EMODEPICTOR : HARNESSING EMOTIONS FROM AUDIO

TARUN KUNKUNURI, SOUMYA SHANIGARAPU,TEJA VINEETH REDDY YERAMAREDDY

ABSTRACT :

In the realm of Speech Emotion Recognition (SER), our pursuit revolves around the automatic classification and understanding of emotional nuances embedded in audio signals. With applications spanning customer service, education, and entertainment, the crux of our objective is to cultivate intelligent systems adept at perceiving and responding to human emotions in spoken language. The imperative nature of SER becomes evident in its potential to elevate customer service experiences through tailored responses based on detected emotions, thereby fostering personalized and empathetic interactions. In the educational landscape, SER emerges as a pivotal tool for gauging student engagement and tailoring teaching methods to create adaptive and effective learning experiences..

Our approach involves a comprehensive exploration of diverse models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Support Vector Machines (SVMs). Each model type contributes unique strengths, with CNNs specializing in spatial feature extraction, RNNs adept at capturing temporal dependencies, and SVMs providing a robust classification framework. The distinctive facet of our work lies in the integration of these models into a cohesive ensemble approach, strategically designed to leverage their complementary strengths. The focal point is to discern the most effective model for the nuanced task of emotional recognition in speech.

Drawing upon the rich diversity of the RAVDESS dataset, our project emphasizes structured emotional categories, a wide spectrum of emotions, and a purposeful design catering to real-world applicability. Our methodology traverses data exploration, preprocessing, feature extraction, and model development, culminating in a rigorous evaluation and comparison of CNN, RNN, and SVM models. The culmination of our efforts aims to unravel insights into the most effective approach for emotion classification, contributing significantly to the ongoing evolution of SER research.

1. INTRODUCTION

The primary challenge addressed in this project is Speech Emotion Recognition (SER), where the goal is to automatically classify and understand the emotional content conveyed through audio signals. This problem is pivotal in applications spanning customer service, education, and entertainment, as it facilitates the development of intelligent systems capable of perceiving and responding to human emotions in spoken language.

Importance of Addressing SER:

Customer Service Enhancement:

The automated recognition of customer emotions during interactions is paramount for enhancing customer service experiences. This capability allows systems to tailor responses based on detected emotions, leading to more personalized and empathetic interactions and ultimately boosting user satisfaction.

Educational Impact:

SER plays a crucial role in the educational landscape by providing a means to gauge student engagement and tailor teaching methods accordingly. Understanding the emotional states of students enables the creation of adaptive and effective learning experiences.

The field of SER has seen various solutions, each leveraging different models and techniques. Previous studies have explored Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Support Vector Machines (SVMs) for emotion classification in audio data. CNNs have demonstrated efficacy in capturing spatial patterns within audio spectrograms, while RNNs excel in modeling temporal dependencies in sequential data. SVMs, on the other hand, offer a robust and interpretable approach to classification tasks. Each of these models has contributed to the understanding and advancement of SER, yet challenges persist in achieving a holistic and robust solution that considers both spatial and temporal features in audio signals.

In our endeavor to address the SER challenge, we have employed a multi-model approach utilizing CNN, RNN, and SVM architectures. Each model is selected for its specific strengths: CNNs for spatial feature extraction, RNNs for capturing temporal dependencies, and SVMs for robust classification. Our unique contribution lies in the integration of these diverse models to create an ensemble approach, aiming to leverage the complementary strengths of each model type. By combining these models, we intend to identify the most effective model for our application. This rigorous comparison will contribute to informed decision-making regarding the selection of the model that best addresses the nuances of emotional recognition in speech.

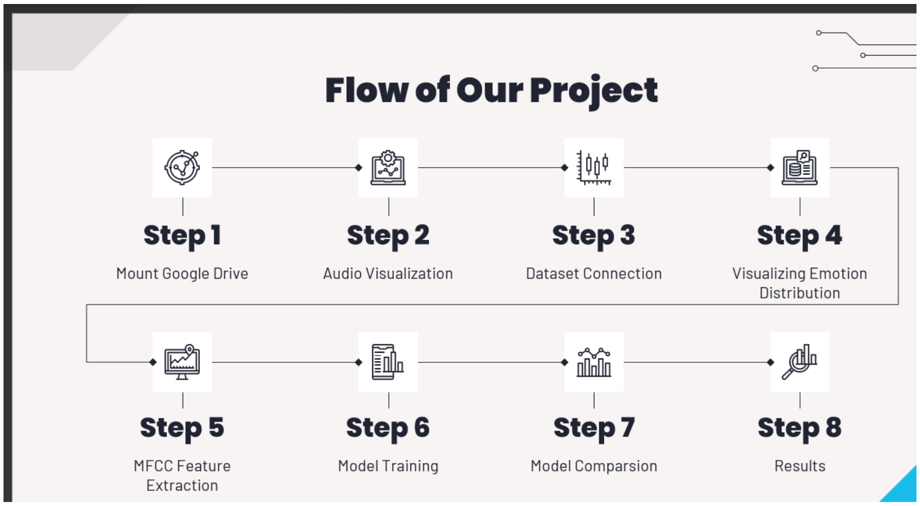
2.DATA DESCRIPTION:

For this project, We have selected the RAVDESS dataset, available on Kaggle [link: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio/]. This dataset serves as a cornerstone in the realm of emotional speech and song analysis, boasting over 1,400 audio clips generously contributed by 24 professional actors. The diversity encapsulated in this dataset, featuring emotions ranging from neutral and calm to happy, sad, angry, fearful, disgusted, and surprised, provides an extensive foundation for training and testing emotion recognition models.

