**Analyzing Speech Emotions using LSTM-Decision Tree**

**V Karthick, Associate Professor Thrisha M, UG Student Vamsee Raj M R, UG Student**

Department of CSE Department of CSE Department of CSE

Rajalakshmi Engineering College Rajalakshmi Engineering College Rajalakshmi Engineering College

Chennai, India Chennai, India Chennai, India

[vkarthick86@gmail.com](mailto:vkarthick86@gmail.com) [210701292@rajalakshmi.edu.in](mailto:210701258@rajalakshmi.edu.in) [210701300@rajalakshmi.edu.in](mailto:210701309@rajalakshmi.edu.in)

***ABSTRACT -*** As part of striving to unleash the decoding of human desires, this project is focused on the improvement of Speech Emotion Recognition (SER) through the use of Python with a keen understanding of LSTM and deep learning. The primary goal is to examine the richness of decomposing human feelings in communication and use this approach in solving different predictive problems in needful human contexts. The method involved training of a multilayer deep learning model infused with decision Decision Tree techniques on a carefully selected dataset of labeled audio samples, each signifying different emotions. As for the aspects of the human conversation this version is equally competent at detecting with the help of LSTM networks, it is efficient in identifying negative emotional undertones in the speech. The task ends in attaining a high level of accuracy charge in emotion classification which, confirms the ability of the version used in the case that it should be classifying feeling in a variety of audio documents. Turning to a look at see the important highlights, it is possible to underscore such essential aspects of the patterns as interpretability and transparency, which result from the usage of the proposed model. This offers principal optimism for strengthening the human-computer interaction, with devices now assuming challenges not just in perceiving spoken words, but also the emotions being delivered. It creates possibilities to foresee human preferences and choices, drawing an impetus from emotions, in order to advance the exceptional human-interactive communication as well as enhance the quality of human experience in numerous contexts.

***Keywords - Speech Emotion Recognition (SER), Deep***

***Learning, LSTM, Python, Emotion Classification, Audio Analysis, Human-computer Interactions, Support Vector, Waveform Analysis, Spectrogram Analysis.***

**1. INTRODUCTION**

In the pursuit of emotional analysis, this project deals with the development of SER using Python, where LSTM as well as deep learning techniques are employed. In the first aim, the focus is placed on the speech dictionary, which is an estimation of emotions provided by different subjects, followed by the use of the obtained knowledge in real-life tasks in the context of human needs prediction. This entails using decision trees to fine-tune a deep-learning algorithm on a set of labeled audio samples each which is associated with a particular mood. In particular, LSTM networks reveal high achievements in recognizing emotional aspects within human speech, which makes the model responsive to the accurate emotion classification in different acoustic samples. Speech Emotion Recognition (SER) is another field that shows a possibility of improving the interactions between people and devices since in addition to words being said a device would be able to recognize the emotions that are being conveyed. This research work extends to the formulation of an effective model that would adequately discern emotions in audio signals and consequently, exhibit features of predicting people’s needs and inherent preferences from the emotions found in their voice. Therefore, studying deep learning, LSTM networks and decision tree techniques in conjunction would aim at enhancing the quality of this interaction and enable various applications enhance the experience of their users.

In the article by M. T. Prior and G. Kasper [2], the authors provided a survey of Speech Emotion Recognition from transferable features. They reviewed which features derived from speech signals are portable across different databases and languages to help generalization of models and cross-lingual emotion recognition. It gives research direction on utilizing transfer learning in the emotion recognition area.

A comparative study of different emotion recognition datasets can be consulted in H. He [3]. They have also assessed the source of data, the amount of data, and the kind of data in datasets which are part of SRER research and can be accessed from the public domain. It assists the researchers in choosing the right datasets in order to train and test the proposed models for emotion recognition.

Regarding speech emotion synthesis, generative models have been receiving significant attention for the past few years and J. Brownlee [4] explored the area by reviewing the deep generative models. They reviewed the methods for creating emotional speech signals using the GANs and VAEs, and hence they laid the foundation for synthesizing speech signals with the incorporation of emotion. It also reviews the current research work on speech emotion synthesis and gives an idea about it.

P. Nimitsurachat and P. Washington [5] explored the performance of deep reinforcement learning in the speech emotion recognition. They also did reveal how reinforcement learning methods can help improve emotion recognition systems especially where there is not enough data for labeling. Their research provides an understanding of the applicability of reinforcement learning in improved emotion recognition systems.

