Current **A**pplied **S**cience and **T**echnology e-ISSN: 2586-9396

Curr. Appl. Sci. Technol. 2025, Vol. 25 (No. 1), e0257184

Research article

Emotion Classification from Speech Waveform Using Machine Learning and Deep Learning Techniques

Smitha Narendra Pai[1], Punnath Balakrishnan Shanthi[2]* and Shivaprasad Hegde[1]

[1]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, 576104, India

[2]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, 576104, India

Received: 4 January 2023, Revised: 7 February 2024, Accepted: 12 Jun 2024, Published:16 October 2024

Abstract

Emotions play a key role in determining the human mental state and indirectly express an individual's well- being. A speech emotion recognition system can extract a person's emotions from his/her speech inputs. There are some universal emotions such as anger, disgust, fear, happiness, pleasantness, sadness and neutral. These emotions are of significance especially in a situation like the Covid pandemic, when the aged or sick are vulnerable to depression. In the current paper, we examined various classification models with finite computational strength and resources in order to determine the emotion of a person from his/her speech. Speech prosodic features like pitch, loudness, and tone of speech, and work spectral features such as Mel Frequency Capstral Coefficients (MFCCs) of the voice were used to analyze the emotions of a person. Although sequence to sequence state of the art models for speech detection that offer high levels of accuracy and precision are currently in use, the computational needs of such approaches are high and inefficient. Therefore, in this work, we emphasised analysis and comparison of different classification algorithms such as multi layer perceptron, decision tree, support vector machine, and deep neural networks such as convolutional neural network and long short term memory. Given an audio file, the emotions that were exhibited by the speaker were recognized using machine learning and deep learning techniques. A comparative study was performed to identify the most appropriate algorithms that could be used to recognize emotions. Based on the experiment results, the MLP classifier and convolutional neural network model offered better accuracy with smaller variations when compared with other models used for the study.

Keywords: speech emotion detection; support vector machine; decision tree; multi-layer perceptron; convolutional neural network; long short-term memory

*Corresponding author: E-mail: shanthi.moorkoth@manipal.edu
https://doi.org/10.55003/cast.2024.257184

1. Introduction

Messages can be communicated by text or speech. Text messages do not convey the mood and tone of the speaker. In speech, the thoughts and feelings of a person are expressed in the form of articulated sounds. Variations in the signal level of speech vary with amplitude, timbre, pitch, duration of the speech and speaker. Signal resolution is better with a higher sampling rate. Speech consists of sequences of articulate sounds. The vibration of the vocal cords helps to categorize speech sounds. During the pandemic, the stress, anxiety, and depression levels of people increased, especially among those with mental health issues. A person's mental health can be detected by noting their behavior, facial expressions (Noroozi et al., 2017), and tone of speech. Music can provide a way of enhancing mood (Kaneria et al., 2021) and providing temporary relief. The style of music can be chosen based on emotional needs. Voice can also be a measure of the mental status of a person. Emotion recognition plays a vital role in the detection of emotions of people. Various deep learning techniques along with traditional models were studied. Different approaches have been carried out to study a person's emotions from their speech. Various deep learning techniques along with traditional models were studied (Khalil et al., 2019; Wani et al., 2021). These papers discussed how pitch, intensity of voice, speaking rate and quality of voice played an important role in measuring various emotions. Speech features were automatically extracted in the layers of a convolution neural network (CNN) and emotions were recognized.

Emotions in speech were recognized using multi-feature and multi-lingual fusion techniques by Wang et al. (2022), who suggested that a fusion of languages and the features using a multimodal approach could improve the accuracy of the results especially in the case of small datasets. The augmentation approach is another method that was used to address the problem of a small data set. Ying et al. (2021) used de-noising, auto encoders and adversarial auto encoders to capture features during training, which assisted in the improvement of the results. Long Short-Term Memory (LSTM) is another approach used for time series data. In the research of Xie et al. (2019), frame level combined with an attention based Long short-term memory recurrent neural network was used for measuring emotions. Hierarchical ConvLSTM was another approach used for studying speech signals (Mustaqeem et al., 2020), in which local features for spatial and temporal cues were used in the speech signals. Global feature weights were used based on the correlation of the input features. Features that have been used for emotion recognition include Mel-Scale Frequency Cepstral Coefficients, zero crossing rate, harmonic to noise rate and energy operator (Aouani et al., 2020). In this work, Auto encoders for feature selection and Support Vector Machine (SVM) for classification is the approach used.

Mel-Frequency Ceptral Coefficient along with deep learning approaches were used for detecting emotions in the signal (de Pinto et al., 2020). The dataset RAVDESS ("Ryerson Audio-Visual Database of Emotional Speech and Song"), was also used in our work. The details of the dataset were available in the literature (Livingstone & Russo, 2018). Another dataset that is used in the current work is TESS ("Toronto emotional speech set") (Pichora-Fuller & Dupuis, 2020). Both these datasets consist of emotions like angry, disgust, fear, happy, pleasant or calm, sad and neutral. It was observed that the effect of reducing the bandwidth using Alex net resulted in a decrease in the accuracy, which suggested the necessity of all the frequencies (Lech et al., 2020). Cepstral averaging and log spectrum were used in audio processing, speech processing, speech recognition and echo detection. Oppenheim & Schafer (2004) provided a deeper insight into the history of

cepstrum and its mathematical background by emphasizing its significant role in many speech recognition systems.

In this study, a relative comparison of machine learning and deep neural network algorithms such as Multilayer perceptron, Decision Tree classifier, Support Vector Machine (SVM), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) was performed. The datasets RAVDESS (Livingstone & Russo, 2018) and TESS (Pichora-Fuller & Dupuis, 2020) were used for implementation. Both these datasets consist of emotions like angry, disgust, fear, happy, pleasant, or calm, sad and neutral.

2. Materials and Methods

2.1 Dataset

The dataset used in this research study was a combination of the RAVDESS (Livingstone & Viriri, 2018) and TESS (Pichora-Fuller & Dupuis, 2020) data sets. The RAVDESS dataset contained a total of 1440 speech data files, which displayed emotions such as neutral, angry, sad, happy, disgust, fear, surprised and pleasant, which were performed by 24 different voice actors. The TESS dataset contained a total of 2800 speech data files that captured the emotions neutral, angry, sad, happy, disgust, fear and pleasant, and were done by voice actors of a range of ages. The voice files ranged in time length from 1.5 to 4.8 s. Figure 1 shows the count of various emotions presented in the dataset (Livingstone & Russo, 2018; Pichora-Fuller & Dupuis, 2020) where the dataset is the combination of both RAVDESS and TESS and data is accessed from both the dataset for processing.