Each audio file within the RAVDESS dataset is meticulously identified by a unique 7-part numerical code, offering detailed insights into the conveyed emotion. The dataset's structured emotional categories, complemented by key features such as pitch, intensity, and spectrogram data, are pivotal for the development and training of models specifically geared towards emotion recognition in speech.

The dataset's structure, featuring 3-second audio clips with varying emotional intensities, vocal channels, and repetitions, is strategically designed to enhance its applicability in real-world scenarios during model training. In essence, the RAVDESS dataset emerges as a robust and diverse collection, positioning itself as an invaluable asset for the progression of research in speech emotion recognition within the scope of this project.

3.FLOWCHART OF OUR PROJECT:



Flow of Our Project

Step 1
Mount Google Drive

Step 2
Audio Visualization

Step 3
Dataset Connection

Step 4
Visualizing Emotion Distribution

Step 5
MFCC Feature Extraction

Step 6
Model Training

Step 7
Model Comparsion

Step 8
Results

FIG 1.1 IMAGE SHOWING THE FLOWCGART OF OUR PROJECT

4.METHODOLOGY FOR EMOTION CLASSIFICATION FROM AUDIO DATA

Our project centers on the complex task of emotion classification from audio data. The primary goal is to develop robust models capable of accurately discerning emotions expressed in diverse audio sources.

4.1DATA EXPLORATION

To gain insights into the dataset, we conducted an initial exploration. Basic statistical analyses and visualizations were performed to understand the distribution of emotions and to identify potential challenges in the data. This phase allowed us to make informed decisions about feature selection and model design.

4.2DATA PREPROCESSING

Our data preprocessing endeavors for audio emotion classification revolved around ensuring data integrity, exploring audio characteristics, and creating a structured dataset for analysis. Here's a detailed breakdown:

IN-DEPTH AUDIO INSPECTION:

Explored a representative audio file (03-01-08-01-02-01-24.wav) using Librosa, gaining insights into its composition.

Employed the Audio function to audibly preview the file, providing a qualitative assessment of its contents.

LIBROSA LIBRARY INSTALLATION:

Ensured availability of the Librosa library for effective audio processing, installing it as needed.

AUDIO WAVEFORM VISUALIZATION:

Loaded the audio file (03-01-08-01-02-01-24.wav) using Librosa, generating a visually informative waveform plot.

Utilized Matplotlib to illustrate amplitude variations over time, enhancing interpretability.

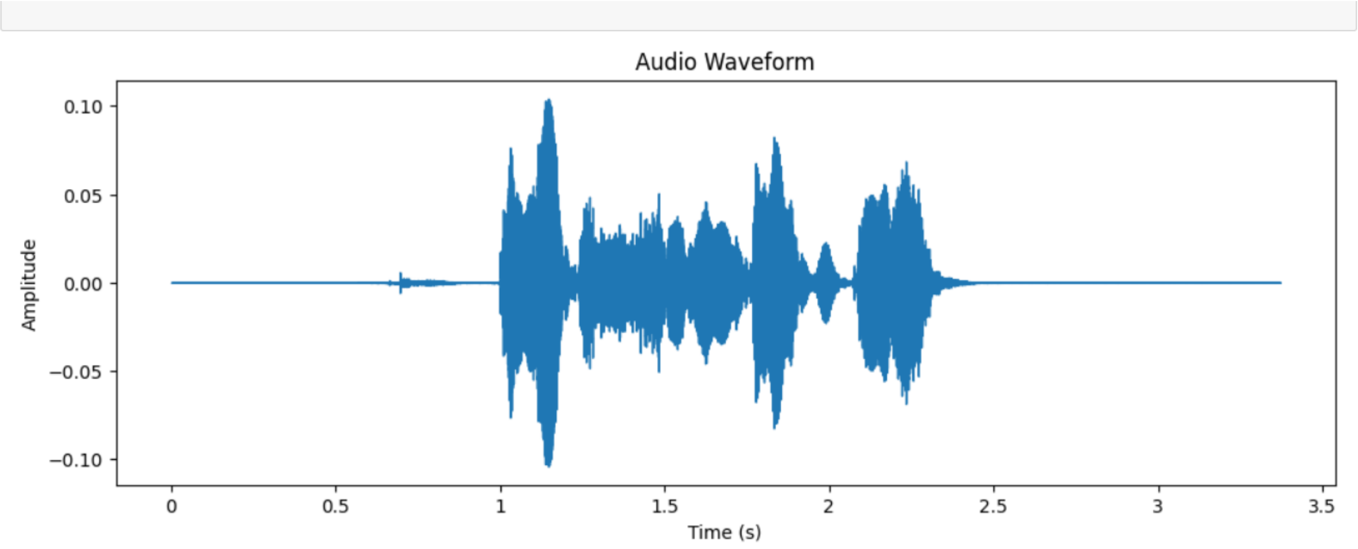


Fig 1.2 Image showing the waveform of one audiofile sample

DATAFRAME CREATION FROM AUDIO FILES:

Systematically processed audio files in the designated directory to extract crucial information.