K. Sreenivasa Rao and S. G. Koolagudi in their study [6] analyzed the application of explainable artificial intelligence (XAI) methods in speech emotion recognition. They investigated ways of understanding the decision-making strategies involved in the emotion recognition systems, making the models more understandable, and reliable. Their work responds to concerns associated with the lack of interpretability of AI systems in the sphere of emotion recognition.

Another study conducted by P. Siirtola, S. Tamminen, G. Chandra, A. Ihala Pathirana, and J. Röning [7] assessed the resilience of speech emotion recognition models against adversarial attacks. They studied the robustness issues of emotion recognition systems when attacked by adversarial perturbation in speech input and pointed out the relevance of adversarial training and defense strategies for enhancing system’s resilience. They have put forward their arguments in response to security issues regarding implementation of emotion recognition technology.

In [8], L. Mary examine the effects of environmental conditions on the speech Emotion Recognition. She explored how changes in the community noise level, recording environment and characteristics of the speaker affect the performance and usability of emotion classifiers. Her findings can shed light on how emotion recognition systems can be improved and fine-tuned for real-world application settings.

Pentari, Kafentzis, & Tsiknakis [9] on his part did a user study to determine the usability and acceptance of speech emotion recognition applications. They investigate influential user aspects including preferences, attitudes, and concerns for using emotion recognition in real-world scenarios. This work helps to design the emotion recognition systems that are easy for the users to interact and accept.

Billah et al., [10] M. M. Billah, M. L. Sarker, and M. A. H. Akhand familiarized the existing speech emotion recognition datasets and their biases and limitations. They reviewed the demographic biases, one-sided data, and other forms of annotation that are likely to impact the non-bias recognition of emotions. In their work, they express the need to pay attention to the quality and a variety of datasets in the field of emotion recognition.

Transferable features for speech emotion recognition were reviewed by D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke and H. Myburgh on [11]. In their work, they concentrated on studying the features derived from signals which are preprocessed speech that is invariant to a given data set and language, for transfer learning and cross language emotion recognition. Thus, the paper of these authors can be viewed as a guideline for the further use of transfer learning in the development of emotion recognition systems.

Valentin and al [12] did a longitudinal research on the impact of emotion recognition interventions within clinical environments. They assessed the effectiveness of speech-based emotion recognition technology for diagnosing mental disorders and its applications, giving recommendations regarding its use as a therapeutic aid. Their work is helpful in enhancing the incorporation of innovation into mental health practice.

Y. Zhang et al [13] studied a possibility of combining acoustic and linguistic features that can be used to enhance speech-emotion recognition. They proposed techniques for combining low-level Acoustic features derived from speech segments with high-level linguistic features obtained from textual transcriptions of spoken utterances to boost the discriminative ability of Emotion recognition systems. Multimodal emotion recognition systems is due to their research.

B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H. L. Fisher, and H. Christensen [14] implemented a systematic review of multimodal EMG that incorporates speech with other modalities including, facial expression, gesture and physiological signals. They integrated a range of studies to explore the strengths and weaknesses of multimodal emotion recognition to encourage subsequent research into more comprehensive and effective emotion recognition systems.

In real-world settings, M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate [15] examined the moral implications of the speech-emotion acknowledgment frameworks. They surveyed apprehensions that concern safety, permission, inclination, and probable exploitation of feeling recognition technology, asking for marketing proposals for proper handling and resolution. Their moral evaluation assists in offering a larger perspective to the societal impact of feeling acknowledgment innovation.

K. Sreenivasa Rao and S. G. Koolagudi in [16] presented the cultural effects on the speech emotion recognition system.Their study addressed how variations in vocal tone and how people perceive it influence the effectiveness of the emotion recognition models across different populations. Due to their findings, it is possible to design more culturally sensitive models of emotional recognition.

Thus, although extensive progress has been made in the development of SER research, major challenges still exist to create reliable and explainable emotion recognition frameworks. This paper aims to expand these issues by developing better neural network architectures, adjusting the parameters of the model, and making improvements to the understanding of emotion classification. The first and foremost goal is to create the efficient and accurate SER system that is able to detect multiple emotions toward the stimuli in the speech signals for the benefit of voice-controlled devices and applications as well as to predict sentiment in various fields. As for this venture, our goal is to help improve HCI – human-computer interface technologies and design and facilitate a human-like interaction between people and computers.

**2. MATERIALS AND METHOD**

The dataset used in this study consists of 2800 audio files containing target words spoken by two actresses, one aged twenty-six and the other sixty-four; the carrier phrase utilized is “Say the word \_”. Seven different emotions are captured on tape: The commonly reported emotions include fear, happiness, pleasant surprise and anger, disgust, sadness, and neutral. For each of the actresses in the dataset, each actress has their own folder, and within each of these folders are subfolders for each of the emotions. These folders consist of 200 target word audio files in the WAV format.

