Journal of Computer Science

Original Research Paper

Multimodal Face Expression Recognition Using Parametric Exponential Linear Unit-Long Short-Term Memory

[1]Kampa Ratna Babu, [2]Akula Suneetha and [3]Kampa Kanthi Kumar

[1]Department of Computer Engineering, M.B.T.S. Government Polytechnic, Guntur, India
[2]Department of Computer Science and Engineering, KKR and KSR
Institute of Technology and Sciences, Guntur, India
[3]Department of Electronics and Communication Engineering, Tirumala Engineering College, Narasaraopet, India

Article history
Received: 17-05-2024
Revised: 15-06-2024
Accepted: 08-07-2024

Corresponding Author:
Kampa Ratna Babu
Department of Computer Engineering, M.B.T.S. Government Polytechnic, Guntur, India
Email: kratnababu@outlook.com

Abstract: Multimodal facial expression recognition combines information from multiple modalities of audio and video to achieve the required accuracy and robustness. By integrating different data sources, multimodal systems capture different aspects of human expression. However, accurately recognizing facial expressions across audio and video modalities causes challenges due to variations in expression representation. In this research, Parametric Exponential Linear Unit-Long Short-Term Memory (PELU-LSTM) is proposed to accurately recognize multimodal facial expressions. Initially, the SAVEE dataset is used to evaluate the performance of the proposed method which contains audio and video frames. In audio pre-processing, a wiener filter is deployed to minimize background noise, while a Gaussian Weighting Function (GWF) is employed to aggregate the entire video into a smaller number of frames which also minimizes the information loss. The Mel-Frequency Cepstral Coefficient (MFCC) is utilized to extract audio features, while the Histogram of Gradient (HOG) and Local Binary Pattern (LBP) are employed for extracting the video features. Then, concatenation is performed to fuse a single feature vector. Finally, PELU-LSTM recognizes the facial emotional expressions accurately. The proposed technique achieves a high accuracy of 99.75%, as compared to the existing techniques like Bi-directional LSTM-Convolution Neural Networks (Bi-LSTM-CNN), attention-based 2D CNN with LSTM and K-means clustering-based Kernel Canonical Correlation Analysis (KMKCCA).

Keywords: Gaussian Weighting Function, Histogram of Gradient, Long Short-Term Memory, Parametric Exponential Linear Unit, Wiener Filter

Introduction

Facial Expression Recognition (FER) is one of the significant visual recognition technologies to identify the emotions of humans (Fard and Mahoor, 2022; Xiao *et al*., 2022). Decision-making and attention management are the major parts dependent on emotions in a human's daily life. Emotionally imbalanced people are less active in performing daily tasks, representing that emotions play a vital role in one's life (Pandey and Seeja, 2022). Human emotion recognition is determined in numerous ways including facial expression, speech data, physiological parameters, body gestures and so on. As each of these modalities is distinct, combining their outcomes in a rich feature representation makes it capable of performing effective emotion recognition (Middya *et al*., 2022; Bodapati *et al*., 2022). It is helpful in various tasks like criminal justice systems, security monitoring, e-learning, customer satisfaction identification, smart card applications, social robots and so on. The primary blocks in the traditional system of emotion recognition are identifying faces, extracting features and categorizing emotions (Chowdary *et al*., 2023). Disgust, happiness, surprise, anger, fear and sorrow are the primary expressions utilized by people for communicating their emotions (Wang *et al*., 2023; Boughida *et al*., 2022). Positive emotions not only make better communication but also maximize the productivity of work-life. Negative emotions cause both physical and mental health, while negatively affecting the human social environment (Aslan, 2022). FER assists in performance-based human emotional intelligence assessment by evaluating the accurate generation of facial expressions for certain emotions and accurate understanding of facial expressions produced by another person (Kim *et al*., 2022). Certain configurations of facial muscle movements provide an impression of people's emotions (Jothimani and Premalatha, 2022; Yu and Xu, 2022). The process of FER has four parts: Image acquisition, pre-processing, feature extraction and classification (Nan *et al*., 2022; Liu *et al*., 2024). Recently, Deep Learning (DL) techniques have been utilized to solve various critical issues.

The benefits of DL come from its capability to learn high-level features (Abdelhamid *et al*., 2022). The existing methods like Bi-directional Long Short-Term Memory with Convolutional Neural Network (BiLSTM-CNN), attention-based 2DCNN with LSTM, K-Means clustering-based Kernel Canonical Correlation Analysis (KMKCCA), 3DCNN-ConvLSTM and MFCC Time-domain feature with IDCNN (MFCCT-1DCNN) have certain benefits, as well as limitations. The benefits of the portfolio of emotions are increased in attention-based 2DCNN with LSTM by combining multiple data sources which leads to higher performances. 3DCNN-ConvLSTM captures both spatial and temporal dependencies which provide dynamic emotion patterns. The limitations are struggling with capturing fine-grained temporal dependencies across audio and visual modalities and sensitivity to noise and varying speech patterns. However, accurately recognizing facial expressions across audio and video modalities causes challenges due to variations in expression representation. These variations arise from differences in individual facial features, occlusions and the dynamic nature of expressions. In order to overcome this issue, PELU-LSTM is proposed for recognizing multimodal facial expressions accurately. PELU's parametrized activation function increases the network's capability to represent various expressions. By adjusting its parameters, PELU manages variations in expressions across different individuals. The LSTM's sequential processing retains temporal content that accurately recognizes FER. Incorporating audio and video modalities generates a better understanding of expressions. PELU-LSTM integrates these modalities effectively by utilizing the strength of both, further enhancing the overall performance of recognition.

