the intractable integral in marginalization through z. Regarding the AE framework, the posterior is related to the encoding (system of inference), and the probability is associated with the decoding (system of generative). Utilizing the inference system as an estimate of the posterior $(E_\phi = q_\phi(z|x) \approx p_\theta(z|x))$ results in the aim of reducing the Kullback–Leibler (KL) divergence $D_{KL}(q_\phi(z|x) \,\|p_\theta(z|x))$ Regarding ϕ. An equal expression of the problem is gained over calculus and relocation of the term of KL. In Formula,

$$\begin{aligned} &\log p_\theta(x) - D_{KL}(q_\phi(z|x)\,\|p_\theta \\ &\quad = -\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\ &\quad + \beta D_{KL}\left(q_\phi(z|x)\,\|\,p_\theta(z)\right) \end{aligned} \tag{18}$$

in which $p_\theta(x)$ is estimated through the generator system for $\beta = 1$. The initial period of the left-hand side of Eq. (3) defines the reconstruction abilities, while the second term guarantees adjustment to standardize the latent vector space. The right-hand side of Eq. (18) characterizes low limits of $\log p_\theta(x)$ and is signified as the evidence of lower bound (ELBO).

During $\beta-$ VAE, the β coefficient controls the power by which encoded instances follow the previous distribution, manipulating disentanglement. A greater β has exposed robust disentanglement but might result

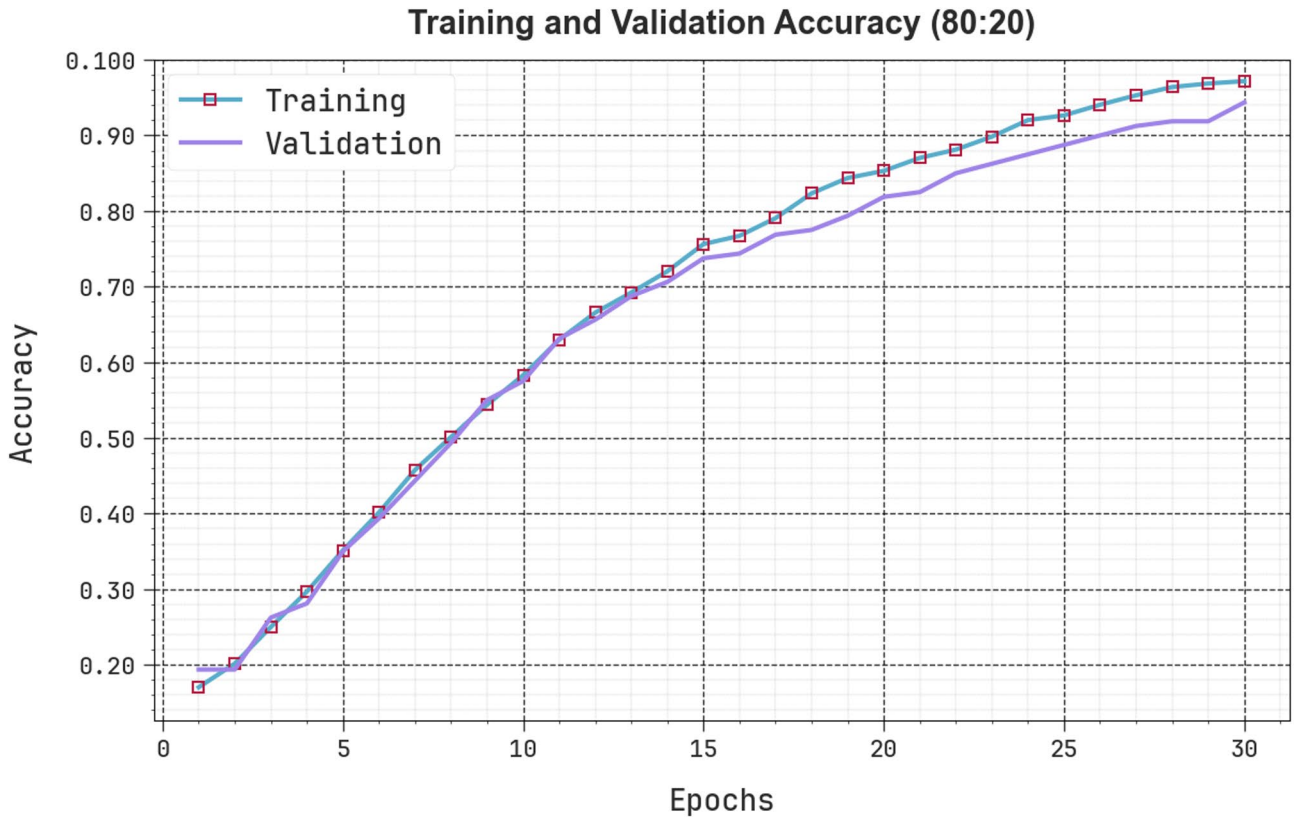

Fig. 7. $Accu_y$ analysis of GRHIP-EDLIBWO approach under 80%TRPH and 20%TSPH.

in reduced reconstruction precision. More decomposing the KL term discloses 3 different mechanisms, as presented, shedding light on the bases of disentanglement, mainly the β complete terms of correlation:

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(x) &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\ &- \alpha I_q(z;x) \\ &- \beta D_{KL}(q_\phi(z)\,|\prod_j q(z_j)) \\ &- \gamma\sum_j D_{KL}(q_\phi(z_j)\,|\mathrm{p}(z_j))\end{aligned} \tag{19}$$

The α parameter controls the mutual data functions I, β the complete term of correlation, and γ the dimensional-wise divergence of KL. The estimation of the distribution $q_\phi(z)$ in Eq. (19) needs the computation of the complete database in every pass of the batches, which may result in a considerable improvement in computational resources and time. During this, an estimate of $q_\phi(z)$ was offered that was utilized as a constant estimator, while N denotes the length of the entire database and B represents the size of the batch:

$$\hat{q}_\phi(z) = \frac{1}{N}q_\phi(z|x) + \frac{N-1}{N(B-1)}\sum_{j\neq j'} q_\phi\left(z^j|x^{j'}\right) \tag{20}$$

The batch size, 512 to 1024, is suggested for a steady estimation and a possible lower bias of Eq. (20).

IBWO-based hyperparameter tuning

Finally, the IBWO model is employed for the hyperparameter tuning of the three ensemble models[34]. This model is chosen due to its biologically inspired mechanism that replicates the cooperative hunting behaviour of beluga whales. Compared to conventional techniques like grid or random search, IBWO presents a more adaptive and robust exploration of the hyperparameter space, mitigating the likelihood of getting stuck in local minima. Its dynamic search ability also ensures improved convergence by balancing exploration and exploitation. The approach is appropriate for complex, non-linear optimization problems where other methods may face difficulty. Furthermore, the flexibility and ability of the IBWO method to scale make it a superior choice in high-dimensional parameter spaces.

