Computers, Materials & Continua
DOI: 10.32604/cmc.2023.028631
Article

Tech Science Press

Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance

Sandeep Kumar[1], MohdAnul Haq[2], Arpit Jain[3], C. Andy Jason[4], Nageswara Rao Moparthi[1], Nitin Mittal[5] and Zamil S. Alzamil[2,*]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, 522502, India
[2]Department of Computer Science, College of Computer and Information Sciences, Majmaah University, 11952, Al-Majmaah, Saudi Arabia
[3]Department of Computer Science and Engineering, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, 244001, India
[4]Department of Electronics and Communication Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, 500068, India
[5]University Centre for Research and Development, Chandigarh University, Mohali, 140413, Punjab, India
*Corresponding Author: Zamil S. Alzamil. Email: z.alzamil@mu.edu.sa
Received: 14 February 2022; Accepted: 24 May 2022

Abstract: Day by day, biometric-based systems play a vital role in our daily lives. This paper proposed an intelligent assistant intended to identify emotions via voice message. A biometric system has been developed to detect human emotions based on voice recognition and control a few electronic peripherals for alert actions. This proposed smart assistant aims to provide a support to the people through buzzer and light emitting diodes (LED) alert signals and it also keep track of the places like households, hospitals and remote areas, etc. The proposed approach is able to detect seven emotions: worry, surprise, neutral, sadness, happiness, hate and love. The key elements for the implementation of speech emotion recognition are voice processing, and once the emotion is recognized, the machine interface automatically detects the actions by buzzer and LED. The proposed system is trained and tested on various benchmark datasets, i.e., Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database, Acoustic-Phonetic Continuous Speech Corpus (TIMIT) database, Emotional Speech database (Emo-DB) database and evaluated based on various parameters, i.e., accuracy, error rate, and time. While comparing with existing technologies, the proposed algorithm gave a better error rate and less time. Error rate and time is decreased by 19.79%, 5.13 s. for the RAVDEES dataset, 15.77%, 0.01 s for the Emo-DB dataset and 14.88%, 3.62 for the TIMIT database. The proposed model shows better accuracy of 81.02% for the RAVDEES dataset, 84.23% for the TIMIT dataset and 85.12% for the Emo-DB dataset compared to Gaussian Mixture Modeling(GMM) and Support Vector Machine (SVM) Model.

Keywords: Speech emotion recognition; classifier implementation; feature extraction and selection; smart assistance

1 Introduction

In everyday life, Speech Emotion Recognition (SER) based devices play an essential role [1,2]. The SER technology has been expanded in sports, e-learning, voice search, and even aircraft cockpit call centers. The main aim of the SER system is to understand the individual emotions of humans [3]. A highly significant feature of the voice recognition system is the reliance on the human voice, i.e., age, language, culture, temperament, environment, etc. The primary issue during speech recognition is more than one emotion expressed in the same vocabulary, so identifying the emotions is very critical [4,5]. There are several traditional methods for feature extraction and classifier implementation models, i.e., KNN (K-Nearest Neighbour), GMM, ANN (Artificial neural network), DNN(Deep neural network), etc. which is used to recognize the emotions from speech but still not optimized [6]. To recognize the emotions from speech, we have proposed a novel SER model. The proposed SER model can detect seven emotions: worry, surprise, neutral, sadness, happiness, hate and love [7,8]. While evaluating the results, the benchmark datasets can be divided into testing and training. Each speech dataset is transferred through the pre-processing function to extract the necessary function vector for features from the data. The vector training set is passed on to the correct classifier and the classifier then forecasts the emotion to validate a model [9]. The identification of speech emotions is carried out in four significant steps to generate speech-based output, i.e., acquisition, processing, output generation and the application of the extracted voice feature. The proposed module is mainly essential for the people treated with social distancing so that the people who are treating them can recognize their emotional condition and treat them well [10]. The proposed methodology also tends to boost the accuracy for better performance of speech emotion detection [11–14].

The rest of the paper is organized as follows: Section 2 discusses the work related to emotion recognition through voice. Further, Section 3 concisely discusses the proposed method and the dataset details. Further, in Section 4, results obtained by the proposed method have been discussed and compared with the existing state-of-the-art methods. The last section concludes along with the future course of action. Let us know more about SER existing systems with their various feature extraction selection methods and classification algorithms in the following literature survey.

2 Literature Work

Much research has been done in the area of emotion recognition through speech. There are many traditional methodologies and evaluation parameters, i.e., accuracy, error, time taken, etc., used by the SER system to recognize the emotion from speech but still exiting methods are the more complex and computational time taken were high. In this section, the literature survey is done based on the various parameters, i.e., approach based, evaluation based, results, databases, and conclusions of various models of SER systems as shown in Tab. 1.

Table 1: Approach, evaluation, and classifiers of existing SER system

S. No.	Author	Methodology		Results	Conclusions	Data-bases
		Approach	Evaluation			
1	Yogesh Kumar et al. [1], 2019	MFCC, LPCC, DELTA, FFT, PLP and RASTA	KNN, SVM, Convolution neural network, Naive Bayes and RNN	The qualified model proposed is tested with a test accuracy of 76.97% in the entire file classification.	Compared to other models, DNN models offer the best results.	Emo-DB and LDC emotional prosody speech database
2	K. Prasada Rao et al. [2], 2019	SSR, PR	MFCC and MSER	The average efficiency of the group is 77%. Happiness and disgust (78%) are feelings with the highest awareness rate.	Provides better efficiency than the uni-modal system.	Emo-DB and Indian Face Database
3	Maryam Imani et al. [3], 2019	Signal Processing and gesture recognition	E-Learning Algorithms	It can be used as a reference for the emotion detection of successful tutoring programs.	It can be used as a reference for the emotion detection of successful tutoring programs.	Science Direct database
4	Wei Jiang et al. [4], 2019	IS10, MFCCs, eGemaps	Heterogeneous Unification Module	Compared to current cutting-edge solutions, 64% of the proposed architecture improves the recognition efficiency.	To improve classification efficiency, use the best multiple and heterogeneous features.	IEMOCAP dataset
5	Nithya Roopa S. et al. [5], 2018	Deep Learning	Inception Net v3	The precision rate is reached by approximately 38 percent.	The highest accuracy rate during data validation is 0.8	IEMOCAP datasets
6	Youddha Beer Singh et al. [6], 2018	MFCC	HMM, KNN, SVM, ANN and GMM	The highest precision of 79.6% for classifier SVM and the lowest accuracy of 54.3% for classification ELM.	Even with different datasets, SVM has recorded the highest precision.	Emo-DB, IITKGP-SESC and Wizard of Oz databases