The main contributions of this research are as follows:

- The Wiener filter is used to minimize the background noise in audio pre-processing, while GWF is used to aggregate the entire video into fewer frames. This process helps in reducing background noise and information loss which enables accurate and reliable facial expressions
- From audio pre-processed data, pitch and harmonics are extracted using MFCC, while HoG and LBP are utilized to capture feature descriptors like shape, edge and texture in images, which further enhance the robustness and accuracy of FER across multiple modalities
- PELU-LSTM is performed to accurately recognize facial expressions. LSTM efficiently learns temporal facial expression evolution, while retaining spatial context. PELU captures a smooth transition among both positive and negative slopes which potentially minimizes the vanishing gradient problems and increases the learning dynamics that provide an accurate recognition in FER

Literature Survey

The related works of FER are discussed with various methods using DL along with their advantages and disadvantages provided below. This analysis assists in recognizing gaps and guiding the development of a more effective and accurate recognition approach.

Sharafi *et al*. (2022) implemented a hybrid technique that contained BiLSTM-CNN for recognizing facial emotions in audio and visual data. The spatial and temporal features were extracted from video frames and fused with MFCC and from audio signals, the energy features were extracted and fed into the BiLSTM output. At last, a SoftMax classifier was utilized to categorize input into a set of target classes. The CNN-BiLSTM-based real-time emotion detection for learning images was adjusted in order to maximize the implemented technique's efficiency and accuracy. However, BiLSTM-CNN struggled with capturing fine-grained temporal dependencies across audio and visual modalities due to the inherent complexity of combining recurrent and convolutional layers.

Singh *et al*. (2023a) suggested an attention-based 2DCNN with LSTM to recognize speech emotion recognition. Initially, normalization and augmentation were employed to normalize the features and to avoid overfitting. Then, the MFCC was utilized to extract the features from pre-processed data. At last, a model was established with four local feature learning blocks depending on 2DCNN-2DCNN with LSTM block for learning long-term dependencies and an attention layer obtained significant data generated by LSTM cell, while the leftover features were dropped out for recognizing emotions. The portfolio of emotions was increased by combining multiple data sources which led to high performances. However, the suggested approach suffered from sensitivity to noise and varying speech patterns due to the model's challenge in distinguishing relevant features.

Chen *et al*. (2022) presented a KMKCCA for recognizing multimodal facial emotions. The multimodal features such as time, frequency domain and grayscale were extracted from facial expressions, where the speech was fused depending on the analysis of kernel canonical correlation. The k-means technique was utilized to choose the features from various modalities and minimize dimensionality. The presented approach increased the heterogenicity between different modalities and enabled complementary multiple modalities to facilitate multimodal emotion recognition. Nonetheless, KMKCCA struggled with noisy or ambiguous data points due to K-means cluster depending on Euclidean distances which were sensitive to noise and outliers.

Singh *et al.* (2023b) introduced a 3DCNN-ConvLSTM to recognize facial emotions in video. Initially, pre-processing generated the video frames as input and transferred facial location landmarks. Then, the framework aligned the identified faces to obtain spatial symmetry depending on the information on facial landmarks. The aligned facial images were cropped and scaled to a standard size. At last, 3DCNN-ConvLSTM generated the pre-processed facial images as input in a fixed duration and learned to classify the expressions. By using CNN and LSTM, it captured both spatial and temporal dependencies which provided dynamic emotion patterns. However, 3DCNN-ConvLSTM suffered from overfitting while managing high-dimensional spatiotemporal data.

Alluhaidan *et al.* (2023) developed an MFCCT-1DCNN to recognize speech expressions. The MFCCT was used to extract the audio features and these extracted features were concatenated utilizing the fusion phase. Then, the CNN technique containing one dimension had activation, max-pooling layers, dropout and Fully Connected (FC) layers for classifying the speech expressions. The developed approach combined both time and frequency domains of audio signal which increased the diversity, robustness, reliabilities and generalization of speech recognition. Still, MFCCT-IDCNN suffered from spectral information loss due to the transformation of the frequency domain to the time domain.

From the overall analysis, it is observed that the existing techniques have limitations of struggles with capturing fine-grained temporal dependencies across audio and visual modalities, sensitivity to noise and varying speech patterns due to the model's difficulty in distinguishing relevant features and combining multiple modalities, thereby leading to complex fusion. To overcome this issue, the PELU-LSTM is proposed to accurately recognize facial expressions. PELU's parametrized activation function increases the network's ability to model different expressions, while the LSTM’s sequential processing retained the temporal content that accurately recognized FER.

Materials and Methods

The PELU-LSTM is proposed to accurately recognize facial expressions. Firstly, the dataset is obtained from SAVEE which has both audio and video frames. For audio pre-processing, a wiener filter is utilized to remove the background noise and the entire video is aggregated into lesser frames using GWF. Then, the audio pre-processed data is extracted using MFCC, while the video pre-processed data is extracted using HOG and LBP. These two extracted techniques' features are concatenated into a single feature vector. Finally, PELU-LSTM is performed to recognize different facial expressions. Figure 1 shows a block diagram for the proposed technique.

