**PONDICHERRY UNIVERSITY**

*(A Central University)*



**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MASTER OF COMPUTER APPLICATION**

NAME : DHEEPAN G

REGISTER NO. : 22352018

COURSE : MCA

SUBJECT : MAIN PROJECT REPORT

**GUIDED BY : Dr. S.L.JAYALAKSHMI**

**SPEECH EMOTION RECOGNITION**

By

**DHEEPAN G**

(Registration Number: 22352018)

**Project report submitted in partial fulfilment of the requirements for the award of the degree of**

**MASTER OF COMPUTER APPLICATION**



**DEPARTMENT OF COMPUTER SCIENCE**

**SCHOOL OF ENGINEERING & TECHNOLOGY**

**PONDICHERRYUNIVERSITY**

**MAY 2024**

**BONAFIDE CERTIFICATE**

This is to certify that this project work entitled “**SPEECH EMOTION RECOGNITION**” is a bonafide record of work done by **Mr. DHEEPAN G** (Reg. Number 22352018) in the partial fulfilment for the degree of **Master of Computer Applications** of Pondicherry University.

This work has not been submitted elsewhere for the award of any other degree to the best of our knowledge.

**INTERNAL GUIDE HEAD OF THE DEPARTMENT**

**DR.S.L.JAYALAKSHMI Dr. S. K. V. JAYAKUMAR**

Assistant Professor Professor/HOD

Department of Computer Science Department of Computer Science

School of Engineering & Technology School of Engineering & Technology

Pondicherry University Pondicherry University

Pondicherry – 605 014 Pondicherry – 605 014

Submitted for the Viva-Voce Examination held on:

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**ACKNOWLEDGEMENT**

Every project, big or small, is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration, support and guidance of all those people who have been instrumental in making this project a success.

I express my heartfelt gratitude to my Project Guide **Dr.S.L.JAYALAKSHMI**, Assistant Professor, Department of Computer Science, Pondicherry University, Pondicherry, for her whole - hearted assistance and direction not only for the duration of the project but for the entire duration of the course. I will always remain grateful to her and whose constant care about me has provided a new direction to work.

I express my gratitude to **Dr. S. K. V. JAYAKUMAR**, Professor and Head, Department of Computer Science, Pondicherry University, Pondicherry for her support and arranging the project in a good schedule.

Finally, I would like to express my regards for all the faculty members of Department of Computer Science and others involved in this project, directly or indirectly

**(DHEEPAN G)**

**TABLE OF CONTENTS**

**TITLE PAGE NO.**

**ACKNOWLEDGEMENT**

**ABSTRACT**

1. **INTRODUCTION**
   1. ABOUT THE PROJECT 1
   2. PROJECT PLAN 2
2. **PROBLEM DEFINITION & FEASIBILITY ANALYSIS**
   1. PROBLEM DEFINITION 3
   2. EXISTING SYSTEM 3
   3. PROPOSED SYSTEM 3
   4. FEASIBILITY STUDY
      1. TECHNICAL FEASIBILITY 5
      2. OPERATIONAL FEASIBILITY 5
      3. ECONOMIC FEASIBILITY 5
3. **SOFTWARE REQUIREMENT** 
   1. HARDWARE REQUIREMENTS 6
   2. SOFTWARE REQUIREMENTS 6
   3. SYSTEM REQUIREMENTS 7
4. **SYSTEM DESIGN**
   1. MODULE DESCRIPTION 9
   2. USE CASE DIAGRAM 10
   3. ACTIVITY DIAGRAM 11
   4. SEQUENCE DIAGRAM 12
5. **IMPLEMENTATION** 13
   1. NOISE REDUCTION
   2. FEATURE EXTRACTION
   3. FEATURE SCALING
   4. DATA SPLITTING
   5. CLASSIFIER SELECTION
   6. MODEL TRAINING
   7. MODEL EVALUATION
6. **SYSTEM TESTING** 
   1. SYSTEM IMPLEMENTATION 14
   2. TESTING
      1. UNIT TESTING 14
      2. VALIDATION TESTING 14
      3. FUNCTIONAL TESTING 15
7. **CONCLUSION**

APPENDIX I: SCREENSHOTS 16

APPENDIX II: OUTPUTS 19

1. **BIBLIOGRAPHY**  25

**ABSTRACT**

The Speech Emotion Recognition (SER) project aims to develop an intelligent system capable of recognizing human emotions from speech signals. Emotion recognition from speech plays a crucial role in various applications such as human-computer interaction, sentiment analysis and psychological research.

In this project, we leverage machine learning techniques and signal processing methods to analyse speech signals and extract features that capture the emotional content of the speech. The project involves several key steps, including data collection, preprocessing, feature extraction, model training and evaluation.

We utilize techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), low-pass filtering, machine learning models like Support Vector Machine(SVM) and deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for emotion classification. The system aims to accurately classify emotions such as happiness, sadness, anger and neutral states from speech signals. Through experimentation and evaluation, the effectiveness of the proposed approach in recognizing emotions from speech is demonstrated.

The SER project holds the potential to contribute to various domains such as human-computer interaction, mental health assessment and affective computing, thereby enhancing our understanding and interaction with human emotions.

1. **INTRODUCTION**
   1. **ABOUT THE PROJECT**

The Speech Emotion Recognition (SER) project attempts to create a sophisticated system capable of effectively detecting and categorizing emotions conveyed through human speech captured in audio recordings. Emotion recognition holds immense promise across diverse domains, including but not limited to human-computer interaction, customer service and healthcare. Leveraging a fusion of machine learning algorithms and advanced signal processing techniques, this project aims to meticulously extract pertinent features from speech signals. These features are then employed to train robust models that can discern and classify emotions with a high degree of accuracy. By accurately discerning the emotional nuances within spoken language, the system stands to offer invaluable insights, thereby elevating user experiences and fostering enhanced communication dynamics.

* 1. **PROJECT PLAN**

Our project aims to develop an advanced system for emotion recognition in human speech using machine learning techniques. Similar to human perception of emotions through voice, we are training computers to achieve this capability. Initially, we collect speech recordings and apply noise reduction techniques to ensure clear data. Subsequently, we employ sophisticated algorithms to analyse various aspects of the recordings, such as speech rate and tone, to extract emotional cues. Leveraging this information, our system learns to classify different emotions, including happiness, sadness and neutral. We utilize a range of techniques to train the system effectively. Following training, rigorous testing is conducted to evaluate the system's ability to accurately identify emotions in new recordings. Ultimately, our objective is to enhance the computer's proficiency in understanding emotions in speech, which holds significant potential for applications such as improving the empathetic capabilities of virtual assistants and assisting therapists in better understanding their clients.

1. **PROBLEM DEFINITION & FEASIBILITY ANALYSIS**

**2.1 PROBLEM DEFINITION**

The problem we aim to address is the accurate recognition of emotions conveyed through human speech. Emotions play a crucial role in communication and being able to detect them accurately can enhance various applications, including virtual assistants, customer service systems and mental health monitoring tools. However, recognizing emotions solely based on audio data poses significant challenges due to the complexity and variability of human speech.

**2.2 EXISTING SYSTEM**

Currently, emotion recognition systems primarily rely on manual analysis or basic rule-based approaches, which often lack accuracy and scalability. These systems struggle to capture subtle nuances in speech that convey different emotions. Additionally, they may require extensive human intervention for training and customization, limiting their practicality and efficiency.