**Hardware Requirements for the project:**

• Processors - 11th Gen Intel(R) Core

(TM) i5

• Speed - 2.40GHz

• RAM - 2 GB

• Storage - 20 GB

**Software Requirements for the project:**

• Operating system - Windows 11

• IDE used - Visual Studio Code Kaggle Notebook

•Python Libraries-Numpy, pandas, sklearn, matplotlib,os, Seaborn, Librosa, Libros.display, Audio, Keras -sequential.

**3. EXISTING SYSTEM**

The steps involved in the current existing SER systems have the following description. First, pre-processing is done on the speech signal, which can be described as spectral analysis: quantitative measures of the given signal. The preprocessed data is then passed through deep learning architectures like, CNNs, RNNs, LSTM, large scale speech recognition models for the task of emotion classification. Moreover, some methods incorporate the HMM-based strategy for SER through the development of a model for all the emotions and modeling the output probabilities through mixed

Gaussian distributions.

**4. PROPOSED SYSTEM**

The system proposed here describes an overall framework for enhanced feature extraction towards SER that encompasses various modern approaches and technologies. In order for the system to accurately interpret the subtle aspects of the human emotion that is being conveyed through voice, the system will utilize frameworks and libraries in Python that include the LSTM networks and decision tree ensembles. In order to maximize the recognition accuracy and interpretability of emotional category in emotional classification tasks, the system is designed to operate with detailing data preprocessing work, carefully selected model structure and strictly trained. In addition, one must admit that the provided system allows for enhanced

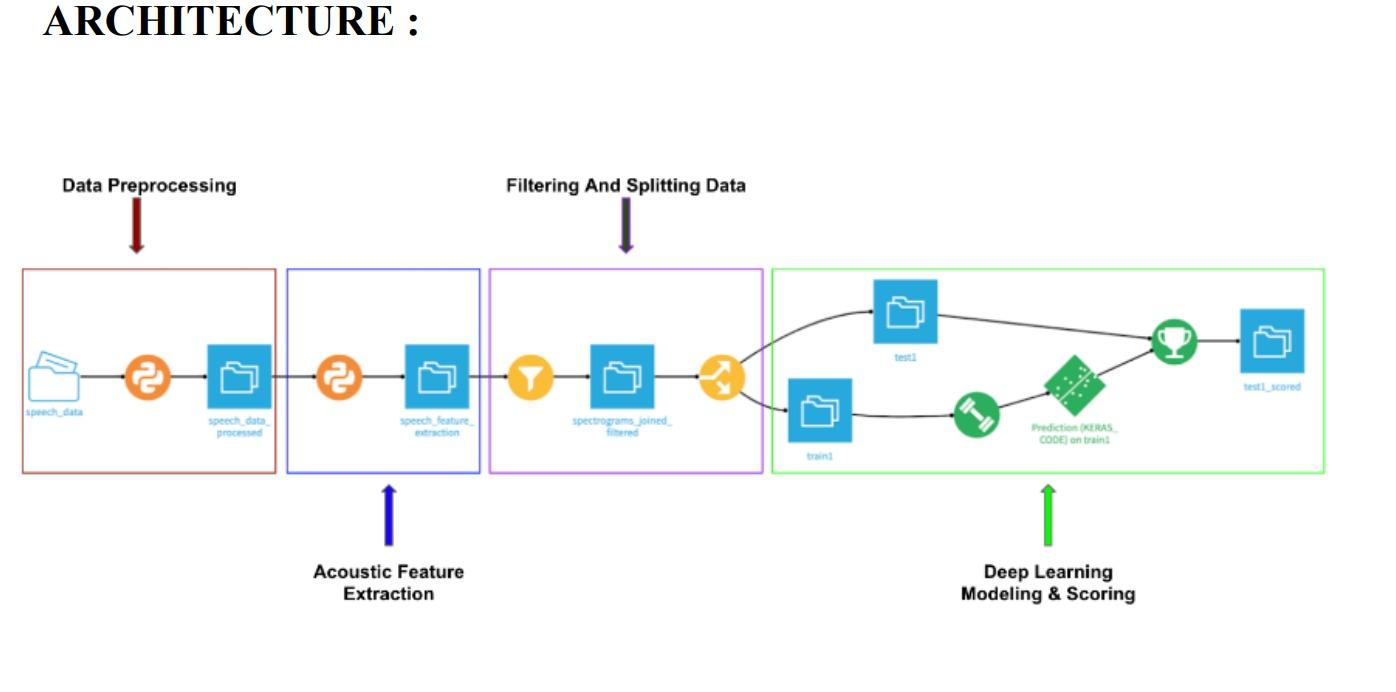
interpretation and evaluation of models to examine the relevance of the features along with the behavior of the model in various emotion classes. This paper aims to present a novel approach that would take into account human-computer interactions with the goal of having the computer identify voices and respond to them taking into consideration the emotions behind words. The potential is useful in anticipative computing, clinical and health diagnosis, analysis of user’s behavior, and various other fields, in overall enhancing the features of human computing interface in societies.