Extracted emotional labels and file paths, creating two separate lists (file_emotion and file_path_list).

Formed two DataFrames (emotion_df and path_df) to capture emotion labels and file paths separately, subsequently concatenated into a comprehensive dataset (dataset_df).

EMOTION DISTRIBUTION VISUALIZATION:

Mapped integer-coded emotions to corresponding labels for clarity.

Employed Matplotlib to craft a bar chart illustrating the distribution of emotions within the dataset, with distinct colors enhancing visual understanding.

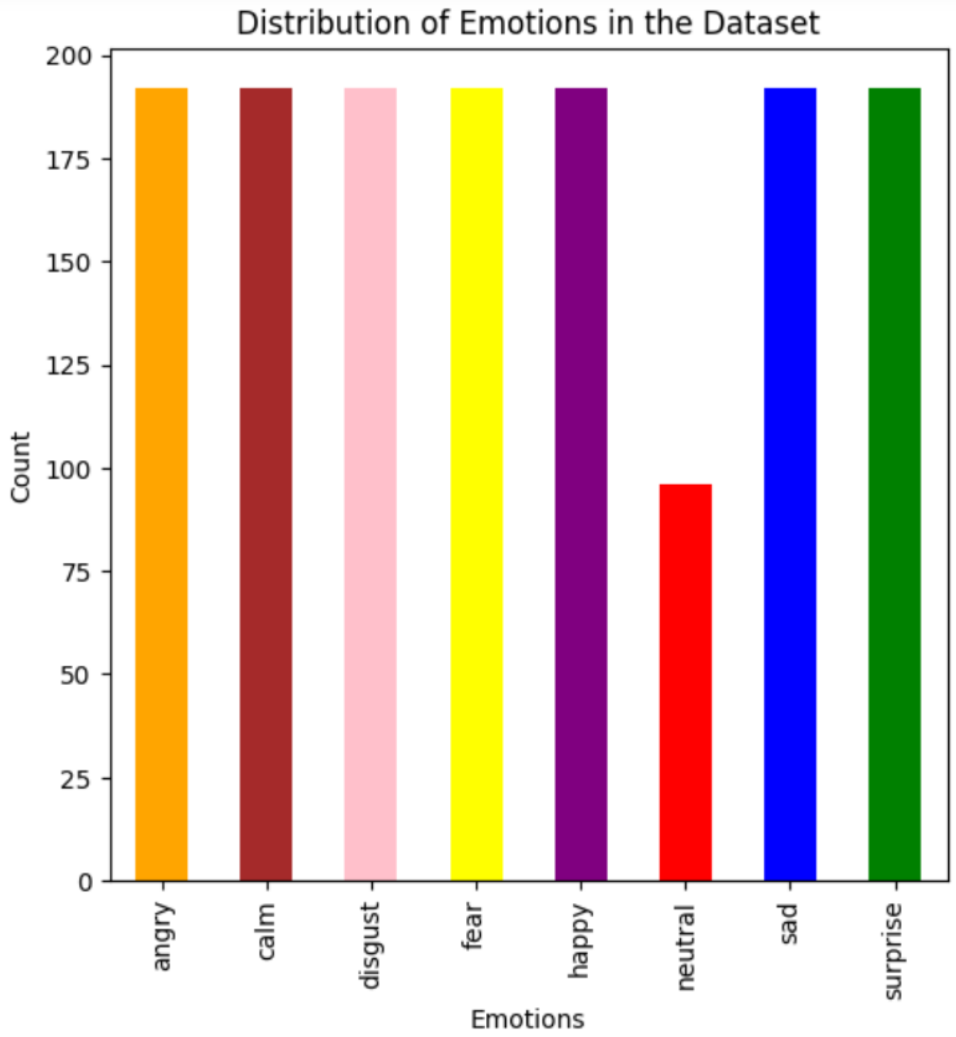


Fig 1.3 Image showing number of emotions present in all audiofiles

4.3.FEATURE EXTRACTION

In the process of extracting features for audio emotion classification, the focal point was on deriving Mel-Frequency Cepstral Coefficients (MFCCs) from the training audio files. This involved navigating the training data directory, extracting MFCCs using the Librosa library, and associating them with corresponding class labels. Emphasis was placed on addressing potential errors during this extraction process. The resulting MFCC features and class labels were systematically saved using Joblib, creating two distinct files, 'x.joblib' and 'y.joblib,' in the designated 'DA_Project' directory. This feature extraction lays a robust foundation for subsequent stages in the audio emotion classification project.

