Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 48

Machine Learning and Deep Learning Techniques for Emotion Recognition from Human Speech using Acoustic Analysis

Anirban Sen
M. Tech
Department of CSE
MANIT
Bhopal, India
anirbansen0053@gmail.com

Dr. Meenu Chawla
Professor
Department of CSE
MANIT
Bhopal, India
meenuchawlamanit@gmail.com

Dr. Namita Tiwari
Assistant Professor
Department of CSE
MANIT
Bhopal, India
namita_tiwari21@rediffmail.com

Abstract—This work investigates the use of machine learning and deep learning techniques for tone recognition in audio data, with the aim of identifying emotions in human speech. Emotion recognition has numerous practical applications, such as in healthcare, human-computer interaction, and customer service. The paper discusses various approaches to emotion recognition including acoustic analysis and Machine Learning algorithms and proposes the best classifier for recognition. RAVDESS and TESS datasets are used in this research which consists of speech recordings labelled with seven emotions: anger, disgust, fear, happiness, sadness, pleasant surprise and neutral. The feature engineering techniques are also discussed thoroughly in this paper. The models experimented are SVM, Decision tree and Random Forest and the accuracy obtained was 84.1%, 88.9% and 93.24% respectively and the experimental results indicate the success of the proposed system. MLP has also been used which gave an accuracy of 93.34%. The paper concludes with a discussion of the potential applications of this technology and future research directions in the field of emotion recognition from audio.

Keywords—Emotion Recognition, MFCC, Mel Spectrogram, Ma-chine Learning, Deep Learning

I. Introduction

Emotion recognition is a fundamental aspect of human commu-nication, allowing us to understand the intentions and feelings of others and respond appropriately. The emotions exhibited by a person provide insights into their psychological state. The ability to recognize emotions in speech has numerous practical applications in fields such as healthcare for mental health diagnosis, human-computer interaction in Alexa, Siri etc, customer service to get an insight about feedback and in security fields to identify suspicious or abnormal behaviour based on changes in emotional states. A few researches have been done to automatically detect emotions from audio signals but there are a few obstacles:

1) Benchmark dataset covering all types of emotions with proper labels needs to be chosen.
2) Mult-lingual datasets are very less.
3) Feature selection is also very challenging.
4) It is necessary to have a dependable classifier and machine learning algorithm that is suitable for the task at hand.

The features extracted from the audio have a significant impact on how well an emotion identification system performs. However, there is no single sound feature that performs optimally in all sound signal processing tasks, and features must be customized to meet the specific demands of the problem being addressed. For this reason, researchers are using the deep learning model because it extracts features of its own [1].In this paper, Mel-Spectrogram and MFCC have been used as the features because they provide a good representation of the spectral content of speech, which can be useful in identifying emotional states [4]. Two baseline databases namely RAVDESS [13] and TESS [5] have been used. Experiments with three Machine learning models were done and Random Forest performed the best among them. Experiments with Multi-layer perceptron(MLP) have also been conducted and a slightly better result is observed.The paper is organized as follows in section 2, The pertinent studies on speech emotion are briefly summarized, and in section 3, the dataset that was used and an EDA is shown. The next section discusses the feature engineering aspects. In Section 5, introspections on the model used, methodology,

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

hyper-parameters and the number of iterations that were made are discussed followed by the results and comparative analysis. The final section outlines the conclusion and potential areas for future research

II. LITERATURE REVIEW

Over the last few decades, many Machine learning techniques and Deep learning techniques have been used to detect emotions from audio. Ye et al. Introduced a temporal Emotional Approach, which they called TIM-net [14] and has achieved the best accuracy of 0.92 on the RAVDESS dataset. Chamiska et al. [2] presents a new method for extracting features from conversational audio data called Bag-of-AudioWords (BoAW) based feature embeddings. The proposed approach also includes a state-of-the-art emotion detection model that uses a Recurrent Neural Network (RNN) to capture both the context of the conversation and the emotional state of the individual parties involved. The model is designed to make real-time predictions of categorical emotions based on the extracted features. They achieved an accuracy of 60% on the IEMOCAP dataset.
Anbalagan et al. proposed SVM algorithm with MFCC feature selection and achieved an accuracy of 89%. Jain et al. [4] introduced PCA on the TESS dataset and achieved an accuracy of 97.86%. Hason et al [11] proposed a feature extractor network based on CNNs and a classifier using a multi-layer perceptron (MLP) along with Mel-spectrogram, Tonnetz and spectrogram for acoustic feature extraction and achieved the best accuracy on Emo-DB Dataset with an accuracy of 92.79% Popova et al [10] used CNN VGG-16 as a classifier and achieved an accuracy of 71%.