(Continued)

Table 1: Continued

S. No.	Author	Methodology		Results	Conclusions	Data-bases
		Approach	Evaluation			
7	Praseetha etal. [7], 2018	FFT, MFCC	DNN, RNN	The accuracy of the DNN model is 89.96%, and the accuracy of the GRU model is 95.82%.	The GRU model works very well for dynamic grouping compared to the DNN model.	IA database
8	Rahhal Errattahi et al. [8], 2018	Evaluation methods of ASR	ASR errors detection and correction techniques	A data set consisting of five separate English articles of about 100 words read by five distinct speakers represented about 2.4% of the proposed technique's error rate.	Further work on the automatic correction of ASR failures is needed, and performance and reliability issues should be addressed.	Nil
9	Sneha Lukose et al. [9], 2017	MFCC and End Point Detection	SVM, GMM	Provides 76.31% positive performance with the GMM model and 81.57% accuracy with the SVM model.	SVM offers more incredible accuracy to extract the emotion from the speech signal.	Emo-DB database
10	David Griol et al. [10], 2017	Feature Extraction and Selection Methods	SVM, PNN and Naive Bayes	The results show that all hypotheses feature classifiers, including recognition and fusion, can be used at every stage.	Concerning precision and training time, ELM delivered the best results.	Images descriptions, UAH and Let's Go corpus
11	Seyed H. Mohammadi et al. [11], 2017	STPK	MLSA	The average score for similarities for top submissions was defined correctly by about 70%.	Some health tests are best incorporated to eliminate listeners who perform below a minimum or inconsistently output threshold.	CMU arctic speech database

(Continued)

Table 1: Continued

S. No.	Author	Methodology		Results	Conclusions	Data-bases
		Approach	Evaluation			
12	S. Lugović et al. [12], 2016	HCI	SER Models	Possibly monitor emotions and actions in different social groups by using emotional recognition in speech.	It improves the efficiency of social technology structures and the benefit to the cost ratio is high as per computers.	DES, BES and SUSAS databases
13	Haihua Jiang et al. [13], 2017	(KNN, GMM, SVM) +INT, PIC, REA	UDD, STEDD	A higher level of precision of 80.30 percent for men and 75.96% for women and an acceptable 75.00% for men and 77.36% for women. for women, a good sensitivity/specific ratio of 75.00%	The highest rating result was shown and both men and women had the best stability.	INT, PIC, IEA databases
14	Isidoros Perikos et al. [14], 2017	FP, ECV, ESV, MEC	Ensemble classifier	The obtained findings suggest satisfactory results concerning the ability to perceive the role of emotions in the text and the emotional polarity of the text.	Ensemble methodology is an effective way to combine different classification algorithms with helping classify textual emotions.	WordNet database
15	Pavol Partila et al. [15], 2014	MFCC	GMM, KNN and ANN	Increased accuracy after training	These three classifiers have shown the highest understanding of emotional indignation.	Emo-DB database

Kumar (2019) et al. [1] obtained an accuracy of 76.98% over the entire classification and concluded that the DNN model provides the best performance compared to other models. In the same year, Prasadarao et al. [2], Imani et al. [3] and Jiang et al. [4] also worked on SSR signal processing, gesture recognition and MFCC models with various evaluation parameters based on MSER (Maximally stable extremal regions) e-learning algorithms resulting in 77% overall classification success on joy as well as disgust emotions. In 2018, Errattahi et al. [6], Singh et al. [7] and Praseetha et al. [8] and worked on MFCC (Mel-frequency cepstral coefficients), FFT(fast fourier transform) and while evaluating the methodology, the highest accuracy achieved 79.6% and the lowest accuracy got

54.3%.Another evaluation parameter, i.e., the error rate, was evaluated using the proposed method on a benchmark database. In contrast with existing state-of-the-art solutions, architecture proposed in other strategies improves recognition effectiveness by 64%. In 2017 Lukose et al. [9], Griol et al. [10], Mohammadi et al. [12] used MFCC, endpoint detection feature selection methods i.e., SVM, ANN, Naive Bayes for SER modules. Finally, 76.31% of devices used the GM model and overall accuracy improved by 1.57% using SVM models. In 2016, Lugovic et al. [13] tested SCR models using HCI (Human-Computer Interaction).

Many authors have worked to improve the performance of the SER based models and but still there is a room of improvement [16–23]. The author concluded the possibility to monitor emotions and actions in different groups by using the emotion recognition module [24–28]. This literature survey formulated the proposed approach with improved overall accuracy.

3 Proposed Methodology

The proposed methodology is wholly based on a multilayer neural network SER system [14–17]. The central aspect of the SER system is to recognize the speech emotions where the speech is given as an input to the system. After that, the multilayer neural network automatically makes feature extraction and selection. The entire proposed work of the speech recognition system is discussed step by step in the proposed algorithm, as shown in Fig. 1.