**2.3 PROPOSED SYSTEM**

Our system automates speech emotion recognition using Python and machine learning techniques. It begins by preprocessing audio data to remove noise and extract features like MFCCs. These features are then scaled and used to train SVM, LSTM and CNN classifiers for emotion classification. Evaluation metrics and confusion matrices are generated to assess classifier performance. The system aims to provide accurate emotion detection in real-time, with potential applications in human-computer interaction and mental health monitoring.

**MFCC**

MFCC is a feature extraction technique widely used in speech and audio processing. MFCCs are used to represent the spectral characteristics of sound in a way that is well-suited for various machine learning tasks, such as speech recognition and music analysis.

In simpler terms, MFCCs are a set of coefficients that capture the shape of the power spectrum of a sound signal.

In our speech emotion recognition system, we employ three different classifiers: Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). Each classifier utilizes different techniques for emotion classification and has its own strengths and weaknesses.

1. **Support Vector Machine (SVM):**
   * SVM is a supervised learning model that analyzes data for classification and regression analysis.
   * It works by finding the hyperplane that best separates different classes in the feature space.
   * SVM is effective in high-dimensional spaces and is robust against overfitting.
   * However, it may not perform well with large datasets and complex nonlinear relationships between features.
2. **Long Short-Term Memory (LSTM):**
   * LSTM is a type of recurrent neural network (RNN) architecture designed to model sequential data.
   * It is well-suited for analyzing time-series data and has memory cells that can maintain information over time steps.
   * LSTM is effective in capturing long-term dependencies in sequential data, making it suitable for analyzing audio signals.
   * However, training LSTM models can be computationally expensive and requires careful tuning of hyperparameters.
3. **Convolutional Neural Network (CNN):**
   * CNN is a deep learning model commonly used for image recognition tasks.
   * It consists of convolutional layers that extract spatial features from input data.
   * CNN can be adapted for analysing one-dimensional data like audio signals by treating them as spectrograms or time-frequency representations.
   * While CNN is efficient in learning hierarchical representations from data, it may require larger datasets for training and can be sensitive to variations in input data.

**2.4 FEASIBILITY STUDY**

**2.4.1 TECHNICAL FEASIBILITY**

From a technical standpoint, our project is feasible as it leverages well-established machine learning frameworks and libraries such as TensorFlow, Keras and scikit-learn. These tools provide comprehensive support for building, training, and evaluating complex models for emotion recognition. Additionally, the availability of open-source datasets and pre-trained models further enhances the technical feasibility of our project.

**2.4.2 OPERATIONAL FEASIBILITY**

Operationally, our system can be integrated into various applications and platforms with ease. Once trained, the emotion recognition model can be deployed as a standalone service or integrated into existing systems through APIs. The system's user-friendly interface allows for seamless interaction, making it accessible to both developers and end-users.

**2.4.3 ECONOMIC FEASIBILITY**

Economically, our project offers significant potential for cost savings and efficiency improvements in various domains. By automating the process of emotion recognition, organizations can reduce the need for manual analysis and intervention, leading to lower operational costs and increased productivity. Additionally, the scalability of our system allows for widespread adoption across different industries, further enhancing its economic feasibility.

**3. SOFTWARE REQUIREMENTS SPECIFICATION**

**3.1 HARDWARE REQUIREMENTS**

1. **Processor (CPU)**
   * Dual-core processor or higher.
   * Recommended: Intel Core i5 or Ryzen 5.
2. **Memory (RAM)**
   * Minimum: 4 GB RAM.
   * Recommended: 8 GB RAM or higher for better performance.
3. **Storage Space**
   * At least 3 GB of free disk space for storing datasets, audio files, and project files.
   * Additional space may be required depending on the size of datasets and generated files.

**3.2 SOFTWARE REQUIREMENTS**

1. **Operating System**

Windows 10, macOS, or Linux-based operating system.

1. **Python**

Ensure you have Python installed on your system. You can download and install Python from the official Python website <https://www.python.org/downloads/>

1. **Text editor**

Vs code

**3.3 SYSTEM REQUIREMENTS**

**PYTHON LIBRARIES**

**pandas**

For data manipulation and handling Excel files.

**scikit-learn**

For machine learning algorithms and evaluation metrics.

**NumPy**

For numerical operations.

**librosa**

For audio feature extraction.

**pydub**

For audio processing and manipulation.

**SciPy**

For signal processing and filtering.

**Keras with TensorFlow backend**

For building and training deep learning models.

**seaborn**

For statistical data visualization based on matplotlib.

**matplotlib**

For creating static, animated, and interactive visualizations in Python.

Install these libraries using:

**pip install pandas scikit-learn numpy librosa pydub scipy keras tensorflow seaborn matplotlib**

**DATASET**

(RAVDESS) Ryerson Audio-Visual Database of Emotional Speech and Song is used as dataset for this

**INTERNET**

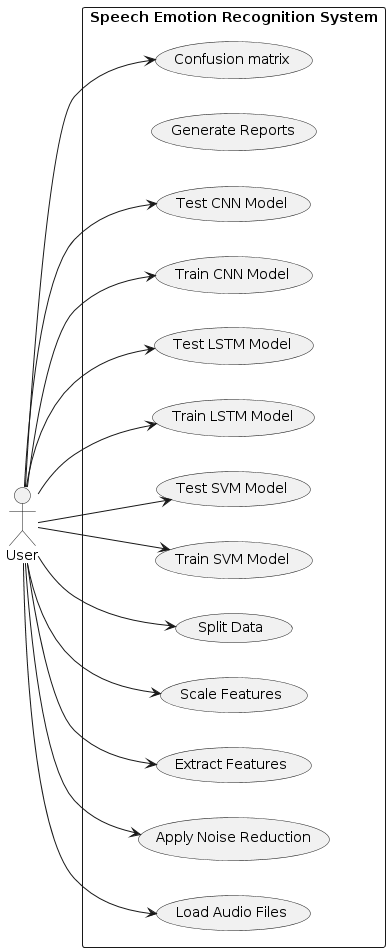
This software will require good internet connection to connect with servers and a good processing system to give best performance

1. **SYSTEM DESIGN**

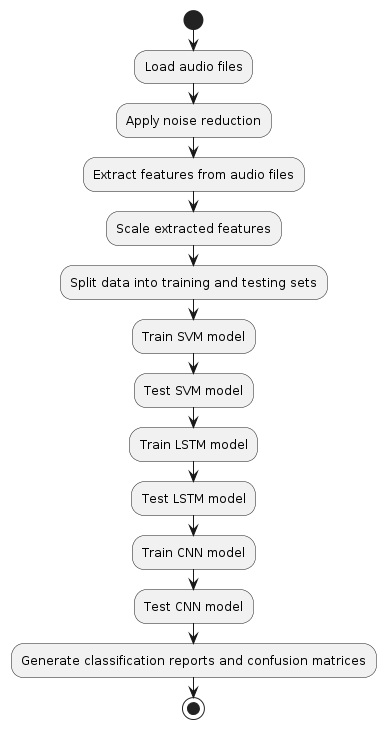
**4.1 MODULE DESCRIPTION**

1. **Data Loading and Preprocessing:**
   * This module handles the loading of audio data from the dataset folder and performs noise reduction to enhance the quality of audio files.
2. **Feature Extraction:**
   * Responsible for extracting relevant features from the audio files using libraries like librosa and pydub.
3. **Feature Scaling:**
   * Scales the extracted features to ensure uniformity and improve model performance during training.
4. **Data Splitting:**
   * Splits the dataset into training and testing subsets for model evaluation.
5. **Audio Classification:**
   * This module comprises three sub-modules:
     + SVM: Utilizes Support Vector Machine classifier for audio classification.
     + LSTM: Implements Long Short-Term Memory neural network for audio classification.
     + CNN: Employs Convolutional Neural Network for audio classification.
6. **Evaluation and Reporting:**
   * Generates classification reports and confusion matrices to evaluate the performance of each classification model.