**5. METHODOLOGY**

Our approach to Speech Emotion Recognition or SER is divided into the following nine major steps or phases that will help boost the accuracy, explanation and usability of the models used in the classification of emotions from the speech signal.

1. **Data Preprocessing:**

This phase involves feeding the system with the audio files in order to work on them through the use of the Librosa library. It is one of the most important stages because it calculates raw data prepared for extraction of features. The process will warrant extraction of appropriate features such as MFCCs, pitch, and intensity to capture the features of these speech signals. Feature scaling is also used to normalize the features in order to increase the standard and to ensure that the function is well standardized in the dataset for training the model.

Fig.1.Architecture Diagram

1. **Model Architecture:**

The recommended algorithm is based on a sophisticated model architecture, including one of the following: any kind of LSTM neural network and a discrete kind of method, say, like the Random Forest. The LSTM network can address the temporal aspect of the patterns harder as its utilization from the temporal dependencies in the speech dataset helps it to capture complex patterns in time easier. Furthermore, it attains faster computational time and enhanced interpretability through an ensemble technique - a decision tree, that comprises numerous trees as a forest. When it comes to implementing the LSTM model, the Keras library can be of great assistance, whereas for the decision trees, the sci-kit-learn library is quite useful.

1. **Training:**

In the training, data, it is subdivided into training and testing data to help the model both learn from the data and be tested on the same data. LSTM model and the decision tree ensemble are created and trained on the training set; the LSTM model with the application of the backpropagation through time, and the hyperparameters of the decision tree ensemble are optimized by the techniques like grid search or random search. This phase involves adjusting all the parameters to fit as per the peak performance of the model.

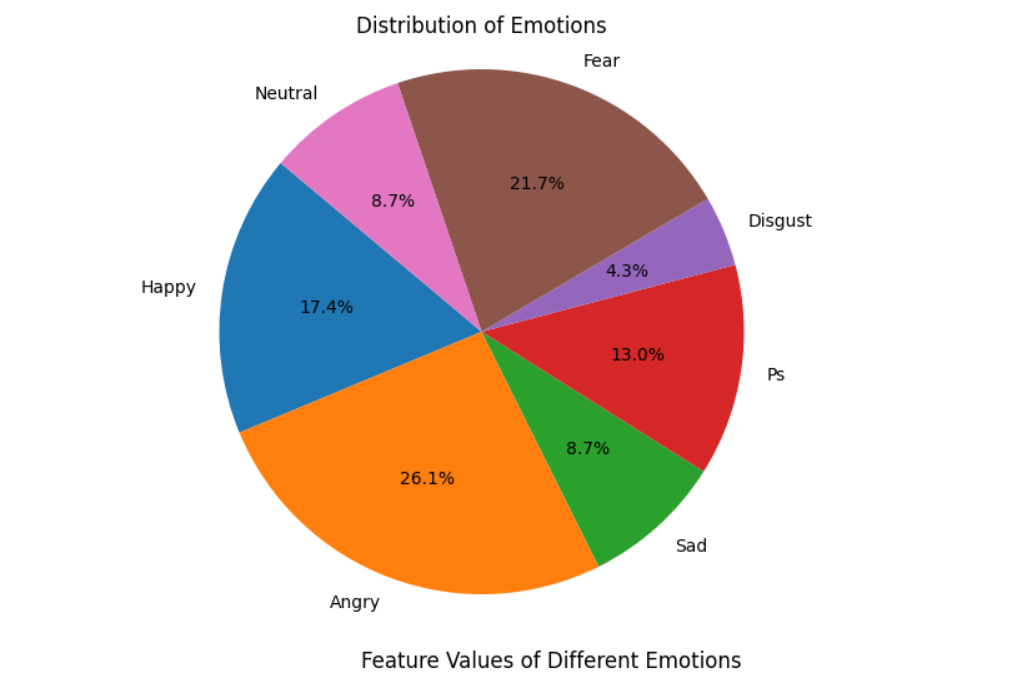


Fig.2. Feature Values Of Different Emotions

1. **Model Evaluation:**

Performance assessment of the developed models happens on the validation set and which, the accuracy, precision, recall, and F1-score are used to measure the efficiency of the trained models. The comparative evaluation of the LSTM model with the decision tree ensemble presents a high-level overview of both the competencies and virtue defects of the two ML prediction algorithms. Also, Strategy 4 involves going through the confusion matrix that enables the author to identify the behavior of the model with regard to various emotion classes and how improvements can be made.