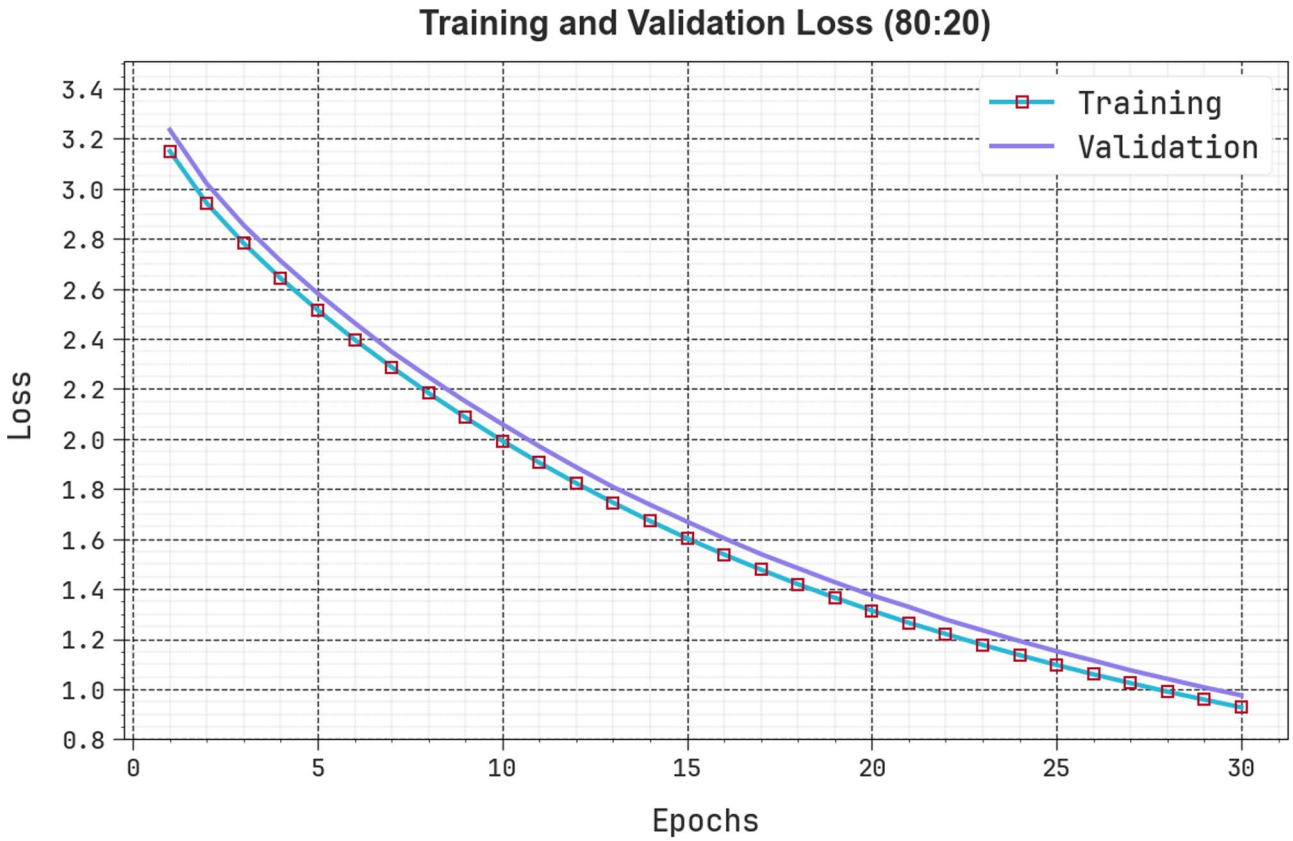

Fig. 8. Loss graph of GRHIP-EDLIBWO approach under 80%TRPH and 20%TSPH.

BWO is a heuristic global optimizer method presented. It is derived from the social behaviour imitation of beluga whales naturally. The normal BWO model establishes strong optimizer abilities but has drawbacks. These comprise an absence of adaptableness vulnerability to local ideals and a variation amongst global and local searching capabilities once presented with intricate optimizer challenges. This work presents the initialization of chaotic and a non-linear wave feature to improve the normal BWO method, thus enhancing its adaptability and exploration capability.

Initially, the chaotic maps were applied to initialize the populations to enhance the local optimizer issue of the normal BWO method. Chaos is a non-linear phenomenon remaining naturally, and an arbitrary point by no reiteration is produced in a particular range and time. Then, chaos maps are presented in this study to make arbitrary points with no reiteration among $(0, 1)$, to gain a regularly dispersed primary population, and to enhance the optimizer efficacy of the method. During this study, logistic chaotic maps, extensively applied in intellectual models and have good optimizer properties, were chosen to enhance the normal BWO. The computation equation of the functional mapping is as shown:

$$x = 4x\left(1 - x_j\right) \tag{21}$$

x_j and x_{j+1} values represent between 0 and 1. The initialization of mapping the location of all individuals by producing complex sequences enhances the exploration capability and diversity of the population and random features, assists the model in preventing local optimal issues, and improves the global searching capability of the method. The selection of Logistic chaotic mapping within the IBWO model contributes significantly to performing the optimum collection of features by guaranteeing a stable and different primary distribution. These diversities avoid premature convergence and improve the model's capability for exploring the feature area widely, which is crucial to identifying the optimum feature subsection globally.

Then, the combination of a non-linear fluctuation feature aids in improving the adaptability and flexibility of the model, thus tackling the flaws related to the normal BWO model, viz., its absence of adaptableness and the imbalance between local and global searching capabilities. This study presents the probability of exploration behaviour $p(T)$ to improve the BWO model. This is stated as shown:

$$p\left(T\right) = 1 - \frac{T}{T_{\max}} \tag{22}$$

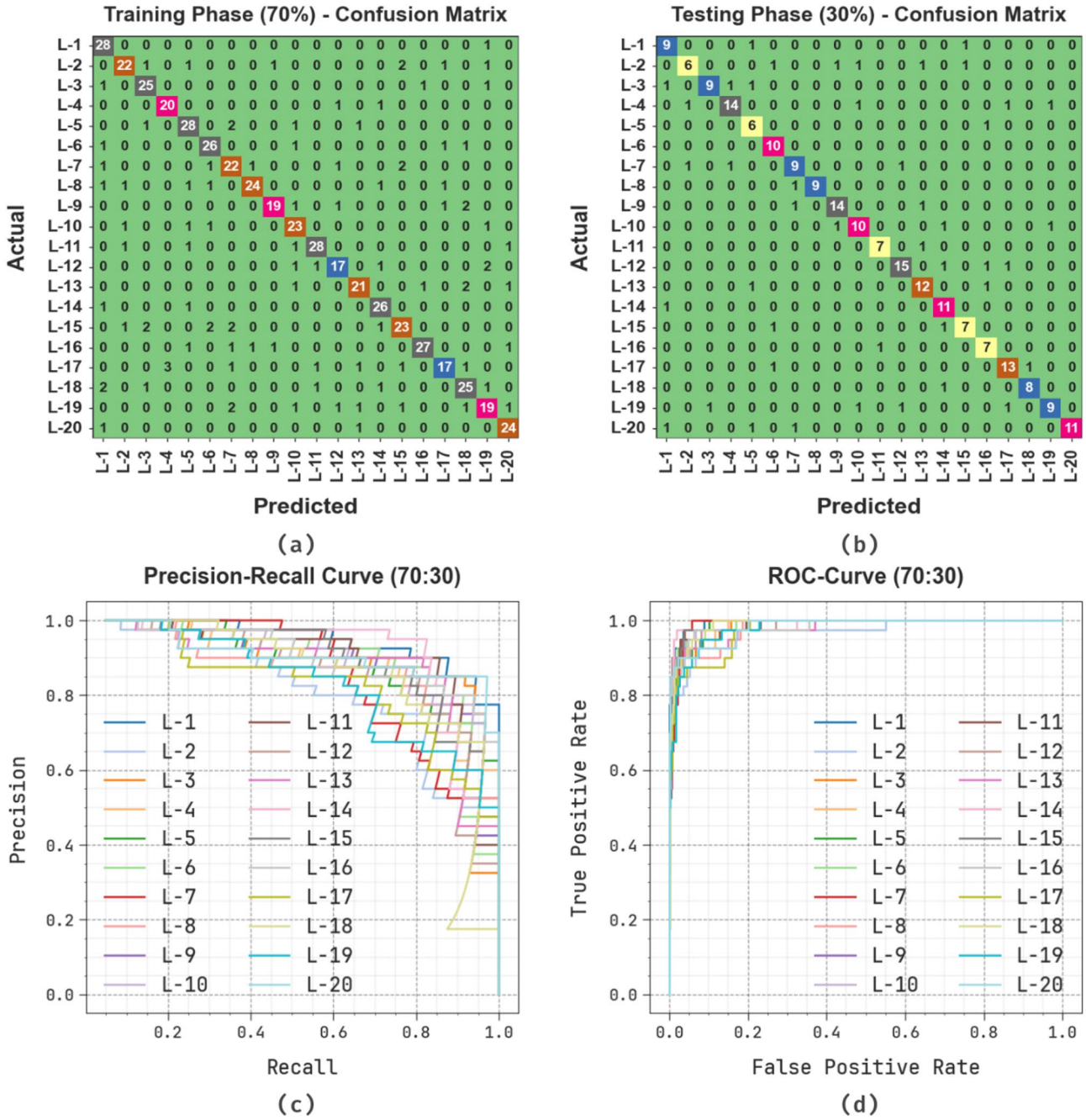

Fig. 9. 70%TRPH and 30%TSPH of (**a**,**b**) of confusion matrix, (**c**,**d**) PR and ROC curves.

whereas $p(T)$ denotes the probability of discovering behaviours at iteration T, T refers to present iteration counts, and $T_{\max}$ signifies maximal iteration counts. $p(T)$ guarantees that exploration probability slowly reduces as time (namely, iteration counts) T improves, permitting the normal BWO method to conduct wide-ranging global hunts in the initial phases and more cautious hunts of local regions in the final stages.

To improve the optimizer capability of local searching behaviour, the model is fine-tuned as demonstrated:

$$\begin{cases} X_i^{T+1} = X_{best}^T + \Psi(T) \times randn(1, D) \\ \Psi(T) = \Psi_{\max}\left(1 - \frac{T}{T_{\max}}\right) \end{cases} \tag{23}$$

In this case, X_i^{T+1} characterizes the upgraded position of a BW i. X_{best}^T symbolizes the optimum position of the BW inside the populations, and Ψ signifies the standard deviation in time applied to fine-tune the perturbation degrees of the solution. Furthermore, $\Psi_{\max}$ represents the first maximal standard deviation, which describes the local searching range at the start of the search. Besides, $randn(1, D)$ makes arbitrary numbers using a D-dimensional standard distribution to present arbitrary perturbations into the solution area. The non-linear

Classes	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$	MCC
TRPH (70%)					
L-1	98.39	77.78	96.55	86.15	85.87
L-2	98.04	84.62	75.86	80.00	79.10
L-3	98.57	83.33	89.29	86.21	85.51
L-4	99.11	86.96	90.91	88.89	88.45
L-5	98.04	82.35	84.85	83.58	82.55
L-6	98.39	83.87	86.67	85.25	84.41
L-7	97.50	73.33	78.57	75.86	74.59
L-8	98.57	92.31	80.00	85.71	85.21
L-9	98.75	90.48	79.17	84.44	84.00
L-10	98.04	76.67	85.19	80.70	79.79
L-11	98.75	90.32	87.50	88.89	88.24
L-12	98.39	80.95	77.27	79.07	78.26
L-13	98.39	84.00	80.77	82.35	81.53
L-14	98.75	83.87	92.86	88.14	87.60
L-15	97.50	79.31	74.19	76.67	75.39
L-16	98.75	93.10	84.38	88.52	87.98
L-17	97.86	80.95	68.00	73.91	73.11
L-18	97.68	78.12	80.65	79.37	78.15
L-19	97.32	73.08	70.37	71.70	70.31
L-20	98.93	85.71	92.31	88.89	88.39
Average	98.29	83.06	82.77	82.72	81.92
TSPH (30%)					
L-1	97.92	75.00	81.82	78.26	77.25
L-2	97.08	75.00	54.55	63.16	62.53
L-3	98.33	90.00	75.00	81.82	81.32
L-4	97.50	87.50	77.78	82.35	81.18
L-5	97.92	60.00	85.71	70.59	70.73
L-6	98.75	76.92	100.00	86.96	87.13
L-7	97.50	75.00	75.00	75.00	73.68
L-8	99.58	100.00	90.00	94.74	94.66
L-9	98.33	87.50	87.50	87.50	86.61
L-10	97.50	76.92	76.92	76.92	75.60
L-11	99.17	87.50	87.50	87.50	87.07
L-12	97.92	88.24	83.33	85.71	84.63
L-13	97.92	80.00	85.71	82.76	81.71
L-14	97.92	73.33	91.67	81.48	80.95
L-15	98.33	77.78	77.78	77.78	76.91
L-16	98.33	70.00	87.50	77.78	77.44
L-17	97.92	81.25	86.67	83.87	82.81
L-18	99.17	88.89	88.89	88.89	88.46
L-19	97.50	81.82	69.23	75.00	73.98
L-20	98.75	100.00	78.57	88.00	88.06
Average	98.17	81.63	82.06	81.30	80.64

Table 3. GR of GRHIP-EDLIBWO method under 70%TRPH and 30%TSPH.