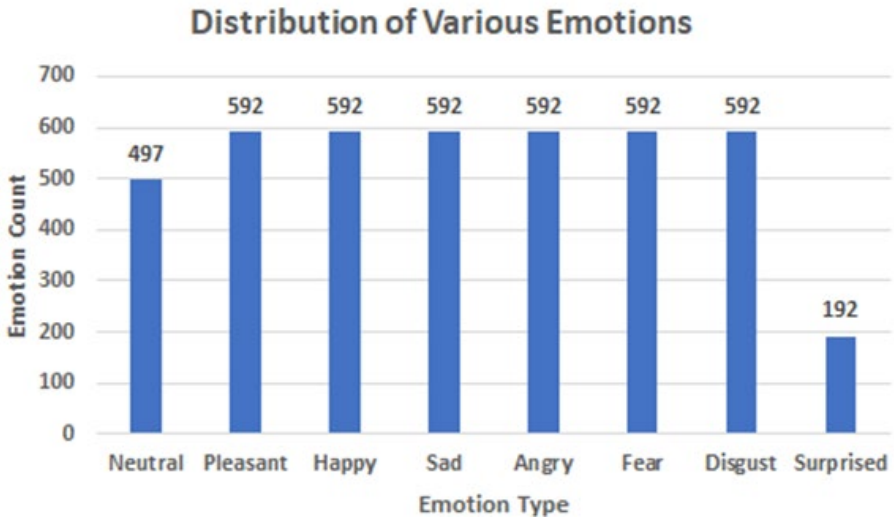

Figure 1. Distribution of various emotions in the combined dataset

2.2 Feature extraction

Feature extraction picks up relevant information from a dataset that can be helpful in classification, prediction, recommendation and finding of correlation between features. An audio signal can be represented using time, amplitude and frequency domain. The features

extracted from audio signals are fed into the model. The extracted features can be pitch, loudness, tone of speech and work spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) (Singh et al., 2012).

It is essential to transform audio signals from the time domain to the frequency domain for better analysis and recognition. The time domain shows signal variation with time. The frequency domain shows the variations of signal within a given band over a range of frequencies. Audio level can be measured using Peak and RMS. Peak measures the highest-level short duration peaks. RMS provides the average level of signal and is more associated with hearing. Zero-crossing is the rate at which the signal changes from positive to negative or vice versa. It aids in distinguishing the voiced and unvoiced sounds of an input speech signal. This feature has been widely used in speech recognition and music information retrieval (Gudmalwar et al., 2019; Abdusalomov et al. (2022). Let 'xi (n), n = 0, 1, … N–1' be the samples of the ith frame, then the zero-crossing rate is calculated as shown in equation 1.

$$Z(i) = \frac{1}{2N}\sum_{n=0}^{N-1} |\ sgn[x_i(n)] - \ \ sgn[x_i(n-1)] \tag{1}$$

where,

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0 \\ -1, & x_i(n) < 0 \end{cases}$$

The total magnitude of the signal is represented by the energy of the signal (Patni et al., 2021). The energy of the signal is defined in equation 2.

$$\text{Energy} = \sum_n |x(n)|^2 \tag{2}$$

Another important parameter that signifies the size of a signal is root mean square (RMS). In audio signals, the signal value is squared, averaged over a period, and then the square root of the result is calculated as shown in Equation 3. This helps to measure the degree to which the speech signals are spread between the central points over a distance (Patni, et al., 2021; Mashhadi & Osei-Bonsu, 2023).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_n |x(n)|^2} \tag{3}$$

The spectral centroid is a commonly used audio feature that gives the centre mass of the spectrum. It identifies the frequency band where most of the energy is concentrated (Choudhury et al., 2018). The value of the spectral centroid of the ith audio frame and A(k) is the spectrum amplitude in a given time window which is shown in equation 4.

$$SC_t = \frac{\sum_{k=1}^{N} k.A(k)}{\sum_{k=1}^{N} A(k)} \tag{4}$$

The rate of change of the spectral bands of a signal is called the cepstrum. The cepstrum is another feature which is expressed as a nonlinear transformation of a spectrum (Madanian et al., 2023). It acts as a tool for analyzing periodic structures in frequency spectrums. The most used cepstrum-based features are 'Mel-frequency cepstral coefficients (MFCCs)' and 'linear predictive cepstral coefficients' (Ancilin & Milton, 2021).

MFCC is a feature widely used in automatic speech recognition. The MFCC (Mel Frequency Cepstral Coefficients) includes a small feature set which describes how the level of sound amplitude changes over time. It is the inverse fourier transform of the logarithm of the estimated signal spectrum (Lalitha et al., 2015; Constantinescu & Brad, 2023). The shape of the vocal tract establishes itself in the envelope as a short time power spectrum, and the role of an MFCC is to precisely identify this envelope. This spectral envelope is the most informative part of the spectrum that contains data with the resonance properties of the vocal tract (Patni et al., 2021; Akinpelu & Viriri, 2023).

The Mel scale relates the perceived frequency, or pitch, of a pure tone to its actual measured frequency. By incorporating the Mel scale, the extracted features of a tone can be harmonized more closely with what humans hear. Mel is known as a unit of perceived fundamental frequency. A frame block is created which is subjected to windowing. After conducting a Fourier transform, the Mel features are extracted. After the use of the MFCC feature extraction method (Nassif et al., 2022), energies in a spectrum are unified by a set of band limited triangular weighting functions called filter banks. Figure 2 shows the steps in extracting the features of the speech signal.

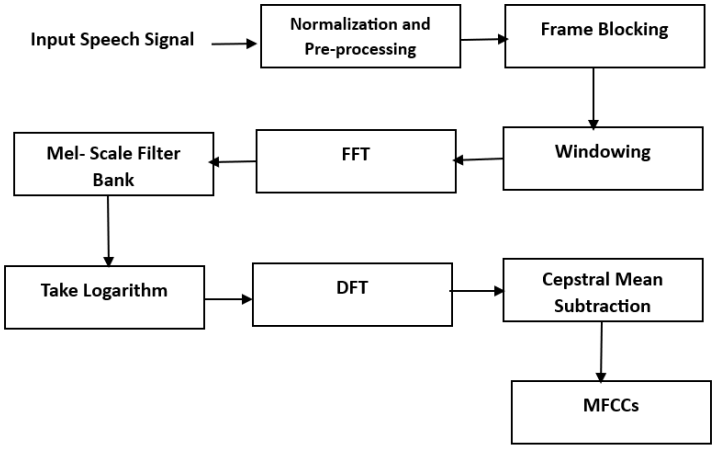

Figure 2. The feature extraction process

The log of combined spectral energies is obtained and projection to cosine bases is performed. Then the MFCC features are derived by computing the DCT of all log Mel spectrums. To reduce differences in the feature representation between speakers, cepstral mean subtraction (CMN) is performed, which helps to mitigate the detrimental influence of the background noise (Nassif et al., 2022).

The band energy ratio (BER) gives the relation between the lower and higher frequency bands (Nassif et al., 2022; Constantinescu & Brad, 2023). It can show the dominance of low frequencies. The band energy ratio for a frame t is computed as in equation 5,

$$BER_t = \frac{\sum_{k=1}^{F-1} A(k)^2}{\sum_{k=F}^{N} A(k)^2} \quad (5)$$

where 'A(k)' denotes spectrum amplitude in a given time window and 'F' denotes the split frequency.