III. DATASETS

In this paper, RAVDESS and TESS datasets have been used and both of them are in the English Language.

- RAVDESS-It stands for Ryerson Audio-Visual Database of Emotional Speech and Song and holds a collection of 7,356 files with a total size of 1.26 GB. It features recordings from 24 professional actors (12 female and 12 male) who spoke two sets of related phrases in a neutral North American accent. The labels of the dataset are neutral, happiness, sadness, anger, fear, surprise, and disgust. All of these recordings are available in three different modalities: only audio with 2 Byte size and 48KHz frequency, audio-video with 720p and 48KHz frequency and only video without sound. [8]
- TESS- It stands for Toronto emotional speech set and is also an open-source dataset and was created by researchers at the University of Toronto containing 2800 audio files, each of which is approximately 2-3 seconds long and it contains labels: neutral, happiness, sadness, anger, fear, surprise, and disgust [9]

In the final model, these 2 datasets have been merged and it is observed that there are 7688 records, containing recordings both of males and females. To ensure diversity in terms of speakers and context, two datasets were utilized since a single dataset was insufficient for predicting real-time audio data accurately. The Exploratory data analysis is shown in the fig.1. It can be seen from fig.1, that the merged dataset is almost label balanced.

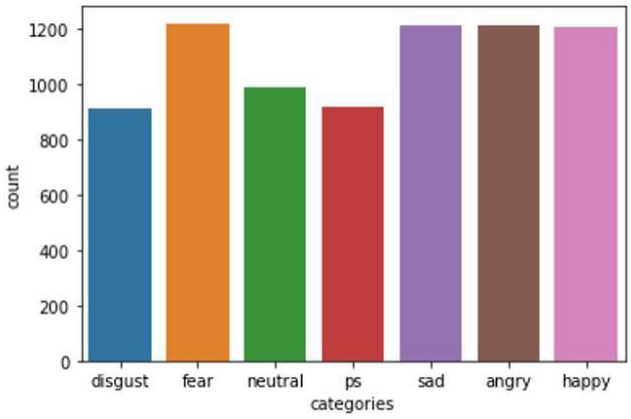

Fig. 1. Exploratory Data Analysis

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

IV. Feature selection

The Librosa package has been used to load the data. The purpose of using the Librosa package is that it converts the audio to mono-channel and preserves a uniform sampling rate of 22Khz. It also normalizes all the discrete values between +1 and -1. For these reasons, the calculations become easier and computations become faster. It is also noticed that there is an increase in accuracy for using the Librosa package instead of the sound file package.

There are three representations of audio features namely, time-domain representation, frequency-domain representation, and time-frequency representation. Among these, time-frequency representation is mostly used for Machine learning pur-poses. Some examples of this are Spectrogram, MFCC, Mel-spectrogram, etc. In this research, MFCC and Mel spectrogram has been used because they provide a good representation of the spectral content of speech. MFCCs extract information about the spectral envelope of the speech signal, while Mel spectrograms provide a visual representation of the spectral content of speech. Both can be used to extract relevant features for emotion recognition, and machine learning algorithms can be trained to identify patterns indicative of particular emotional states.

- MFCC- It stands for Mel-Frequency Cepstral Coefficients

  [6] and is widely used in audio signal processing for feature extraction. Through this technique, knowledge about the envelope is attained [6]. At first, the waveform in the time domain is converted to a log-Amplitude spectrum by applying DFT on the waveform. Then MEL scaling is done with the help of Mel filter banks and at last Discrete cosine transformation is applied to get the MFCCs. After a lot of empirical analysis, it was decided that the number of MFCCs will be 40. The formula for computing the Cepstrum is the equation 1

$$C(x(t)) = F^{-1}[\log(F(x(t)))] \quad (1)$$

  The above equation reveals that on applying the inverse Fourier transformation on the Spectrum, Cepstrum will be obtained and it is the building block of MFCC.