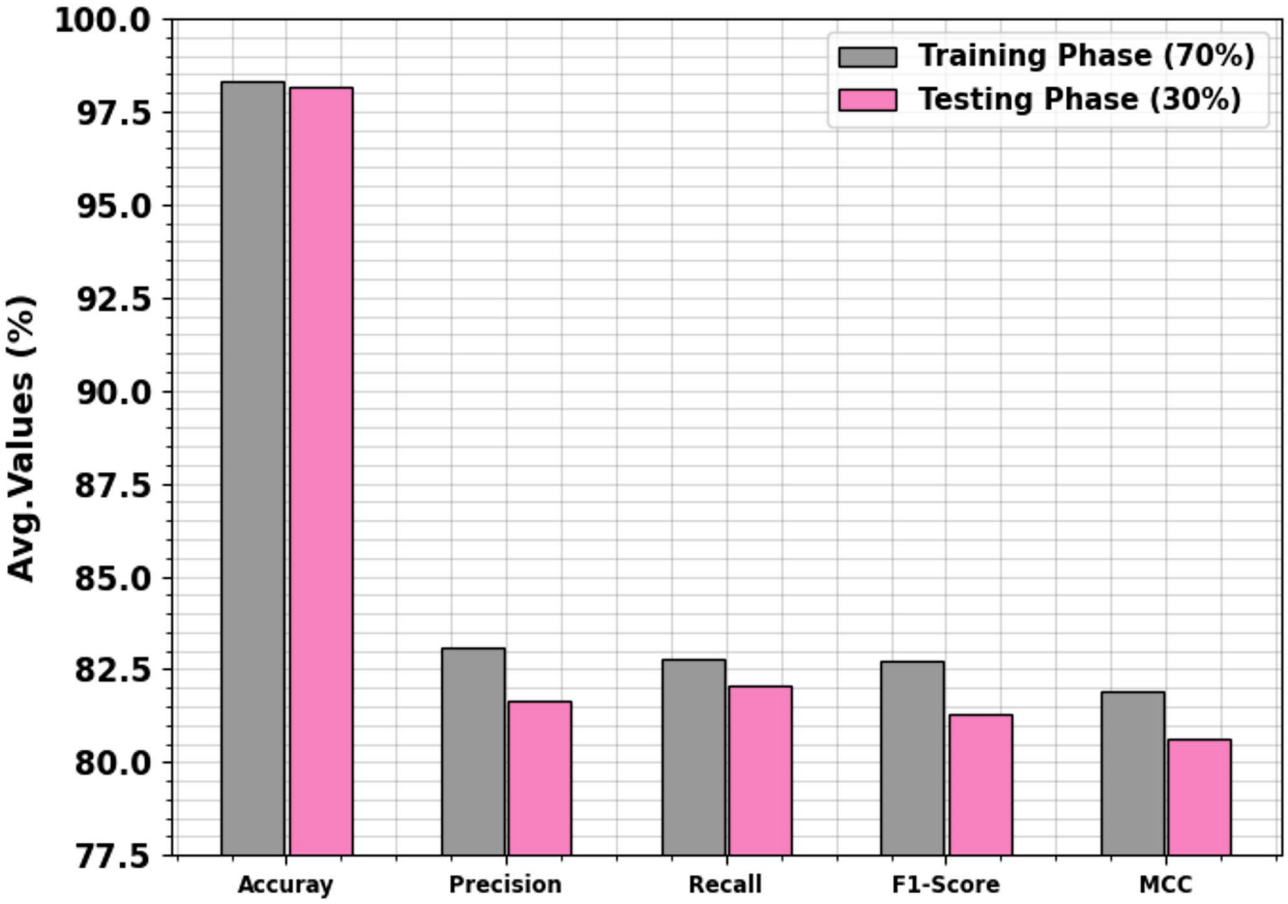

Fig. 10. Average of GRHIP-EDLIBWO model under 70%TRPH and 30%TSPH.

reducing standard deviation modification model enables the slow convergence of random interruptions from wide to limited ranges, thus permitting the model to gain a modified search for the global optimum solutions. This development of the normal BWO method enables a balance between the development and exploration phases, permitting the recognition of the optimum grouping of features and, subsequently, the improved optimizer effects.

The IBWO model originates from a fitness function (FF) to attain boosted classification performance. It outlines an optimistic number to embody the better outcome of the candidate solution. In this work, the classifier error rate reductions were reflected as FF. Its mathematical formulation is represented in Eq. (24).

$$\begin{aligned} fitness\,(x_i) &= Classifier\ Error\ Rate\,(x_i) \\ &= \frac{No.\ of\ misclassified\ samples}{Total\ no.\ of\ samples} \times 100 \end{aligned} \tag{24}$$

Experimental analysis

The performance evaluation of the GRHIP-EDLIBWO technique is verified under an ISL dataset[35]. The dataset covers 800 images under 20 different phrases, each represented by 40 images, as depicted in Table 1. Figure 4 illustrates the sample images.

Figure 5 presents the classifier results of the GRHIP-EDLIBWO methodology under 80%TRPH and 20%TSPH. Figure 5a,b illustrates the confusion matrix with correct recognition and classification of all classes. Figure 5c demonstrates the PR curve, signifying superior performance over all class labels. At the same time, Fig. 5d depicts the ROC analysis, demonstrating proficient outcomes with better ROC analysis for dissimilar class labels.

Table 2 and Fig. 6 represent the GR of the GRHIP-EDLIBWO approach under 80%TRPH and 20%TSPH. The outcomes imply that the GRHIP-EDLIBWO approach correctly identified the samples. With 80%TRPH, the GRHIP-EDLIBWO technique presents an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 98.72%, 87.37%, 87.14%, 87.04%, and 86.48%, correspondingly. Additionally, with 20%TRPH, the GRHIP-EDLIBWO technique presents an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 98.44%, 84.70%, 84.01%, 83.46%, and 83.13%, respectively.