Datasets

The SAVEE dataset (Dataset link, 2015) is used to evaluate the proposed PELU-LSTM technique. It has 480 video clips from four male researchers and students at the Centre for Vision, Speech and Signal Processing (CVSSP), Surrey University. Each speaker speaks 120 phonetically balanced English sentences in seven emotional classes which are, happy, fear, anger, disgust, surprise, neutral and sadness. The collected data are fed into audio and video pre-processing, so as to reduce the background noise in audio and to aggregate fewer frames in an entire video.

Pre-Processing

After obtaining data, the pre-processing stage is split into two types audio and video pre-processing. For audio pre-processing, the wiener filter is used to minimize the background noise and GWF is employed for video processing which is discussed.

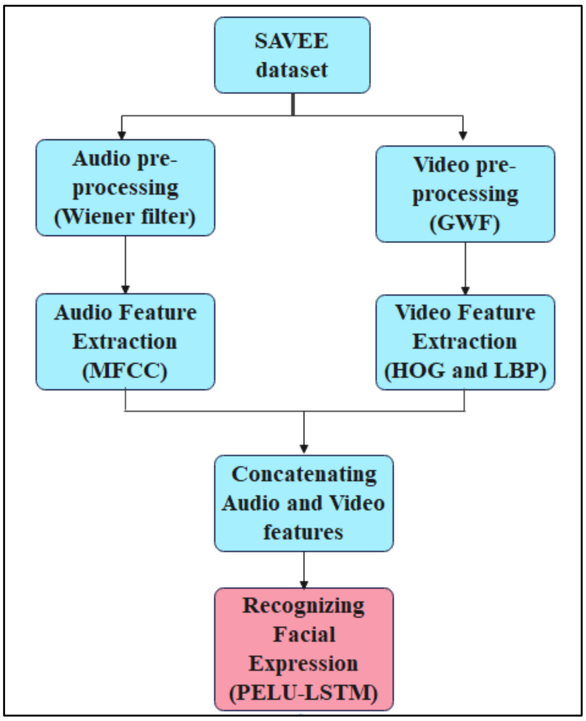
SAVEE dataset
Audio pre-processing (Wiener filter)
Video pre-processing (GWF)
Audio Feature Extraction (MFCC)
Video Feature Extraction (HOG and LBP)
Concatenating Audio and Video features
Recognizing Facial Expression (PELU-LSTM)

Fig. 1: Block diagram for the proposed technique

Audio Pre-Processing

The wiener filter is deployed to attenuate the background noise in audio signals for recognizing the multimodal facial expressions by estimating clean speech signals depending on the observed noisy signal and the characteristics of noise. The mathematical formula for the wiener filter is expressed in Eq. (1):

$$S(f_1, f_2) = \frac{H(f_1,f_2)S_{yy}(f_1,f_2)}{H(f_1,f_2)^2 S_{yy}(f_1,f_2)+S_{nn}(f_1,f_2)} \tag{1}$$

where, $S_{yy}(f_1, f_2)$ represents the power spectrum's original image, $S_{nn}(f_1, f_2)$ indicates additive noise power spectrum and $H(f_1, f_2)$ determines filter with a blurring factor. Using this technique effectively increases the audio signal quality by minimizing the background noise which produces accurate and reliable facial expressions. The output of pre-processing is passed through the audio feature extraction process.

Video Pre-Processing

The GWF (Basha *et al*., 2022) is used to aggregate the entire video into a lesser number of frames which decreases the information loss in FER. While applying video data, GWF assigns higher weights to frames that have more essential visual content and lower weights to frames with less appropriate information. It is assumed that $(I_n\}_{n\epsilon N}$ is an exhaustive sequence of non-overlapping (collection of every frame of a video), as expressed in Eq. (2):

$$(I_n\} = \{I_1, I_2, \dots, I_k, \dots\} \tag{2}$$

where, $(I_k\}$ indicates the k^{th} subsequence of $(I_n\}$ and $k < n$. The Gaussian weighting function G for I_k subsequence is formulated in Eq. (3):

$$G\ (I_k, W) = \sum_{j=1}^{M} I_{k_j} * \frac{W_j}{\sum_{j=1}^{M} W_j} \tag{3}$$

where, function G considers a sub-sequence $\{I_k\}$ and W indicates the Gaussian weight vector as input. This enables the aggregation of the data into one frame. The W_j determines the j^{th} element of the weight vector in Gaussian W and M illustrates the Gaussian weight vector size. This process assists in integrating data from multiple modalities which provide an understanding of multimodal facial expression. Overall, the resulting aggregated frames maintain appropriate multimodal facial expressions which minimizes the information loss and enables more efficient and accurate facial expressions. After video pre-processing, a feature extraction technique is employed to extract the features in the video for FER.

Feature Extraction

The two types of feature extraction techniques are established after the audio and video pre-processing. In audio feature extraction, MFCC is utilized to extract features like pitch, harmonics, etc. In video feature extraction, HOG and LBP are used to capture the feature descriptors like shape, edge and texture. Detailed information about these techniques is explained as given below.

Audio Feature Extraction

The audio-preprocessed data is fed into input to extract the features using MFCC to recognize facial expressions. MFCC (Phan *et al*., 2023) effectively captures spectral characteristics of audio signals which help in recognizing emotional cues. By extracting MFCCs from speech facial expressions, the models understand better and interpret emotional states. This method increases facial expression recognition by representing facial features in a discriminative manner. By this method, the frequency band is split into sub-bands by employing the MEL scale and then the coefficients of cepstral are extracted depending on the Discrete Cosine Transform (DCT). Additionally, it contains five phases: Pre-emphasis, frame blocking and windowing, Discrete/Fast Fourier Transform (DFT/FFT), cepstrum (inverse DCT) and Mel-frequency warping, as explained.