$$x(t) = e(t) * h(t) \quad (2)$$

$$X(t) = E(t).H(t) \quad (3)$$

$$\log(X(t)) = \log(E(t)) + \log(H(t)) \quad (4)$$

In equation 2, it is shown that Speech is generated through the interaction of the glottal pulse and the resonant frequencies of the vocal tract, resulting in the production of sound. In Equation 3 Fourier transformation has been applied and the logarithmic function is applied to get the final coefficients from Equation 4. In fig. 2, there are 40 MFCCs and the lower bands are the glottal pulse coefficient and contain more information about the glottal pulse whereas the higher bands are the vocal tract coefficients and contain more information about the vocal tract frequencies. The lowest band in fig. 2 shows that it contains the most information about the Glottal pulse.

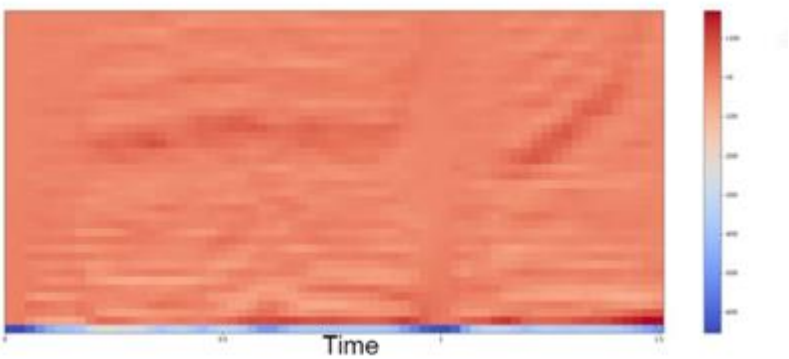

Fig. 2. Result of the MFCCs

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

- Mel-spectrogram- Normal spectrogram is linear in nature but the Human ear perceives sound in a logarithmic nature and that's where Mel-spectrogram is beneficial. Here, the term Mel comes from Melody and Mel scale is a frequency scale which is logarithmic in nature and is just like how the human ear perceives a sound. To create a Mel-spectrogram, the first step is to transform the frequency scale of the spectrogram from linear to Mel scale with the help of Mel- filter banks, which are evenly spaced filters that are triangular in shape [12]. In fig. 3, it can be seen that the differences between the Mel frequencies are same and the higher the Mel frequency larger the radius of the triangle which reveals that more frequencies in Hz are covered by higher Mel bands giving an essence of the logarithmic nature.

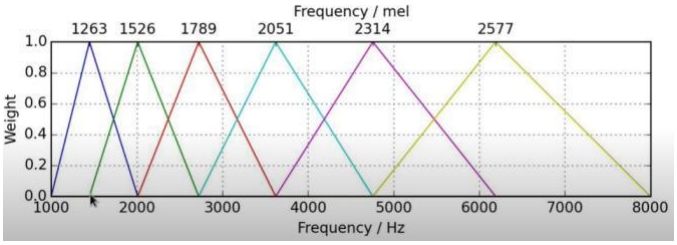

Fig. 3. Mel filter banks

The energy of a particular frequency range is captured by each filterbank output and these outputs are merged to generate the Mel-spectrogram. Further, to compress the dynamic range, Logarithm is applied on Mel-spectrogram to get a log-Mel-Spectrogram. The below formula has been used to find the Mel frequency.

$$m = 2595.\log(1 + \frac{f}{500}) \qquad (5)$$

40 Mel banks have been used in this research work and the below figure shows the Mel Spectrogram.
In fig. 4, the Mel spectrogram of an angry voice is shown and it gives a clear understanding of the time-frequency components. Mel spectrograms of different emotions are different from each other and are an excellent feature to train the model to detect emotions.

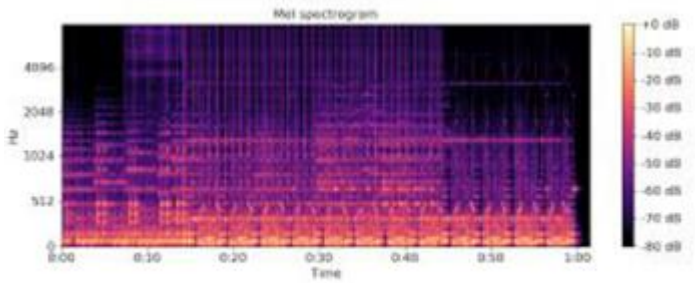

Fig. 4. Mel-Spectrogram

40 MFCCs and 40 Mel-bands are used resulting in 80 features in total. In the encoding technique, Label encoder has been used and experiments showed that using a Label encoder in place of One hot encoding, there is an increase in accuracy of 5% to 6% in most of the algorithms.