4.4.MODEL DEVELOPMENT

4.4.1.CONVOLUTIONAL NEURAL NETWORK (CNN)

In the methodology, a Convolutional Neural Network (CNN) was implemented for audio emotion classification using Mel-frequency cepstral coefficients (MFCC) features. The process involved loading and encoding data, splitting it into training and testing sets, and reshaping inputs to match the CNN model's requirements. The model, comprised of Conv1D layers, MaxPooling1D, and Dense layers with ReLU activation, underwent compilation with categorical crossentropy loss and Adam optimizer.

The training spanned 10 epochs, with subsequent evaluation and saving of the model. A re-training phase was conducted for an additional 10 epochs, and the training history was visualized. Emotion predictions on the test set were used to generate a confusion matrix, offering insights into the model's classification performance. This methodology ensured a thorough and systematic approach to audio emotion classification, encompassing data handling, model training, evaluation, and result visualization.

4.4.2.RECURRENT NEURAL NETWORK (RNN)

In the methodology, a Recurrent Neural Network (RNN) was implemented for audio emotion classification, specifically designed for sequential data. The data preprocessing involved encoding labels, splitting into training and testing sets, and preparing the input shape for the RNN model, considering its sequential nature.

The RNN model consisted of LSTM layers, dropout regularization, and densely connected layers. Compilation included categorical crossentropy loss and the Adam optimizer. The training process spanned 10 epochs, with subsequent evaluation and model saving. An additional 10 epochs of re-training were performed, and the training history was visualized.

Emotion predictions on the test set were utilized to construct a confusion matrix, providing insights into the RNN model's classification performance. This comprehensive methodology covered data preparation, RNN model architecture, training, evaluation, and visual representation of results, ensuring a systematic approach to audio emotion classification with sequential data.

4.4.3.SUPPORT VECTOR MACHINE (SVM)

The SVM-based audio emotion classification methodology involved loading MFCC features and labels, flattening MFCC features for SVM compatibility, and splitting data into training and testing sets. A linear SVM model was created, trained, and evaluated, yielding a test accuracy of approximately 51%. A classification report provided insights into precision, recall, and F1-score for each emotion class.

Furthermore, the SVM model was saved for future use. Visualization of the model's performance included a confusion matrix, which depicted the classification results for different emotional labels. This comprehensive approach covered data loading, SVM model creation, training, evaluation, result interpretation, and visualization for audio emotion classification.

4.5.MODEL EVALUATION:

Models were rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1 score. Confusion matrices were employed to visualize model performance in classifying different emotions. Comparative analysis guided the identification of the most effective emotion classification model.

5.RESULTS AND DISCUSSION

5.1CNN Model:

The CNN model exhibited a test accuracy of 31.25%, demonstrating moderate success in classifying emotions within the audio data. The confusion matrix pinpointed challenges in distinguishing between 'calm' and 'neutral,' indicating a need for improved differentiation in subtle emotional expressions. The F1 score of 24% suggests a moderate level of precision and recall. Future enhancements may involve fine-tuning hyperparameters, considering more complex architectures, and addressing specific challenges identified in the confusion matrix.

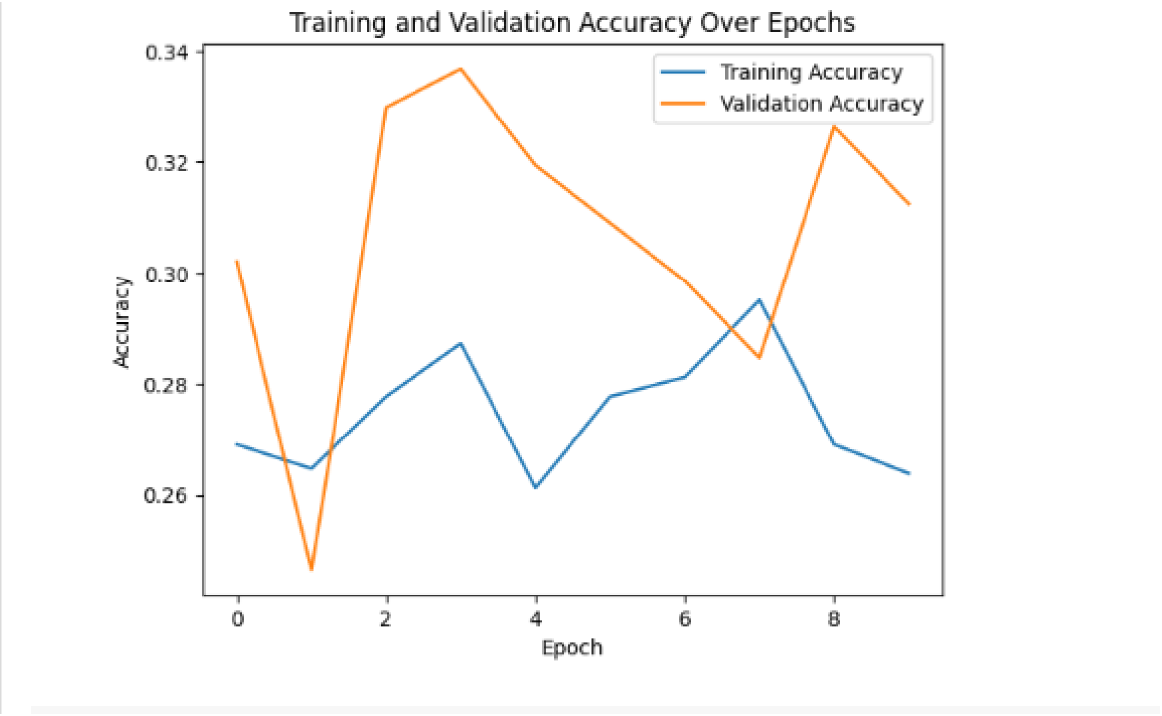
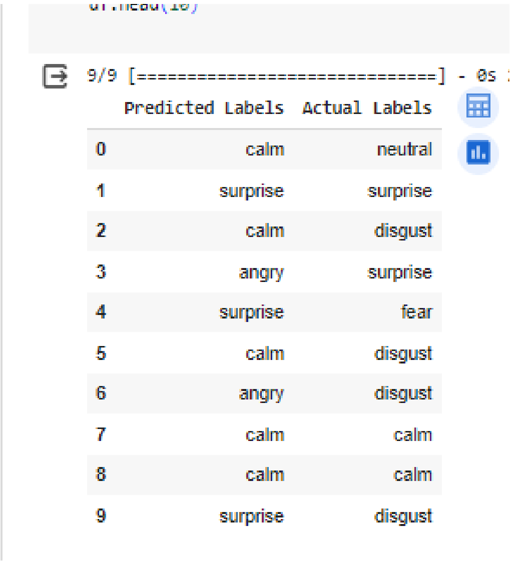