Pre-emphasis: The MFCC's initial phase is a pre-emphasis where the energy magnitude is amplified in high-frequency elements of the audio signal. It is the procedure of passing a signal via a first-order High Pass Filter (HPF). The main goal of this method is to minimize noise during the sound capture procedure and the following function is utilized as a filter in this technique which is numerically expressed in Eq. (4):

$$y[n] = x[n] - \propto * x[n-a] \tag{4}$$

where, x denotes the input audio signal, y indicates the output audio signal and $\propto \in (0,1)$ determines the pre-emphasis constant filter. The n represents the present time index or sample number in a discrete time signal and a denotes the delay applied to the signal.

Frame blocking and windowing: Likewise, in other analyzing audio signal techniques, MFCC is employed on short data intervals as an audio signal that divides into overlapped segments (frames) because the speech signals are quasi-stationary over short periods. Every frame has numerous audio sample data while 2 consecutive frames have certain common samples. Therefore, the Hamming or Hanning window technique is applied to enable the increase in harmonics and smooth edges and to minimize the edge influence when calculating FFT/DFT on the signal, which assists in a comprehensive understanding of facial expressions across multiple modalities.

FFT: In this phase, each windowed frame is transformed into spectrum magnitude utilizing FFT. It rapidly calculates the DFT by converting each frame from a time domain into a frequency domain. The mathematical formula for this technique is denoted in Eq. (5):

$$x[n] = \sum_{k=0}^{N-1} x(k) e^{-2\pi jkn/N} \tag{5}$$

Here, $n = 0, 1, \ldots, N-1$ and $j = \sqrt{-1}$ and N represents the amount of points utilized to calculate FFT and the output of this phase is a spectrum. This technique enables the representation of facial expressions in the frequency domain that generates temporal and spatial characteristics of expressions across video modalities.

Mel-frequency wrapping: According to mel-scale, the magnitude spectrum is fed into a triangular filter that approximates human auditory perception. It is calculated by passing Fourier signal transform via Band Pass Filter (BPF), called Mel-filter bank. Spectrogram values are squared to obtain the DFT power spectrum. Then, Mel-scale BPF is employed on every frequency range. The resultant value of every filter is the frequency energy band, covered by a filter known as the Mel-scale power spectrum. These spectra represent the energy distribution across frequency bands that provide the extraction of relevant features for the analysis of multimodal facial expressions.

Cepstrum: In this phase, the Mel spectrum is transformed into a time domain utilizing DCT that enhances the representation of audio features and provides a more effective analysis, leading to the extraction of MFCC as a final output.

By performing these 5 phases, MFCC extracts the features from audio components of data by capturing significant characteristics of speech that correlate with facial expressions. MFCCs make effective representations of audio features by encoding spectral characteristics and removing unwanted data. The variability in emotional expressions maximizes the accuracy and robustness of FER systems.

Video Feature Extraction

The video pre-processed data are passed through a feature extraction process using HOG and LTP. The detailed process of both techniques is explained below.

Histogram of Gradient (HOG)

HOG effectively captures spatial gradients of facial texture and identifies local intensity gradients and their orientations that represent the facial structure and texture patterns. It encodes information about facial contours, shapes and edges that make robust recognition of facial expressions. The main goal is to simplify image representations by extracting the appropriate information from the image and eliminating redundant parts (Kumari and Anand, 2023). This method detects objects by estimating feature descriptors utilizing vertical and horizontal gradients. The image divides into 8×8 cells, with every cell pixel having magnitude and direction that assist in establishing a compact representation. This method calculates each pixel gradient in both directions, while the image is in a discrete function. The vector having direction and magnitude is expressed in Eq. (6):

$$\theta = arctan(\frac{f_x}{f_y}) \tag{6}$$

$$mag = \sqrt{(f_x^2 + f_y^2)} \tag{7}$$

A nine-bit histogram ranging from 0-180 degrees quantizes the gradient orientations with values in the 0-255 range. Every histogram is combined to generate a feature vector H. The final HOG vector is represented in Eq. (8), generating shape regions or point information within an image:

$$F_{HOG} = \{f_{h1}, f_{h2}, f_{h3}, \ldots, f_{hn}\} \tag{8}$$

The Laplacian filter is deployed to sharpen the edges and minimize noise. The image intensity values are rescaled to enhance the contrast of an image. This method effectively captures the underlying facial features for accurate expression recognition across video frames.

Local Binary Pattern (LBP)

LBP captures dynamic texture variations in facial expressions across frames. Analyzing LBP patterns (Vu and Nguyen, 2022) within various regions of facial images, makes detection of subtle changes in texture related to different expressions. It is a statistical technique to determine the image features. The operator establishes a binary value $S(f_p - f_c)$ for every pixel and then the center pixel f and surrounding pixels $f_p = (P = 0, 1, \ldots, 7)$ over 3×3. It performs via binarization between pixel neighbors in every image, as expressed in Eq. (9):

$$LBP_{P,R}(x_c) = \sum_{p=0}^{p-1} \mu(x_p - x_c) 2^p, \mu(y) = \{\begin{matrix} 1, & y \geq 0 \\ 0, & y < 0 \end{matrix} \tag{9}$$

In $LBP_{P,R}$, the R determines the radius and distance of species from neighbor pixels, P represents the neighbor pixel count, while the uniform patterns evaluate textures are corners, edges and spots. By encoding local texture patterns within video frames, *LBP* provides the extraction of discriminative facial expression features. Its ability to capture dynamic texture information across frames increases the robustness and accuracy of facial expression recognition.