The waveform shows the amplitude of the sine wave but not the frequency or pitch changes over time. A sequence of phonemes constitutes the speech signal. Figure 3 shows the variation in sound for different emotions that were obtained from the RAVDESS (Livingstone & Viriri, 2018) and TESS (Pichora-Fuller & Dupuis, 2020) data sets. In the frame level, the frequency domain representation of the speech signal is the short-term Fourier transform. The spectrogram provides a visual illustration of the frequency domain sound representation. It shows the intensities of frequencies over time. It is the squared magnitude of the short time Fourier transform (STFT). It gives the magnitude of frequency for each bin for a frame at various times.

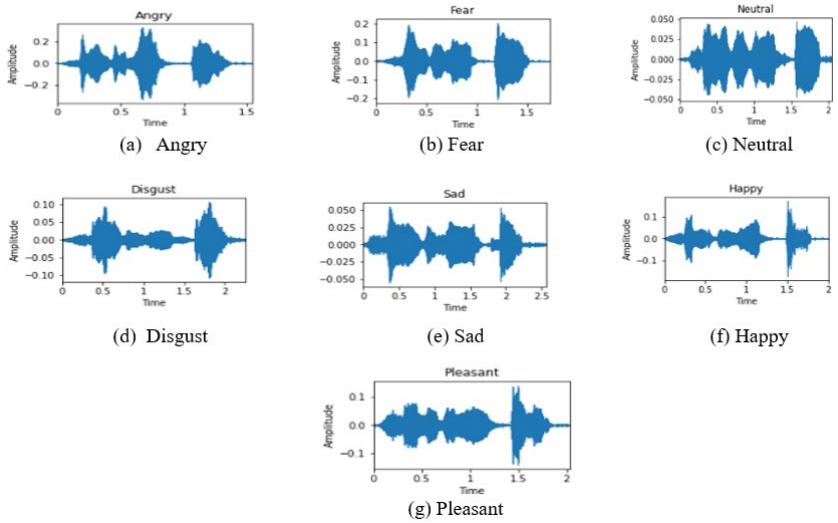

Figure 3. Variation in sound waveform for different emotions

Chroma features represent audio by projecting the entire spectrum into 12 chroma of octaves with 12 bins. At a gap of one octave, the similarity is noticed. With the chroma distribution, and even without the absolute frequency or original octave, the audio information can be obtained. This can assist in finding similarity in audio signals which might not be apparent in the original spectra. A visual representaion of the angry emotion is presented in Figure 4. Based on the survey of speech emotion analysis among the features extracted, MFCC plays an effective role in manitaining the general form of audio waveform and it is utilized in a wide variety of speech processing applications. So, in our proposed convolutional model, the MFCC features are used as input for speech emotion classification (Nassif et al., 2022; Akinpelu & Viriri, 2023).

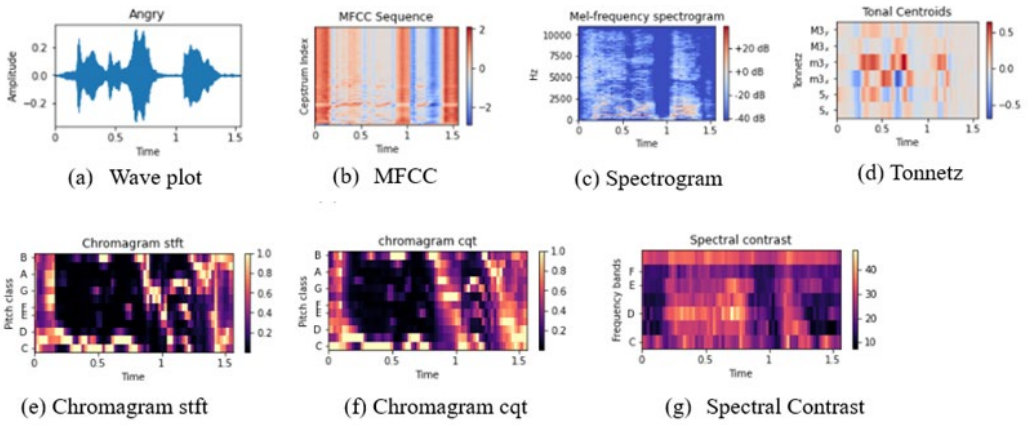

Figure 4. Visualization of angry emotion

2.3 Machine learning techniques

Machine learning techniques are used for classification and clustering techniques. In the current work, various machine learning algorithms were used for multiclass classification problems. An explanation of each of these techniques is given below.

2.3.1 Support vector machine (SVM)

In the SVM algorithm, features form the coordinates when features are plotted in n-dimensional space. The classes are differentiated by drawing a hyperplane in the n-dimensional space, which is shown in Figure 5. SVM consists of a kernel function that takes low dimensional input value and converts it to higher dimensional values. Given a set of labelled training examples, the SVM algorithm finds the points that have the least distance between the classes and these points are called support vectors (Jain et al., 2020). The margins for these classes are found by computing the distances between the line and the support vectors points. The hyperplane that has a maximum margin is considered as optimal in the current case. The parameters used in SVM are as follows.

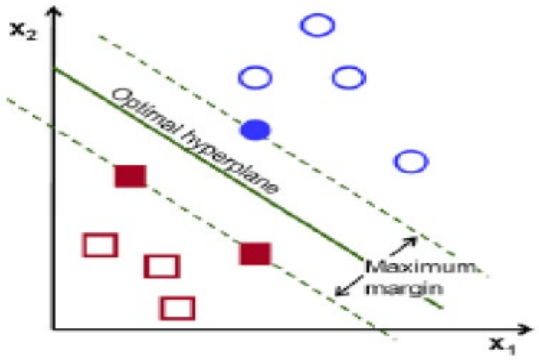

Figure 5. SVM classifier

C control, which is a hyper parameter, is used to regulate the error in the SVM. It provides a balance between margin and training error and aids in minimizing misclassification on training data. Therefore, the values of C are varied to get a perfectly balanced curve and to try not to over fit. In this research study, C=1 and C=10 was chosen.

Entropy is an approach which is used to measure the goodness or the correctness of a split at each node. For 'n' equally probable outcomes, the probability of each outcome is 1/n. Gamma, which is a hyper parameter in the SVM classifier, is only used in the case of an RBF Gaussian kernel. The gamma parameter is set before the training of the model has begun, and it helps to decide the curvature of the decision boundary (Samantaray et al. 2015; Jain et al., 2020). For higher gamma values, the spread of points is lower throughout the boundary and the curvature is highly bounded. For lower values of gamma, the spread of points in the hyperplane is greater in the boundary, and the curvature is lower. The kernel function is used as a shortcut method to operate on higher dimensional data in cases of non-separable samples. The class of decision functions are simplified by mapping the samples from the input space into a high-dimensional feature vector. The kernel functions allow SVM classifiers to accomplish partings even with very complex boundaries. Many kernel functions, including polynomial, sigmoid, and radial basis functions (RBFs) are available. In this study, all kernel functions were compared to check for maximum accuracy (Milton, et al., 2013; Samantaray et al., 2015).