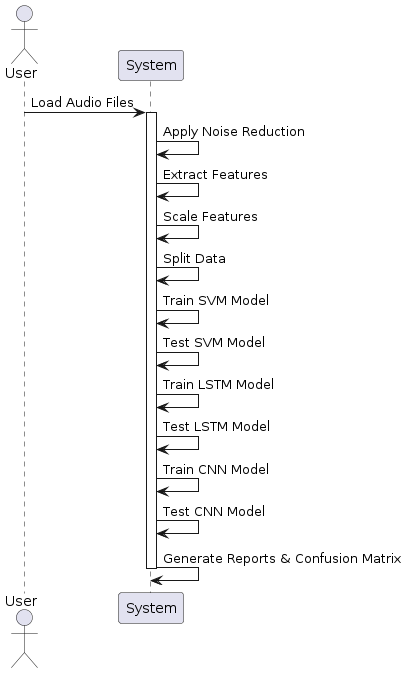
**4.2 USE CASE DIAGRAM**

****

**4.3 ACTIVITY DIAGRAM**

****

**4.4 SEQUENCE DIAGRAM**



1. **IMPLEMENTATION**

**5.1 NOISE REDUCTION**

This module preprocesses audio data by applying a low-pass filter to reduce background noise, improving the quality of audio recordings.

**5.2 FEATURE EXTRACTION**

Extracts relevant features from audio signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), which are used as input for emotion classification.

**5.3 FEATURE SCALING**

Normalizes or scales extracted features to ensure consistency and improve the performance of machine learning models.

**5.4 DATA SPLITTING**

Splits the dataset into training and testing sets to evaluate model performance.

**5.5 CLASSIFIER SELECTION**

Chooses appropriate classifiers such as Support Vector Machines (SVM), Convolutional Neural Networks (CNNs), or Long Short-Term Memory (LSTM) networks for emotion classification.

**5.6 MODEL TRAINING**

Trains the selected classifier using the training data to learn patterns and relationships between input features and emotion labels.

**5.7 MODEL EVALUATION**

Evaluates the trained model's performance using the testing data, calculating metrics such as accuracy, precision, recall and F1-score.

**6. SYSTEM TESTING**

**6.1 SYSTEM IMPLEMENTATION**

* Implement each module of the project according to the design specifications.
* Ensure that the modules are developed using appropriate programming languages and libraries.
* Perform rigorous testing during the implementation phase to catch any bugs or errors early on.

**6.2 TESTING**

**6.2.1 UNIT TESTING**

- Conduct unit testing for each module individually to ensure that they perform as expected.

- Verify the correctness of functions and methods within each module.

- Test boundary cases and edge conditions to check for robustness.

- Use testing frameworks like pytest or unittest to automate the testing process.

**6.2.2 VALIDATION TESTING**

- Validate the system against the user requirements and specifications.

- Ensure that the system meets the intended objectives and functionalities.

- Verify that the system accurately detects and classifies emotions from speech recordings.

- Evaluate the system's performance in terms of accuracy, speed, and resource utilization.

**6.2.3 FUNCTIONAL TESTING**

- Perform functional testing to validate the end-to-end functionality of the system.

- Test each feature and functionality to ensure they work as intended.

- Conduct scenario-based testing to simulate real-world usage scenarios.

- Verify the integration of different modules and components within the system.

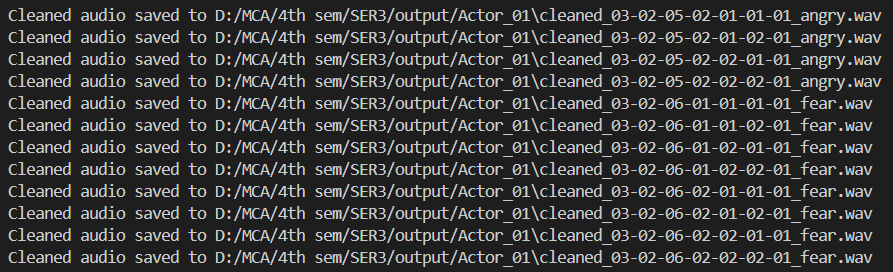
- Identify and address any discrepancies or deviations from the expected behavior.

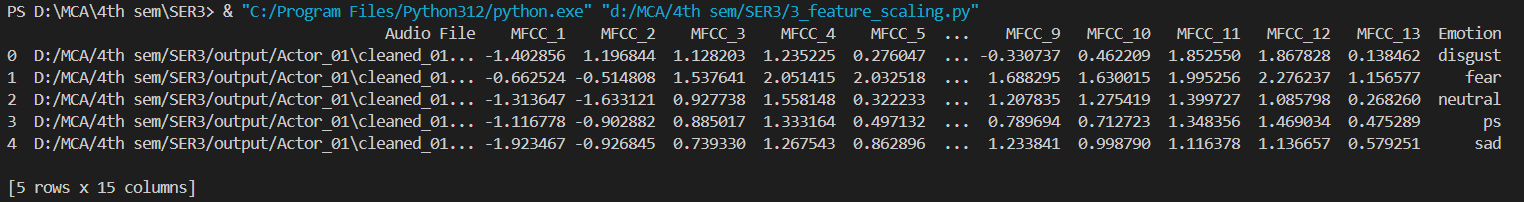
1. **CONCLUSION**

In our scenario, the accuracy of SVM is 0.75, LSTM is 0.25, and CNN is 0.5. Based on these results, SVM appears to be the most accurate classifier for your speech emotion recognition task. SVM's high accuracy indicates that it effectively separates different emotional states in the feature space and performs well on our dataset. However, it's essential to consider other factors such as computational efficiency and scalability when choosing the best classifier for your application.