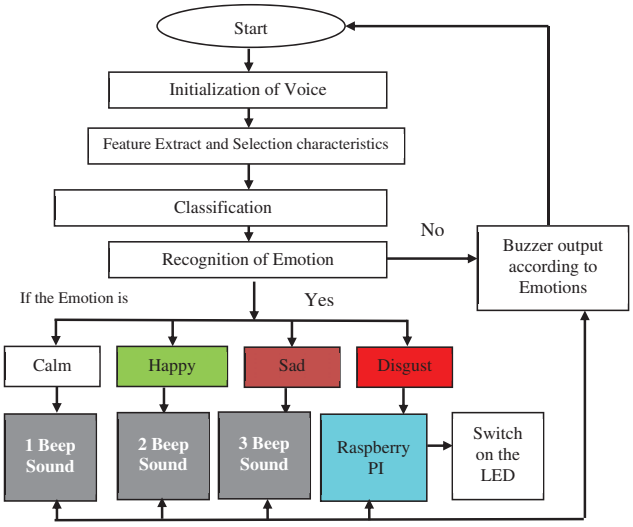


Figure 1: Flow chart of smart assistant based on SER

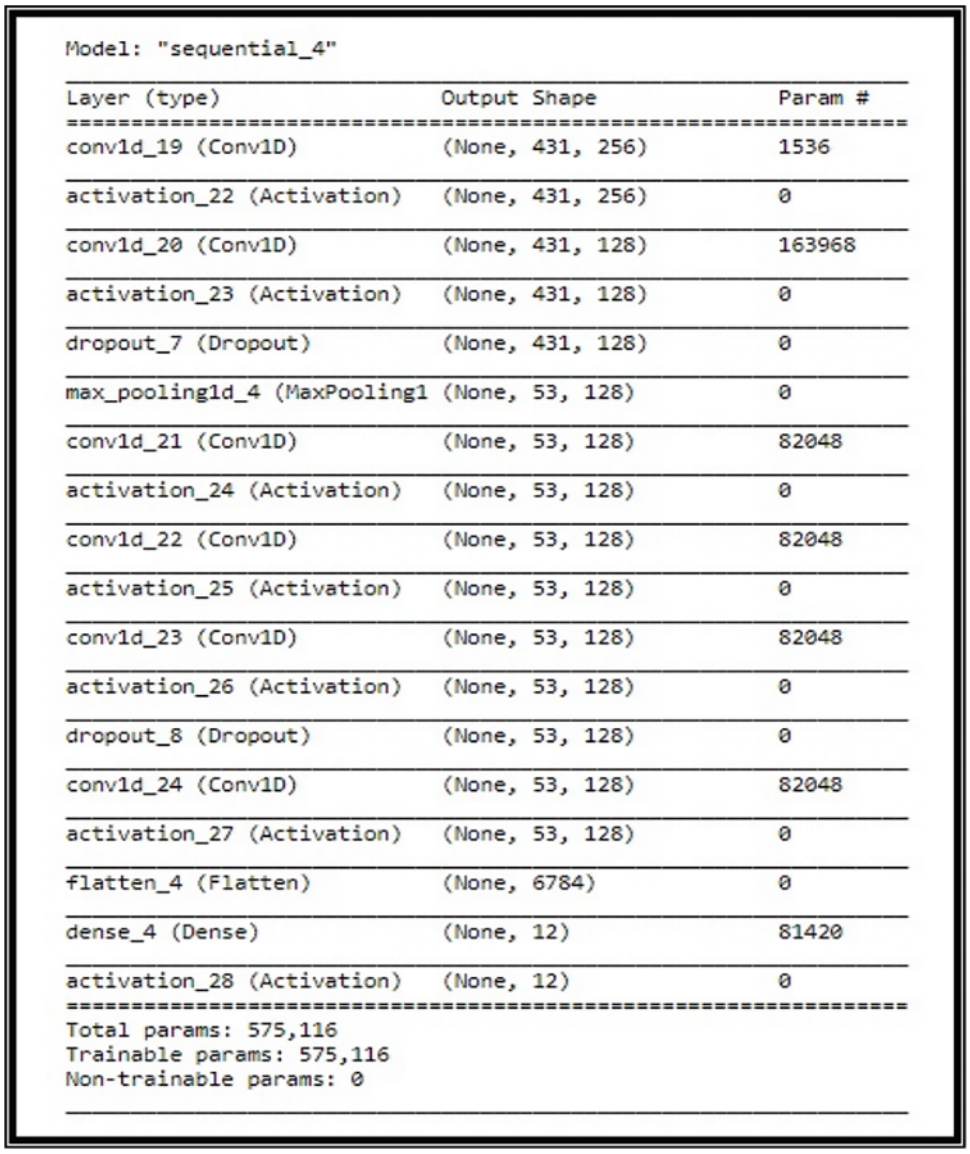
3.1 Voice Acquisition

In the first step of speech recognition, is voice sample has been taken from benchmark datasets for further process.

3.2 Feature Extraction and Selection

Feature extraction is a process of extracting the characteristics of the input sample to perform the classification task. In this model, the novel algorithm has been proposed and Mel Frequency Cepstral Coefficients is used in the speech recognition system for feature extraction. MFCC generates a discreet

cosine transformation (DCT) of a natural short-term energy logarithm on the Mel frequency scale as shown in Fig. 2 and also specifies no of output samples that are considered for trainable parameters from a dataset. There are some advantages of choosing the functionalities and benefits of performing role selection before designing the data of MFCC.



```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_19 (Conv1D)           (None, 431, 256)          1536
_________________________________________________________________
activation_22 (Activation)   (None, 431, 256)          0
_________________________________________________________________
conv1d_20 (Conv1D)           (None, 431, 128)          163968
_________________________________________________________________
activation_23 (Activation)   (None, 431, 128)          0
_________________________________________________________________
dropout_7 (Dropout)          (None, 431, 128)          0
_________________________________________________________________
max_pooling1d_4 (MaxPooling1 (None, 53, 128)           0
_________________________________________________________________
conv1d_21 (Conv1D)           (None, 53, 128)           82048
_________________________________________________________________
activation_24 (Activation)   (None, 53, 128)           0
_________________________________________________________________
conv1d_22 (Conv1D)           (None, 53, 128)           82048
_________________________________________________________________
activation_25 (Activation)   (None, 53, 128)           0
_________________________________________________________________
conv1d_23 (Conv1D)           (None, 53, 128)           82048
_________________________________________________________________
activation_26 (Activation)   (None, 53, 128)           0
_________________________________________________________________
dropout_8 (Dropout)          (None, 53, 128)           0
_________________________________________________________________
conv1d_24 (Conv1D)           (None, 53, 128)           82048
_________________________________________________________________
activation_27 (Activation)   (None, 53, 128)           0
_________________________________________________________________
flatten_4 (Flatten)          (None, 6784)              0
_________________________________________________________________
dense_4 (Dense)              (None, 12)                81420
_________________________________________________________________
activation_28 (Activation)   (None, 12)                0
=================================================================
Total params: 575,116
Trainable params: 575,116
Non-trainable params: 0
_________________________________________________________________
```

Figure 2: Features extracted based on NN

- Eliminates over fittings.
- Enhances Accuracy: Modeling accuracy increases with less misleading results.
- Reduces training time: Fewer data points increase the algorithm complexity and learn quicker algorithms.

$$\mathrm{C}\,[\mathrm{n}] = \frac{x\hat{}[n] + x\hat{}[n]}{2} \tag{1}$$

where C[n] is real ceptron and x[n] is the real input signal

The extracted features are considered validation points to calculate the SER model's accuracy, time, and error rate. The Classifier is implemented to classify the emotions in speech after feature extraction and collection of voice samples so that the functions of the classifier understand the feelings. The emotions are categorized and remembered by training the dataset of various audio files for SER.