Two vectors 'yi' and 'yj' with σ representing the kernel parameter are shown in equation 6. This function transforms the input data into a higher dimensional space that allows to capture non-linear relationships between the input data.

$$F(y_i, y_j) = (\sigma y_i^T y_j + k)^d , \ \sigma > 0 \quad (6)$$

An RBF kernel uses an exponential function to find the distance, as shown in equation 7. It helps in the handling of non-linear relations between the classes and the attributes by mapping them into higher dimensional space. The RBF model is represented as:

$$F(yi, yj) = \exp (-\sigma \, ||yi - yj \, || \, 2), \ \sigma > 0 \quad (7)$$

A sigmoid kernel defines that the kernel should be positive definite and symmetric by definition, i.e. ($K=K^T$). A sigmoid kernel does not exhibit the positive semi definite nature for some values of the parameter and is represented as follows in equation 8:

$$K(y_i, y_j) = \tanh (\sigma y \, Ti \ \ y_j + r) \quad (8)$$

2.3.2 Multi-layer perceptron (MLP)

This is a feed forward artificial neural network with multiple layers, and all layers are connected. The network uses a BackPropagation algorithm for training the model which comes under the class of deep learning (Rathor et al., 2021). Back propagation is a technique that is used to find the loss that occurs in the model. In the current work, a total 64 hidden layers were used where each node was assigned a weight for computation.

An activation function is used to specify if the node information should be considered for further computation or not. The nodes which are present in the model are assigned with some weights which are later computed, and the activation function is applied on each level. Rectified linear units (ReLU) are one such activation function. ReLU is a linear function which interprets the positive part of its arguments, if the function receives positive value and returns zero for negative value. Figure 6 shows the ReLU functionality. ReLU overcomes the vanishing gradient problem by not considering all neurons for computation (Poojary et al., 2021; Yuan et al., 2022).

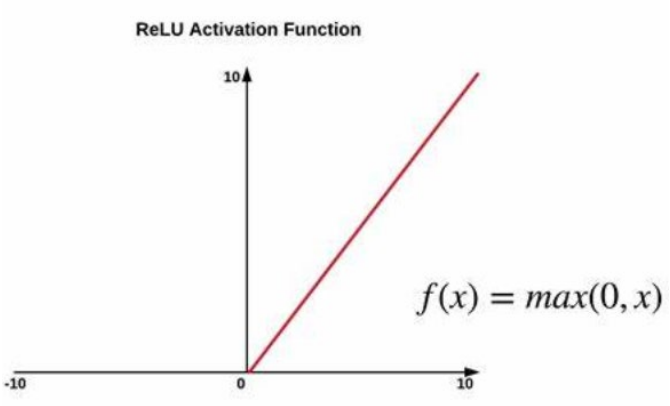

Figure 6. ReLU function

2.3.3 Decision tree classifier

Decision tree classifier is a predictive modelling approach that involves building a model in the form of a tree structure. The model defines the set of rules and generates a tree, from a given dataset. It breaks down the dataset into smaller subsets until the leaf node is reached. Each node is further split based on some condition on the node. Figure 7 shows the structure of a decision tree (Liu et al., 2018; Sun et al. 2019).

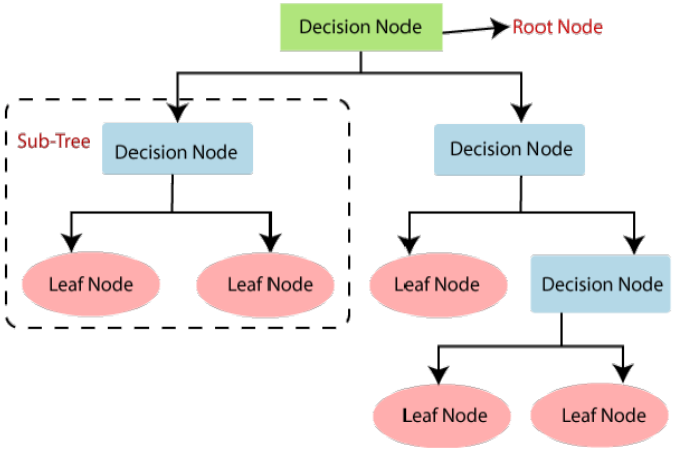

Figure 7. Structure of decision tree

The criteria used for creating a tree includes a splitting attribute with each node of the decision tree. Max depth is another parameter that defines the maximum depth between the root node and leaf node. Splitting criterion can define the condition by which the splitting at a node takes place based on the attribute which is chosen, and is called the splitting criterion at that node (Sun et al., 2019).

Entropy and Gini are the common criteria used for measurement of splitting attributes. Entropy is an approach which is used to measure the goodness or the correctness of a split at each node. If there are n equally probable outcomes, then the probability of each outcome is 1/n. Probability distribution and entropy are represented in equations 9 and 10, repectively.

$$P = (p1, p2, ... pn) \tag{9}$$

$$Entropy(P) = - p1 \log(p1) - ... - pn \log (pn) \tag{10}$$

The Gini index is used to evaluate the goodness of a split. It has a maximum value of 1, and a minimum value of 0. If the value of the Gini index is maximum, there is an equal distribution of classes. If the Gini index is low, a single class will have higher chance. The best splitter is the one that decreases the diversity of the record sets (Liu et al., 2018). The Gini index is a diversity measure used to evaluate the goodness of a split. If a dataset contains n classes, and p represents the probability of being recognized in the class, the Gini index is defined using equation 11.

$$Gini = 1 - \sum_{i=1}^{n} {p_i}^2 \tag{11}$$

2.4 Deep learning techniques

Deep learning techinques are used in speech emotion recognition. Multi- layered neural networks are used to capture and classify emotional states in speech signals. Automatic recognition of speech and emotion has become an important research area and has been used for different purposes in fields such as education, communication, automobiles, and health.

2.4.1 Convolutional neural network (CNN)

Emotion identification using deep learning has recently received much attention because it can classify a nonlinear problem, and also because it can handle sequential characteristics of speech signals. To extract features automatically from raw audio data, deep neural networks based on generalized discriminant analysis (GerDA) are used (Abdu et al., 2021). The main difficulty faced in speech recognition is to extract the appropriate features related to speech and to identify the best classification model. Deep learning approaches use unstructured audio representation and some of the commonly used features include Mel-frequency cepstral coefficients, chromagram, Mel-scale spectrogram, Tonnetz representation, and spectral contrast features. These features are extracted from sound files and used as input into the best convolutional model for recognition of emotions (Wani et al., 2020; Abdu et al., 2021).

In our proposed work using CNN, raw audio images were preprocessed using noise, stretching, shifting, and pitching. Spectogram is considered as a useful representation of speech that visualizes many pertinent features of speech signals.

Therefore, among the features extracted, MFCC features were used to determine the classification of the emotions. The proposed CNN model was composed of three convolution layers, three pooling layers, a fully connected layer, and included activation functions such as ReLU and SoftMax. Figure 8 shows the structure of the CNN. The initial layer of the model had 256 filters with a kernel size of 8x8 with stride 1. The ReLU activation function introduced non-linearity, and ConvNet and was used to achieve higher convergence rates. The pooling layers aided in reducing the dimensionality of each feature map without losing the important information. Then, a maximum pooling layer was added to reduce the spatial size of the representation, which aided in the reduction of the amount of parameters and computation in the network. Next, convolutional layers with 128 filters and kernel sizes of 8 were added followed by the addition of a third layer with 64 filters and 3 x 3 kernel size. A fully connected layer is a flattened matrix that combines each neuron in one layer with neurons in other layers. The flattening layer helps in the reduction of features to a single value. SoftMax unit is used for multiclass classification problems that solve the issue of assigning an instance to one class when the number of possible classes is larger than two. This was the last activation function of the neural network and was used to normalize the output of the network to a probability distribution over predicted output class.