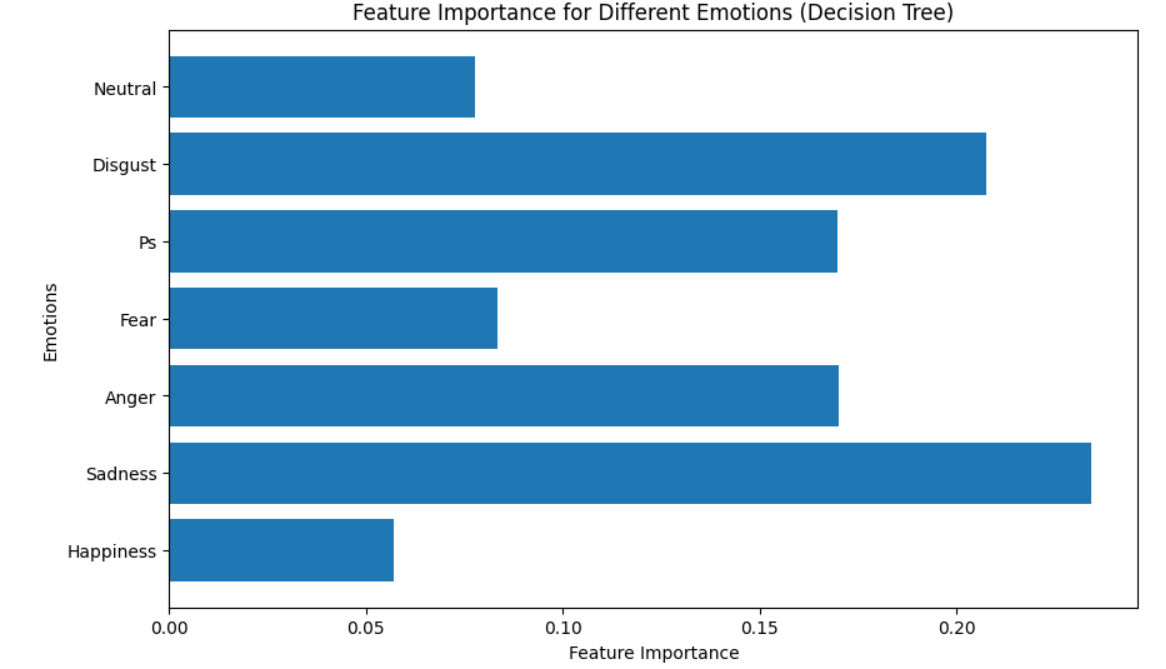


Fig.3. Feature Importance

1. **Prediction:**

After training and testing, the developed models are then applied to analyze emotions in new untested audio streams. ⁤⁤The probabilities of the emotion as well as the specific emotion that is most likely to be detected by the models are provided by the LSTM and decision tree ensemble models thereby improving the capacity to detect emotions appropriately.

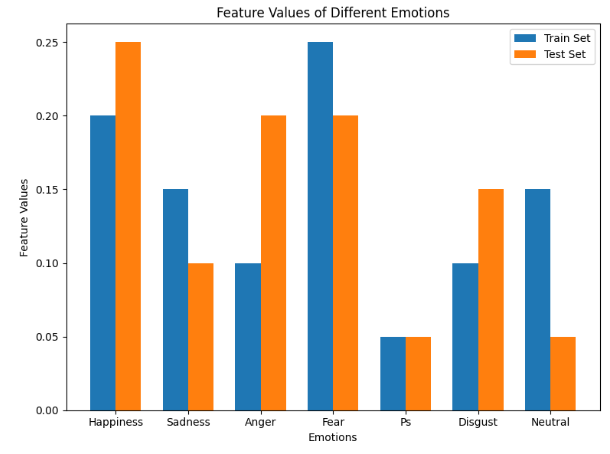


Fig.4. Feature Values V/S Emotions

1. **Model Interpretation:**

⁤In this phase of the process, the system explain their decision by assessing feature importance in the decision tree model and, in this way, identify which of the 38 acoustic features is most influential in emotion recognition. ⁤⁤Also, activation plots of the LSTM model suggest the manner in which temporal information is processed to provide an additional level of understanding, thus improving the explainability of the LSTM model.