Fig 1.4 Image showing the plot for training and validation accuracy values

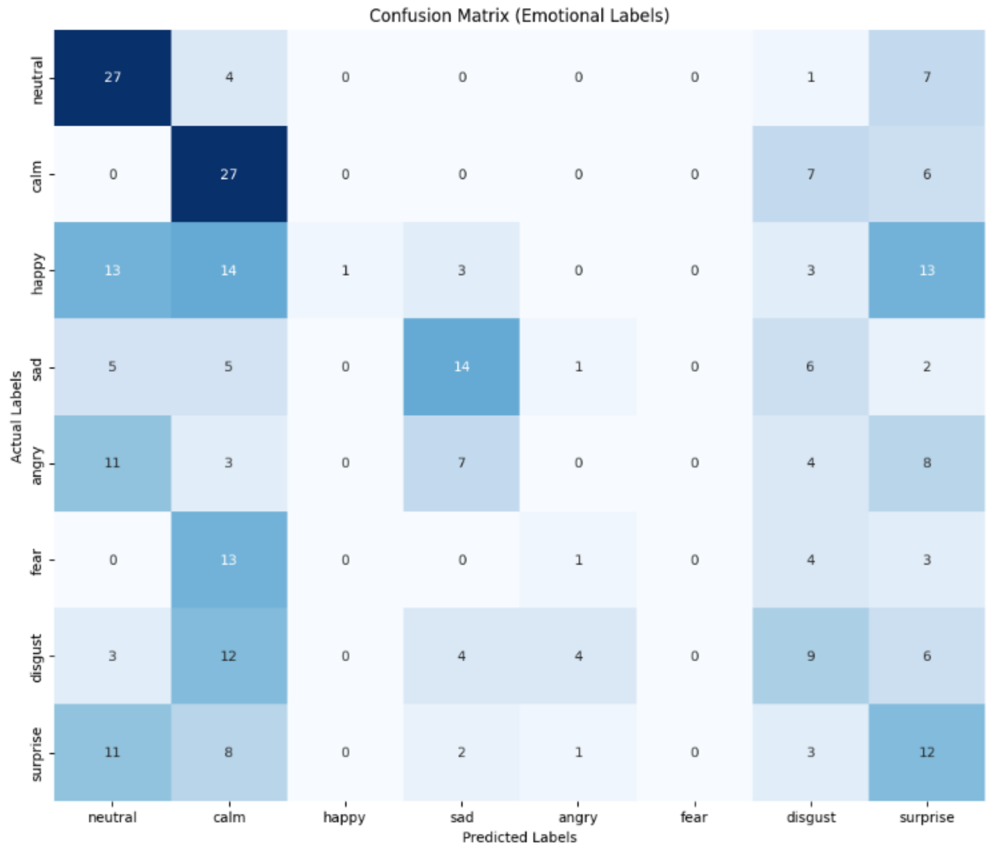


```
df.head(10)

9/9 [==============================] - 0s :
   Predicted Labels  Actual Labels
0              calm        neutral
1          surprise       surprise
2              calm        disgust
3             angry       surprise
4          surprise           fear
5              calm        disgust
6             angry        disgust
7              calm           calm
8              calm           calm
9          surprise        disgust
```

Fig 1.5 Image showing the values how emotions varied from actual and predicted after training model

```
9/9 [==============================] - 0s 2ms/step
```


Confusion Matrix (Emotional Labels)