V. Proposed Methodology

In the first step, the Labelled dataset is acquired from Kaggle and both the datasets (RAVDESS and TESS) are merged. These are benchmark dataset and contains preprocessed data, therefore preprocessing is not needed. It has been seen that after

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

merging, the final dataset becomes almost label balance. Mel spectrogram and MFCC have been used as audio features. 40 features from the Mel spectrogram and 40 features from MFCCs have been used. 80% of the data has been used for training, while 20% has been allocated for testing.

After numerous experiments, it has been found that Random Forest is performing best in this problem and an accuracy of 93.24% was achieved. Random forest being an ensemble learning technique uses bagging under the hood and therefore it prevented the problem of overfitting and handled the categorical values well. The number of decision trees in the Random Forest has been kept at 80, which is equal to the number of features. Gini impurity as the criterion has been used in this work and since in Random Forest, random sampling is done and the base estimators are weak learners so pruning was not done. The max_feature hyperparameter was considered to be "sqrt". It is a hyperparameter that specifies the maximum number of features that can be considered at each split. A flow chart of the proposed methodology is shown in fig. 5.

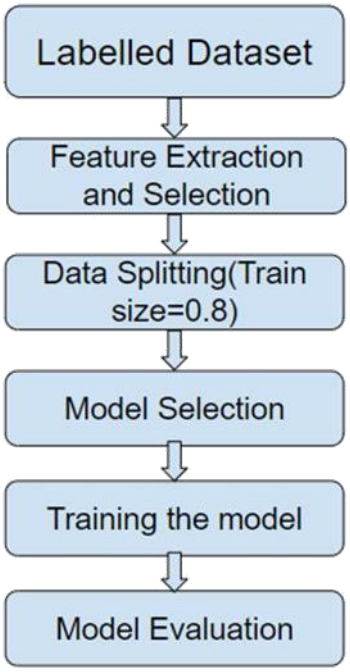

Fig. 5. Proposed Methodology

Each Decision tree in the random forest algorithm is trained on smaller datasets and fewer features are chosen randomly from the main dataset to determine the best feature to split on. The final prediction is then made by taking the majority voting algorithm.

Testing was also performed with an Artificial Neural network with 80 inputs, 2 hidden layers, 3 dropout layers and 7 output layers and the total trainable parameters were 11,647. To build the neural network model for the research work, the Rectified Linear Unit (ReLU) activation function has been used in all layers except for the output layer. The output layer, on the other hand, was activated using the Softmax function. Relu stands for Rectified Linear unit and is as described the equation 6.

$$f(x) = \max(0, x) \quad (6)$$

It has been used for its simplicity, effectiveness and solves the problem of the vanishing gradient. [14] The softmax function is used to obtain a probability distribution and therefore is best suited for categorical values. [14]. It is described in equation 7.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (7)$$

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

where z_i is the input to the i^{th} output neuron, and K is the total number of output neurons. The model summary is shown in fig. ??. A total of 100 epochs were done to test the result.

Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 80)	6480
dropout_3 (Dropout)	(None, 80)	0
dense_5 (Dense)	(None, 40)	3240
dropout_4 (Dropout)	(None, 40)	0
dense_6 (Dense)	(None, 40)	1640
dropout_5 (Dropout)	(None, 40)	0
dense_7 (Dense)	(None, 7)	287

Total params: 11,647
Trainable params: 11,647

Fig. 6. Model summary of the MLP

The "Early stopping" mechanism has been used for faster convergence.

VI. Results and Analysis

In this experiment, the effectiveness of traditional machine learning models such as Random Forest, Decision Tree, and SVM have been analysed. The GridSearchCV package has been utilized to perform hyperparameter tuning and to find the most appropriate model with optimal hyperparameters. Before testing these algorithms in the main merged dataset, analysis was also done in both of these datasets individually. Since the datasets were appended one after the other, so k fold cross-validation technique in the random Forest and SVM have been used to reduce the bias but it was found that the accuracy is getting reduced because of the average function.