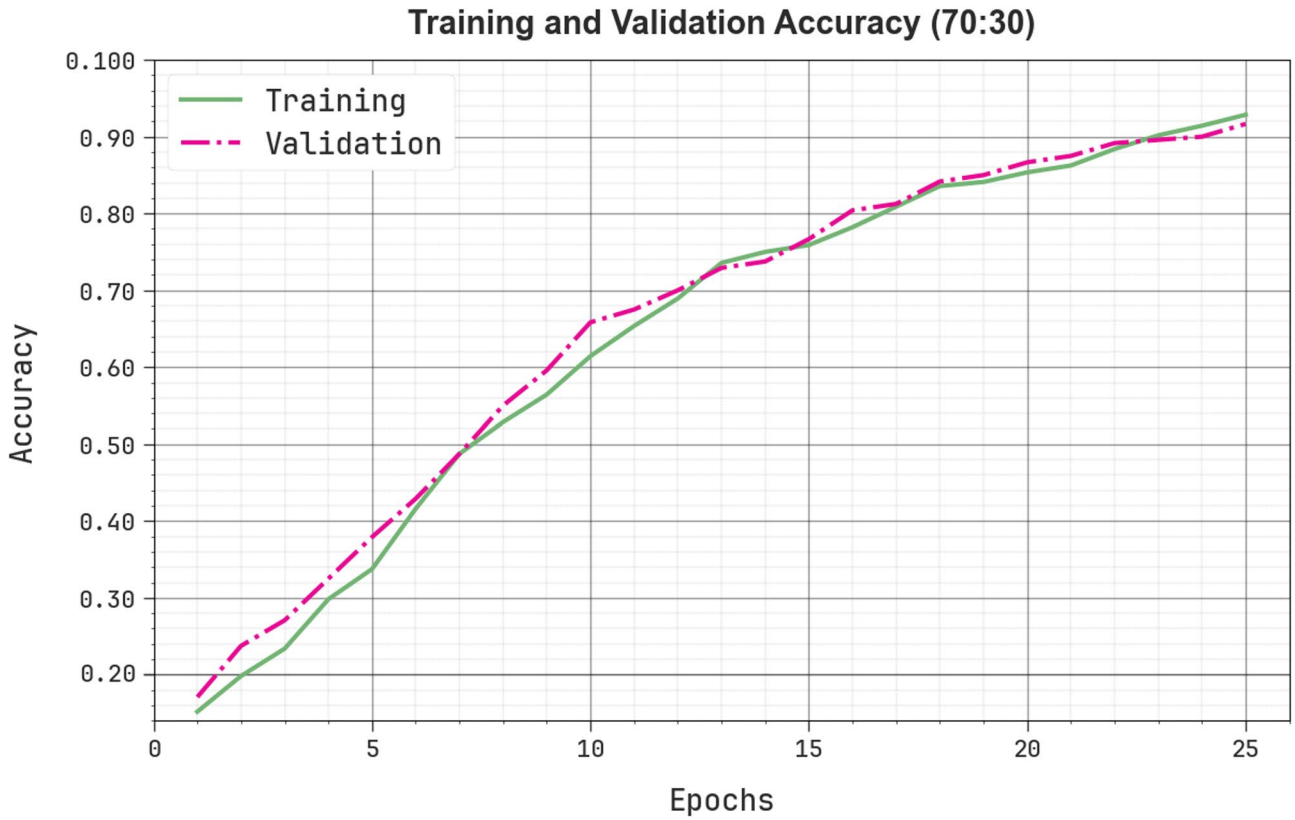

Fig. 11. $Accu_y$ analysis of GRHIP-EDLIBWO model under 70%TRPH and 30%TSPH.

Figure 7 illustrates the training (TRA) $accu_y$ and validation (VAL) $accu_y$ analysis of the GRHIP-EDLIBWO methodology at 80%TRPH and 20%TSPH. The $accu_y$ analysis is computed across the 0–30 epoch counts range. The figure highlights that the TRA and VAL $accu_y$ values display increasing tendencies, which informs the capacity of the GRHIP-EDLIBWO methodology with maximal outcomes over several iterations. Likewise, the TRA and VAL $accu_y$ is closer across the epochs, which identifies inferior overfitting and presents the higher performance of the GRHIP-EDLIBWO model, ensuring reliable prediction on unseen instances.

Figure 8 shows the TRA loss (TRALOS) and VAL loss (VALLOS) curves of the GRHIP-EDLIBWO technique under 80%TRPH and 20%TSPH. The loss values are computed over 0–30 epoch counts. It is denoted that the TRALOS and VALLOS values exemplify decreasing tendencies, informing the capabilities of the GRHIP-EDLIBWO model to balance a trade-off between generality and data fitting. The constant fall in loss values ensures the GRHIP-EDLIBWO system's maximum performance and tunes the prediction results in time.

Figure 9 represents the classifier outcomes of the GRHIP-EDLIBWO approach under 70%TRPH and 30%TSPH. Figure 9a,b illustrates the confusion matrices with correct classification and identification of each class label. Figure 9c exhibitions the PR analysis, representing superior outcomes across all class labels. Simultaneously, Fig. 9d illustrates the ROC values, demonstrating proficient results with better ROC values for dissimilar classes.

Table 3 and Fig. 10 signify the GR of GRHIP-EDLIBWO methodology under 70%TRPH and 30%TSPH. The outcomes suggest that the GRHIP-EDLIBWO methodology correctly identified the samples. With 70%TRPH, the GRHIP-EDLIBWO methodology presents an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 98.29%, 83.06%, 82.77%, 82.72%, and 81.92%, correspondingly. In addition, with 30%TRPH, the GRHIP-EDLIBWO approach presents an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 98.17%, 81.63%, 82.06%, 81.30%, and 80.64%, respectively.

Figure 11 demonstrates the TRA $accu_y$ and VAL $accu_y$ analysis of the GRHIP-EDLIBWO methodology under 70%TRPH and 30%TSPH. The $accu_y$ analysis is computed within the 0–25 epoch counts range. The figure highlights that the TRA and VAL $accu_y$ analysis exhibits an increasing trend, which notified the capacity of the GRHIP-EDLIBWO methodology with maximum outcome across multiple iterations. Besides, the TRA and VAL $accu_y$ remain adjacent across the epoch counts, which indicates inferior overfitting and demonstrates maximum outcomes of the GRHIP-EDLIBWO method, guaranteeing constant prediction on unseen instances.

Figure 12 establishes the TRALOS and VALLOS analysis of the GRHIP-EDLIBWO methodology under 70%TRPH and 30%TSPH. The loss values are computed over 0–25 epoch counts. The TRALOS and VALLOS values exemplify a reducing trend, notifying the capacity of the GRHIP-EDLIBWO methodology to balance an exchange between generality and data fitting. The continual reduction in loss values pledges more significant outcomes for the GRHIP-EDLIBWO method and tunes the prediction results in time.

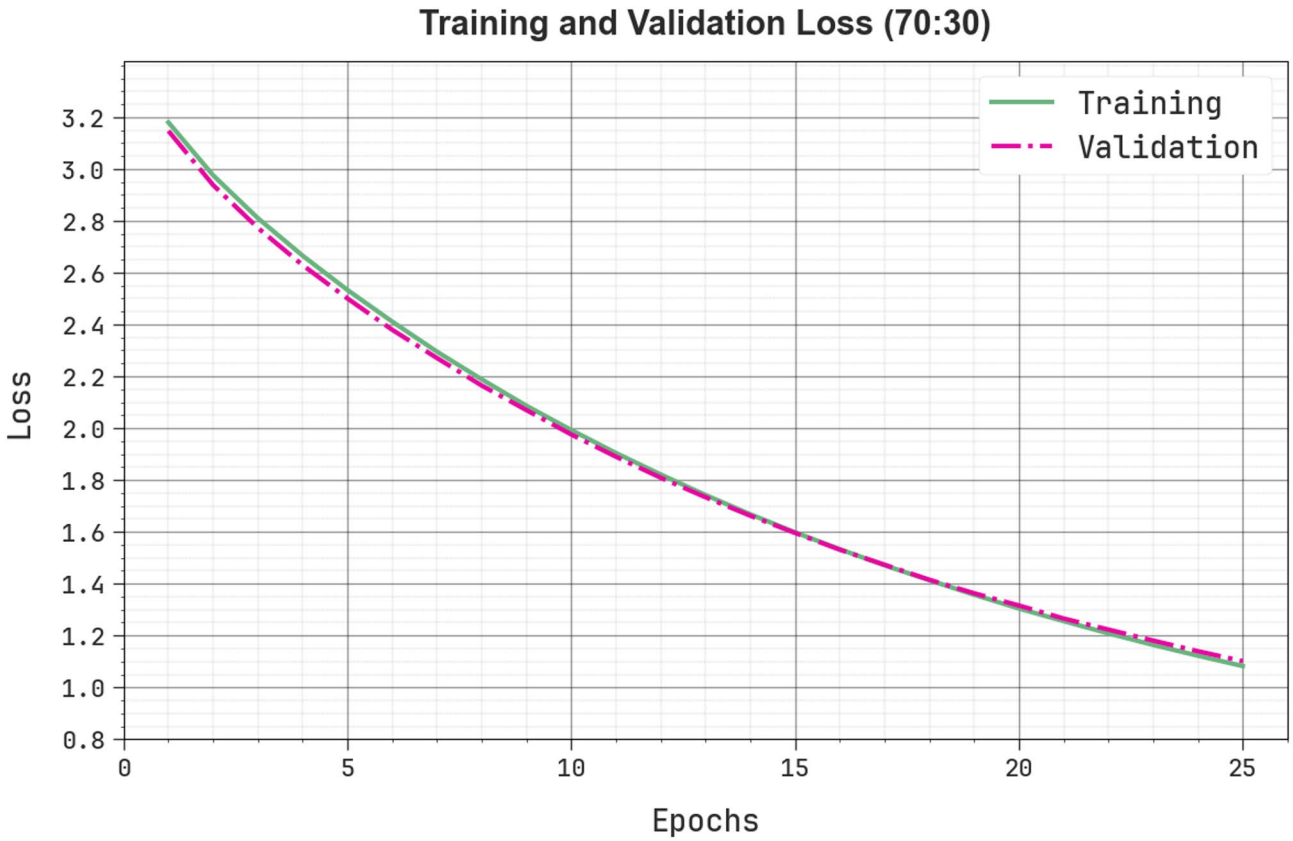

Fig. 12. Loss graph of GRHIP-EDLIBWO technique under 70%TRPH and 30%TSPH.

