

Emo-Track



Team Members	Roll Number
Pradip Ghosh	12200220013 (21)
Deep Kumar Goenka	12200220038 (25)
Aishi Paul	12200220039 (45)
Abhishek Chaudhury	12200220056 (09)

Mentor – Dr. Arijit Ghosal

Group No.: 3

Contribution

Method – 1

Aishi Paul

- Integrated feature extraction methods into a single MATLAB file.
- Contributed in Method 1 in analyzing TESS and CREMA-D datasets using different classifiers.

Deep Kumar Goenka

- Implemented Feature Selection using Extra Trees Classifier.
- Contributed in Method 1 in analyzing RAVDESS, SAVEE and EmoDB datasets using different classifiers.

Method – 2

Abhishek Chaudhury

- Extracted features using MATLAB.
- Contributed in Method 2 in analyzing RAVDESS datasets using different classifiers.

Pradip Ghosh

- Extracted features using MATLAB.
- Contributed in Method 2 in analyzing RAVDESS datasets using different classifiers.



Contents



1. Abstract
2. Objective
3. Literature Review
4. Introduction
5. Project Plan
6. Dataset
7. System Block Diagram
8. Feature Extraction
9. Feature Selection
10. Model
11. Accuracy
12. Classification Report
13. Confusion Matrix
14. Application
15. Conclusion
16. Future Work

Abstract

Emo-Track: Sentiment Analysis model through Speech.

Datasets Used: RAVDESS, TESS, SAVEE, EmoDB, CREMA-D.

Emotions Used: Angry, Disgust, Fearful, Happy, Sad.

Feature Extraction: Spectral and Temporal features.

Feature Selection: Extra Trees Classifier algorithm.

Classifiers Tested:

- Support Vector Machine (SVM)
- Random Forest
- Logistic Regression
- Naive Bayes
- Multilayer Perceptron
- Stack Ensemble

Contribution: Advances emotion recognition, enhancing human computer interaction and affective computing.

Objective

- Develop a speech emotion recognition platform to aid individuals with mental health issues.
- Utilize emotion detection technology to analyze the emotional state based on speech patterns.
- Aim to enhance wellbeing across various domains.
- Ensure the platform is user-friendly and accessible to a wide range of individuals.
- Promote speech emotion recognition for multiple applications.

Literature Review

1. J. Guru Monish Amartya, S. Magesh Kumar, "Speech Emotion Recognition in Machine Learning to Improve Accuracy using Novel Support Vector Machine and Compared with Random Forest Algorithm", 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), pp.862-866, 2022.

This study compares SVM and RF for emotion recognition. SVM achieves 90% accuracy with 104 samples, surpassing RF's 71%, highlighting SVM's superiority and the importance of algorithm selection.

2. Rumagit, R. Y., Alexander, G., & Saputra, I. F. (2020). Model Comparison in Speech Emotion Recognition for Indonesian Language. In 5th International Conference on Computer Science and Computational Intelligence. Jakarta, Indonesia: Bina Nusantara University.2020.

This study compares MLP, SVM, and LR for emotion recognition using MFCCs, highlighting the need for tailored approaches for linguistic diversity and the relevance of acoustic-based features in capturing emotional cues.

3. N. P. Poojary, S. G. S. Kumar, and A. K. B. H. (2021). Speech Emotion Recognition Using MLP Classifier. International Journal of Scientific Research in Science and Technology, pp. 218-222, July. 2021.

The study used CNNs for speech emotion recognition, achieving 68% accuracy on the RAVDESS dataset. It highlights CNNs' effectiveness in capturing emotional nuances and the potential of deep learning to enhance human-machine interactions.

4. S. G. Shaila, A. Sindhu, L. Monish, D. Shivamma, and B. Vaishali, "Speech Emotion Recognition Using Machine Learning Approach," in ICAMIDA 2022, ACSR, vol. 105, pp. 592–599, 2023.