3.3 Classification

The key and most important classification aspect is the Multilayer Perceptron (MLP) classification, which we used for our SER module. Multilayer Perceptron is an artificial neural feed-forward (ARN) type that consists of 3 node layers: one input, one hidden and one output layer, as shown in Fig. 3. According to Fig. 3, the input layer values are denoted as $x_1, x_2 \ldots \ldots x_n$ output layer values as $y_1, y_2 \ldots \ldots y_n$, and hidden layer values as $h_1, h_2 \ldots \ldots h_n$. Each layer is fully connected to the next with the activation function forward feeding. We evaluate it based on the following equations. For each training sample, d, do: Propagate the input forward through the network and calculate network output for d's input values. Propagate the errors backward through a neural network and for each network output unit j

$$\delta_j = O_j.\left(1 - O_j\right) * \left(t_j - O_j\right) \tag{2}$$

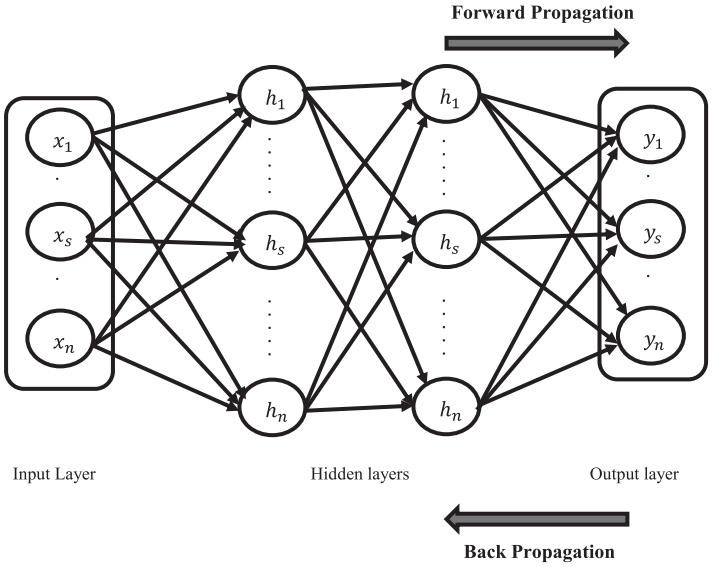


Figure 3: Neural network of MLP

For each hidden unit j

$$\delta_j = O_j.\left(1 - O_j\right) * \sum \delta_k.W_{jk} \tag{3}$$

Update weights (w_{ji}) by back-propagating error and using learning rule

$$W_{ij}\,(new) = \Delta W_{ij} + W_{ij}\,(old)\,;\ where\ \ \Delta W_{ij} = \eta_{.j}.\ O_i \tag{4}$$

where Δw=Predicted desired output, d=The learning rate is usually less than 1, η=Input data. After summation of the input values substitute the value in the sigmoid function

$$\text{Sigmoid } f(x) = \frac{1}{1 + e^{-s}} \tag{5}$$

MLP uses a supervised training method for backpropagation characterized by linear perceptron multi-layered and non-linear activation functions. MLP cannot linearly separate distinguishable information, but this process is evaluated in real-time. The training data is stored in the target folder as an audio sample format compared with the input data. MLP recognizes the emotion based on the previous outcomes to improve the accuracy by learning rate (0.9) from the previous data, as shown in Fig. 4. The algorithm recognizes the emotion as previous when similar binary data is found. If not, it stores the input data and learns from it. In our proposed methodology, we trained our dataset based on voice modulations of the speaker, voice input and prediction sentences, as shown in Fig. 5.

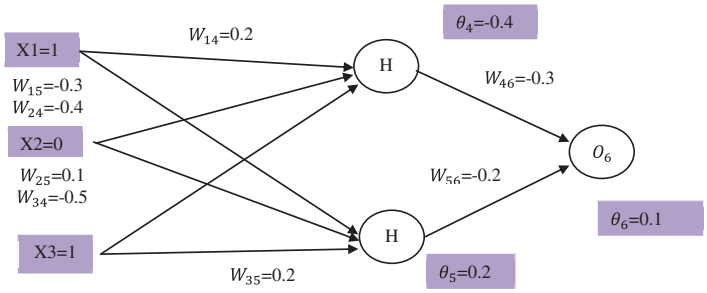


Figure 4: Training of the input values based on MLP

In this way, the MLP-based neural network predicts emotion based on the input audio signal and a few output samples on various emotions are shown in Fig. 6. The neural network's performance depends on the number of layers concealed in the network. As the number of hidden layers' increases, predicting emotions will increase and help the neural network recognize emotions more accurately. Our neural network consists of 17 hidden layers that analyze speech characteristics based on the extraction and selection process. After the classification is done, we can recognize the following seven emotions based on the ANN by giving samples of voice messages, as shown in Fig. 6.

3.4 Hardware Module

The proposed module receives the input data through a microphone, and output is observed in the form of alerts created by a buzzer and LED's to the user, as shown in Figs. 7a and 7b. Firstly, hardware components are used as raspberry pi3 model B+, which looks like a small card-sized electronic board, monitor/system as a screen to show efficiency [18,19]. A compilation of code & display results, statements & to start the process, connecting cables to connect the system to raspberry pi3 & buzzer and LED and interface with raspberry, mic for the real-time speech recognition with high quality, wires, raspberry and for a system power supply should be connected or if laptop no need of extra power supplier.