Concatenating Audio and Video Features

After extracting the MFCC coefficient from audio and HOG, the LBP features from video frames employ a concatenation process performed to combine a single feature vector. Concatenating is essential for combining different modalities to leverage data which increases the overall recognition performance. This fusion captures both spectral audio information from MFCC and spatial texture details from HOG, while the LBP provides a comprehensive set of features to recognize facial expressions. Also, the concatenating helps manage variability in expression representation across various modalities. The integrated feature vector enables a better understanding of facial expressions by including both audio and video features. Moreover, fusion increases the model's ability to generalize across various conditions which leads to improved performance. The concatenated feature vector serves as input to the DL technique which raises the need to use complementary information from both audio and visual modalities in recognition performance.

Recognizing Facial Expression

After performing the fusion process, the PELU-LSTM is used to classify the facial emotion expressions. The input data is temporal and has spatial data in recognizing expressions. LSTM effectively captures both temporal and spatial dependencies within the sequences of data in facial expressions. By generating sequential input data, LSTM efficiently learns temporal facial expression evolution while retaining spatial context which increases the recognition accuracy. Even if Recurrent Neural Networks (RNN) and Gated Recurrent Units (GRU) perform similar processes, the RNNs exhibit memory issues. Unlike GRU, LSTM employs gating mechanisms but LSTM has separate input, output and forget gates which generate more flexibility in modeling complex temporal dependencies. PELU is used as an activation function that solves dying neurons.

LSTM is an RNN architecture that learns long-term dependencies effectively. It has one cell state that is located in a horizontal line at the structure's top and has 3 gates input, output and forget gate which write and read the cell state. Gates are made up of various neurons that are trained to decide which data to remember or forget and retrieve, for obtaining the output depending on the prior and present outputs. The forget gate is used to determine what data is ignored or remembered from the prior cell state. It is performed by the sigmoid function which considers $h_t - 1$ from -1 to1 for each number in $the\ C_t - 1$ cell state. Hence, the output f_t of the forget gate is expressed in Eq. (10):

$$f_t = \sigma(W_f.[h_t - 1, X_t] + b_f \quad (10)$$

where, W_f denotes the weight matrix, $h_t - 1$ indicates a hidden state, X_t represents input and b_f determines bias. The σ represents the sigmoid activation function as numerically formulated in Eq. (11):

$$\sigma = \frac{1}{1+e^{-x}} \quad (11)$$

The sigmoid function output is restricted between 0 and 1. The next phase decides what new data is stored in the cell state. This is performed utilizing two phases: The initial phase employs the sigmoid function in the input gate for updating and the second phase utilizes the $tanh$ function for constructing new vector candidate values C'_t which are included in the state. Then, the integration of two vectors is employed for constructing the update value in the state. A result of both steps' input and candidate states is expressed in Eq. (12):

$$i_t = \sigma(W_i.[h_t - 1, X_t] + b_i \quad (12)$$

$$C'_t = tanh(\,W_c.[h_t - 1, X_t] + b_c) \quad (13)$$

where, $tanh$ is expressed in Eq. (14):

$$tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (14)$$

Following an evaluation of the old cell state and relevant decisions on what data need to be forgotten or remembered, $C_t - 1$ values are updated into a new cell state C_t. It is evaluated by multiplying $C_t - 1$ old state by factor f_t and then including to $i_t * C'_t$ value as expressed in Eq. (15):

$$C_t = f_t * C_t - 1 + i_t * C'_t \quad (15)$$

The last phase is to evaluate output data depending on the cell state. A computed value is first filtered utilizing the sigmoid function for deciding what cell state value is to be eliminated or reported as shown in Eq. (16). Then, an additional tanh function is employed to push output values among -1 and 1, as mathematically expressed in Eq. (17):

$$O_t = \sigma(W_O.[h_t - 1, X_t] + b_O \quad (16)$$

$$h_t = O_t\, tanh\, C_t \quad (17)$$

where, h_t represents a new hidden state, O_t indicates the output gate, i_t determines the input gate, C'_t denotes candidate values, C_t denotes new cell state and b_O, b_i and b_f state bias terms for output, input and forget gate. $\mathrm{W_O}$, W_i and W_f refer to the weight terms for output, input and forget gate, respectively. $PELU$ captures a smooth transition among both positive and negative slopes, potentially minimizing the vanishing gradient problems and increasing the learning dynamics, as mathematically expressed in Eq. (18):

$$PELU\,(x) = \{\begin{matrix} x, & for\ x > 0 \\ \propto.(exp\left(\frac{x}{\propto}\right) - 1, & for\ x \le 0 \end{matrix} \quad (18)$$

where, x represents the input to the $PELU$ function $\propto$ and indicates a learnable parameter that controls the negative branch behavior, initialized to a small positive value like 0.5 or 1. Therefore, this activation function provides a more stable and effective training. PELU's parametrized activation function increases the network's ability to model different expressions, while the LSTM's sequential processing retains temporal content for accurate recognition. PELU's adaptive activation function captures non-linear patterns and LSTM that have the ability to retain long-term temporal dependencies. This combination proves that both spectral and temporal data are effectively preserved. PELU-LSTM provides a robust method for FER which achieves high performance by effectively integrating spatial and temporal information.