**APPENDIX I: SCREENSHOTS**

**CLEANED AUDIO SAVED IN ACTOR\_01**

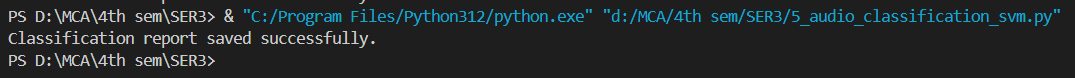
****

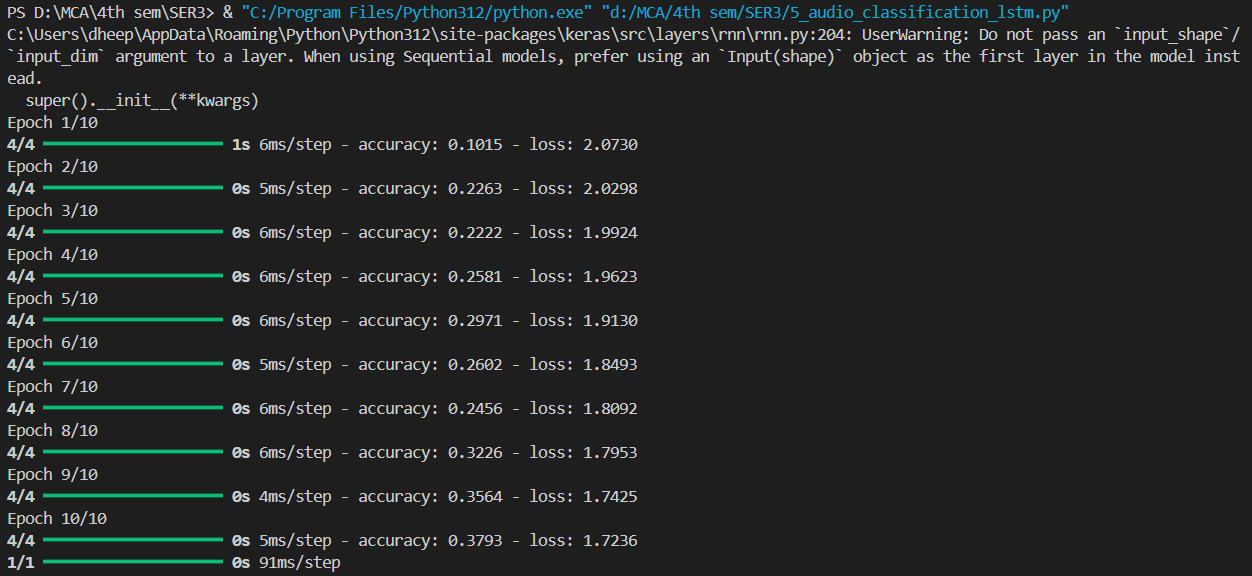
**FEATURE SCALING**

**DATA SPLITTING**

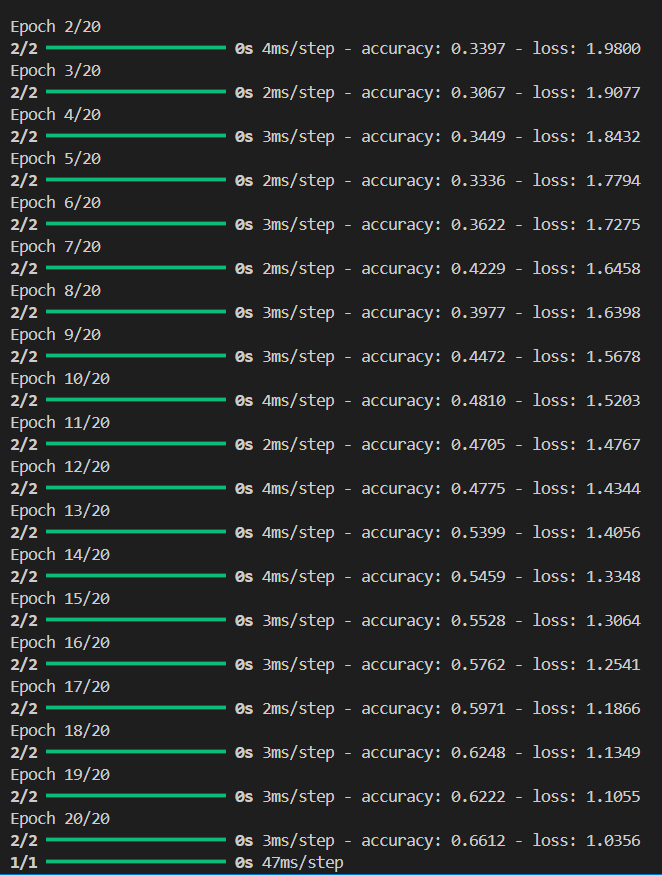
****

**AUDIO CLASSIFICATION SVM**

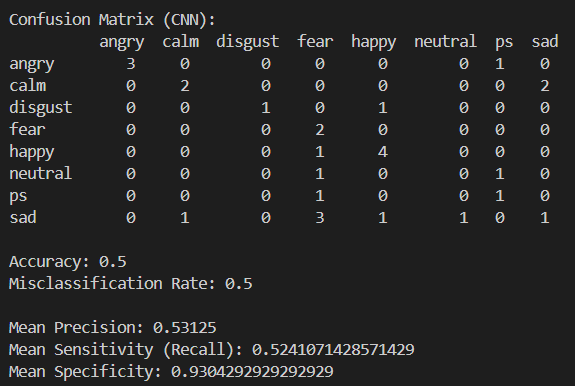
**AUDIO CLASSIFICATION LSTM**

****

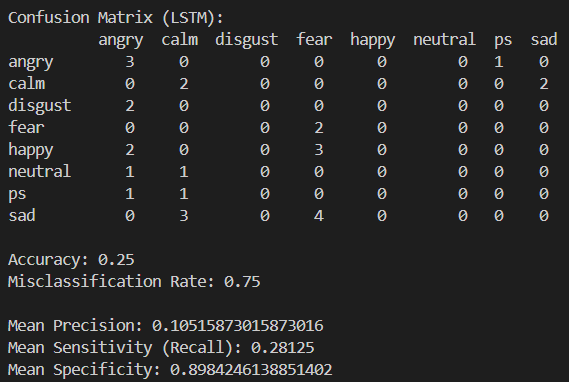
**AUDIO CLASSIFICATION CNN**

****

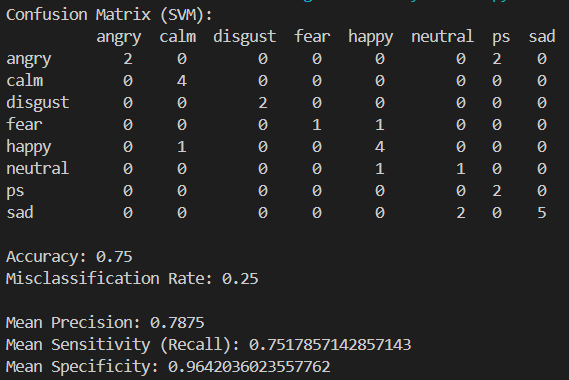
**CONFUSION MATRIX CNN**

****

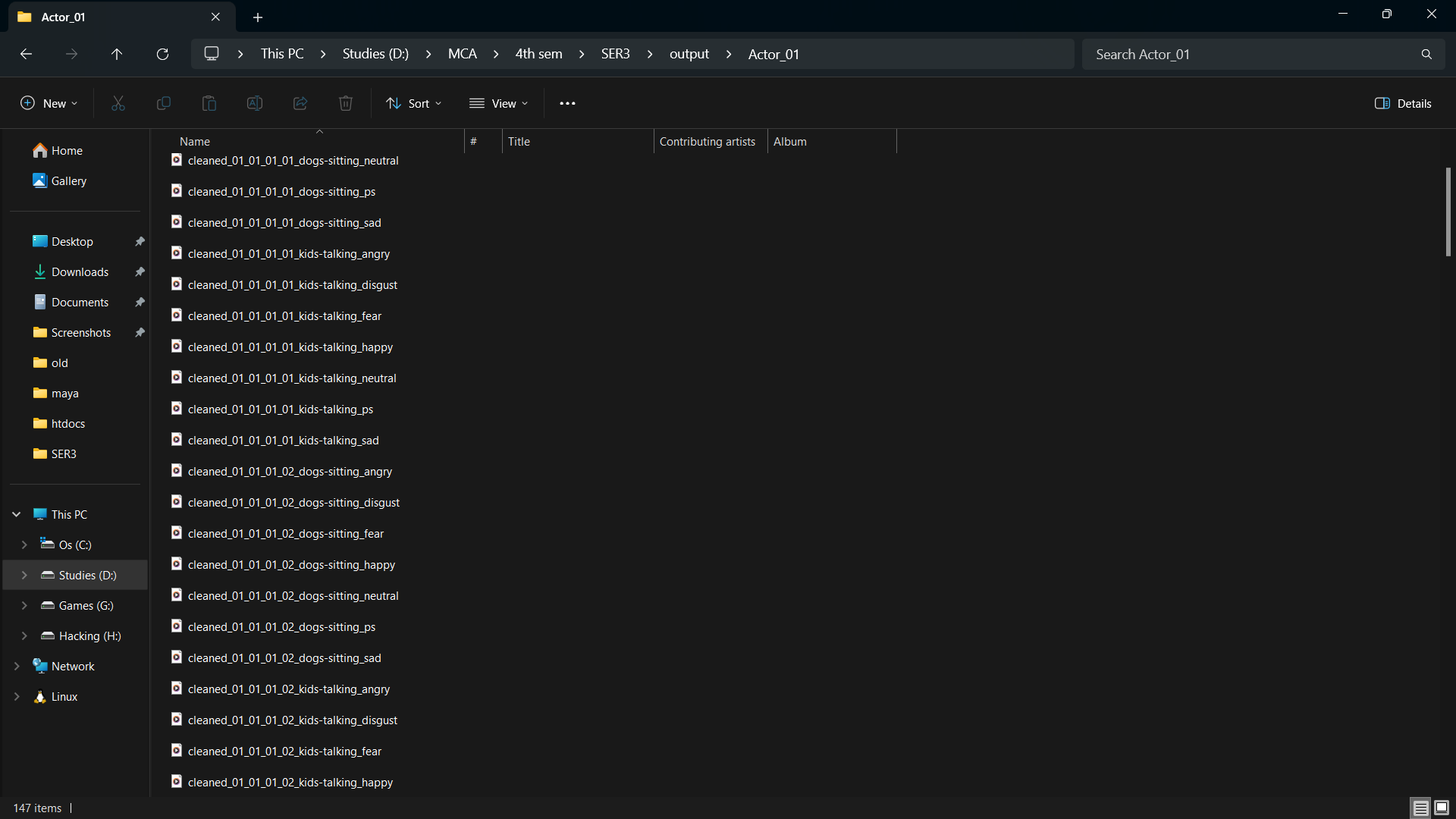
**CONFUSION MATRIX LSTM**

****

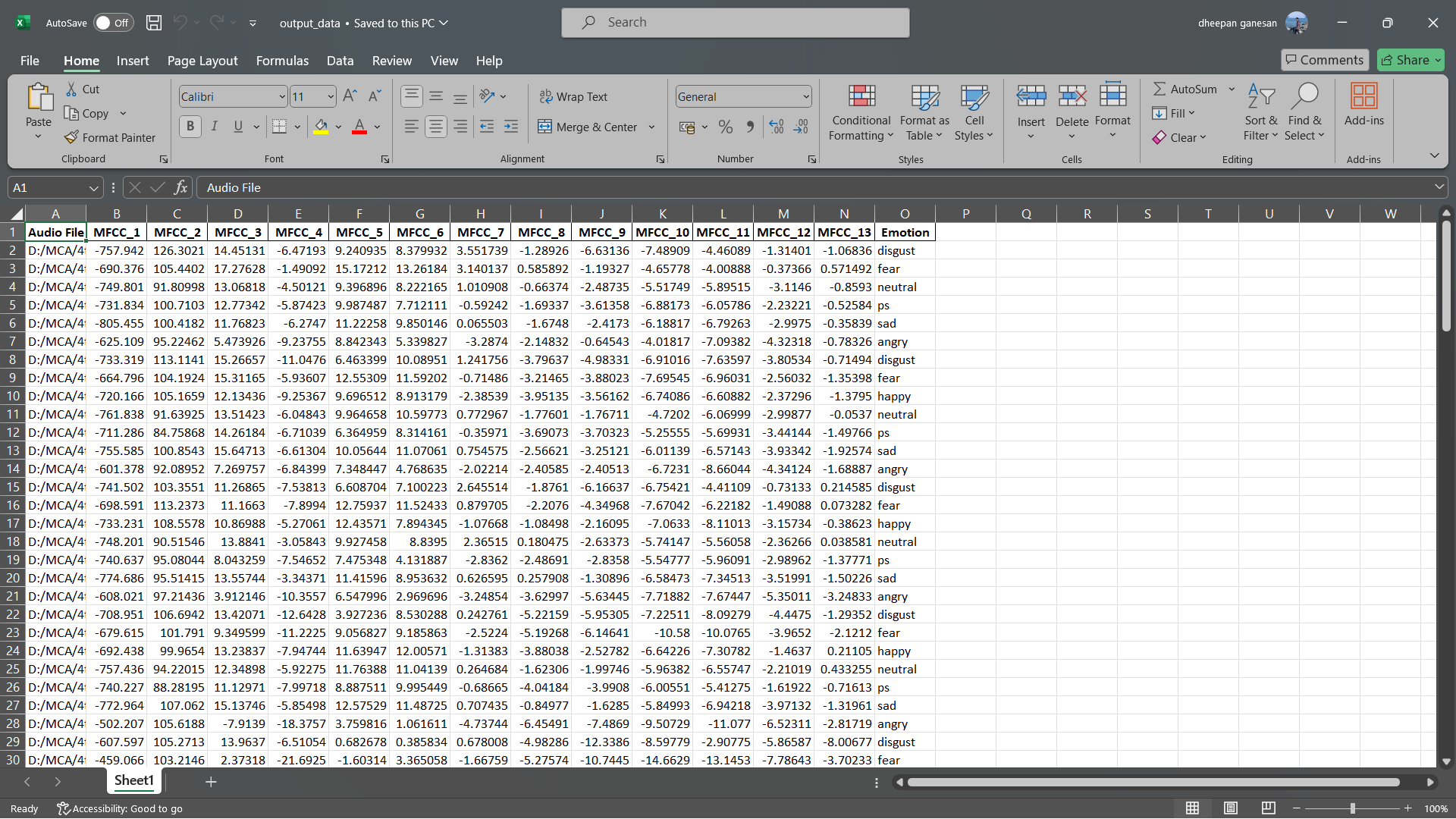