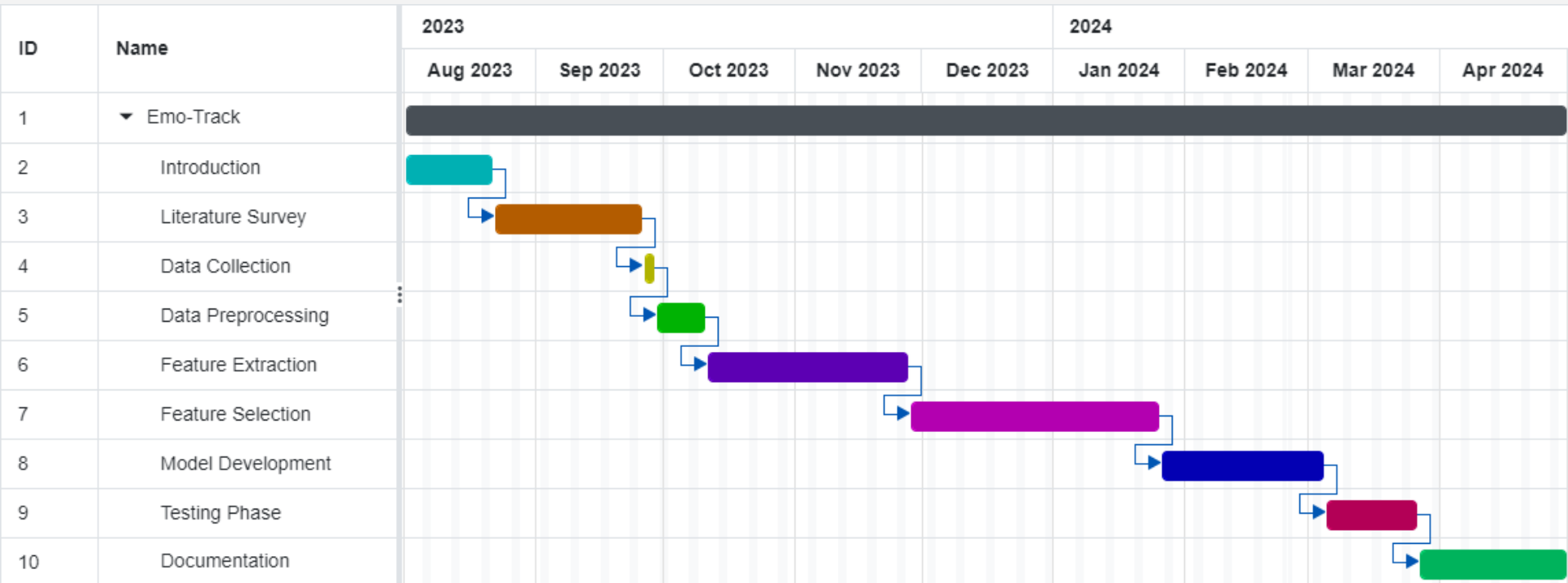
The study highlights the importance of speech emotion recognition (SER) for improving Human-Computer Interaction (HCI). Using the RAVDESS dataset, it addresses challenges like noise removal and classification with models such as Random Forest, MLP, SVM, CNN, and Decision Tree, showing promising accuracy and the potential for more empathetic HCI systems.



Introduction



- Sentiment Analysis through Speech using diverse datasets.
- Utilized RAVDESS, TESS, SAVEE, EmoDB, and CREMA-D Datasets.
- Emotions Recognized are Angry, Disgust, Fearful, Happy, and Sad.
- Extracted Spectral and Temporal features using MATLAB and stored features in a CSV file.
- Employed Extra Trees Classifier algorithm for Feature Selection and selected the top 30 most relevant features.
- Used multiple classifiers like Support Vector Machine (SVM), Random Forest, Naive Bayes, Logistic Regression, Multilayer Perceptron, Stack Ensemble.
- Potential Applications: Human-Computer Interaction, Affective Computing, Mental Health Monitoring.