Results

The proposed PELU-LSTM is simulated by employing MATLAB R2020b with an i5 intel processor, 16 GB RAM and Windows 10 operating system. The PELU-LSTM uses performance metrics of accuracy, F1-score, recall and precision. Accuracy is defined as the number of correct predictions by the total number of predictions. F1-score is a combination of recall and precision. Recall is the number of correctly identified positive instances (TP) from all actual positive samples. Precision defines the fraction of correctly predicted instances among the ones predicted as positives. The mathematical formula for these metrics is represented in Eqs. (19-22):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

$$F1 - Score = \frac{2\times TP}{2\times TP+FP+FN} \times 100 \tag{20}$$

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

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

where, TP determines True Positive, TN represents True Negative and FP illustrates False Positive, FN states False Negative.

Performance Analysis

The performance evaluation of the proposed PELU-LSTM is analyzed in Tables 1-2. Table 1 indicates different feature extraction techniques using the SAVEE dataset. The performance of LTP, MFCC, HOG and LBP are compared with the MFCC + HOG + LBP method. Figure 2 represents a graphical representation of different feature extraction techniques. The MFCC + HOG + LBP achieves a better accuracy of 99.95% due to its comprehensive feature fusion. Integrating spectral features from MFCC with spatial texture from HOG and LBP captures both spatial and spectral data which effectively enhances the recognition accuracy.

Table 1: Different feature extraction techniques using the SAVEE dataset

Metrics (%)	LTP	MFCC	HOG	LBP	MFCC+ HOG+ LBP
Accuracy	82.67	83.99	85.72	86.97	99.95
F1-score	80.55	81.25	82.06	82.64	98.12
Recall	81.05	82.97	84.36	85.01	97.58
Precision	81.98	82.58	84.25	85.13	97.05

Table 2: Performance analysis of the classification technique

Metrics (%)	PELU-CNN	PELU-RNN	PELU-DNN	PELU-GRU	PELU-LSTM
Accuracy	86.25	87.66	90.25	93.52	99.95
F1-score	84.08	86.37	88.09	90.55	98.12
Recall	88.96	90.24	92.28	93.05	97.58
Precision	89.22	90.25	91.87	92.33	97.05

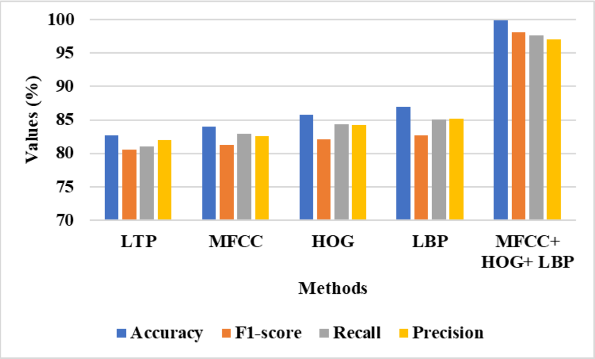

Fig. 2: Graphical representation of different feature extraction using the SAVEE dataset

Table 2 determines the performance analysis of classification using the saved dataset. The performance of PELU-CNN, PELU-RNN, PELU-DNN and PELU-GRU is compared with that of the PELU-LSTM technique. Figure 3 shows a graphical representation of classification analysis on the same dataset. The proposed technique achieves a better accuracy of 99.95% due to the LSTM capturing long-term dependencies effectively. The integration of PELU's adaptive non-linearity with LSTM sequential processing provides a more accurate understanding of facial expression patterns.

Table 3 displays a different activation function for classification techniques using the same dataset. The performance of ReLU-LSTM, Leaky ReLU-LSTM and ELU-LSTM are compared with the proposed PELU-LSTM technique. Figure 4 indicates a graphical representation of different activation functions in the classification technique. The results represent that PELU-LSTM attains a high accuracy of 99.95% due to the PELU's parametrized activation function, therefore increasing the network's ability to model different expressions, while the LSTM's sequential processing retains a temporal content for accurate recognition which leads to superior performance.

Table 4 represents a performance analysis of audio + video data for the proposed PELU-LSTM approach. The performance measures of accuracy, F1-score, recall and precision are analyzed for the proposed approach. The audio + video data obtains 99.95%, as opposed to individual audio and video data.

Comparative Analysis

Table 5 represents a comparative analysis of existing methods using SAVEE datasets. The existing techniques like Bi-LSTM-CNN (Sharafi *et al*., 2022), attention-based 2DCNN with LSTM (Singh *et al*., 2023a-b), KMKCCA (Chen *et al*., 2022), 3DCNN-Conv LSTM and MFCCT-IDCNN (Alluhaidan *et al*., 2023) are used to compare with the PELU-LSTM approach. The obtained results show that the proposed approach achieves a high accuracy of 99.95% using SAVEE datasets. Due to PELU's parametrized activation function, the network's ability to model different expressions is increased, while the LSTM's sequential processing retains a temporal content for accurate recognition, resulting in superior performance.