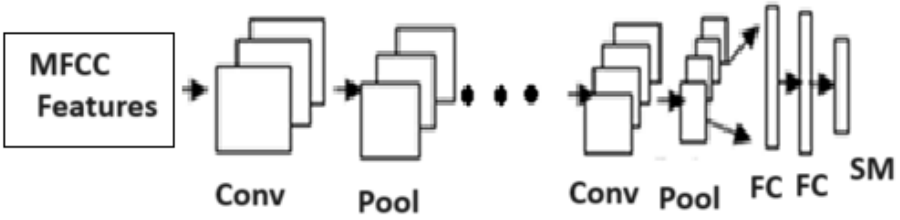

Figure 8. Structure diagram of a CNN

2.4.2 Long short-term memory (LSTM)

LSTM (long short-term memory) is an extension of recurrent neural networks (RNNs). LSTM, which is depicted in Figure 9, uses gates to control the memorizing process and provides prolonged short-term memory. By using LSTM, the vanishing gradient problem is almost removed, and the LSTM network can also handle noise, distributed representations, and continuous values. It provides a large range of parameters such as learning rates, and input and output biases. In the LSTM shown in Figure 9, 'X' represents scaling information, '+' represents adding information, 'σ' represents the sigmoidal layer, 'tanh' represents the tanh layer, 'h(t-1)' represents the output of last LSTM unit, 'c(t-1)' is the memory from last LSTM unit, 'c(t)' is the new updated memory, 'X(t)' is the current input, and 'h(t)' is the current output. To overcome the vanishing gradient problem, the tanh function, whose second derivative can be sustained for a long range before going to zero, is used (El Maghraby et al., 2021).

In the proposed model, 180 MFCC features were extracted using Librosa library and inputted into the model. The model was executed with the Keras library by including two LSTMs stacked sequentially with 512 and 256 parameters. The output dense layered with 7 nodes is shown in Figure 10. The activation function used was Softmax.