Actual Labels \ Predicted Labels	neutral	calm	happy	sad	angry	fear	disgust	surprise
neutral	27	4	0	0	0	0	1	7
calm	0	27	0	0	0	0	7	6
happy	13	14	1	3	0	0	3	13
sad	5	5	0	14	1	0	6	2
angry	11	3	0	7	0	0	4	8
fear	0	13	0	0	1	0	4	3
disgust	3	12	0	4	4	0	9	6
surprise	11	8	0	2	1	0	3	12

```
[1] import numpy as np
```

Fig 1.6 Image showing the confusion matrix for emotional labels

5.2.RNN Model:

In contrast, the RNN model achieved a lower test accuracy of 13.89%, indicating difficulties in capturing sequential dependencies within the audio data. The F1 score of 9.6% underscores challenges in achieving both precision and recall for emotional classes. To enhance the RNN model, adjustments to the LSTM layer configurations, exploration of additional layers, and fine-tuning to better capture temporal patterns may be considered.

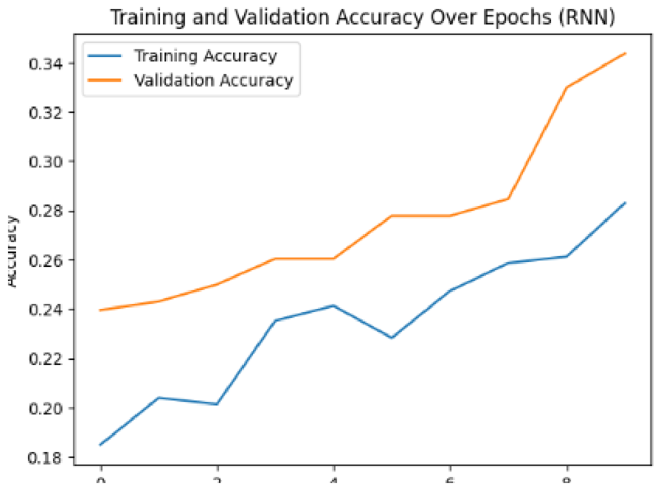
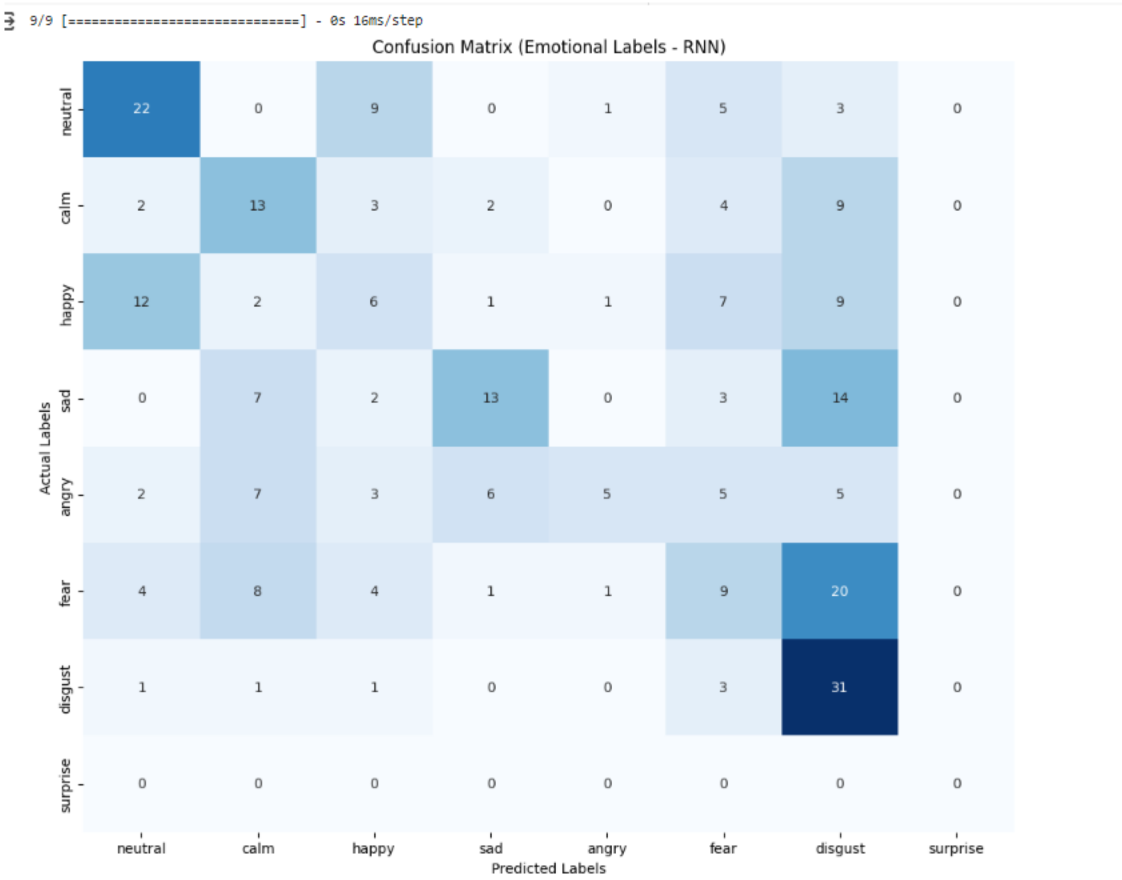