Project Plan

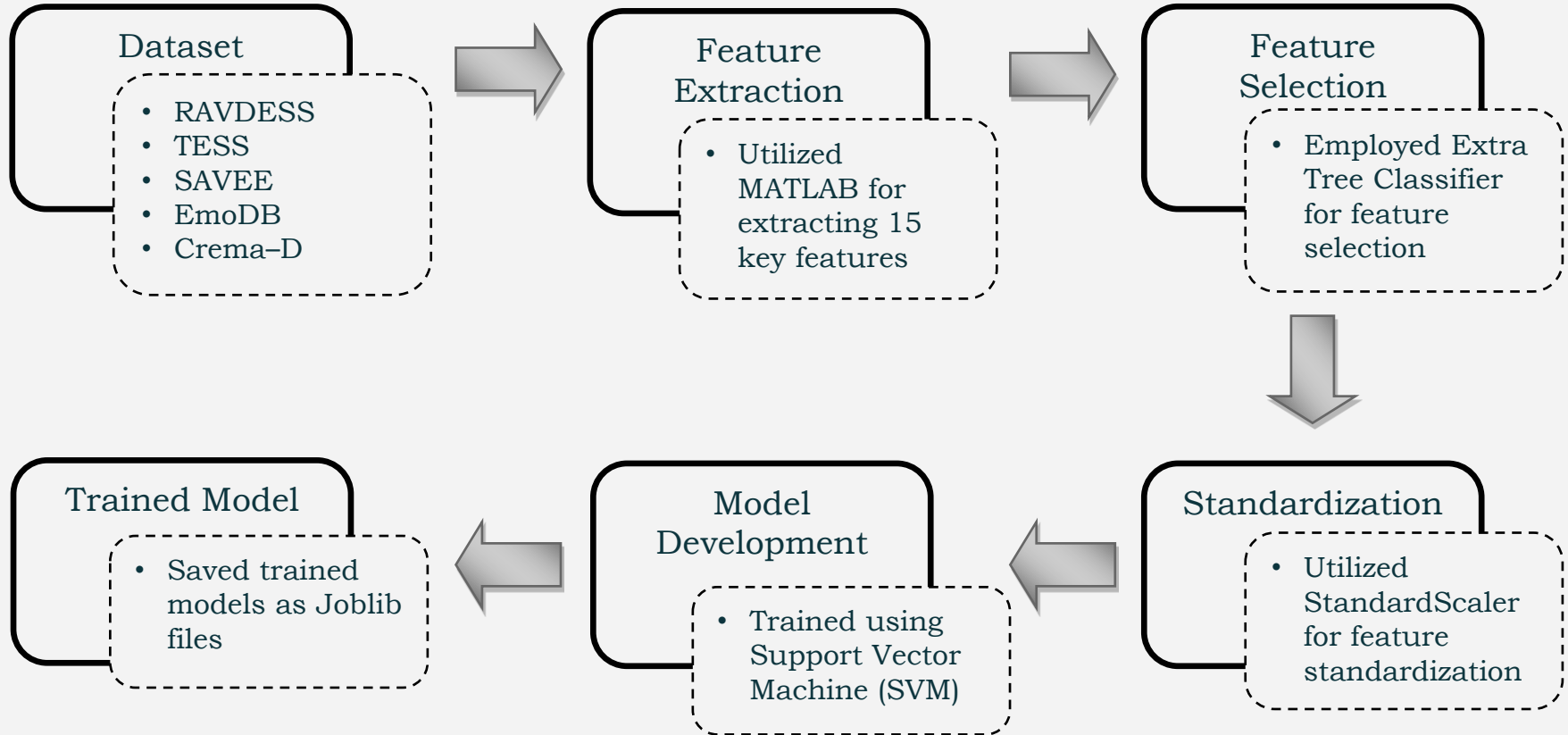


Dataset

1. Ryerson Audio – Visual Database of Emotional Speech and Song (RAVDESS)
 - Emotional Speech Audio dataset
 - 1440 files: 60 trials per actor x 24 actors.
 - 24 professional actors: 12 female, 12 male.
 - Emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust.
 - Emotional intensity: Normal, Strong (except for Neutral).
2. Toronto Emotional Speech Set (TESS)
 - High-quality audio dataset for emotion classification.
 - Total of 2800 WAV format audio files.
 - Consists of recordings by two actresses.
 - Emotions: Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, Neutral.

- 
- 
3. Surrey Audio-Visual Expressed Emotion (SAVEE)
 - Total of 480 WAV format audio files.
 - Recorded from four native English male speakers.
 - Emotions: Anger, disgust, fear, happiness, sadness, surprise, neutral.
 4. Berlin Database of Emotional Speech (EmoDB)
 - Freely available German emotional database.
 - 535 utterances in total.
 - Ten professional speakers: Five males, five females.
 - Emotions: Anger, Boredom, Anxiety, Happiness, Sadness, Disgust, Neutral.
 5. Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)
 - Variety of races and ethnicities represented.
 - Contains 7,442 original audio clips.
 - 91 actors: 48 male, 43 female
 - Emotions: Anger, Disgust, Fear, Happy, Neutral, Sad.
 - Emotion levels: Low, Medium, High, Unspecified.

System Block Diagram



Feature Extraction

From the audio files, 15 features were extracted using MATLAB with a Window size of 2048 and Overlap of 1024. These features include:

- Mel-Frequency Cepstral Coefficients (MFCC): Captures the spectral envelope of a signal.
- Mel Spectrogram: Represents the power spectral density of a signal on a mel scale.
- Spectral Flux: Measures the change in spectral shape over time.
- Spectral Skewness: Describes the asymmetry of the spectral distribution.
- Spectral Slope: Indicates the rate of change of spectral magnitude.
- Spectral Entropy: Measures the amount of information in the spectral distribution.
- Spectral Rolloff: Represents the frequency below which a certain percentage of the total spectral energy lies.
- Chromagram: Captures the energy distribution of musical notes.

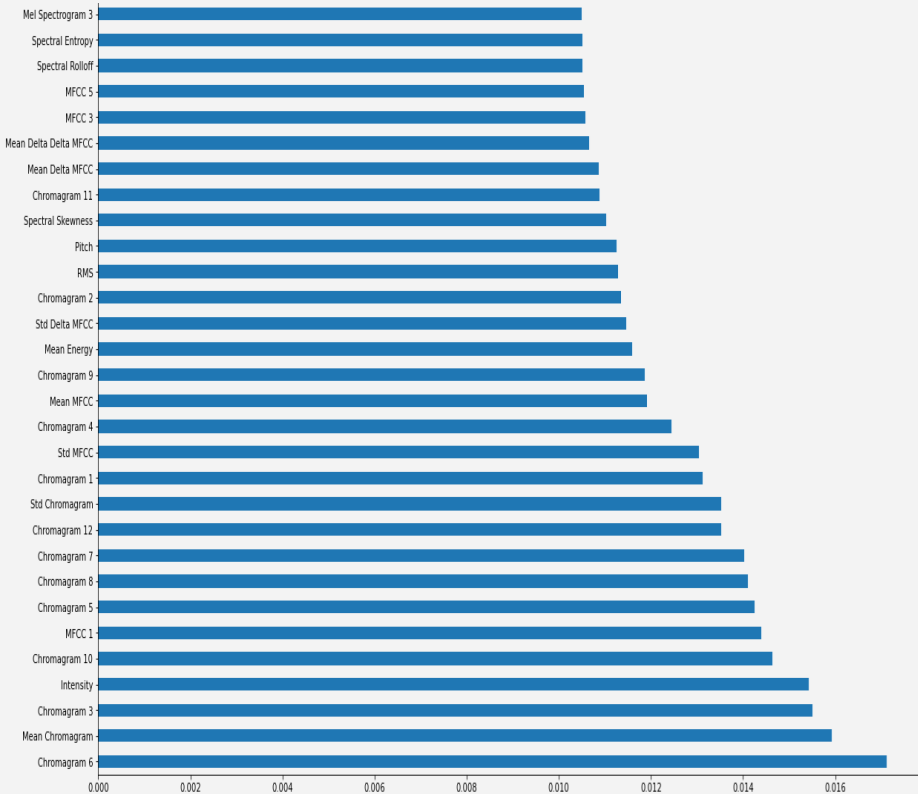
- Linear Predictive Coding (LPC): Models the spectral envelope of a signal using linear prediction.
- Zero Crossing Rate (ZCR): Represents the rate at which the signal changes its sign.
- Energy: Represents the signal's overall energy content.
- Pitch: Indicates the perceived frequency of the signal.
- Intensity: Represents the loudness of the signal.
- Harmonic-to-Noise Ratio (HNR): Measures the ratio of harmonic components to noise in the signal.
- Root Mean Square (RMS): Represents the average power of the signal over time.