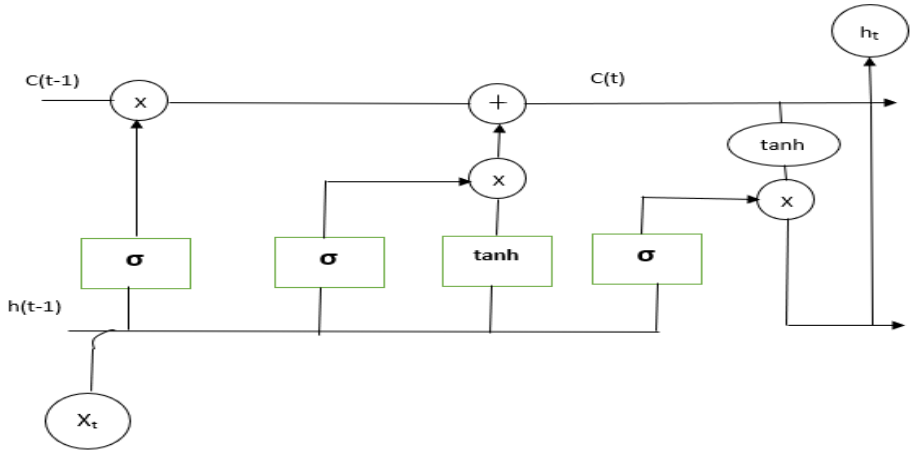

Figure 9. Structure diagram of LSTM

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, None, 512)         1419264
_________________________________________________________________
lstm_1 (LSTM)                (None, 256)               787456
_________________________________________________________________
dense (Dense)                (None, 7)                 1799
=================================================================
Total params: 2,208,519
Trainable params: 2,208,519
Non-trainable params: 0
```

Figure 10. LSTM model specification

The sigmoid layer can output 0 or 1, which are used to forget or remember the old output information. The sigmoid layer takes the input x(t) and h(t-1) and decides which part of the old output should be removed by outputting a 0. The sigmoid layer decides which of the new information should be updated or ignored. The tanh layer creates a vector of all the possible values from the new input. These two are multiplied to update the new cell state (Wani et al., 2020; El Maghraby et al., 2021). This new memory is then added to old memory c(t-1) to give c(t). This model has been widely used in speech recognition, text prediction and sentiment analysis because it can effectively exploit temporal dependencies in acoustic data and produce effective classification accuracy.

3. Results and Discussion

Analysis of various classification algorithms including multi layer perceptron, decision tree, support vector machine, and deep neural networks such as convolutional neural network and long short term memory were performed and the results are presented. Emotions considered in the the study were angry, disgust, fear, happy, pleasant, sad and neutral. MFCC represented the vocal tract and extracted features from the audio signal. The audio signal was divided into smaller windows and each window captured features of the signal represented as MFCC. One hundred and eighty features were extracted and used for this study. The combined dataset used for the study comprised 4048 files of which 2833 (70%) were used for training and 1215 (30%) for testing. Based on the experiment results, deep neural networks were used to the improve the accuracy of the speech emotion recognition.

3.1 Evaluation metrics

Confusion matrix is one of the tools and techniques that has been used for analyzing the performance of the classifier. Evaluation metrics such as 'precision', 'recall', 'accuracy', and 'F1 score' were used to analyze the performance of the classifier and the relevant equations are shown in Equations 12, 13, 14, 15 and 16. True positive (tP) is an outcome where the model retrieves relevant instances, false positive (fP) indicates retrieval of instances that are not relevant, true negative (tN) is an outcome whereby the model suitably predicts negative class, and false negative (fN) is where the test result incorrectly indicates the presence of condition or disease.

$$\text{Error rate} = \frac{fP+fN}{tP+tN+fP+fN} \tag{12}$$

$$\text{Accuracy} = \frac{tP+tN}{tP+tN+fP+fN} \tag{13}$$

$$\text{Precision} = \frac{tP}{tP+fP} \tag{14}$$

$$\text{Recall} = \frac{tP}{tP+fN} \tag{15}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

3.2 MLP classifier model

The multi-layer perceptron model is used to recognize emotions in fixed areas of the speech signal, each of which is characterized by a Mel-frequency cepstral coefficient (MFCC). Figure 11 shows the accuracy graph for the MLP classifier model over the number of iterations. The GridSearchCV approach was used to find the best hyper parameter. The values assigned were alpha with value 0.01, batch size 256, 64 hidden layers, ReLU activation function, and a maximum iteration of 100. Adaptive learning rates were used; this kept the learning rate constant and aided in controlling the step-size in updating the weights. Experiments for different values of hidden layer, alpha value, activation function

Figure 11. Accuracy plot for MLP classifier

and batch size were carried out. The hyperparameters used for the study gave marginally better results. With lower values of batch size, the accuracy slightly increased and the computational time increased with increased number of hidden layers. Table 1 depicts the various metrics for comparison of results. These values were chosen after experimentation.

Table 1. Performance evaluation using MLP classifier

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	94	90	94	93
Disgust	90	87	90	88
Fear	85	90	83	86
Happy	80	83	80	82
Pleasant	84	83	84	84
Sad	87	82	87	85
Neutral	80	82	80	81
Overall	85.71	85.28	85.42	85.57

3.3 Decision tree classifier

Figure 12 demonstrates the accuracy plot for decision tree classifier over the range of maximum depth of a tree from 2-25, by taking split criterion as Entropy and minimum sample split as 4.

Figure 13 depicts the accuracy plot for decision tree classifier over the range of maximum depth of a tree from 2-25, by taking Gini as the split criterion and minimum sample split as 4. The optimal depth of 12 was chosen for computation based on the results. Table 2 shows the evaluation results for various emotions using decision tree with criterion Entropy. Table 3 shows the evaluation result for various emotions using decision tree with Gini as criterion.

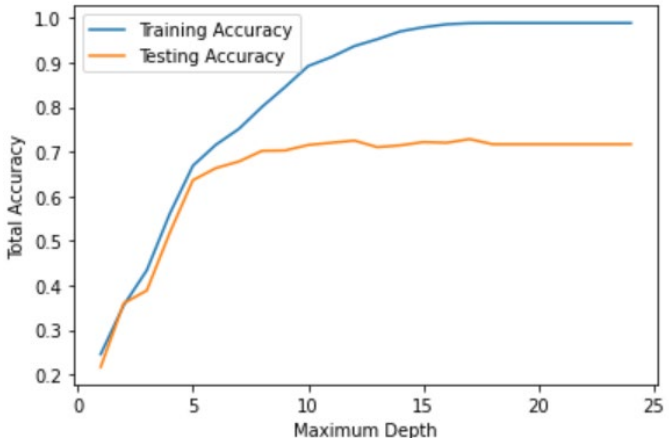

Figure 12. Accuracy plot (criterion= “Entropy”)

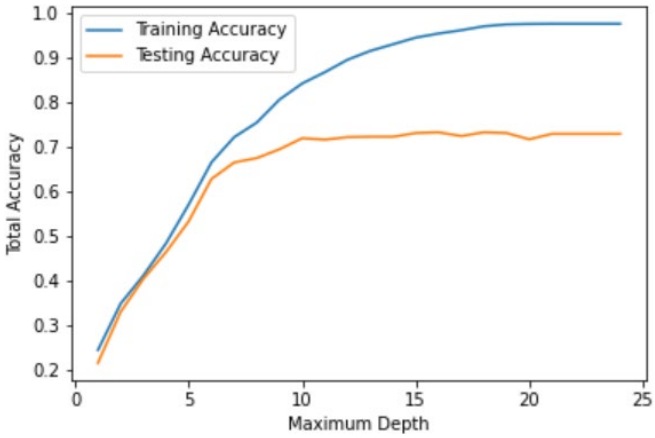

Figure 13. Accuracy plot (criterion= “Gini”)

Table 2. The results for decision tree classifier with depth=12, criterion= “Entropy”

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	77	73	76	75
Disgust	74	69	74	71
Fear	69	74	69	71
Happy	68	70	67	69
Pleasant	82	75	82	78
Sad	74	66	74	70
Neutral	66	80	65	72
Overall	72.85	72.42	72.42	72.28

Table 3. The results for decision tree classifier with depth=12, criterion= “Gini”

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	80	72	80	76
Disgust	72	57	72	64
Fear	73	75	72	73
Happy	65	70	65	67
Pleasant	82	81	82	81
Sad	70	76	69	72
Neutral	66	79	65	71
Overall	72.26	72.85	72.1	72

3.4 SVM Classıfıer

The SVM classifier was used with various values of C and gamma for the kernel as shown in Table 4. Table 5 shows the results obtained using C=1 and gamma as auto. The accuracy graph for the SVM classifier is shown in Figure 14. From the experiment, it can be concluded that SVM classifier in polynomial kernel function with C=1 and gamma value as “auto” gave the highest accuracy.

Table 4. Accuracy using various values for gamma and different kernel functions

Accuracy Measure in %	Gamma Values			
Kernel Function	**Auto**		**Scale**	
	C=1	**C=10**	**C=1**	**C=10**
Polynomial	81.56	81.48	68.97	78.02
RBF	64.86	67.49	66.67	78.19
Sigmoid	13.58	13.58	25.58	12.02