Methodology	$Accu_y$	$Prec_n$	$Recal$	$F1_{score}$
2DCNN and LSTM-LMS	89.50	82.67	83.09	81.28
CNN-2layer LSTM	94.00	76.25	73.91	86.63
3DCNN-SL-GCN	98.00	78.61	74.36	76.15
CNN + RNN	75.11	69.46	71.13	70.21
Pose Estimation + LSTM	94.99	79.90	86.87	83.02
LiST-LFCISLT	96.92	84.91	78.35	80.32
ANN Algorithm	89.45	74.44	86.00	81.80
MLP Model	90.04	76.97	84.16	73.29
hDNN-SLR	97.10	74.99	81.69	75.55
Bi-LSTM	92.84	81.55	82.01	84.09
HNN	91.58	75.78	85.70	85.55
VA-E	92.05	73.00	85.54	78.21
GRHIP-EDLIBWO	98.72	87.37	87.14	87.04

Table 4. Comparative outcomes of GRHIP-EDLIBWO methodology with existing models[20,21,36–38].

Table 4 and Fig. 13 describe the comparative analysis of GRHIP-EDLIBWO methodology with existing techniques[20,21,36–38]. The table values indicated that the proposed GRHIP-EDLIBWO methodology has attained effectual performance. The results highlighted that the 2DCNN and LSTM-LMS, CNN-2layer LSTM, CNN + RNN, Pose Estimation + LSTM, ANN, and MLP approaches have reported worse performance. Meanwhile, 3DCNN-SL-GCN and LiST-LFCISLT methodologies have reached closer outcomes. Moreover, hybrid Deep Neural Net with SL recognition (hDNN-SLR), Bidirectional LSTM (Bi-LSTM), HNN, and Variational Auto-Encoders (VA-E) methodologies exhibited slightly reduced results. Besides, the GRHIP-EDLIBWO approach reported superior performance with maximal $accu_y$ of 98.72%, $prec_n$ of 87.37%, $recal$ of 87.14%, and $F1_{score}$ of 87.04%.

Table 5 and Fig. 14 show the computational time (CT) analysis of the GRHIP-EDLIBWO technique over existing models. The results show that 2DCNN and LSTM-LMS take the longest at 19.63 s, followed by Pose Estimation LSTM at 20.88 s. Methods like CNN 2layer LSTM at 13.63 s, 3DCNN SL GCN at 12.96 s, and CNN

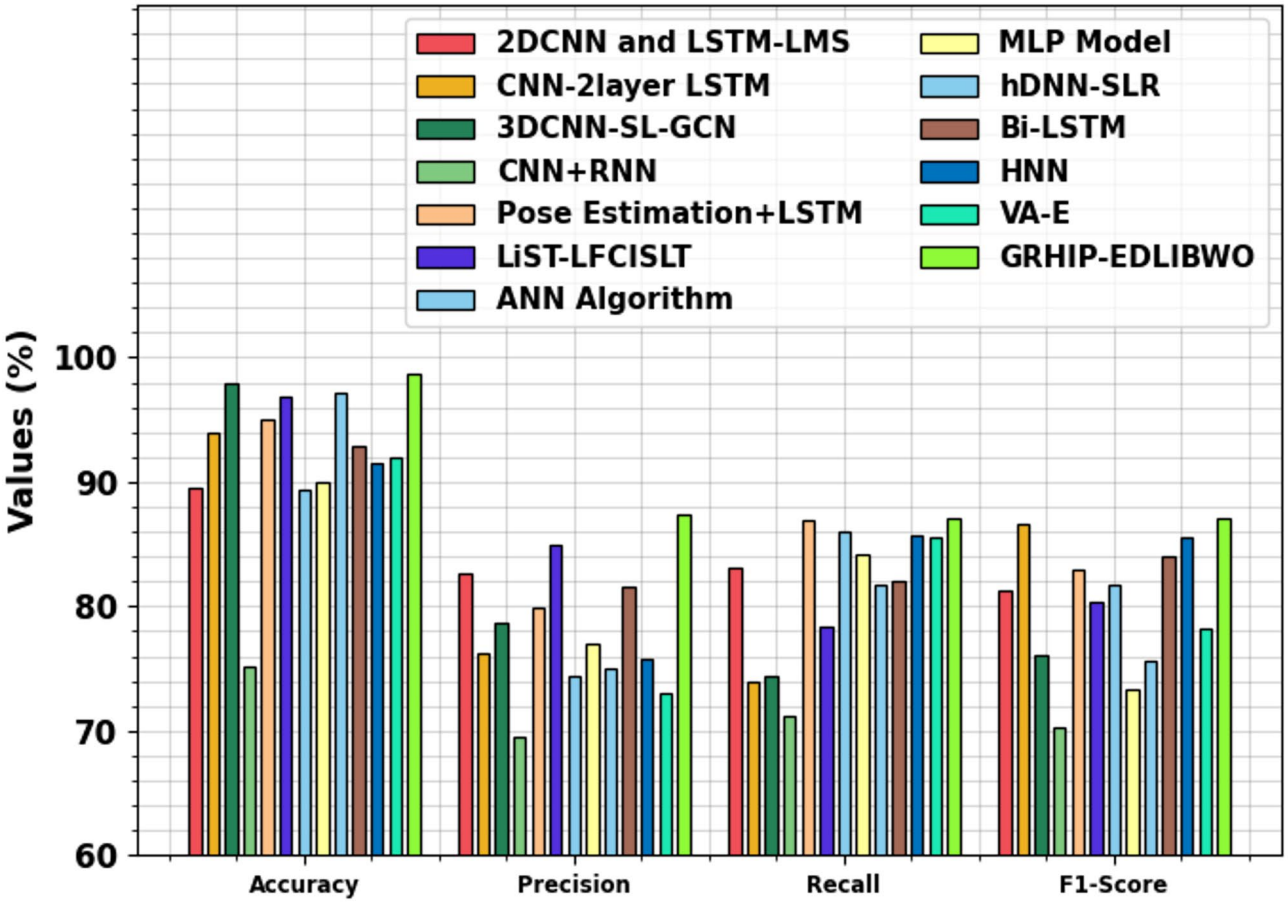

Fig. 13. Comparative analysis of GRHIP-EDLIBWO methodology with existing models.