Method – 1



Feature Selection

- Extra Tree Classifier is used for Feature Selection.
- Principle: Information Gain determines feature importance.
- Higher Information Gain = More important feature.
- Lower Information Gain = Less important feature.
- Selection of best 30 features based on importance scores.
- StandardScaler applied for uniform scaling of selected features.



Model

- Supervised Learning paradigm was followed.
- Data partitioned into training (80%) and testing (20%) subsets.
- Utilized Scikit-learn, a Python library, for classification tasks.
- Employed multiple classifiers to get the highest accuracy:
 - Support Vector Machine (SVM)
 - Random Forest
 - Naïve Bayes
 - Logistic Regression
- SVM achieved the highest accuracy among all classifiers across datasets.
- SVM's robustness and non-linear boundary handling makes it the better in speech emotion recognition.
- Trained model was saved in a joblib file format for future use.
- Performance evaluation conducted on the testing dataset to assess classifier efficacy.

Accuracy of all Classifiers

	RAVDESS Dataset	TESS Dataset	SAVEE Dataset	EmoDB Dataset	CREMA-D Dataset
Support Vector Machine (SVM)	85.4%	100%	93.3%	95.7%	60.3%
Naive Bayes	50.5%	83.5%	53.3%	76.1%	41.9%
Random Forest	68.2%	99.75%	78.3%	78.3%	52.7%
Logistic Regression	63.0%	99.25%	63.3%	82.6%	53.4%