Table 5. The results for SVM classifier with C=1 and gamma= “auto”

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	89	79	88	83
Disgust	87	81	86	84
Fear	82	80	81	81
Happy	73	80	72	76
Pleasant	87	83	87	85
Sad	81	81	80	81
Neutral	74	84	74	78
Overall	81.56	81	81.14	81.2

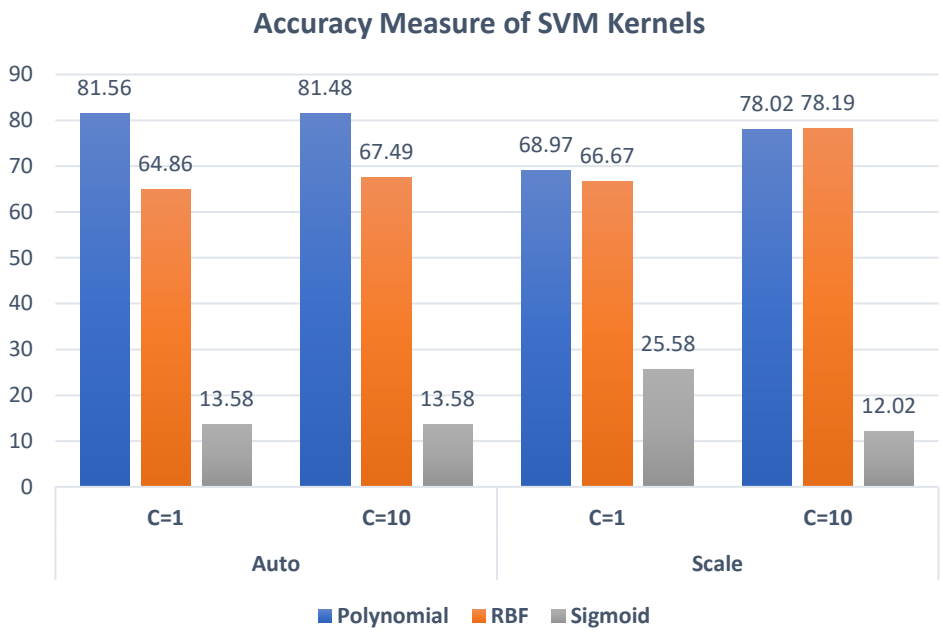

Figure 14. Accuracy graph for SVM classifier

3.5 Convolutional neural network

The strength of deep neural networks is that it can be used to learn high level features without relying on hand crafted features. It improved the performance of speech signal processing to a great extent. The results using CNN are shown in Table 6. The accuracy plot of the CNN classifier is shown in Figure 15.

Table 6. Accuracy, precision, recall and F1 score metrics using CNN

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	92	92	92	92
Disgust	83	81	93	86
Fear	84	84	88	86
Happy	90	89	80	84
Pleasant	84	87	77	82
Sad	83	83	94	88
Neutral	96	97	81	89
Overall	87.42	87.57	86.42	86.71

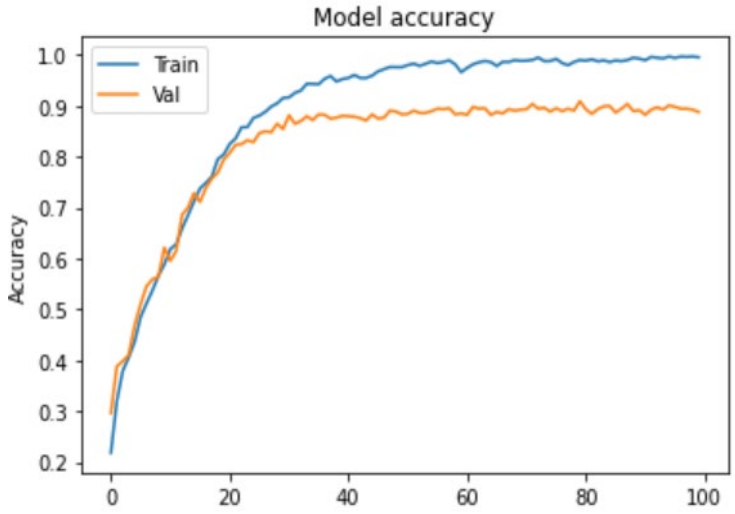

Figure 15. Accuracy plot for CNN

3.6 Long short-term memory

LSTM layer is used to learn long-term dependencies from local learned functions. To verify the performance ability of the developed LSTM network, output was recorded for the training and verification sets. The results using LSTM are shown in Table 7 and the accuracy plot of the LSTM classifier is shown in Figure 16.

Table 7. Accuracy, precision, recall and F1 score metrics using LSTM

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Angry	86	90	86	88
Disgust	83	91	83	87
Fear	77	91	75	82
Happy	83	76	82	79
Pleasant	77	82	77	79
Sad	77	78	77	77
Neutral	91	65	91	76
Overall	82	81.85	81.57	81.14

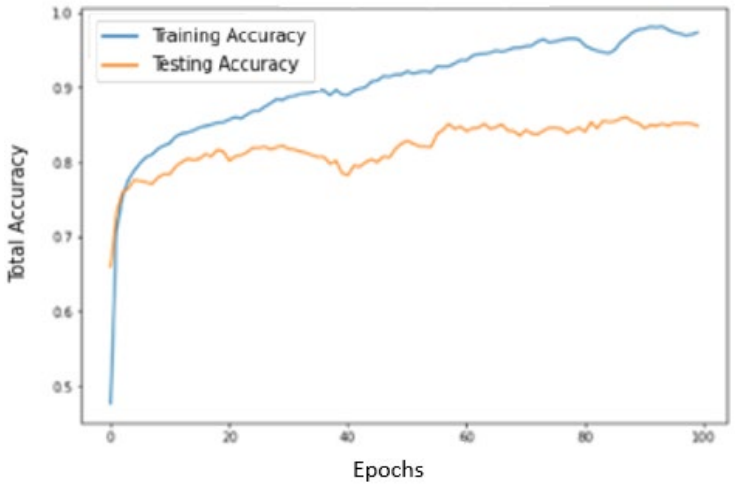

Figure 16. Accuracy plot for LSTM

3.7 Comparative study of the models

In our study, we experimented with speech emotion, with different models, including traditional machine learning and deep learning algorithms. Initially, multi-layer perceptron was used and an overall accuracy of 85.71% was achieved. With the use of decision tree, the maximum optimal depth was identified as 12. When the criteria “entropy” and “gini” were used to measure the quality of split, a maximum accuracy of 72% was achieved. Using SVM, an accuracy of 81.56% with C=1 and gamma= “auto” was achieved. To improve the accuracy, deep convolutional neural network was used and the highest accuracy of 87% was achieved. As LSTM used gates and provided elongated short-term memory, we used this network and an accuracy of 82% was achieved. As a part of future work, a LSTM variant called ‘gated recurrent unit’ can be explored to improve the accuracy of the classification and to reduce the complexity of the model.

The overall comparison of various models using k fold cross validation is summarized in Table 8. After a number of experiments, the changes made in various parameters were alpha with 0.01, batch size as 256, hidden layer size as 64 and using ReLU as activation function with learning rate as adaptive and maximum number of iterations as 200. For decision tree, maximum depth was 25, the criterion used was entropy, and the minimum sample split was 4. For SVM, various parameters used were with gamma value as auto, kernel type polynomial with degree 2 and C value of 1. We performed K fold validations for all machine learning models used in the study. In the experimentation, MLP showed good accuracy when compared to SVM. K fold cross validation was more appropriate for simple models with few parameters. For CNN and LSTM, however, increasing the number of layers in the neural network introduced thousands of parameters. Therefore, it did not show good accuracy. Instead of k fold validation, we thus performed experiments with learning rate, batch size, dropout and batch normalization, which produced acceptable results.

Table 8. Comparison of various models for average accuracy, precision, F1 score, recall, and SD using 5 and 10fold cross validation

	5-fold (%)					10-fold (%)				
Model	**Acc**	**Prec**	**Recall**	**F1 Score**	**SD**	**Acc**	**Prec**	**Recall**	**F1 Score**	**SD**
SVM	80.9	83.2	80.8	81.4	0.02	81.3	83.4	81.1	81.6	0.02
MLP	**86.8**	**87.3**	**86.4**	**86.4**	**0.01**	**87.1**	**87.8**	**86.8**	**87.3**	**0.02**
DT	73.9	74.2	74.0	74.0	0.02	74.8	75.0	75.0	74.9	0.03
CNN	**87.02**	**87.0**	**87.0**	**86.0**	**0.03**	**89.36**	**90**	**89.0**	**89.0**	**0.03**

Acc = accuracy, Prec = precision, SD = standard deviation

4. Conclusions

This study presented different ways to detect emotions from speech with the help of MFCC feature vectors. Moreover, machine learning and deep learning techniques such as multi-layer perceptron, decision tree classifier, support vector machine, convolutional neural networks and long short-term memory classifier were used. The aims of this study were to find optimal supervised parameters and feature representation. Among the machine learning techniques used for the study, multi-layer perceptron and the SVM classifier with polynomial kernel performed better than other models. Moreover, the activation function of MLP classifier improved the accuracy of the model. However, finding out the effective and optimal activation and kernel function for a given learning task is still an unsolved problem. In deep learning techniques, convolutional neural networks performed well in our study when compared with LSTM. CNN showed good accuracy as it learned the features deeply to predict the emotions. As LSTM used gates to control the memorizing process, provided prolonged short-term memory, and exploited temporal dependencies in acoustic data effectively, it then seemed to be a more appropriate model for speech recognition. By increasing the sample size of the data, the model can be made less prone to overfitting problems. Among the studies made on various machine learning and deep learning models, MLP and CNN generally provided better accuracy with smaller variation. For future work, other deep neural network models should be utilized to more effectively classify emotions with improved accuracy.

5. Acknowledgements

We are grateful to all authors mentioned in the reference section for providing insights to perform and execute the study.

6. Conflicts of Interest

The authors declare that they have no conflicts of interest.

ORCID
Smitha Narendra Pai https://orcid.org/0000-0002-3258-2688
Shanthi Punnath https://orcid.org/0000-0003-3276-7073

References

Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion*, 76, 204-226. https://doi.org/10.1016/j.inffus.2021.06.003

Abdusalomov, A. B., Safarov, F., Rakhimov, M., Turaev, B., & Whangbo, T. K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. *Sensors*, 22(21), Article 8122. https://doi.org/10.3390/s22218122