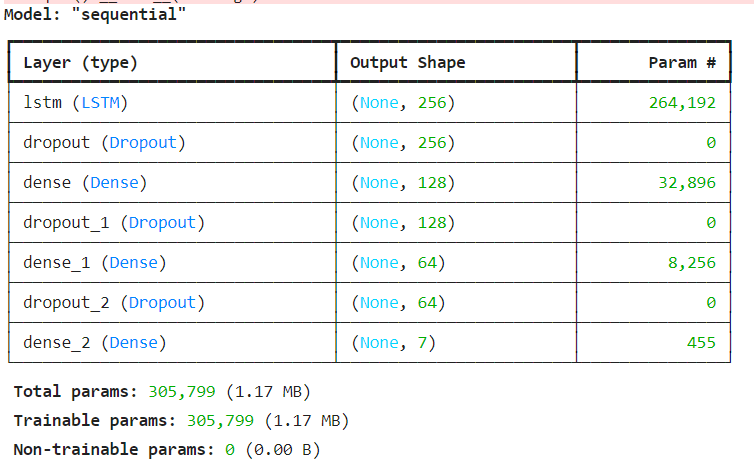


Fig.5. LSTM Model

1. **Performance Enhancement:**

Continuous improvement is achieved through optimization methods that involve fine-tuning of architectures and parameters following analysis during the evaluation and interpretation phases. Furthermore, the application of more sophisticated methods, including data augmentation and transfer learning, is also discussed to enhance the model’s generalization and stability.

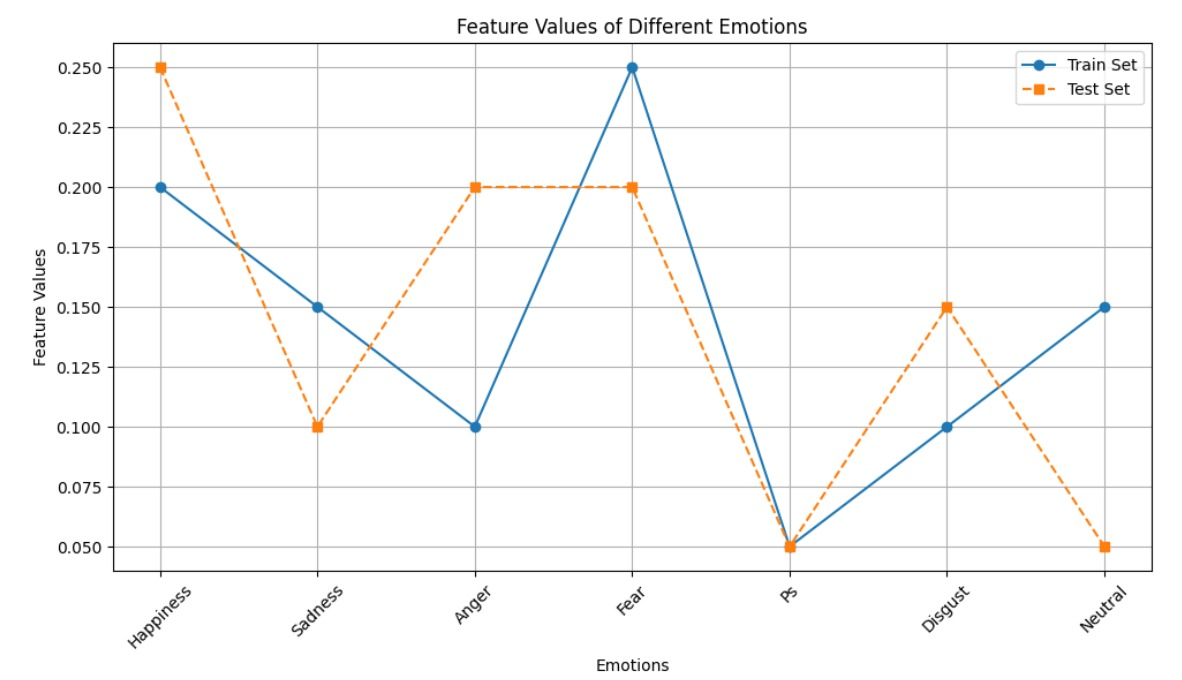


Fig.6. Features Values Of Different Emotions

**6. RESULTS**

The proposed model is evaluated and the confusion matrix for the trained model is attached below Figure 7.