Methodology	CT (s)
2DCNN and LSTM-LMS	19.63
CNN-2layer LSTM	13.63
3DCNN-SL-GCN	12.96
CNN + RNN	14.35
Pose Estimation + LSTM	20.88
LiST-LFCISLT	14.02
ANN Algorithm	10.12
MLP Model	18.79
hDNN-SLR	20.11
Bi-LSTM	15.33
HNN	17.22
VA-E	14.02
GRHIP-EDLIBWO	9.21

Table 5. CT analysis of GRHIP-EDLIBWO technique with existing models.

RNN at 14.35 s are relatively faster. LiST LFCISLT and VA E each need 14.02 s, while the ANN method performs in 10.12 s. MLP Model takes 18.79 s, and hDNN SLR requires 20.11 s. Bi LSTM and HNN require 15.33 and 17.22 s, respectively, with GRHIP-EDLIBWO being the fastest at 9.21 s. In conclusion, GRHIP-EDLIBWO is the most effective, while Pose Estimation LSTM and hDNN SLR are among the slowest methods.

Conclusion

In this study, the GRHIP-EDLIBWO model is proposed. The main intention of the GRHIP-EDLIBWO model framework for GR is to assist as a valuable tool for developing accessible communication systems for hearing-

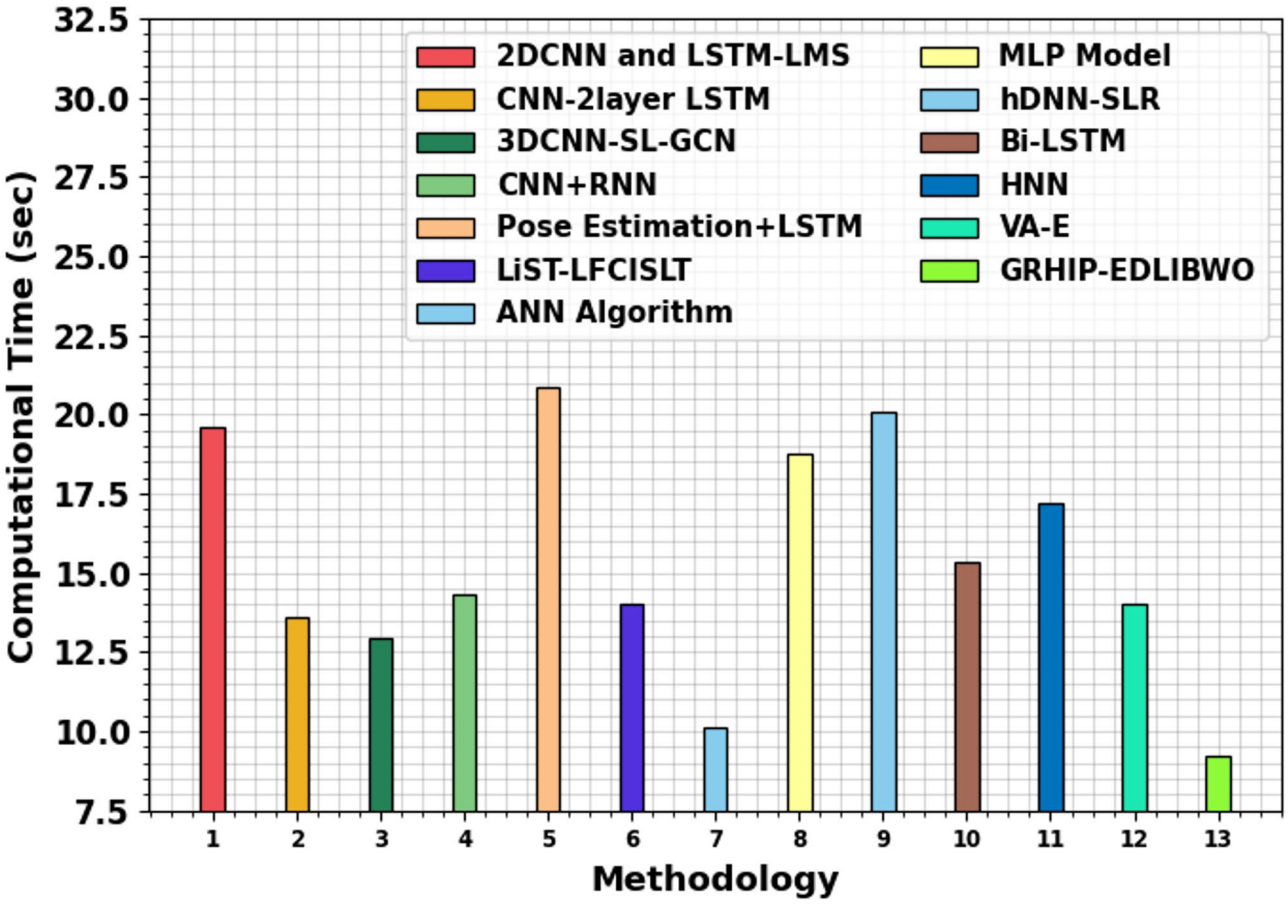

Fig. 14. CT analysis of GRHIP-EDLIBWO technique with existing models.

impaired individuals. To accomplish that, the GRHIP-EDLIBWO method initially performs image preprocessing using SF to enhance edge detection and extract critical gesture features. The SE-CapsNet effectively captures spatial hierarchies and complex relationships within gesture patterns for feature extraction. In addition, an ensemble of classification processes such as BiLSTM**,** BiGRU, and VAE techniques is employed. Eventually, the IBWO method is implemented for the hyperparameter tuning of the three ensemble models. Extensive simulations are conducted on an ISL dataset to achieve a robust classification result with the GRHIP-EDLIBWO approach. The performance validation of the GRHIP-EDLIBWO approach portrayed a superior accuracy value of 98.72% over existing models. The GRHIP-EDLIBWO approach's limitations include reliance on a relatively small dataset, which may affect the generalizability of the results to larger, more diverse populations. The model's performance may also degrade with noisy or incomplete data in real-world conditions. This approach also lacks scalability for handling multilingual or multimodal data, which limits its applicability in diverse linguistic and environmental contexts. Furthermore, the model's real-time performance and computational efficiency could be improved. Future work should expand the dataset, integrate multilingual and multimodal capabilities, improve real-time processing, and optimize the model for improved scalability and robustness in real-world applications.

Data availability

The data supporting this study's findings are openly available in a repository at https://data.mendeley.com/datasets/w7fgy7jvs8/2, reference number[35].

Received: 17 December 2024; Accepted: 10 June 2025
Published online: 01 July 2025

References

1. Padmanandam, K., Rajesh, M. V., Upadhyaya, A. N., Chandrashekar, B. & Sah, S. Artificial intelligence biosensing system on hand gesture recognition for the hearing impaired. *Int. J. Oper.Res. Inf. Syst. (IJORIS)* **13**(2), 1–13 (2022).
2. Bangaru, S. S., Wang, C., Zhou, X., Jeon, H. W. & Li, Y. Gesture recognition–based smart training assistant system for construction worker earplug-wearing training. *J. Constr. Eng. Manag.* **146**(12), 04020144 (2020).
3. Hisham, B. & Hamouda, A. Supervised learning classifiers for Arabic gesture recognition using Kinect V2. *SN Applied Sciences* **1**(7), 768 (2019).
4. Jeyanthi, P., Ajees, A., Kumar, A.P., Revathy, S. & Gladence, M. Interactive hand gesture recognition with audio response. In *Multidisciplinary Applications of AI and Quantum Networking* (pp. 195–212). IGI Global (2025).