Akinpelu, S., & Viriri, S. (2023). Speech emotion classification using attention based network and regularized feature selection. *Scientific Reports*, 13(1), Article 11990. https://doi.org/10.1038/s41598-023-38868-2

Ancilin, J., & Milton, A. (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, 179, Article 108046. https://doi.org/10.1016/j.apacoust.2021.108046

Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. *Procedia Computer Science*, 176, 251-260. https://doi.org/10.1016/j.procs.2020.08.027

Choudhury, A. R., Ghosh, A., Pandey, R., & Barman, S. (2018). Emotion recognition from speech signals using excitation source and spectral features. In *2018 IEEE Applied signal processing conference (ASPCON)* (pp. 257-261). IEEE. https://doi.org/10.1109/ASPCON.2018.8748626

Constantinescu, C., & Brad, R. (2023). An Overview on Sound Features in Time and Frequency Domain. *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*, 13(1), 45-58. https://doi.org/10.2478/ijasitels-2023-0006

de Pinto, M. G., Polignano, M., Lops, P., & Semeraro, G. (2020). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In *2020 IEEE conference on evolving and adaptive intelligent systems (EAIS)* (pp. 1-5). IEEE. https://doi.org/10.1109/EAIS48028.2020.9122698

El Maghraby, E. E., Gody, A. M., & Farouk, M. H. (2021). Audio-visual speech recognition using LSTM and CNN. *Recent Advances in Computer Science and Communications,* 14(6), 2023-2039. http://doi.org/10.2174/2666255813666191218092903

Gudmalwar, A. P., Rao, C. V. R., & Dutta, A. (2019). Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology*, 22, 521-531. https://doi.org/10.1007/s10772-018-09576-4

Jain, M., Narayan, S., Balaji, P., Bharath, K. P., Bhowmick, A., Karthik, R., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*. https://doi.org/10.48550/arXiv.2002.07590

Kaneria, A. V., Rao, A. B., Aithal, S. G., & Pai, S. N. (2021). Prediction of song popularity using machine learning concepts. In *Smart sensors measurements and instrumentation: Select proceedings of CISCON 2020* (pp. 35-48). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-16-0336-5_4

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE access*, 7, 117327-117345. https://doi.org/10.1109/ACCESS.2019.2936124

Lalitha, S., Geyasruti, D., Narayanan, R., & Shravani, M. (2015). Emotion detection using MFCC and cepstrum features. *Procedia Computer Science*, 70, 29-35. https://doi.org/10.1016/j.procs.2015.10.020

Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2, Article 14. https://doi.org/10.3389/fcomp.2020.00014

Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J. P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273, 271-280. https://doi.org/10.1016/j.neucom.2017.07.050

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS One*, 13(5), Article e0196391. https://doi.org/10.1371/journal.pone.0196391

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. *Intelligent systems with applications*, 20, Article 200266. https://doi.org/10.1016/j.iswa.2023.200266

Mashhadi, M. M. R., & Osei-Bonsu, K. (2023). Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *Plos One*, 18(11), Article e0291500. https://doi.org/10.1371/journal.pone.0291500

Milton, A., Roy, S. S., & Selvi, S. T. (2013). SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications*, 69(9), 34-39. https://doi.org/ 10.5120/11872-7667

Mustaqeem, & Kwon, S. (2020). CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. *Mathematics*, 8(12), Article 2133. https://doi.org/10.3390/math8122133

Nassif, A. B., Shahin, I., Elnagar, A., Velayudhan, D., Alhudhaif, A., & Polat, K. (2022). Emotional speaker identification using a novel capsule nets model. *Expert Systems with Applications*, 193, Article 116469. https://doi.org/10.1016/j.eswa.2021.116469

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1), 60-75. https://doi.org/10.1109/TAFFC.2017.2713783

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine*, 21(5), 95-106. https://doi.org/10.1109/MSP.2004.1328092

Patni, H., Jagtap, A., Bhoyar, V., & Gupta, A. (2021). Speech emotion recognition using MFCC, GFCC, chromagram and RMSE features. In *2021 8th international conference on signal processing and integrated networks (SPIN)* (pp. 892-897). IEEE. https://doi.org/10.1109/SPIN52536.2021.9566046

Pichora-Fuller, M. K., & Dupuis, K. (2020). *Toronto emotional speech set (TESS)*. https://doi.org/10.5683/SP2/E8H2MF

Poojary, N. N., Shivakumar, G. S., & Akshath, K. B. H. (2021). Speech emotion recognition using MLP classifier. *International Journal of Science Research in Computer Science, Engineering and Information Technology*, 7(4), 218-222. https://doi.org/10.32628/CSEIT217446

Rathor, S., Kansal, M., Verma, M., Garg, M., & Tiwari, R. (2021). Use of artificial intelligence in emotion recognition by ensemble based multilevel classification. *IOP Conference Series: Materials Science and Engineering*, 1116, Article 012196. https://doi.org/10.1088/1757-899x/1116/1/012196

Samantaray, A. K., Mahapatra, K., Kabi, B., & Routray, A. (2015). A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages. In *2015 IEEE 2nd international conference on recent trends in information systems (ReTIS)* (pp. 372-377). IEEE. https://doi.org/10.1109/ReTIS.2015.7232907

Singh, N., Khan, R. A., & Shree, R. (2012). MFCC and prosodic feature extraction techniques: a comparative study. *International Journal of Computer Applications*, 54(1), 9-13. https://doi.org/10.5120/8529-2061

Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, 115, 29-37. https://doi.org/10.1016/j.specom.2019.10.004

Wang, C., Ren, Y., Zhang, N., Cui, F., & Luo, S. (2022). Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimedia Tools and Applications*, 81(4), 4897-4907. https://doi.org/10.1007/s11042-021-10553-4

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9, 47795-47814. https://doi.org/10.1109/ACCESS.2021.3068045

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. (2020). Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. In *2020 6th international conference on wireless and telematics (ICWT)* (pp. 1-6). IEEE. http://doi.org/10.1109/ICWT50448.2020.9243622

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11), 1675-1685. https://doi.org/10.1109/TASLP.2019.2925934

Ying, Y., Tu, Y., & Zhou, H. (2021). Unsupervised feature learning for speech emotion recognition based on autoencoder. *Electronics*, 10(17), Article 2086. https://doi.org/10.3390/electronics10172086

Yuan, X., Wong, W. P., & Lam, C. T. (2022). Speech emotion recognition using multi-layer perceptron classifier. In *2022 IEEE 10th international conference on information, communication and networks (ICICN)* (pp. 644-648). IEEE. https://doi.org/10.1109/ICICN56848.2022.10006474