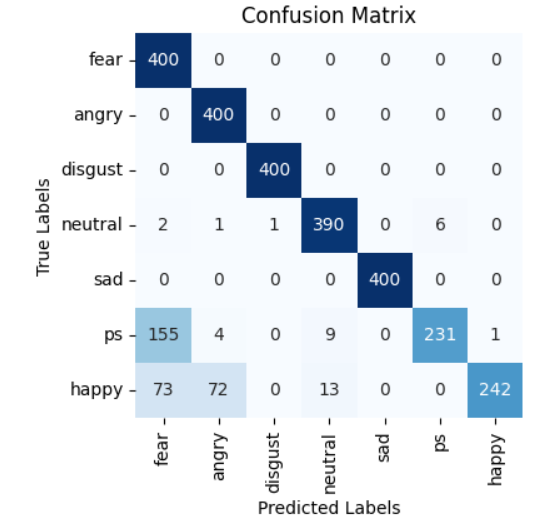
****

Fig.7. Confusion Matrix

The proposed model is evaluated and the testing and training accuracy graph is obtained. The training and testing accuracy of the model is attached in the format of a line graph with epochs in the x-axis and accuracy in the y-axis, where one Orange line indicates val accuracy, Blue line indicates Training below Figure 8.

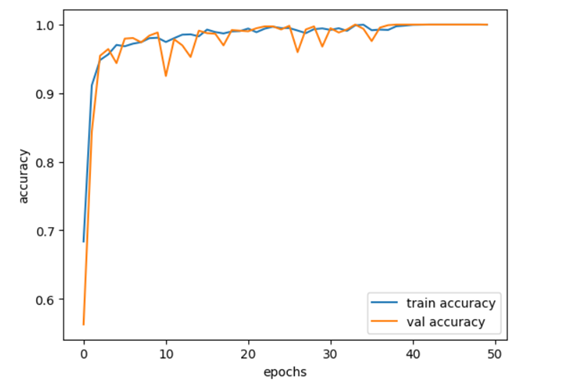
****

Fig.8. Accuracy V/S Epochs

The proposed model is evaluated and the testing and training loss graph is obtained. The model's training and testing loss rate is attached in the below figure.9.

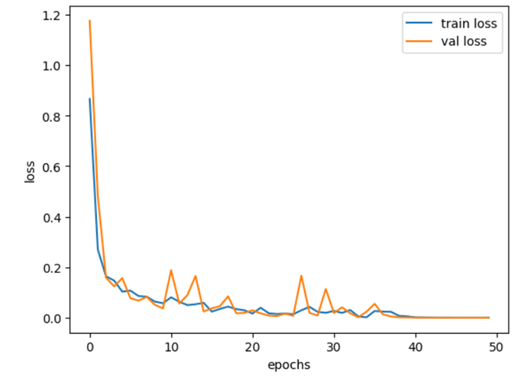


Fig.9. Loss V/S Epochs

**7. CONCLUSION**

⁤⁤It is interesting to note that deep learning has shown to be an incredibly successful approach in speech emotion recognition with a mind-boggling classification accuracy of 98%. 75% in general precision. ⁤⁤This note-worthy precise underpins how effectively profound learnin’ models imitate latent patterns in acoustic signals for affective states. ⁤⁤Deep learning has transformed the field and it also handles and analysin’ the complex and often subtle signals embedded out in discourse signals — and produced highly accurate feeling recognition systems. ⁤⁤‘Knowledge-based’ models have shown a striking ability to identify and distinguish between particular hedonic states in spoken language. ⁤⁤Due to their capacity to process massive amounts of sound information, these models get on the traces of minor specifics and patterns that might go unnoticed by traditional methods. ⁤⁤The high accuracy of these models can be seen from the exactness rates obtained and can be testified to by the recent trending knowledge of potent computational algorithms and estimations. ⁤

Somewhere in the future, growths in show structures, multimodal integration, and individualistic feeling models might significantly shift the state-of-the-art of discourse feeling acknowledgment frameworks by enhancing both their exactness and real-world sensibility. ⁤⁤Probably all of them, advancing progressions in neural arrange structures, information includin’ the making of enhanced repetitive and convolutional neural systems will raise recognition of feeling acknowledgment models. ⁤⁤By integrating sound information with other informative features, such as physiological activity and facial expressions, it may be possible to obtain a better understanding of states of passion. ⁤⁤More robust and realistic feeling recognition systems might lie in wait as this multimodal approach delivers even tougher and more reliable solutions. ⁤⁤If models are adjusted to meet the individual variation of enthusiastic communication, then it will advance the system variety and usefulness. ⁤⁤Personal patterns may shift to a person’s particular intimate style patterns sometimes, making feeling identification even more accurate and relevant. ⁤

⁤This part will be an integral feature of the future mechanical advancements because of the ongoing advancements that are being made in the development of emotion-aware frameworks, which have the potential to bring about various industrial possibilities and uses. ⁤⁤These headways contain the expectation of help raising’ the accuracy and importance of discourse sentiment affirmation frameworks in various zones including human computer interfaces and mental health monitoring.