Pseudo Code

For H_4

$A_4 = -0.7$

$H_4 = f(A_4) = f(-0.7) = 0.332$

For H_5

$A_5 = -0.7$

$H_5 = f(A_5) = f(0.1) = 0.525$

When we pass these two values to the output layer, we get

For O_6

$A_6 = -0.105$

$O_6 = f(A_6) = f(0.105) = 0.474$

Since the actual value is 1 and we got 0.474 then the error will be

Error=$y_{target} - y_6 = 1 - (0.474) = 0.526$

For output unit:

$$\delta_6 = y_6(1 - y_6)(y_{target} - y_6)$$

$= 0.476*(1-0.474)*(1-0.474)=0.1311$

For hidden unit:

$$\delta_5 = y_5(1 - y_5)w_{56} * \delta_6$$

$= 0.525*(1-0.525)*(-0.2*0.1311)=-0.0065$

$$\delta_4 = y_4(1 - y_4)w_{46} * \delta_6$$

$= 0.332*(1-0.332)*(-0.3*0.1311)=-0.0087$

$\Delta W_{46} = \eta.\delta_6.y_4$=0.9*0.1311*0.332 = 0.03917

$W_{46}(new) = \Delta W_{46} + W_{46}(old) = 0.0391 + (-0.3) = -0.261$

$\Delta W_{14} = \eta.\delta_4.x_1$=0.9*-0.0087*1 = −0.0078

$W_{14}(new) = \Delta W_{14} + W_{14}(old) = -0.0078 + 0.2 = -0.192$

ΰ $\theta_6 = \eta.\delta_4$=0.9*0.1311 = 0.1179

$\theta_6(new) =$ ΰ $\theta_6 + \theta_6$=0.1179+0.1=0.218

For O_6

$A_6 = 0.061$

$O_6 = f(A_6) = f(0.061) = 0.515$

now

Error=$y_{target} - y_6 = 1 - (0.515) = 0.484$

Figure 5: PSEUDO code

Here we used python version 3.8.3 software as it was advanced and better. This software was inbuilt in the raspberry PI and is worked by using a VNC viewer (Virtual Network Computing). The program of speech and recognition will run in this VNC viewer software. This VNC viewer gives a cryptographic representation of your pi and we use SSH (Secure Shell), which is standard to support encrypted data transfer between two computers and gives access to the pi via terminal [20,21].

4 Experimental Setup and Database

All tests were performed on Python 3.8 using Intel Core I5 system specification, 8 GB RAM, RADEON Graphics Card, etc. This software was inbuilt in the raspberry PI and is worked by using a VNC viewer (Virtual Network Computing). The program of speech and recognition will run in this VNC viewer software. The entire proposed module works on speech emotion recognition where the input is given through a microphone and output is observed in the form of alerts created by peripherals as assistants for the user. The main component in the module is raspberry Pi which carries the complete control of the module and will assign work to each peripheral and give a response according to the speech input is given to the system.

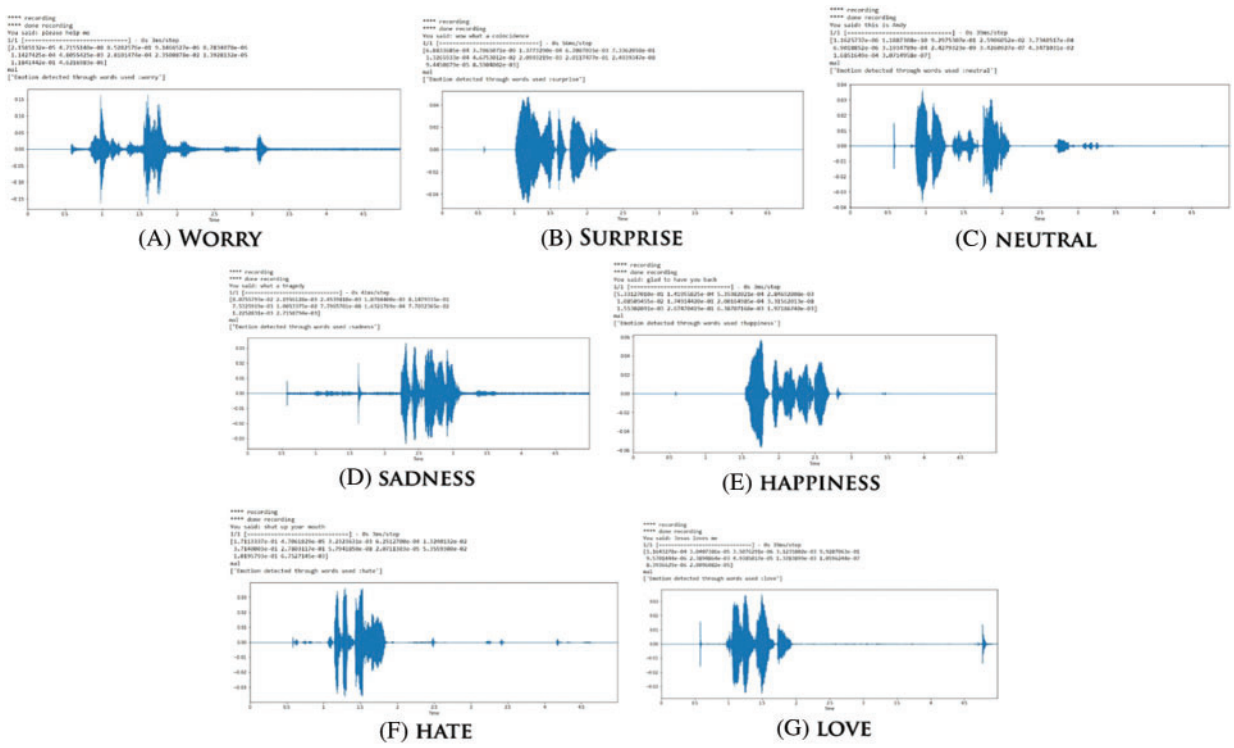


Figure 6: Sample images of the outputs of recognizing the emotions (Sadness, Worry, Happiness, Surprise, Love, Neutral, Hate) based on the audio signal where the x-axis represents the period taken to record the audio signal and the y-axis represents the frequency of speech signal in decibels (DB)

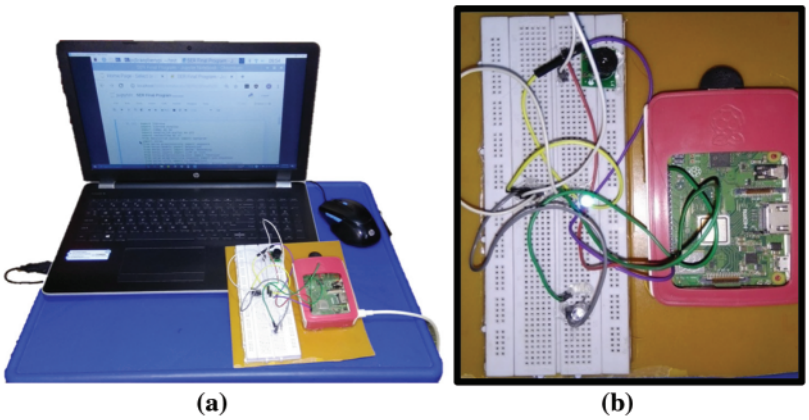


Figure 7: (a) Complete hardware setup (b) Smart assistant kit

4.1 Standard Datasets

For evaluation of the proposed work, three benchmark datasets were used, i.e., RAVDNESS, TIMIT Corpus and Emo-DB and description of all three datasets as shown in Tab. 2.