Accuracy of SVM Classifier

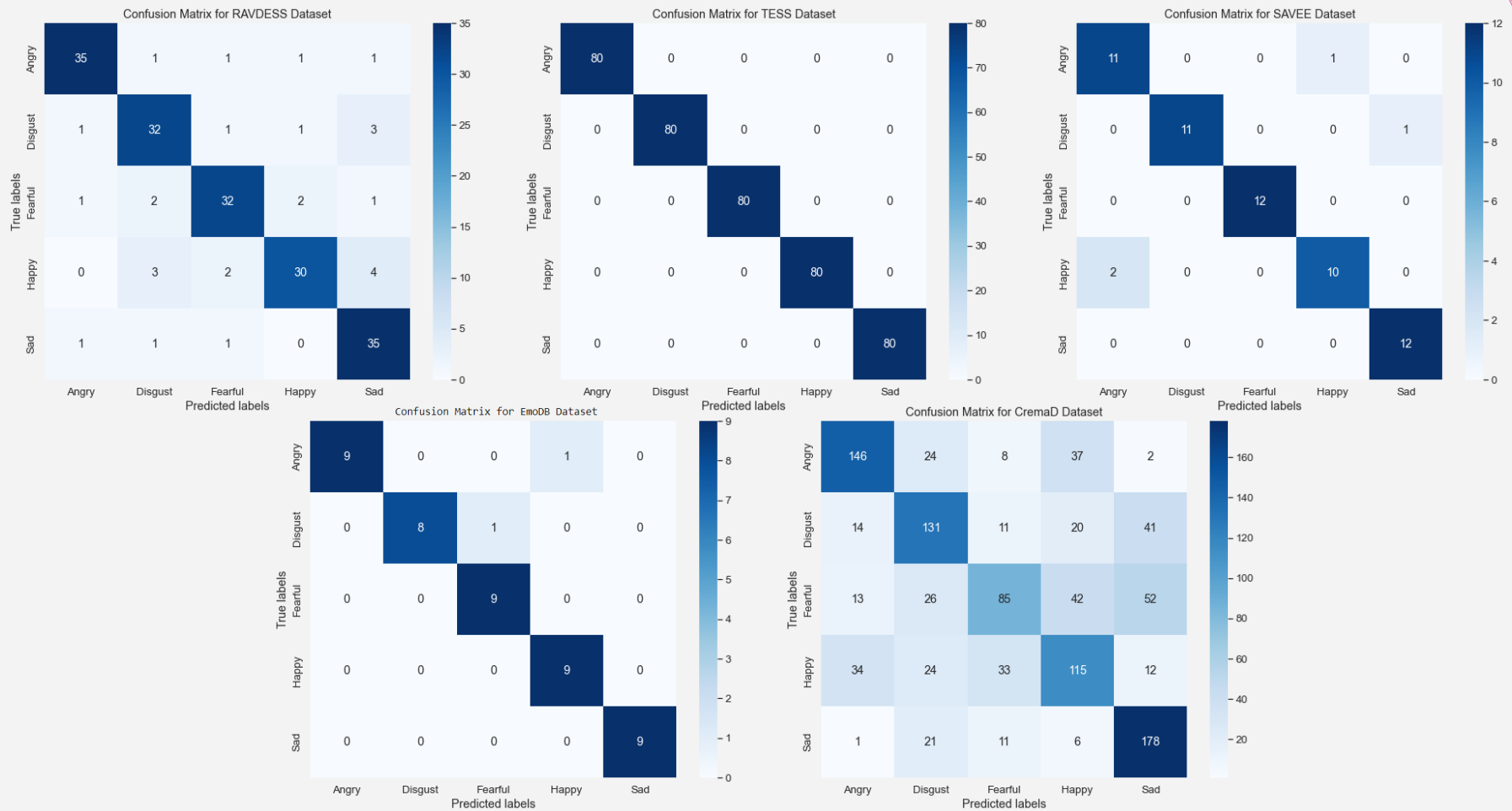
DATASET	ACCURACY (%)
RAVDESS	85.4
TESS	100
SAVEE	93.3
EmoDB	95.7
CREMAD	60.3

Classification Report

Classification Report of RAVDESS Dataset					Classification Report of TESS Dataset					Classification Report of SAVEE Dataset				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
Angry	0.92	0.90	0.91	39	Angry	1.00	1.00	1.00	80	Angry	0.85	0.92	0.88	12
Disgust	0.82	0.84	0.83	38	Disgust	1.00	1.00	1.00	80	Disgust	1.00	0.92	0.96	12
Fearful	0.86	0.84	0.85	38	Fearful	1.00	1.00	1.00	80	Fearful	1.00	1.00	1.00	12
Happy	0.88	0.77	0.82	39	Happy	1.00	1.00	1.00	80	Happy	0.91	0.83	0.87	12
Sad	0.80	0.92	0.85	38	Sad	1.00	1.00	1.00	80	Sad	0.92	1.00	0.96	12
accuracy			0.85	192	accuracy			1.00	400	accuracy			0.93	60
macro avg	0.86	0.85	0.85	192	macro avg	1.00	1.00	1.00	400	macro avg	0.94	0.93	0.93	60
weighted avg	0.86	0.85	0.85	192	weighted avg	1.00	1.00	1.00	400	weighted avg	0.94	0.93	0.93	60

Classification Report of EmoDB Dataset					Classification Report of CremaD Dataset				
	precision	recall	f1-score	support		precision	recall	f1-score	support
Angry	1.00	0.90	0.95	10	Angry	0.70	0.67	0.69	217
Disgust	1.00	0.89	0.94	9	Disgust	0.58	0.60	0.59	217
Fearful	0.90	1.00	0.95	9	Fearful	0.57	0.39	0.46	218
Happy	0.90	1.00	0.95	9	Happy	0.52	0.53	0.53	218
Sad	1.00	1.00	1.00	9	Sad	0.62	0.82	0.71	217
accuracy			0.96	46	accuracy			0.60	1087
macro avg	0.96	0.96	0.96	46	macro avg	0.60	0.60	0.60	1087
weighted avg	0.96	0.96	0.96	46	weighted avg	0.60	0.60	0.60	1087