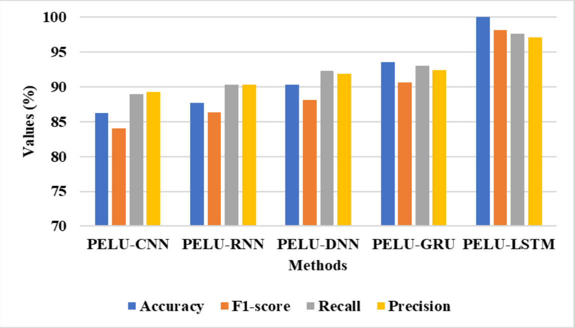

Fig. 3: Graphical representation of classification techniques

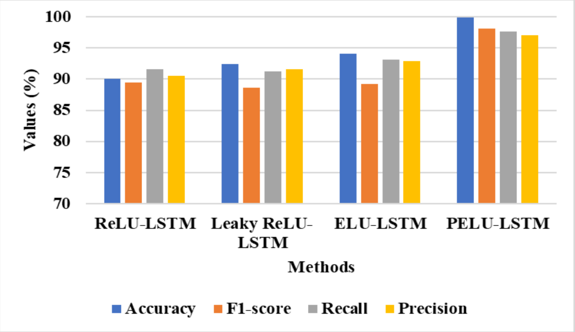

Fig. 4: Graphical representation of different activation functions for classification technique

Table 3: Different activation functions for classification technique

Metrics (%)	ReLU-LSTM	Leaky ReLU-LSTM	ELU-LSTM	PELU-LSTM
Accuracy	90.05	92.47	94.08	99.95
F1-score	89.46	88.67	89.25	98.12
Recall	91.57	91.24	93.08	97.58
Precision	90.57	91.57	92.87	97.05

Table 4: Performance analysis of audio + video data for PELU-LSTM

Metrics (%)	Audio	Video	Audio + video
Accuracy	90.25	92.58	99.95
F1-score	89.77	88.66	98.12
Recall	88.27	87.57	97.58
Precision	88.05	87.68	97.05

Table 5: Comparative analysis of existing techniques using save dataset

Methods	Accuracy (%)
Bi-LSTM-CNN Sharafi *et al*. (2022)	99.75
Attention-based 2DCNN with LSTM Singh *et al*. (2023b)	57.50
KMKCCA Chen *et al*. (2022)	94.16
3DCNN-Conv LSTM Singh *et al*. (2023a)	98.83
MFCCT-IDCNN Alluhaidan *et al*. (2023)	92.6
PELU-LSTM	99.95

Discussion

The advantages of the proposed PELU-LSTM and the disadvantages of existing techniques are discussed in this section. The existing method Bi-LSTM-CNN (Sharafi *et al*., 2022) faces limitations like the struggle with capturing fine-grained temporal dependencies across both modalities due to the inherent complexity of combining recurrent and convolutional layers. Attention-based 2DCNN with LSTM (Singh *et al*., 2023a) suffers from sensitivity to noise and varying speech patterns due to the model’s difficulty in distinguishing relevant features which leads to reduced accuracy and generalization. KMKCCA (Chen *et al*., 2022) struggles with noisy or ambiguous data points due to K-means clusters depending on the Euclidean distances, sensitivity to noise and outliers. MFCCT-IDCNN (Alluhaidan *et al*., 2023) suffers from spectral information loss due to the transformation of the frequency domain to the time domain for various signal-processing tasks.

The proposed technique overcomes these limitations. LSTM manages sequential data by maintaining a memory of past data via their cell states and gates. This enables the model to capture long-range patterns and temporal dependencies in data which overcomes the Bi-LSTM-CNN problem. PELU enables the network to better capture and distinguish appropriate features from noise which improves the model’s robustness to variations in speech patterns which overcomes the Attention-based 2DCNN with LSTM and KMKCCA. PELU’s adaptive activation function captures non-linear patterns and LSTM has a superior ability to retain (Mohana *et al*., 2023) long-term temporal dependencies. This combination renders both spectral and temporal data effectively preserved, thereby overcoming MFCCT-IDCNN. PELU solves vanishing gradient issues and enhances learning dynamics by introducing learnable parameters that adaptively scale and shift output. This improves gradient flow and enables the network to learn effectively from deep layers. By solving all these limitations, the PELU-LSTM achieves a better accuracy of 99.95%.

Conclusion

In this research, the PELU-LSTM is proposed to accurately recognize facial emotional expressions. PELU effectively solves the dying neuron problems and produces fast convergence speed during training. PELU's parametrized activation function increases the network's capability to represent various expressions, while the LSTM captures long-term dependencies in sequential data, ultimately resulting in accurate recognition. Hence, PELU-LSTM provides a robust method for FER which achieves high performance by effectively integrating spatial and temporal information. The wiener filter minimizes the background noise in audio pre-processing, while GWF aggregates the entire video into fewer frames. This process facilitates reducing background noise and information loss, also enabling accurate and reliable facial expressions. The pitch and harmonics are extracted using MFCC, while HoG and LBP are employed to capture feature descriptors like shape, edge and texture in images, which further enhance the robustness and accuracy of FER across multiple modalities. By performing this process, the PELU-LSTM accomplishes a commendable accuracy of 99.75% on the SAVEE dataset, as opposed to the existing techniques, Bi-LSTM-CNN, attention-based 2DCNN with LSTM and KMKCCA. A limitation of PELU-LSTM is that it suffers from overfitting issues due to the model's increased capacity to learn complex patterns. In the future, effective classification techniques will be used to further increase the model performances.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the effort of the editorial team in reviewing and editing our work and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Kampa Ratna Babu: Formal Analysis resources, methodology, software, written, draft preparation.

Akula Suneetha: Written-reviewed and visualization, edited.

Kampa Kanthi Kumar: Investigation, validation, supervision.