Table 2: Details of benchmark datasets

Sr. No.	Dataset Name	Remarks
1	RAVDNESS	RAVDESS contains 7356 files (total size: 24.8 GB). Datasets consist of 24 professional actors (12 females and 12 males). Speech includes expressions of calm, happiness, sad, anger, fear, surprise, and disgust expressions.
2	TIMIT Corpus	TIMIT contains broadband recordings of 630 speakers of eight significant dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions, and a 16-bit, 16 kHz speech waveform file for each utterance.
3	Emo-DB	EMO-DB is another database used in our project for comparing accuracy with our database. It contains about 500 utterances spoken by actors in a happy, angry, anxious, fearful, bored and disgusted way, as well as a neutral version. You can choose utterances from 10 different actors and ten different texts.

4.2 Evaluation Parameter

As we already mentioned, the proposed module evaluated the various parameters, i.e., Accuracy, Error Rate and Time Taken, on three benchmark datasets.

4.2.1 Error Rate

$$\bullet \text{ Word Error Rate (WER)} \cdot = \frac{C}{N} \text{ for discrete speech,} \tag{6}$$

$$\frac{N-(1+D+S)}{N} \text{ for continuous speech.}$$

$$\cdot \text{ Authentication Accuracy (AA)} = \frac{SA}{TA} \tag{7}$$

where SA is the number of successful authentication attempts and TA is the total several authentication attempts.

$$\cdot \text{ Command Error rate (CER)} = \frac{CC}{TC} \text{ for commands with no attribute} \tag{8}$$

$$\frac{CC+CA}{TC+TA} \text{ for commands with attribute} \tag{9}$$

where, TC is the total number of commands issued.

CC is the number of correctly carried out command's beta software.

CA is the total number of attributes correctly interpreted by the software.

TA is the total number of attributes issued.

$$\cdot \text{ Rejection rate (RR)} = \frac{NRR}{TRR} \tag{10}$$

where NRR number of rejected unwanted sounds.

TRR total number of unwanted sounds that should be rejected.

- Accuracy testing for varying sound pressure level values (SPL)

$$SPL_{variance} = \frac{WER_2 + AA_2 + CER_2 + RR_2}{WER_1 + AA_1 + CER_1 + RR_1} = \frac{T_2A}{T_1A} \tag{11}$$

where T_2A is the accuracy-test at a greater distance.

T_1A is the accuracy-test at a short distance.

- Accuracy testing for wearing signal to noise ratios (SNR)

$$SNR_{variance} = \frac{WER_3 + AA_3 + CER_3 + RR_3}{WER_1 + AA_1 + CER_1 + RR_1} = \frac{T_3A}{T_1A} \tag{12}$$

T_3A is the accuracy test with background noise.

T_1A is the accuracy test without background noise.

4.2.2 Accuracy

- SPAB (Speech Processing Accuracy Benchmark)

$$T_1A + SPL_{variance} + SNR_{variance} \tag{13}$$

4.2.3 Time Taken

Time Taken is calculated based on the output generated after the compilation of the proposed methodology is shown in Fig. 8.

5 Results and Discussion

5.1 Results

The total no of speech samples acquired by the standard databases consists of various male and female voice samples. The MLP based proposed work classified them based on emotions: worry, surprise, neutral, sadness, happiness, hate, and love, as shown in Tab. 3. Here we have used all three datasets for training and testing. The K-fold cross-validation technique was used in our module. During testing of the module, the trained data is saved in .csv format, which is 70% and testing data which is 30% of total samples of all datasets based on test train and split validation technique.

```
**** recording
**** done recording
You said: please help me
1/1 [==============================] - 0s 3ms/step
[2.1585132e-05 4.7155140e-08 8.5282576e-01 9.1466527e-06 8.7834078e-06
 1.1427425e-04 4.8055425e-03 2.8101474e-04 2.3500878e-02 1.3928132e-05
 1.1841442e-01 4.6216983e-06]
mal
['Emotion detected through words used :worry']
```

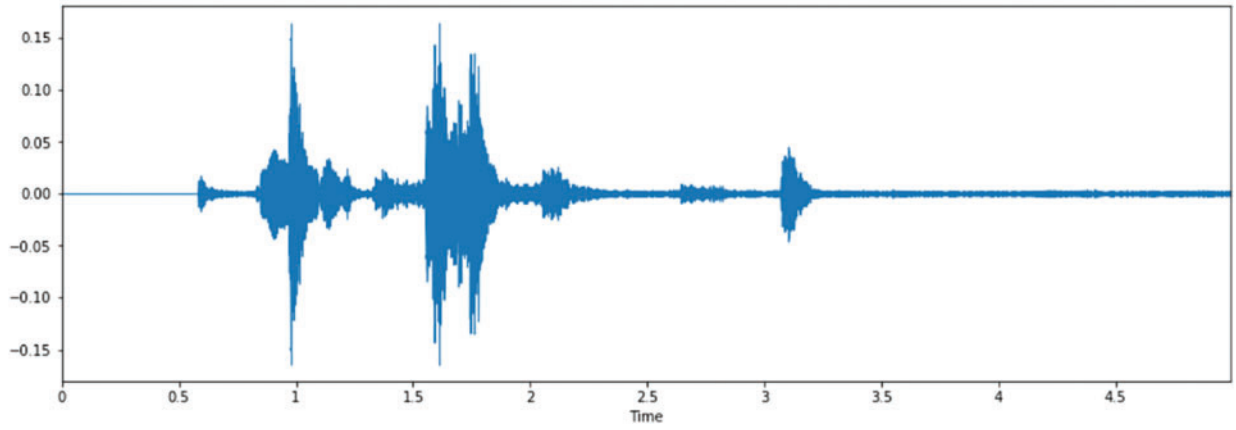


Figure 8: Emotion recognized within time

Table 3: No of samples based on TIMIT, Emo-DB and RAVDESS datasets

Emotions	Number of Speech Samples		
	Emo-DB	TIMIT	RAVDESS
Hate/Anger	127	91	1052
Worry	81	63	1073
Love	46	76	1093
Surprise	69	118	1085
Happiness	71	102	1021
Sadness	62	83	1029
Neutral	79	97	1003
Total	535	630	7356