**CONFUSION MATRIX SVM**

****

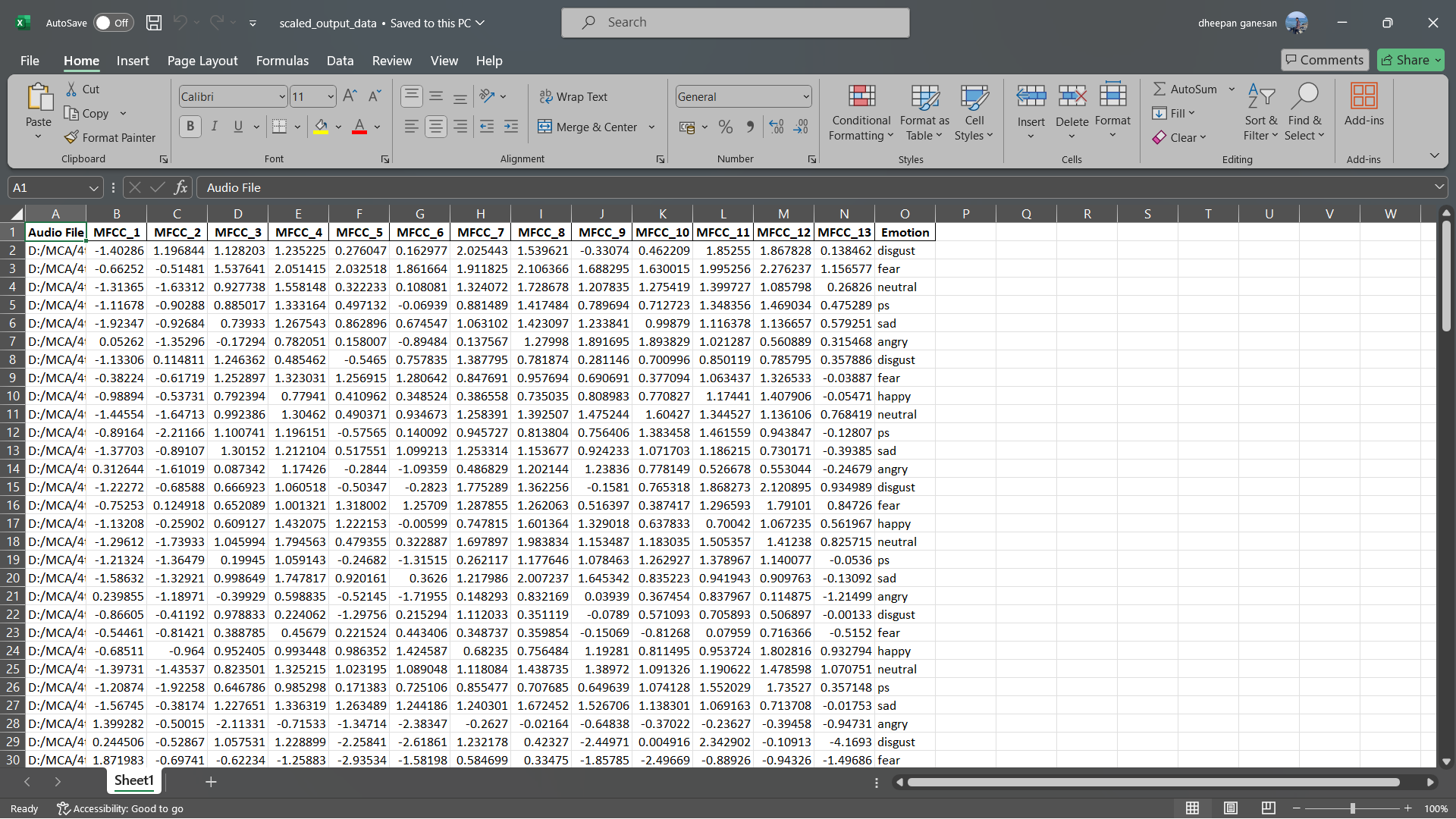
**APPENDIX II: OUTPUTS**

**CLEANED AUDIO IS SAVED IN AUDIO\_01 FOLDER**

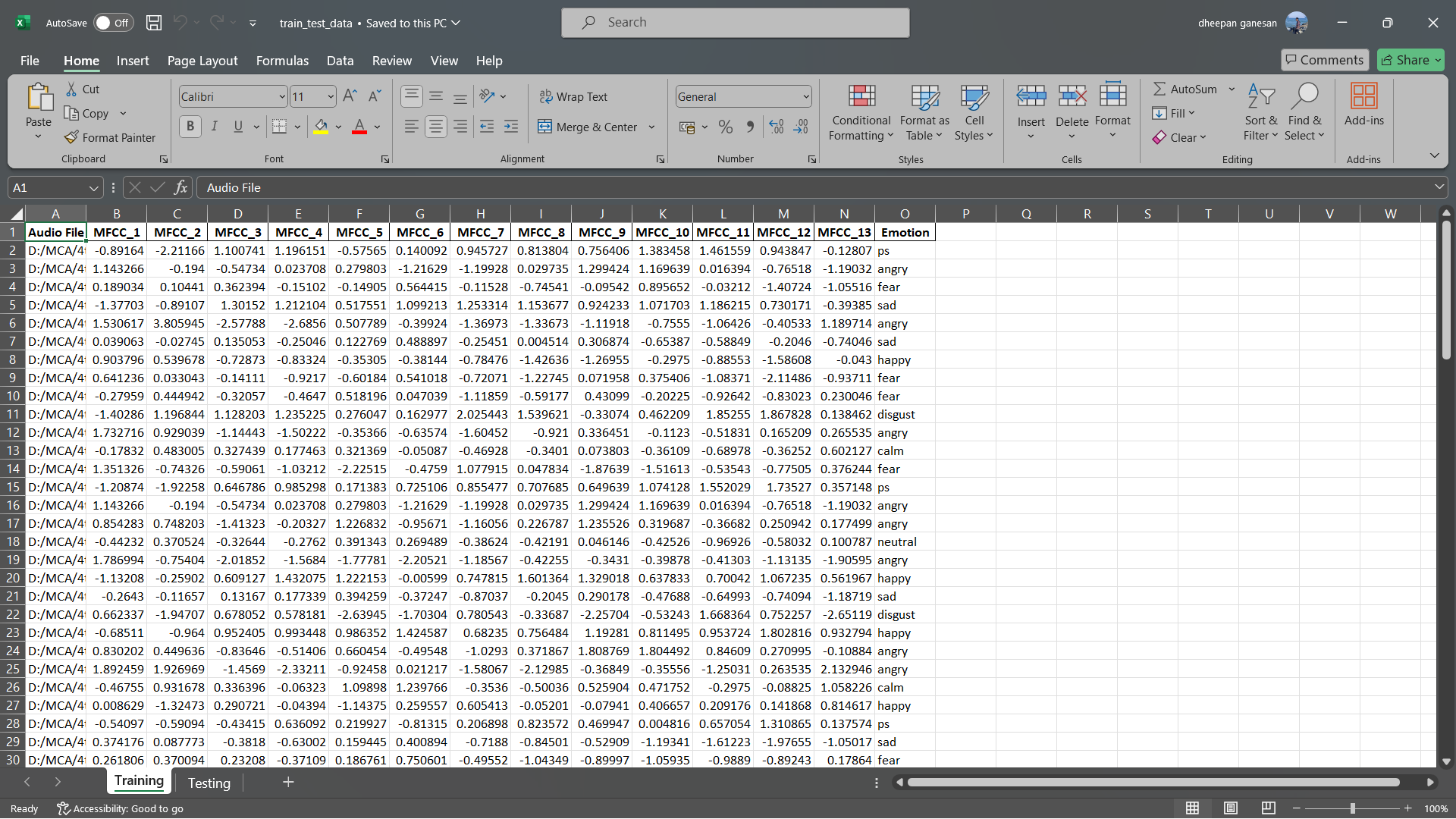
**FEATURES EXTRACTED AND SAVED AS OUTPUT\_DATA.XLSX**

****

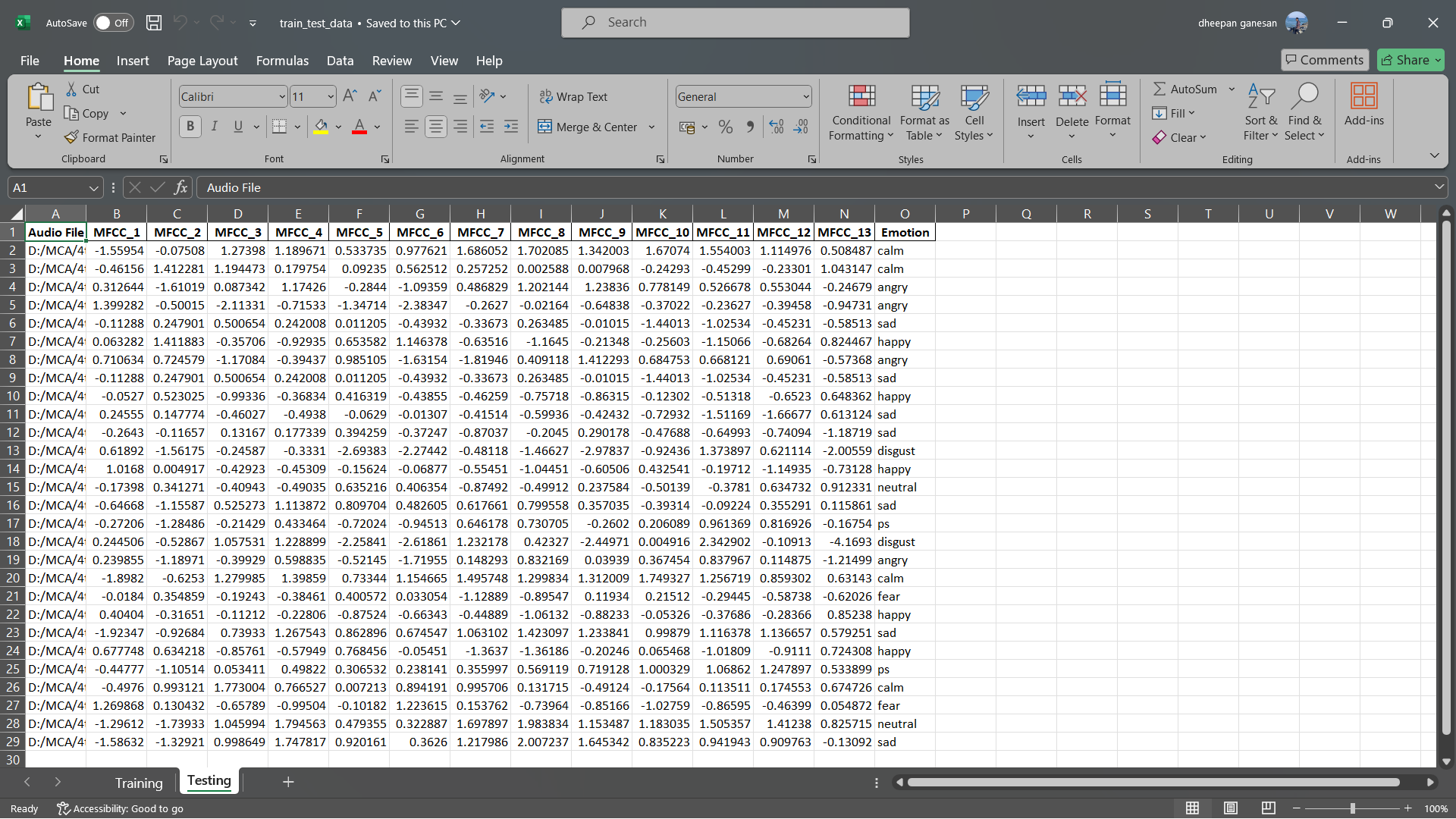
**FEATURE SCALING DONE AND SAVED AS SCALED\_OUTPUT\_DATA.XLSX**

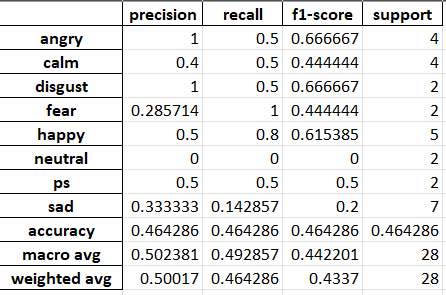
****

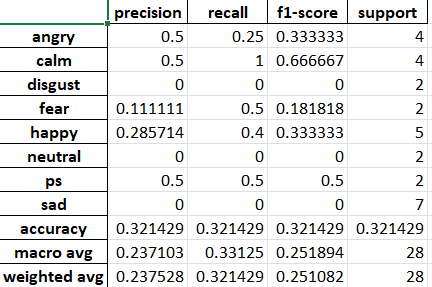
**DATA SPLITTING**

**TRAINING DATASET**

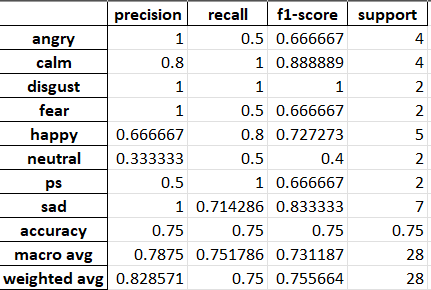
**TESTING DATASET**

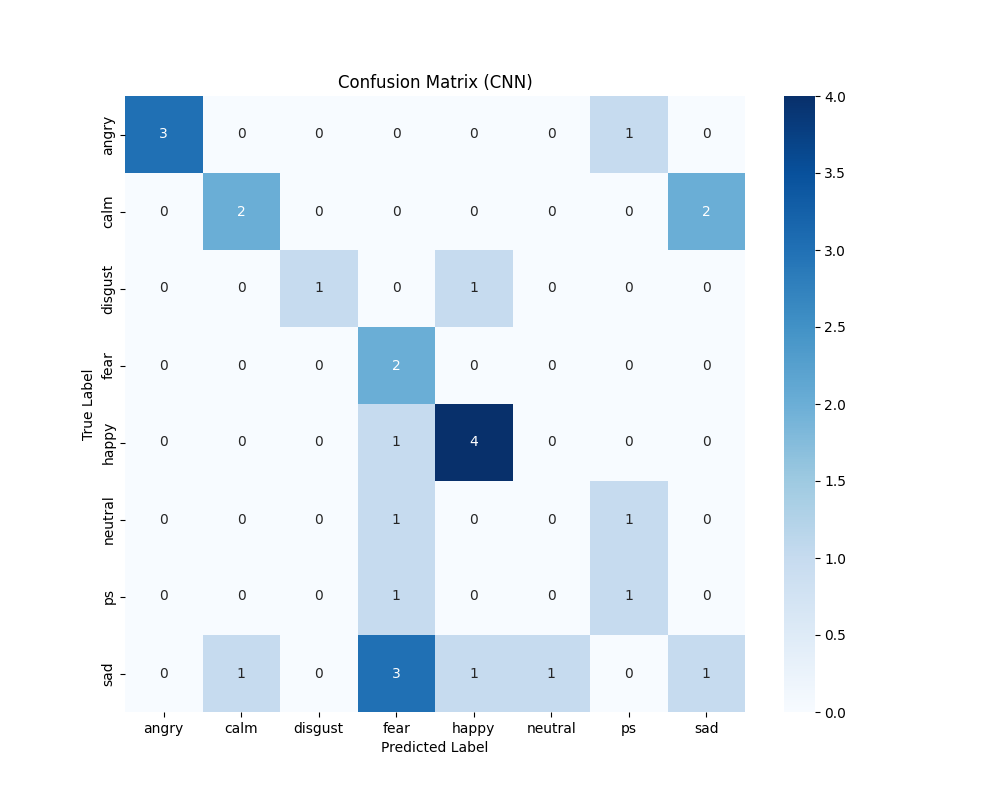
****

**CLASSIFY REPORT CNN**

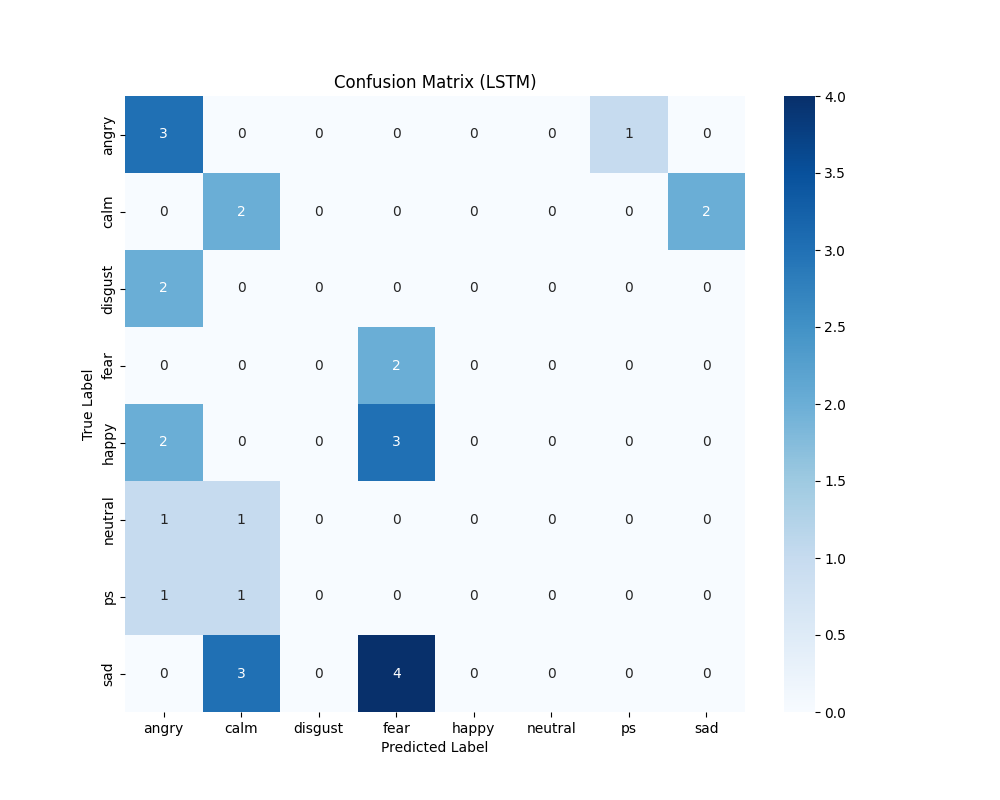