In table I, it is evident that the performance of all the models deteriorates when tested on the RAVDESS dataset. This can be attributed to label imbalance in the dataset. The decision tree algorithm, in particular, exhibits a high variance and low bias, which indicates the presence of overfitting. In an attempt to address this issue, post-pruning and k-fold cross-validation techniques were employed, but they did not yield a significant improvement in accuracy. The best results are shown in the table after performing several iterations and tuning the hyperparameters. The SVM has the least accuracy due to the constraint optimization problem and Decision Tree was facing the problem of high Variance. Random Forest is an ensemble learning technique and uses weak learners, therefore, it reduces the bias. The number of trees has been kept at 80 after empirical analysis. Therefore, Random Forest is performing best in all the scenarios. The confusion matrix is shown in fig.7. A comparative analysis has been shown in Table 1.

TABLE I Results of the experiments

Author	Method	Dataset	Accuracy
Random Forest	n_estimator=80, criterion="Gini"	Ravdess Tess Ravdess+Tess	87.75 99.73 93.24
Decision Tree	n_estimator=80, criterion="Gini", Pruning=None	Ravdess Tess Ravdess+Tess	78.7 89.64 88.9
SVM	C=60, Kernel=Poly	Ravdess Tess Ravdess+Tess	83.2 99.64 84.1

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

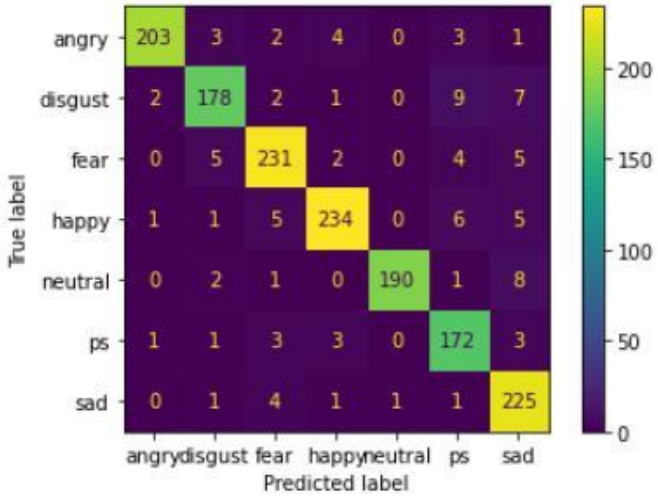

Fig. 7. Confusion Matrix of the Random Forest

Fig. 7 reveals that a majority of the labels were accurately classified. However, the emotions with the "disgust" and "neutral" labels exhibited a higher number of misclassifications. Specifically, out of all the misclassified audio samples, nine with the "disgust" label were incorrectly classified as "pleasant surprises," while eight with the "neutral" label were erroneously identified as "sad." Fig. 8 shows the accuracy graph of the ANN model and Fig. 9 shows the loss graph of the MLP model.

Fig. 8 is an accuracy graph and it is a chart that displays a model's performance over time or as a function of a specific parameter. The x-axis represents the independent variable, such as time or hyperparameter values, while the y-axis displays the

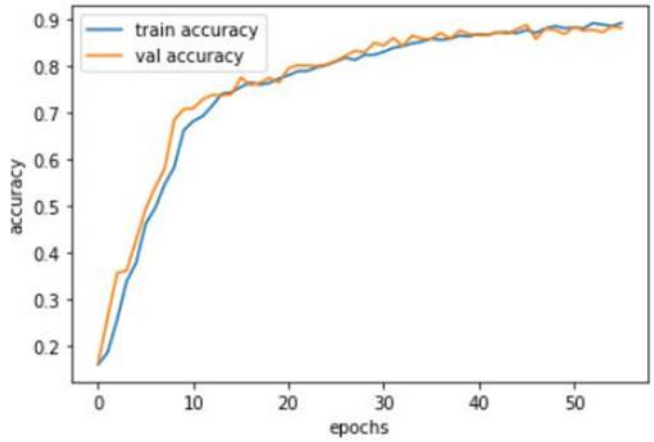

Fig. 8. Accuracy Graph of the MLP

accuracy score. This graph is valuable for understanding how a model's performance changes with different settings, and it can help identify trends and patterns in accuracy. Researchers and practitioners can use the accuracy graph to compare models and variations of the same model, evaluate and optimize machine learning models, and make informed decisions about model selection and parameter tuning. An accuracy of 93.34% was achieved. "Early Stopping" mechanism for faster convergence and in fig. 8 it can be seen that the algorithm converges at the 52nd iteration. The accuracy could have been better if a larger dataset had been used while training the Multi-layer perceptron. Fig. 9 shows a loss graph and it is a visual representation of a mathematical function called the loss function, which measures the difference between the predicted output and the actual