5.1.1 Software Part

In our proposed methodology, using MLP classifier, the overall efficiency increased inaccuracy, time taken and the error rate reduced. The recognition rate was acquired as 81.02% for the RAVADEES dataset, 86.71% for Emo-DB and 84.23% for the TIMIT dataset as shown in Tab. 4 and graphical representation of our proposed work on three benchmark datasets are shown in Fig. 9. The error rate is calculated using equation six and the result acquired is a decrease of error rate by 19.79% for the RAVDEES dataset, 15.77% for the TIMIT dataset, and 14.88% for the Emo-DB dataset. Similarly, the time taken is also evaluated from the output sample as shown in Tab. 5, which 5.13 s decreases for

the RAVDEES dataset, 3.62 s for the TIMIT dataset and 0.01 s for the Emo-DB dataset. After the precise analysis of the results, it is established that the proposed model is giving better results than the existing state-of-art-methodology and could be a big boon for human life.

Table 4: SER accuracy based on the SPAB score in (%)

	WER	AA	CER	RR	SPL	SNR	SPAB	Accuracy (%)
EMO-DB	0.9642	1	0.966	0.87	0.942	0.959	5.173	86.2
TIMIT	0.9464	1	0.9	0.75	0.912	0.955	5.124	85.43
RAVDESS	0.9136	0.933	0.889	0.625	0.862	0.891	4.812	80.21

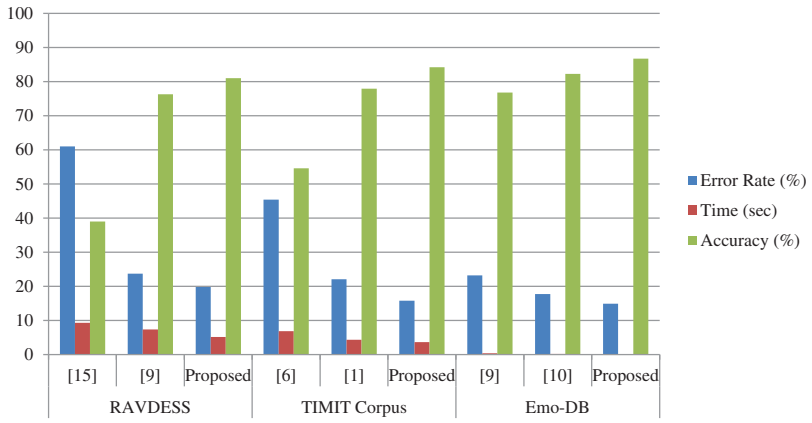


Figure 9: Comparison of accuracy with the state-of-art- methods (%)

Table 5: Comparison of accuracy with the state-of-art- methods (%)

Dataset	Reference	Methodology	Error Rate (%)	Time (sec)	Accuracy (%)
RAVDESS	[15]	GMM	61	9.31	39
	[9]	SVM	23.68	7.35	76.32
	Proposed	MLP based on ANN	19.79	5.13	81.02
TIMIT Corpus	[6]	GMM	45.39	6.82	54.61
	[1]	SVM	22.07	4.32	77.93
	Proposed	MLP based on ANN	15.77	3.62	84.23
Emo-DB	[9]	GMM	23.18	0.32	76.82
	[10]	SVM	17.7	0.19	82.30
	Proposed	MLP based on ANN	14.88	0.01	86.71

5.1.2 Hardware Part

After recognition, when the output comes under these recognitions, i.e., sadness, worry, hate and love the buzzer is used as an alert assistant. The recognition of emotions is categorized based on the number of beep sounds made by the buzzer. If the emotion is recognized as love, the buzzer sounds one beep, two beeps for worry, three beeps for hate, and four for love. When the emotion is recognized as happiness and surprise, the green LED glows, and the white LED glows to detect neutral emotion. This hardware component helps the people who are kept alone for social distancing to avoid contagious infections and also the staff who are treating the patient could monitor the patient's emotional condition and treat him well. In this way, buzzers and LEDs alert the following emotions, as shown in Fig. 7. When the proposed module cannot recognize the emotion from speech, the system will display a message that it could not recognize speech then automatically; the system goes to recording mode.

5.2 Discussion

From the deep analysis of the proposed module, results show that the GMM model produces an accuracy of 36%, 54.61% and 76.82% and when it comes to SVM Model, which gives 76.32%, 77.93% and 82.30% for RAVDEES, TIMIT and Emo-DB datasets. The other evaluation parameters, error rate and time, is taken were mentioned in Tab. 5. Many authors have worked on the same dataset and used various evaluation parameters per the analysis. In the RAVDESS dataset GMM method has produced a higher error and the proposed method has made minor errors comparatively, which is 4% less than the existing one. At the same time, execution time is also reduced by 2.22 s and accuracy is improved by 4.70%, which is a good improvement in accuracy and time. Comparative analysis is performed on the TIMIT Corpus dataset and again, GMM has produced the highest error rate of 45.39% and at the same time, the proposed method has reduced the error rate by 6.30% execution time is reduced by 0.70 s which does not show the significant difference between the existing and proposed methodology. But at the same time, accuracy is improved by 6.30%. Emo-DB is also used for the performance analysis of the proposed methodology and once again GMM method doesn't perform well and produces a higher error rate. The error rate is decreased by 2.82% using the proposed methodology. Execution time is not improved much through the proposed methodology, but still, it shows some improvement. The accuracy is also enhanced by 4.41% and it's a good improvement. Finally, we can conclude that our proposed work outperforms all three benchmark datasets. As emotion recognition through speech is a growing and exciting area of research, it could be helpful for human beings in many ways. The proposed method assists the people through LED alerts and buzzers. The key elements for the implementation of Speech Emotion Recognition (SER) are voice processing, and once the emotion is recognized, the machine interface automatically detects the actions by Buzzer and LED. But still, there is a scope of improvement in terms of accuracy because, as of now, we have achieved the highest accuracy of 86.71%. Hence accuracy can be improved in future work.