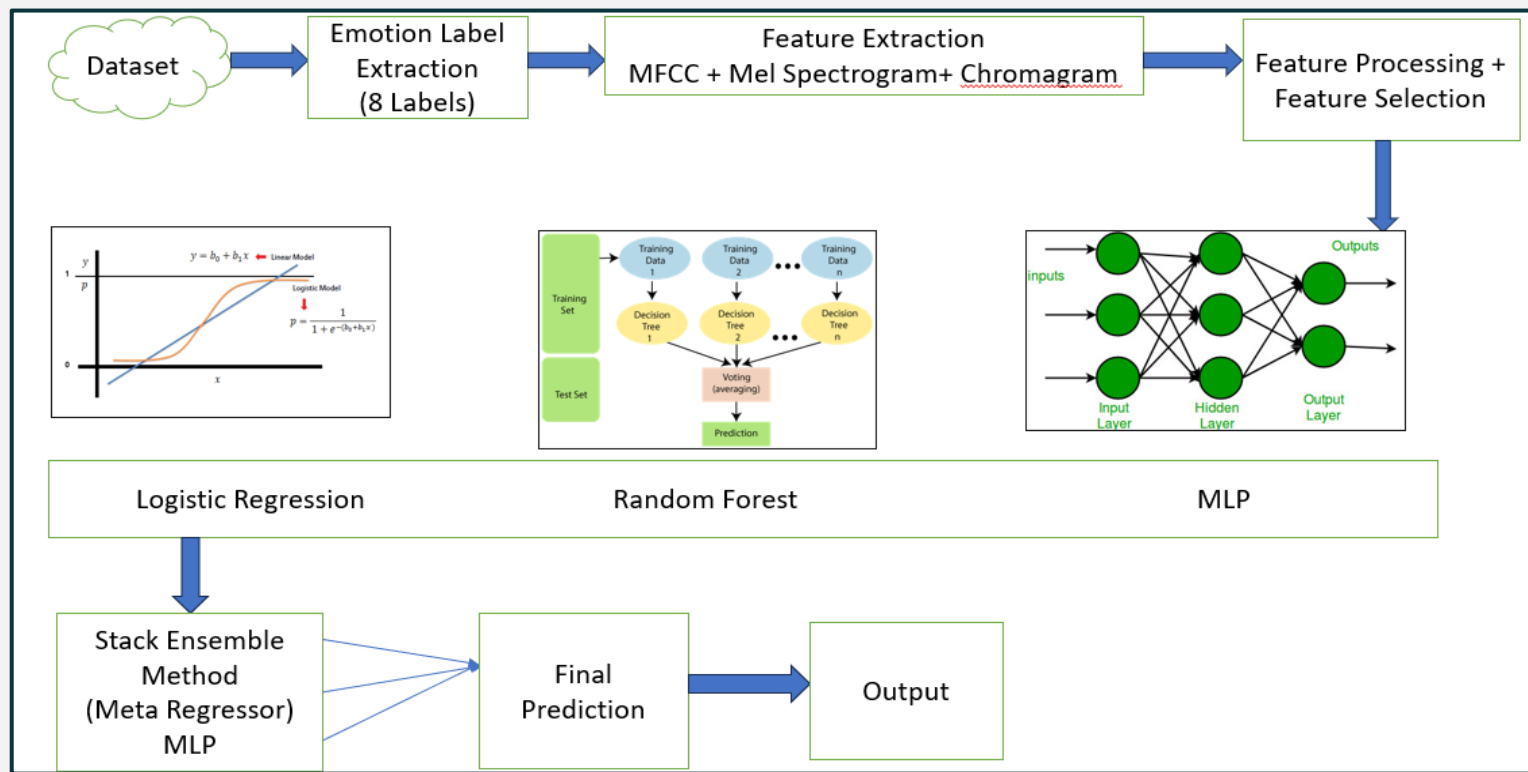
Confusion Matrix



Method – 2



➤ Flow Chart:



➤ Feature Selection:

The feature selection process is based on the few points.

- 1. Relevance:** Features selected should capture essential information pertinent to the task, avoiding noise or unnecessary complexity.
- 2. Discriminative Power:** Chosen features must effectively differentiate between various categories in the dataset, enhancing classification accuracy.
- 3. Robustness:** Selected features should remain informative despite variations or noise, ensuring consistent performance across different conditions.

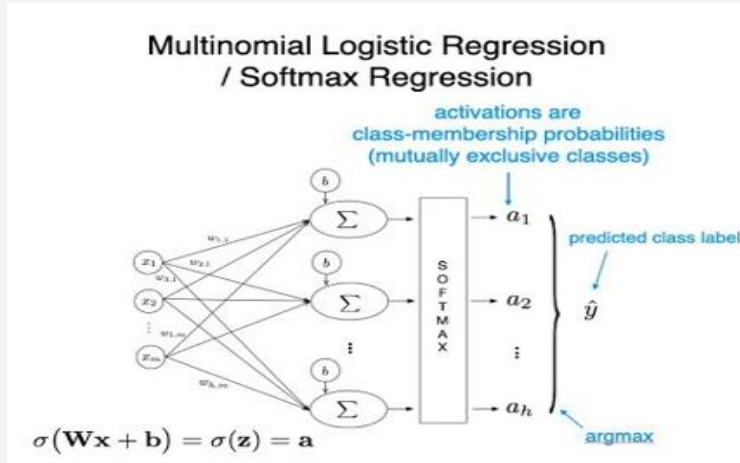
Our project is based on supervised learning, where the input variable are the feature matrix, the numerical value (the features extracted from the audio files in RAVDESS) and output variable is the emotion labels, the categorical value. So in the various feature selection method, wrapper and embedded method are useful. Kendall's rank coefficient (nonlinear), Recursive Feature Elimination, Sequential Floating Forward Search, Backward elimination are helpful in this case.

After the feature selection we have selected 3 features (MFCC, Mel Spectrogram, Chromagram) from all other features based on the importance and contribution of each feature in classification of emotion classes.

➤ Classifiers:

>>Logistic Regression (Multinomial):

- 1. One-vs-Rest Strategy:** Trains individual models for each emotion class (e.g., happy vs. all other emotions, sad vs. all other emotions).
- 2. Feature Transformation & Probability Estimation:** Each model calculates the probability of a data point belonging to its specific emotion class based on the features and learned weights.
- 3. Maximum Probability Wins:** The emotion class with the highest predicted probability across all models is chosen as the final classification result for the data point.