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

output of a machine learning model. The goal of training a model is to minimize this function by adjusting the model's weights and biases through an optimization process. The loss graph shows the value of the loss function over time or as a function of a specific parameter during the training process, with the y-axis representing the value of the loss function and the x-axis representing the number of iterations or epochs. As the training progresses, the values of the loss function should decrease, indicating improved predictions. A loss graph is a valuable tool for evaluating and optimizing machine learning models, monitoring the training process, and making informed decisions about model selection and parameter tuning. From fig. 9 it can be seen that after the 20th epoch, the value of the loss function gradually decreases and converges to a very small value. In the field of machine learning, the progress of a model's performance during training, spanning multiple epochs, is often visualized through an accuracy vs. epoch graph. An epoch is defined as a complete iteration through the entirety of the training dataset. Accuracy is a performance measure, expressed as a percentage, that quantifies the model's ability to accurately predict the target labels for the input data. It is a commonly used evaluation metric for classification tasks, where the objective is to classify input data into predefined categories. The accuracy vs. epoch graph serves as a graphical representation of the model's performance over the course of training. It provides insights into how the accuracy of the model evolves over time as it learns from the training data. Initially, the accuracy may be low as the model's parameters are being adjusted. However, with continued training, the accuracy may improve as the model's parameters are fine-tuned and it becomes more proficient at making accurate predictions. Monitoring the accuracy vs. epoch graph during training is a standard practice among researchers and practitioners to assess the effectiveness of the model's learning process and identify potential areas for improvement, such as adjusting hyperparameters, refining model architecture, or optimizing training strategies. This graphical representation serves as a valuable tool in analyzing and interpreting the performance of machine learning models and is commonly included in research papers as a means of presenting experimental results.

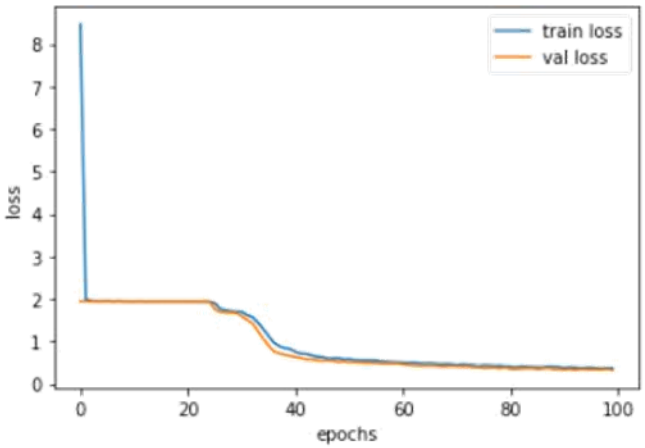

Fig. 9. Loss graph of the MLP

In the context of machine learning research, a loss graph is a graphical representation of the change in the value of the loss function during model training, typically with the value of the loss on the y-axis and the number of epochs on the x-axis. The loss function quantifies the discrepancy between the model's predicted outputs and the actual target outputs for the training data. The loss graph provides a visual depiction of how the loss evolves over the course of model training. A comparative analysis is shown in Table II. Since all the cited studies in this research have utilized the same dataset and employed identical audio features for emotion recognition from human speech, a comparative analysis can be conducted to evaluate the performance of various machine learning and deep learning techniques in this context. A bar graph for better visualization has been shown in fig. 10.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 7-Issue 1, January 2024
Paper: 48

TABLE II COMPARATIVE ANALYSIS

Author	Model	Dataset	Accuracy
Ye et al. [14]	TIM-net(Temporal-aware bI-direction Multi-scale Network)	Ravdess+Tess	92.08
Jain et al. [4]	SVM with PCA	TESS	97.86
Luna et al. [7]	pre-trained xlsr-Wav2Vec2.0 transformer	Ravdess Tess Ravdess+Tess	87 94 91
Dolka et al. [3]	ANN	Ravdess+Tess	88.72
Proposed Work	MLP	Ravdess Tess Ravdess+Tess	87.75 99.73 93.34