5. Alnaim, N. *Hand gesture recognition using deep learning neural networks* (Doctoral dissertation, Brunel University London) (2020).
6. Ascari, R. E. S., Pereira, R. & Silva, L. Computer vision-based methodology to improve interaction for people with motor and speech impairment. *ACM Trans. Access. Comput. (TACCESS)* **13**(4), 1–33 (2020).
7. Bohra, T., Sompura, S., Parekh, K. & Raut, P. Real-time two way communication system for speech and hearing impaired using computer vision and deep learning. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 734–739). IEEE (2019).
8. Kareem, D. A. & Rajesh, D. Enhancing WBAN performance with cluster-based routing protocol using black widow optimization for healthcare application. *J. Intell. Syst. Internet Things* **14**(1) (2025).
9. Tateno, S., Liu, H. & Ou, J. Development of sign language motion recognition system for hearing-impaired people using electromyography signal. *Sensors* **20**(20), 5807 (2020).
10. Allehaibi, K. H. Artificial Intelligence based automated sign gesture recognition solutions for visually challenged people. *J. Intell. Syst. Internet Things* **2**, 127–227 (2025).
11. Sümbül, H. A novel mems and flex sensor-based hand gesture recognition and regenerating system using deep learning model. *IEEE Access* (2024).
12. Hossain, S. S., Das, P. & Bhattacharya, I. Hand Gesture recognition using deep learning for deaf and dumb community. In *International Conference on Frontiers in Computing and Systems* (pp. 443–455). Singapore: Springer Nature Singapore (2023).
13. Vyshnavi, S. L., Chandana, N., Ramya, N. N. S. & Suvarna, B. GestureSense: A deep learning-based gesture language translator using VGG1. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)* (pp. 484–488). IEEE (2024).
14. Ravinder, M. et al. An approach for gesture recognition based on a lightweight convolutional neural network. *Int. J. Artif. Intell. Tools* **32**(03), 2340014 (2023).
15. Shinde, S., Mahalle, P., Panchal, S., Mahalle, S., Pandit, A. & Tonpe, P. Sign language recognition using deep learning. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–5). IEEE (2024).
16. Barbhuiya, A. A., Karsh, R. K. & Jain, R. ASL hand gesture classification and localization using deep ensemble neural network. *Arab. J. Sci. Eng.* **48**(5), 6689–6702 (2023).
17. Kavitha, M. N., Saranya, S. S., Prasad, M., Kaviyarasu, S., Ragunath, N. & Rahul, P. An ensembled real-time hand-gesture recognition using CNN. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–5). IEEE (2024).
18. Izzalhaqqi, M. Y. D. Gesture recognition in Indonesian Sign language using hybrid deep learning models. In *2023 International Workshop on Intelligent Systems (IWIS)* (pp. 1–6). IEEE (2023).
19. Ramadan, A. F. & Abd-Alsabour, N. A novel control system for a laptop with gestures recognition. *J. Trends Comput. Sci. Smart Technol.* **6**(3), 213–234 (2024).
20. Rajalakshmi, E. et al. Static and dynamic isolated Indian and Russian sign language recognition with spatial and temporal feature detection using hybrid neural network. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **22**(1), 1–23 (2022).
21. Rajalakshmi, E. et al. Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture. *IEEE Access* **11**, 2226–2238 (2023).
22. Shanthi, R., Azeez, S. A. & Kumar, N. D. Virtual painting through hand gestures: A machine learning approach. *J. Ubiquitous Comput. Commun. Technol.* **6**(1), 39–49 (2024).
23. Ramkarthik, D. & Benita, A. Integrating quantum networking with explainable AI and ensemble learning approaches for enhanced sign language recognition: Indian Sign language (ISL), convolutional neural networks (CNNs). In *Multidisciplinary Applications of AI and Quantum Networking* (pp. 213–226). IGI Global (2025).
24. Narayan, S. & Jain, V. K. Enhanced hand gesture recognition using image transformer model and particle swarm optimization. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (pp. 1–9). IEEE (2024).
25. Hasan, K. R. & Adnan, M. A. EMPATH: MediaPipe-aided ensemble learning with attention-based transformers for accurate recognition of Bangla Word-Level Sign Language. In *International Conference on Pattern Recognition* (pp. 355–371). Springer, Cham (2025).
26. Marzouk, R., Aldehim, G., Al-Hagery, M. A., Hilal, A. M. & Alneil, A. A. Automated gesture recognition using artificial rabbits optimization with deep learning for assisting visually challenged people. *Fractals* 2450131 (2024).
27. Tounsi, M., Ali, H., Azar, A. T., Al-Khayyat, A. & Ibraheem, I. K. Comprehensive Learning salp swarm algorithm with ensemble deep learning-based ECG signal classification on internet of things environment. *Eng. Technol. Appl. Sci. Res.* **15**(1), 19492–19500 (2025).
28. Manoharan, J. & Sivagnanam, Y. Enhanced hand gesture recognition using optimized preprocessing and VGG16-based deep learning model. In *2024 10th International Conference on Communication and Signal Processing (ICCSP)* (pp. 1101–1105). IEEE (2024).
29. Wang, X. et al. Static gesture segmentation technique based on improved Sobel operator. *J. Eng.* **2019**(22), 8339–8342 (2019).
30. Bian, L., Zhang, L., Zhao, K., Wang, H. & Gong, S. Image-based scam detection method using an attention capsule network. *IEEE Access* **9**, 33654–33665 (2021).
31. Siami-Namini, S., Tavakoli, N. & Namin, A. S. The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)* (pp. 3285–3292). IEEE (2019).
32. Zhang, L. & Xue, G. Short-Term Heat Load Forecasting Based on Ceemd and a Hybrid Idbo-Tcn-Bigru Network. *Available at SSRN 5030114*.
33. Kapsecker, M., Möller, M. C. & Jonas, S. M. Disentangled representational learning for anomaly detection in single-lead electrocardiogram signals using variational autoencoder. *Comput. Biol. Med.* **184**, 109422 (2025).
34. Wang, J., Kong, Z., Shan, J., Du, C. & Wang, C. Corrosion rate prediction of buried oil and gas pipelines: A new deep learning method based on RF and IBWO-optimized BiLSTM–GRU combined model. *Energies* **17**(23), 5824 (2024).
35. https://data.mendeley.com/datasets/w7fgy7jvs8/2
36. Kothadiya, D. et al. Deepsign: Sign language detection and recognition using deep learning. *Electronics* **11**(11), 1780 (2022).
37. Poonia, R. C. LiST: a lightweight framework for continuous indian sign language translation. *Information* **14**(2), 79 (2023).
38. Kothadiya, D. R., Bhatt, C. M., Kharwa, H. & Albu, F. Hybrid InceptionNet based enhanced architecture for isolated sign language recognition. *IEEE Access* (2024).

Author contributions

Mohammad Assiri: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Conceptualization, Project administration. Mahmoud Selim: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Investigation, Formal analysis. Data curation, Methodology, Formal analysis.

Funding
The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group no KSRG-2024-064.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.