**CLASSIFY REPORT LSTM**

**CLASSIFY REPORT SVM**

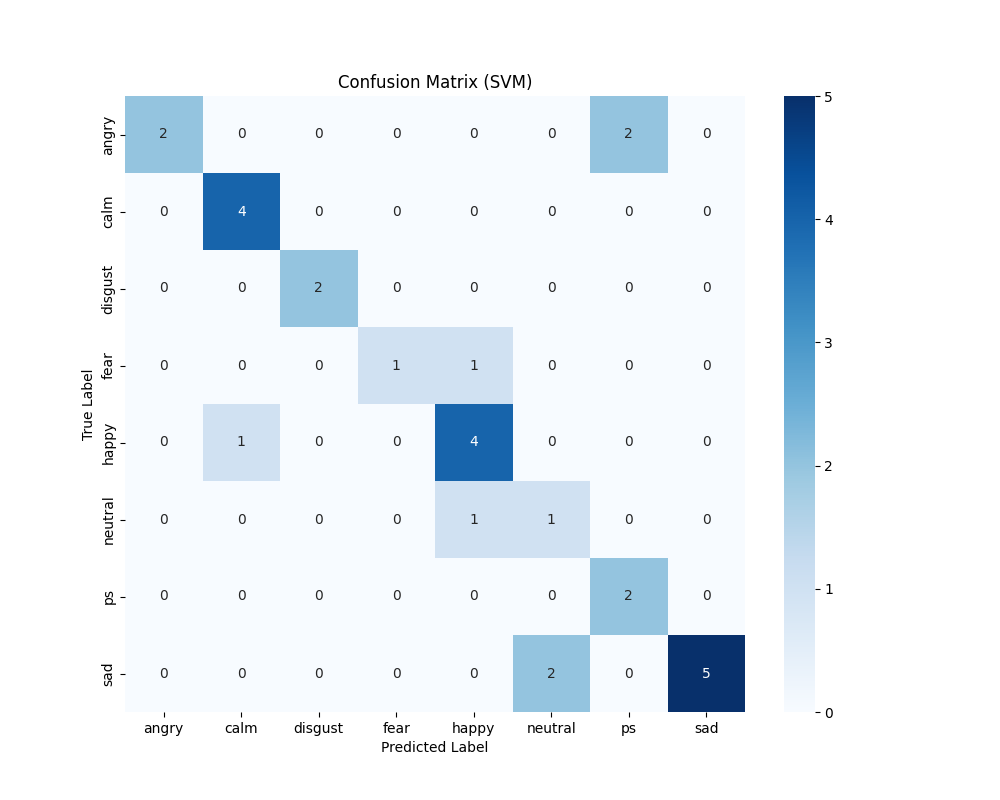


**CONFUSION MATRIX CNN**

**CONFUSION MATRIX LSTM**



**CONFUSION MATRIX SVM**



1. **BIBLIOGRAPHY**

**REFERENCE**

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, „„EERA-ASR: An energy- efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing,‟‟ IEEE Access, vol. 6, pp. 52227– 52237, 2018.

[2] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, „„Animage- based deep spectrum feature representation for the recognition of emotional speech,‟‟ in Proc. 25th ACM MultimediaConf. (MM), 2017, pp. 478–484.

[3] Mustaqeem and S. Kwon, „„A CNN-assisted enhanced audio signal processing for speech emotion recognition,‟‟ Sensors, vol. 20, no. 1, p. 183, 2020.

[4] J. Huang, B. Chen, B. Yao, and W. He, „„ECG arrhythmia classification using STFT- based spectrogram and convolutional neural network,‟‟ IEEE Access, vol. 7, pp. 92871– 92880,2019.

[5] K. Simonyan and A. Zisserman, „„Very deep convolutional networks for large-scale image recognition,‟‟ 2014, arXiv:1409.1556. [Online].

[6] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, „„Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,‟‟ IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 77–86,Jan. 2020

[7] Rao, K. Sreenivasa, et al. "Emotion recognition from speech." International Journal of Computer Science and Information Technologies 3.2 (2012): 3603-3607.

[8] Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001..

[9] Pfister, Tomas. "Emotion Detection from Speech." 2010(2010).

[10] Sapra, Ankur, Nikhil Panwar, and Sohan Panwar. "Emotion recognition from speech." International journal of emerging technology and advanced engineering 3 (2013): 341-345.

[11] Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." International Journal of Applied Information Syatems (IJAIS) (2013): 5-8.

[12] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition44.3 (2011): 572-587.