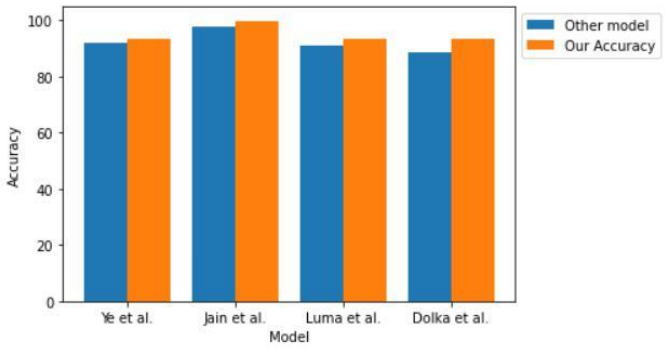

Fig. 10. Comparative Analysis

VII. CONCLUSION AND FUTURE WORK

Method to automatically detect emotion from audio has been conducted in this research and an accuracy of 93.24% has been achieved using Random Forest Classifier and 93.34% accuracy has been achieved with 3 layers neural network and has the best-achieved accuracy among all the previous works. Robust and accurate emotion recognizer can be a boon to society and has the potential to revolutionize various industries by providing insights into the emotional states of individuals and enhancing the quality of human-computer interaction, healthcare, education, marketing, entertainment and security. Although, future works can be extended to languages other than English and researchers can also explore other audio features that can help to improve the accuracy. Larger datasets are needed to deploy Deep neural network techniques. This work can also be incorporated with Sentiment analysis which can work in conjunction with text sentiment analysis and Emotion recognition of the speaker where both tone and context of the speaker can be gathered to analyse the Sentiment of the speaker.

REFERENCES

[1] Sudipta Bhattacharya, Samarjeet Borah, Brojo Kishore Mishra, and Atreyee Mondal. Emotion detection from multilingual audio using deep analysis. Multimedia Tools and Applications, 81(28):41309–41338, 2022.

[2] Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Damminda Ala-hakoon, Daswin De Silva, Naveen Chilamkurti, and Vishaka Nanayakkara. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 81(24):35173–35194, 2022.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 48

[3] Harshit Dolka, Arul Xavier VM, and Sujitha Juliet. Speech emotion recognition using ann on mfcc features. In 2021 3rd international conference on signal processing and communication (ICPSC), pages 431–435. IEEE, 2021.

[4] Kabir Jain, Anjali Chaturvedi, Jahnvi Dua, and Ramesh K Bhukya. In-vestigation using mlp-svm-pca classifiers on speech emotion recognition. In 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pages 1–6. IEEE, 2022.

[5] Kabir Jain, Anjali Chaturvedi, Jahnvi Dua, and Ramesh K Bhukya. In-vestigation using mlp-svm-pca classifiers on speech emotion recognition. In 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pages 1–6. IEEE, 2022.

[6] MS Likitha, Sri Raksha R Gupta, K Hasitha, and A Upendra Raju. Speech based human emotion recognition using mfcc. In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pages 2257–2260. IEEE, 2017.

[7] Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M Montero, and Fernando Fernández-Martínez. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. Applied Sciences, 12(1):327, 2021.

[8] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. PloS one, 14(8):e0221152, 2019.

[9] Ameya Ajit Mande, Sukrut Dani, Shruti Telang, and Zongru Shao. Emotion detection using audio data samples. International Journal of Advanced Research in Computer Science, 10(6), 2019.

[10] Anastasiya S Popova, Alexandr G Rassadin, and Alexander A Pono-marenko. Emotion recognition in sound. In Advances in Neural Computation, Machine Learning, and Cognitive Research: Selected Papers from the XIX International Conference on Neuroinformatics, October 2-6, 2017, Moscow, Russia 19, pages 117–124. Springer, 2018.

[11] David Hason Rudd, Huan Huo, and Guandong Xu. Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition. In Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II, pages 392–404. Springer, 2022.

[12] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. arXiv preprint arXiv:1912.10458, 2019.

[13] Jiaxin Ye, Xincheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. arXiv preprint arXiv:2211.08233, 2022.

[14] Jiaxin Ye, Xincheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. arXiv preprint arXiv:2211.08233, 2022.