Ethics

We declare that the work submitted for publication is original, previously unpublished in English or any other languages and not under consideration for publication elsewhere.

References

Abdelhamid, A. A., El-Kenawy, E. S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., & Eid, M. M. (2022). Robust speech emotion recognition using CNN + LSTM based on stochastic fractal search optimization algorithm. *IEEE Access*, *10*, 49265-49284. http://dx.doi.org/10.1109/ACCESS.2022.3172954

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, *13*(8), 4750. https://doi.org/10.3390/app13084750

Aslan, M. (2022). CNN based efficient approach for emotion recognition. *Journal of King Saud University-Computer and Information Sciences*, *34*(9), 7335-7346. https://doi.org/10.1016/j.jksuci.2021.08.021

Basha, S. S., Pulabaigari, V., & Mukherjee, S. (2022). An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos. *Multimedia Tools and Applications*, *81*(28), 40431-40449. https://doi.org/10.1007/s11042-022-12856-6

Bodapati, J. D., Srilakshmi, U., & Veeranjaneyulu, N. (2022). FERNet: A deep CNN architecture for facial expression recognition in the wild. *Journal of The Institution of Engineers (India): series B*, *103*(2), 439-448. https://doi.org/10.1007/s40031-021-00681-8

Boughida, A., Kouahla, M. N., & Lafifi, Y. (2022). A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evolving Systems*, *13*(2), 331-345. https://doi.org/10.1007/s12530-021-09393-2

Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, *70*(1), 1016-1024. https://dx.doi.org/10.1109/TIE.2022.3150097

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, *35*(32), 23311-23328. https://doi.org/10.1007/s00521-021-06012-8

Dataset link. (2015). Surrey Audio-Visual Expressed Emotion (SAVEE) Database. http://kahlan.eps.surrey.ac.uk/savee/Database.html

Fard, A. P., & Mahoor, M. H. (2022). Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, *10*, 26756-26768. http://dx.doi.org/10.1109/ACCESS.2022.3156598

Jothimani, S., & Premalatha, K. (2022). MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. *Chaos, Solitons & Fractals*, *162*, 112512. https://doi.org/10.1016/j.chaos.2022.112512

Kim, J. C., Kim, M. H., Suh, H. E., Naseem, M. T., & Lee, C. S. (2022). Hybrid approach for facial expression recognition using convolutional neural networks and SVM. *Applied Sciences*, *12*(11), 5493. https://doi.org/10.3390/app12115493

Kumari, D., & Anand, R. S. (2023). Fusion of Attention-Based Convolution Neural Network and HOG Features for Static Sign Language Recognition. *Applied Sciences*, *13*(21), 11993. https://doi.org/10.3390/app132111993

Liu, S., Huang, S., Fu, W., & Lin, J. C. W. (2024). A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. *International Journal of Machine Learning and Cybernetics*, *15*(1), 19-35. https://doi.org/10.1007/s13042-022-01681-w

Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, *244*, 108580. https://doi.org/10.1016/j.knosys.2022.108580

Mohana, M., Subashini, P., & Krishnaveni, M. (2023). Emotion recognition from facial expression using hybrid CNN–LSTM network. *International Journal of Pattern Recognition and Artificial Intelligence*, *37*(08), 2356008. https://doi.org/10.1142/S0218001423560086

Nan, Y., Ju, J., Hua, Q., Zhang, H., & Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, *61*(6), 4435-4444. https://doi.org/10.1016/j.aej.2021.09.066

Pandey, P., & Seeja, K. R. (2022). Subject independent emotion recognition from EEG using VMD and deep learning. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1730-1738. https://doi.org/10.1016/j.jksuci.2019.11.003

Phan, T. T. H., Nguyen-Doan, D., Nguyen-Huu, D., Nguyen-Van, H., & Pham-Hong, T. (2023). Investigation on new Mel frequency cepstral coefficients features and hyper-parameters tuning technique for bee sound recognition. *Soft Computing*, *27*(9), 5873-5892. https://doi.org/10.1007/s00500-022-07596-6

Sharafi, M., Yazdchi, M., Rasti, R., & Nasimi, F. (2022). A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. *Biomedical Signal Processing and Control*, *78*, 103970. https://doi.org/10.1016/j.bspc.2022.103970

Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S. (2023a). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology*, *15*(4), 1819-1830. https://doi.org/10.1007/s41870-023-01183-0

Singh, J., Saheer, L. B., & Faust, O. (2023b). Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, *20*(6), 5140. https://doi.org/10.3390/ijerph20065140

Vu, H. N., Nguyen, M. H., & Pham, C. (2022). Masked face recognition with convolutional neural networks and local binary patterns. *Applied Intelligence*, *52*(5), 5497-5512. https://doi.org/10.1007/s10489-021-02728-1

Wang, S., Qu, J., Zhang, Y., & Zhang, Y. (2023). Multimodal emotion recognition from EEG signals and facial expressions. *IEEE Access*, *11*, 33061-33068. https://doi.org/10.1109/ACCESS.2023.3263670

Xiao, H., Li, W., Zeng, G., Wu, Y., Xue, J., Zhang, J., ... & Guo, G. (2022). On-road driver emotion recognition using facial expression. *Applied Sciences*, *12*(2), 807. https://doi.org/10.3390/app12020807

Yu, W., & Xu, H. (2022). Co-attentive multi-task convolutional neural network for facial expression recognition. *Pattern Recognition*, *123*, 108401. https://doi.org/10.1016/j.patcog.2021.108401