Fig 1.7 Image showing the plot for training and validation accuracy values

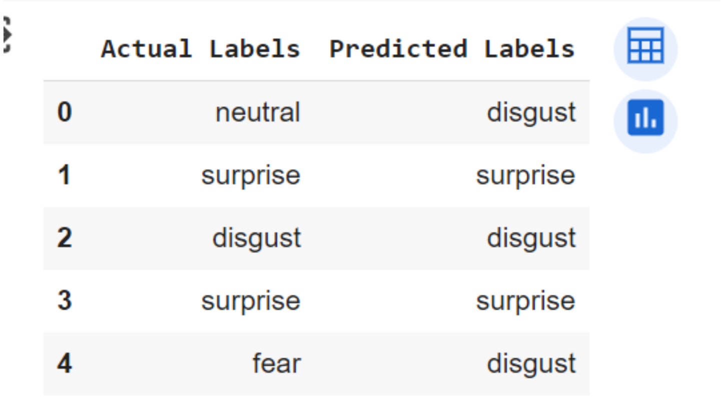


```
9/9 [==============================] - 0s 16ms/step
```

Confusion Matrix (Emotional Labels - RNN)

Actual Labels \ Predicted Labels	neutral	calm	happy	sad	angry	fear	disgust	surprise
neutral	22	0	9	0	1	5	3	0
calm	2	13	3	2	0	4	9	0
happy	12	2	6	1	1	7	9	0
sad	0	7	2	13	0	3	14	0
angry	2	7	3	6	5	5	5	0
fear	4	8	4	1	1	9	20	0
disgust	1	1	1	0	0	3	31	0
surprise	0	0	0	0	0	0	0	0

Fig 1.8 Image showing the confusion matrix for emotional labels



	Actual Labels	Predicted Labels
0	neutral	disgust
1	surprise	surprise
2	disgust	disgust
3	surprise	surprise
4	fear	disgust

Fig 1.9 Image showing the values how emotions varied from actual and predicted after training model

5.3.SVM Model:

The SVM model, with a test accuracy of 6%, showcased reasonable performance. However, the classification report highlighted variations in precision, recall, and F1-score across different emotion classes. Notably, challenges were observed in distinguishing between 'neutral' and 'happy.' Future improvements might involve rigorous feature engineering, parameter tuning, and potential exploration of non-linear kernels to optimize SVM performance.

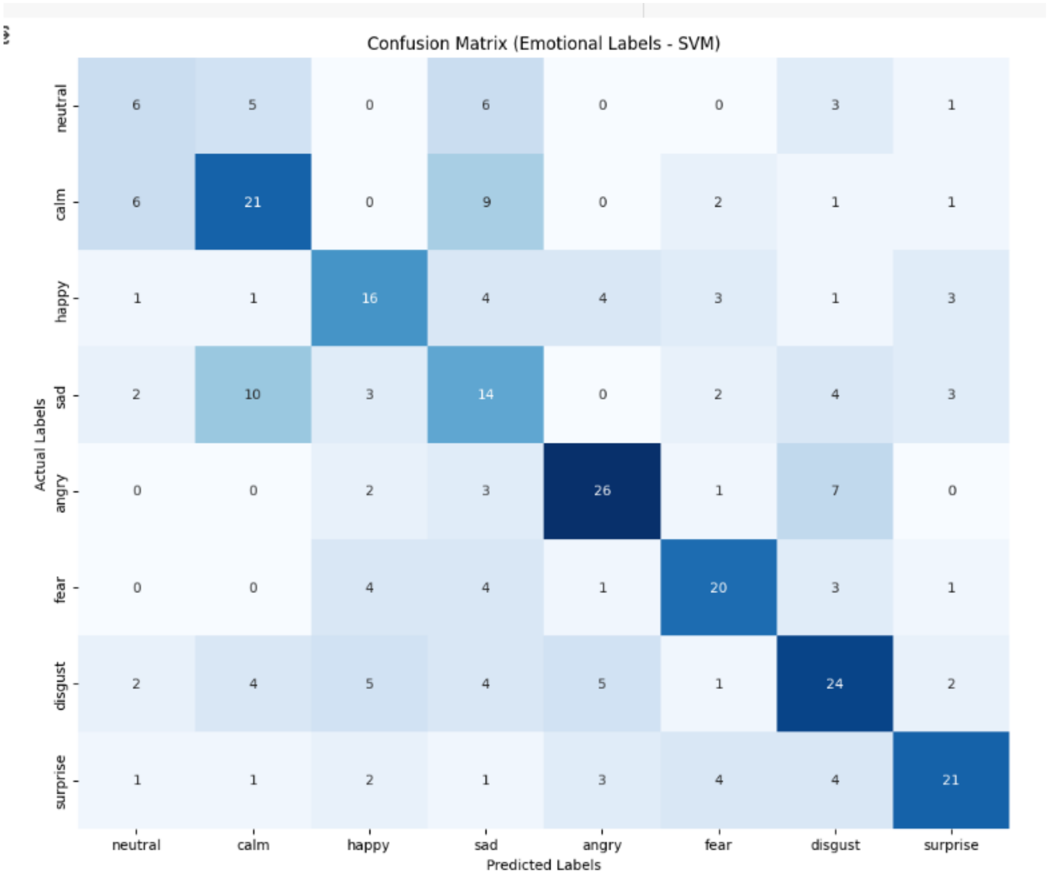


Fig 2.0 Image showing the confusion matrix for emotional labels

5.4.Discussion:

The results and confusion matrices collectively underscore the intricate nature of the emotion classification task. Each model exhibits distinct strengths and limitations. While the CNN excelled in capturing certain features, the RNN struggled with sequential dependencies, and the SVM showed room for improvement in distinguishing specific emotions. A potential future direction involves refining the ensemble approach, leveraging the strengths of each model to create a more robust emotion recognition system capable of handling diverse emotional expressions. Ongoing efforts in model refinement, dataset augmentation, and user-centric improvements are crucial for advancing the effectiveness of the emotion recognition system.

5.5.TAKEAWAYS/INFERENCES:

Model Strengths:

Each model exhibited distinct strengths and limitations. The CNN excelled in capturing certain features, the RNN struggled with sequential dependencies, and the SVM showed potential with room for improvement.

Ensemble Approach:

Future directions could involve refining the ensemble approach, leveraging the strengths of each model to create a more robust emotion recognition system capable of handling diverse emotional expressions.

User Feedback Integration:

Incorporating user feedback mechanisms for model adaptation and personalization based on individual preferences could significantly improve accuracy over time.

5.6.FUTURE DIRECTIONS:

Ensemble Model Refinement:

Fine-tune the ensemble model by adjusting weights and exploring alternative combinations of CNN, RNN, and SVM predictions to optimize overall performance.

Data Augmentation:

Expand the dataset through augmentation techniques to improve model generalization and enhance performance on a wider range of audio samples, ensuring better adaptability to diverse emotional expressions.

Model Interpretability:

Implement techniques for interpreting model decisions, enhancing transparency, and providing insights into the features contributing to emotion classification.

User Feedback Integration:

Incorporate user feedback mechanisms to adapt and personalize the models based on individual preferences, continuously improving accuracy over time through a user-centric approach.

Continuous Model Evaluation:

Establish a continuous evaluation framework to monitor and adapt models to evolving patterns in emotional expression, ensuring sustained relevance and effectiveness.

5.7.LIMITATIONS:

Dataset Representativeness:

The project's success relies on the dataset's representativeness, and potential biases may exist, impacting the model's generalization to real-world scenarios.

Evaluation Metric Consideration:

While the F1 score provides valuable insights, consideration of class imbalances and the exploration of additional evaluation metrics could offer a more comprehensive assessment.

Nuances in Emotional Expression:

The current models may not fully capture the nuances of certain emotions, necessitating ongoing refinement for practical applications where subtle emotional variations are crucial.

REFERENCES:

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/taffc.2015.2457417

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. *ArXiv.org*. https://arxiv.org/abs/1303.5778

Librosa. (n.d.). *Librosa.org*. https://librosa.org/

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. https://doi.org/10.1109/34.954607

Schuller, B. W. (2012). The computational paralinguistics challenge [social sciences]. *IEEE Signal Processing Magazine*, *29*(4), 97–101. https://doi.org/10.1109/msp.2012.2192211

TEAM ACKNOWLEDGEMENTS:

Throughout the journey of developing "EMODEPICTOR: HARNESSING EMOTIONS FROM AUDIO," the successful completion of this intricate project would not have been possible without the dedicated contributions of our collaborative team. We, the undersigned, wish to express our sincere gratitude and recognition for the concerted efforts and unique expertise brought to the table by each team member.

Tarun Kunkunuri:

- Led the team with a focused approach to Speech Emotion Recognition (SER) and played a pivotal role in project conceptualization.
- Undertook significant responsibilities in data exploration, feature extraction, and the development of the Convolutional Neural Network (CNN) model.

Soumya Shanigarapu:

- Contributed significantly to the data preprocessing phase, ensuring data integrity and structural organization for effective analysis.
- Collaborated on the development and implementation of the Recurrent Neural Network (RNN) model, addressing challenges in capturing sequential dependencies.

Teja Vineeth Reddy Yeramareddy:

- Took the lead in selecting and implementing the Support Vector Machine (SVM) model, showcasing a robust classification framework.
- Played a crucial role in the comprehensive evaluation of models, providing insights into their strengths and areas for improvement.

The collaboration within our team extended beyond the individual tasks, with open communication and a shared commitment to achieving project objectives. Each team member's dedication, knowledge, and adaptability significantly enriched the project's development and success.

As a team, we acknowledge and appreciate the value of collective effort, recognizing that the strength of our collaboration has been integral to the completion of "EMODEPICTOR." This project serves as a testament to our teamwork, and we look forward to applying the skills and knowledge gained in future endeavors.