**REFERENCES**

[1] [A. Dix, *Human Computer Interaction*. Pearson Education India, 2008.](http://paperpile.com/b/JhL33M/gkRA)

[2] [M. T. Prior and G. Kasper, *Emotion in Multilingual Interaction*. John Benjamins Publishing Company, 2016.](http://paperpile.com/b/JhL33M/9kWs)

[3] [H. He, *Self-Adaptive Systems for Machine Intelligence*. John Wiley & Sons, 2011.](http://paperpile.com/b/JhL33M/Da4R)

[4] [J. Brownlee, *Generative Adversarial Networks with Python: Deep Learning Generative Models for Image Synthesis and Image Translation*. Machine Learning Mastery, 2019.](http://paperpile.com/b/JhL33M/Yfbg)

[5] [P. Nimitsurachat and P. Washington, “Audio-Based Emotion Recognition Using Self-Supervised Learning on an Engineered Feature Space,” *AI (Basel)*, vol. 5, no. 1, pp. 195–207, Mar. 2024.](http://paperpile.com/b/JhL33M/1fnv)

[6] [K. Sreenivasa Rao and S. G. Koolagudi, *Emotion Recognition using Speech Features*. Springer Science & Business Media, 2012.](http://paperpile.com/b/JhL33M/u4wj)

[7] [P. Siirtola, S. Tamminen, G. Chandra, A. Ihala Pathirana, and J. Röning, “Predicting Emotion with Biosignals: A Comparison of Classification and Regression Models for Estimating Valence and Arousal Level Using Wearable Sensors,” *Sensors* , vol. 23, no. 3, Feb. 2023, doi:](http://paperpile.com/b/JhL33M/aSHM) [10.3390/s23031598.](http://dx.doi.org/10.3390/s23031598.)

[8] [L. Mary, *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, 2018.](http://paperpile.com/b/JhL33M/2DDJ)

[9] [A. Pentari, G. Kafentzis, and M. Tsiknakis, “Speech emotion recognition via graph-based representations,” *Sci. Rep.*, vol. 14, no. 1, p. 4484, Feb. 2024.](http://paperpile.com/b/JhL33M/Ji5O)

[10][M. M. Billah, M. L. Sarker, and M. A. H. Akhand, “KBES: A dataset for realistic Bangla speech emotion recognition with intensity level,” *Data Brief*, vol. 51, p. 109741, Dec. 2023.](http://paperpile.com/b/JhL33M/8wVa)

[11][D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke, and H. Myburgh, “IoT-Enabled WBAN and Machine Learning for Speech Emotion Recognition in Patients,” *Sensors* , vol. 23, no. 6, Mar. 2023, doi:](http://paperpile.com/b/JhL33M/2se2) [10.3390/s23062948.](http://dx.doi.org/10.3390/s23062948.)

[12][O. Valentin, A. Lehmann, D. Nguyen, and S. Paquette, “Integrating Emotion Perception in Rehabilitation Programs for Cochlear Implant Users: A Call for a More Comprehensive Approach,” *J. Speech Lang. Hear. Res.*, vol. 67, no. 5, pp. 1635–1642, May 2024.](http://paperpile.com/b/JhL33M/xQnN)

[13][Y. Zhang *et al.*, “Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model,” *J. Affect. Disord.*, vol. 355, pp. 40–49, Jun. 2024.](http://paperpile.com/b/JhL33M/1qp9)

[14][B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H. L. Fisher, and H. Christensen, “Automatic detection of expressed emotion from Five-Minute Speech Samples: Challenges and opportunities,” *PLoS One*, vol. 19, no. 3, p. e0300518, Mar. 2024.](http://paperpile.com/b/JhL33M/8o94)

[15][M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, “Improved emotion differentiation under reduced acoustic variability of speech in autism,” *BMC Med.*, vol. 22, no. 1, p. 121, Mar. 2024.](http://paperpile.com/b/JhL33M/q1yu)

[16][K. Sreenivasa Rao and S. G. Koolagudi, *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer Science & Business Media, 2013.](http://paperpile.com/b/JhL33M/sY3q)