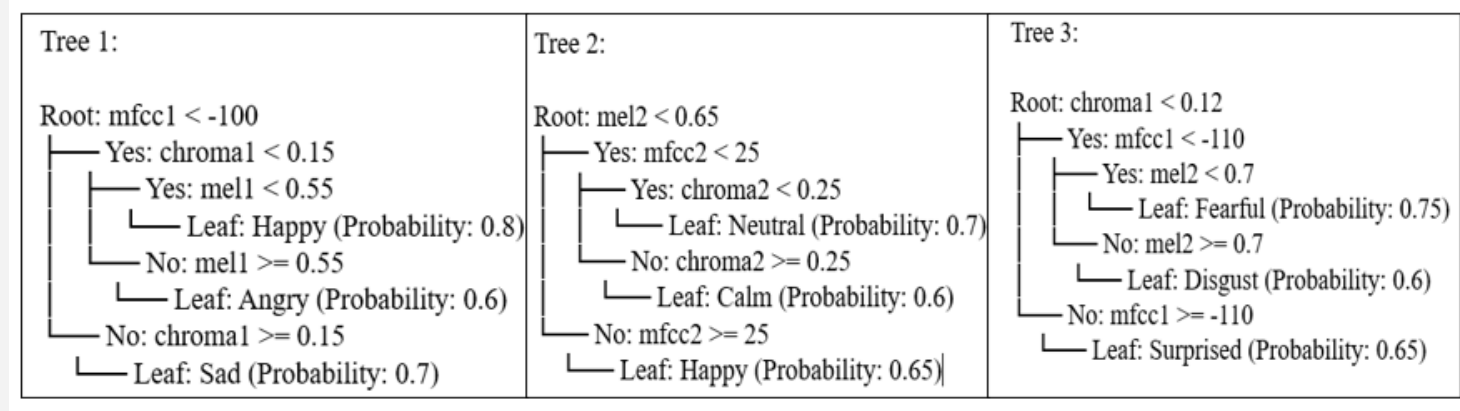


$$a_i = P(y = i | X) = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}}$$

$$P(\text{Emotion} | X) = \frac{e^{\wedge \text{Emotion}}}{e^{\wedge \text{Neutral}} + e^{\wedge \text{Calm}} + e^{\wedge \text{Happy}} + e^{\wedge \text{Sad}} + e^{\wedge \text{Disgust}} + e^{\wedge \text{Angry}} + e^{\wedge \text{Fearful}} + e^{\wedge \text{Surprised}}}$$

>>Random Forest:

- 1. Ensemble of Decision Trees:** Each decision tree uses a random subset of features and split points to create classification rules.
- 2. Feature Importance & Splitting:** Trees consider random features at each split, capturing diverse decision boundaries and reducing overfitting.
- 3. Majority Vote for Prediction:** The emotion class with the most votes from all trees is chosen as the final prediction, improving accuracy and reducing the influence of any single tree.



>>Multi-layer Perceptron:

1. **Feature Learning & Non-Linear Relationships:** MLPs learn complex, non-linear relationships between features (Mel Spectrogram, MFCCs, etc.) and emotions, capturing intricate data patterns.

$$\text{Weighted Sum(Input Layer)} = z_j = \sum (w_{ji} \cdot x_i) + b_j$$

$$\text{Output layer} = O_1 = (A_1 \times V_{11}) + (A_2 \times V_{12}) + \dots + (A_h \times V_{1h}) + c_1$$

2. **Hidden Layers & Activation Functions:** Multiple hidden layers with activation functions (e.g., ReLU) introduce non-linearity, enabling the network to learn complex decision boundaries.

$$\text{Activation Function (ReLU): } A_1 = \max(0, Z_1)$$

3. **Softmax Output & Probability Distribution:** The final layer uses a softmax function to assign probabilities to each emotion class. The class with the highest probability is chosen as the predicted emotion.

Softmax Calculation: Convert logits O_k to probabilities:

$$P_1 = \frac{e^{O_1}}{1 + e^{O_1} + e^{O_2} + \dots + e^{O_m}}$$

$$\text{Final Prediction: } P_k = \text{Max}(P_1, P_2, P_3, \dots, P_n)$$

>>Stack Ensemble Method:

1. Two-Stage Learning:

Stage 1 (Base Regressors):

- a. Individual Models
- b. Predict Probabilities

Stage 2 (Meta-Classifer):

- a. Separate Meta-Classifier
- b. New Features from Predictions
- c. Higher-Level Representation
- d. Combining Features

2. Leveraging Base Regressor Strengths: Meta-classifier leverages Random Forest's ability to capture complex decision boundaries and MLP's proficiency in learning non-linear relationships, enhancing prediction by combining their strengths.

3. Improved Generalizability: Meta-classifier learns from combined outputs, potentially improving generalization on unseen data.

➤ **Accuracy:**

Random Forest Test Accuracy: 52%

Logistic Regression Test Accuracy: 55%

MLP Test Accuracy: 59%

Stacking Test Accuracy: 66.6%

1. Promising Results with Stacking

2. Systematic Improvement: Sequentially improved classification accuracy with different classifiers

3. Data-Driven Approach Informs Future Steps: Data-driven model selection led to the identification of Stacking as the most suitable approach for enhanced emotion recognition on the RAVDESS dataset, enabling focused refinement efforts.

➤ Classification Report:

	precision	recall	f1-score	support
Angry	0.63	0.90	0.74	29
Calm	0.70	0.62	0.66	42
Disgust	0.65	0.46	0.54	48
Fearful	0.59	0.59	0.59	37
Happy	0.46	0.51	0.49	37
Neutral	0.60	0.38	0.46	24
Sad	0.47	0.45	0.46	38
Surprised	0.49	0.70	0.57	33
accuracy			0.57	288
macro avg	0.58	0.58	0.57	288
weighted avg	0.59	0.59	0.58	288

Logistic Regression

	precision	recall	f1-score	support
Angry	0.70	0.90	0.79	29
Calm	0.63	0.69	0.66	42
Disgust	0.75	0.62	0.68	48
Fearful	0.69	0.65	0.67	37
Happy	0.58	0.57	0.58	37
Neutral	0.45	0.42	0.43	24
Sad	0.68	0.55	0.61	38
Surprised	0.59	0.73	0.65	33
accuracy			0.64	288
macro avg	0.59	0.60	0.61	288
weighted avg	0.60	0.62	0.62	288

MLP

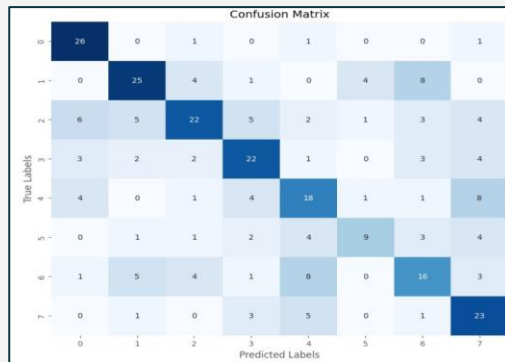
	precision	recall	f1-score	support
Angry	0.58	0.66	0.61	29
Calm	0.63	0.88	0.73	42
Disgust	0.62	0.54	0.58	48
Fearful	0.59	0.35	0.44	37
Happy	0.44	0.46	0.45	37
Neutral	0.47	0.29	0.36	24
Sad	0.48	0.29	0.36	38
Surprised	0.40	0.67	0.50	33
accuracy			0.53	288
macro avg	0.52	0.52	0.50	288
weighted avg	0.53	0.53	0.51	288