6 Conclusion and Future Scope

In this paper, proposed methodology provides a better result for the speech emotion recognition system over the seven emotions by MLP classification for all benchmark datasets which are considered for the research. While analyzing the results, it is observed that the MLP classifier has a high accuracy rate as compared to other state-of-art-method for detecting emotions from the speech signal. This proposed methodology leads us to conclude that speech recognition plays a vital role in better supporting individuals than other speech aids. In addition, the proposed module can help us to track

our loved ones and at the same time we can alert them. This will be a small contribution to us to our present situation of corona virus and its victims, where it could monitor the health condition of the people by maintaining social distance and provide better support to the people, alerts them through buzzer and LEDs which should be located in the audible and visible range.

In future, model could be trained and tested with real-time datasets to improve the performance of the system. Including camera modules and other peripherals could make the proposed methodology more efficient to serve people and it could also use in the defense for security purpose.

Acknowledgement: Zamil S. Alzamil would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-166.

Funding Statement: Zamil S. Alzamil would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-166.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Kumar and M. Mahajan, "Machine learning-based speech emotions recognition system," *International Journal of Scientific & Technology Research,"* vol. 8, no. 7, pp. 722–729, 2019.

[2] R. K. Prasada, M. S. Rao and N. H. Chowdary, "An integrated approach to emotion recognition and gender classification," *Journal of Visual Communication and Image Representation*, vol. 60, no. 1, pp. 339–345, 2019.

[3] M. Imani and G. Ali Montazer, "A survey of emotion recognition methods with emphasis on e-learning environments," *Journal of Network and Computer Applications*, vol. 147, no. 2, pp. 102423, 2019.

[4] W. Jiang, Z. Wang, J. S. Jin, X. Han and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors*, vol. 19, no. 12, pp. 2730, 2019.

[5] S. N. Roop, M. Prabhakaran and P. Betty, "Speech emotion recognition using deep learning," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 247–250, 2018.

[6] R. Errattahi, A. E. Hannani and H. Ouahmane, "Automatic speech recognition errors detection and correction: A review," *International Conference on Natural Language and Speech Processing*, vol. 128, pp. 34–37, 2018.

[7] Y. B. Singh and S. Goel, "Survey on human emotion recognition: Speech database, features and classification," in *Int. Conf. on Advances in Computing, Communication Control and Networking*, Greater Noida, India, pp. 298–301, 2018.

[8] V. M. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *Journal of Computer Science*, vol. 14, no. 11, pp. 1577–1587, 2018.

[9] S. Lukoseand and S. S. Upadhya, "Music player based on emotion recognition of voice signals," in *Int. Conf. on Intelligent Computing, Instrumentation and Control Technologies*, Kerala, India, pp. 1751–1754, 2017.

[10] D. Griol, J. M. Molina and Z. Callejas, "Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," *Neurocomputing*, vol. 326, no. 1, pp. 132–140, 2017.

[11] H. Jiang, B. Hu, Z. Liu, L. Yan and T. Wang, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, no. 2, pp. 39–46, 2017.

[12] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. 1, pp. 65–82, 2017.

[13] S. Lugović, I. Dunđer and M. Horvat, "Techniques and applications of emotion recognition in speech," in *39th Int. Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia, pp. 1551–1556, 2016.

[14] I. Perikos, and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Engineering Applications of Artificial Intelligence*, vol. 51, no. 1, pp. 191–201, 2016.
[15] P. Partila, J. Tovarek, M. Voznak and J. Safarik, "Classification methods accuracy for speech emotion recognition system," *Nostradamus 2014: Prediction, Modeling and Analysis of Complex Systems Prediction*, vol. 289, no. 3, pp. 439–447, 2014.
[16] C. A. Jason and S. Kumar, "An appraisal on speech and emotion recognition technologies based on machine learning," *International Journal of Recent Technology and Engineering*, vol. 8, no. 5, pp. 211–228, 2020.
[17] S. Kumar, S. Singh and J. Kumar, "Gender classification using machine learning with multi-feature method," in *IEEE 9th Annual Computing and Communication Workshop and Conference*, Las Vegas, NV, USA, pp. 0648–0653, 2019.
[18] S. Kumar, S. Singh and J. Kumar, "Live detection of face using machine learning with multi-feature method," *Wireless Personal Communications*, vol. 103, no. 3, pp. 2353–2375, 2018.
[19] S. Kumar,S. Singh and J. Kumar, "Automatic live facial expression detection using genetic algorithm with haar wavelet features and SVM," *Wireless Personal Communications*, vol. 103, no. 3, pp. 2435–2453, 2018.
[20] S. Kumar, S. Singh and J. Kumar, "Multiple face detection using hybrid features with SVM classifier," in *Data and Communication Networks, Data and Communication Networks*, Singapore: Springer, pp. 253–265, 2019. Online Available: https://link.springer.com/chapter/10.1007/978-981-13-2254-9_23.
[21] Z. Liu, M. Wu, W. Cao, J. Mao, J. Xu *et al.,* "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, no. 2, pp. 253–265, 2017.
[22] M. A. R. Khan and M. K. Jain, "Feature point detection for repacked android apps," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1359–1373, 2020.
[23] N. Binti, M. Ahmad, Z. Mahmoud and R. M. Mehmood, "A pursuit of sustainable privacy protection in big data environment by an optimized clustered-purpose based algorithm," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1217–1231, 2020.
[24] R. Shilpa, K. Lakhwani and S. Kumar, "Three-dimensional wireframe model of medical and complex images using cellular logic array processing techniques," in *Int. Conf. on Soft Computing and Pattern Recognition*, Switzerland, pp. 196–207, 2020.
[25] R. Shilpa, K. Lakhwani and S. Kumar, "Three dimensional objects recognition & pattern recognition technique; related challenges: A review," *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 17303–17346, 2022.
[26] X. R. Zhang, W. F. Zhang, W. Sun, X. M. Sun and S. K. Jha, "A robust 3-D medical watermarking based on wavelet transform for data protection," *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1043–1056, 2022.
[27] X. R. Zhang, X. Sun, X. M. Sun, W. Sun and S. K. Jha, "Robust reversible audio watermarking scheme for telemedicine and privacy protection," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3035–3050, 2022.
[28] S. Choudhary, K. Lakhwani, and S. Agrwal, "An efficient hybrid technique of feature extraction for facial expression recognition using AdaBoost classifier," *International Journal of Engineering Research & Technology, v*ol. 8, no. 1, pp. 30–41, 2012.