Random Forest

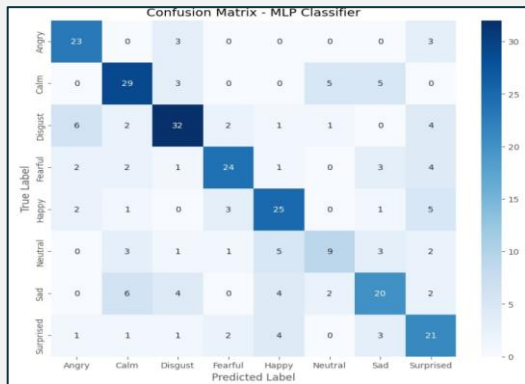
	precision	recall	f1-score	support
Angry	0.70	0.90	0.79	29
Calm	0.74	0.74	0.74	42
Disgust	0.68	0.67	0.67	48
Fearful	0.75	0.65	0.70	37
Happy	0.56	0.51	0.54	37
Neutral	0.50	0.42	0.45	24
Sad	0.60	0.47	0.53	38
Surprised	0.50	0.70	0.58	33
accuracy			0.64	288
macro avg	0.66	0.66	0.65	288
weighted avg	0.67	0.67	0.66	288

Stack Ensemble

➤ Confusion Matrix:

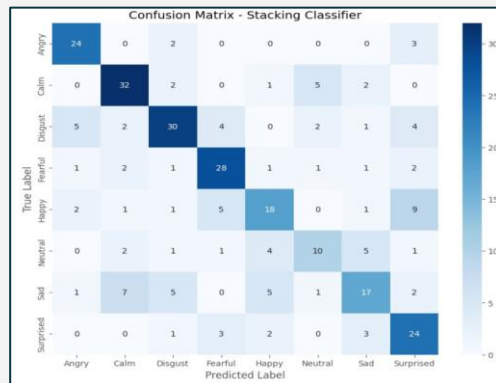


Random Forest



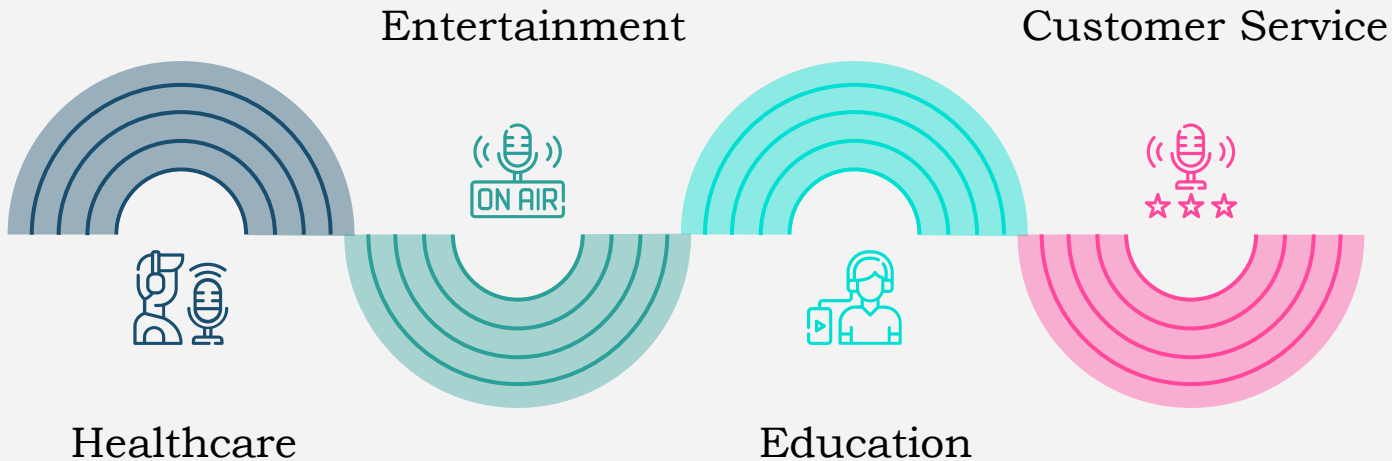
MLP

Logistic Regression



Stack Ensemble

Applications



Conclusion

- The "Emo-Track" project is a pioneering endeavor in sentiment analysis through speech, employing advanced machine learning methodologies.
- State-of-the-art techniques facilitate accurate classification of emotional content.
- The project holds promise for diverse applications, including real-time customer sentiment analysis and mental health monitoring.
- It underscores the importance of feature extraction in unlocking emotional insights within speech data.
- Emo-Track continues as a trailblazer, enriching emotional wellbeing and user experiences across domains.

Future Work

Shortcomings

- Data mainly consists of American and European accents.
- Mix of English and Spanish recordings affects system performance for speakers with different speaking styles.

Improvements

- Training model on user's audio input can mitigate issues to some extent.





Thank You