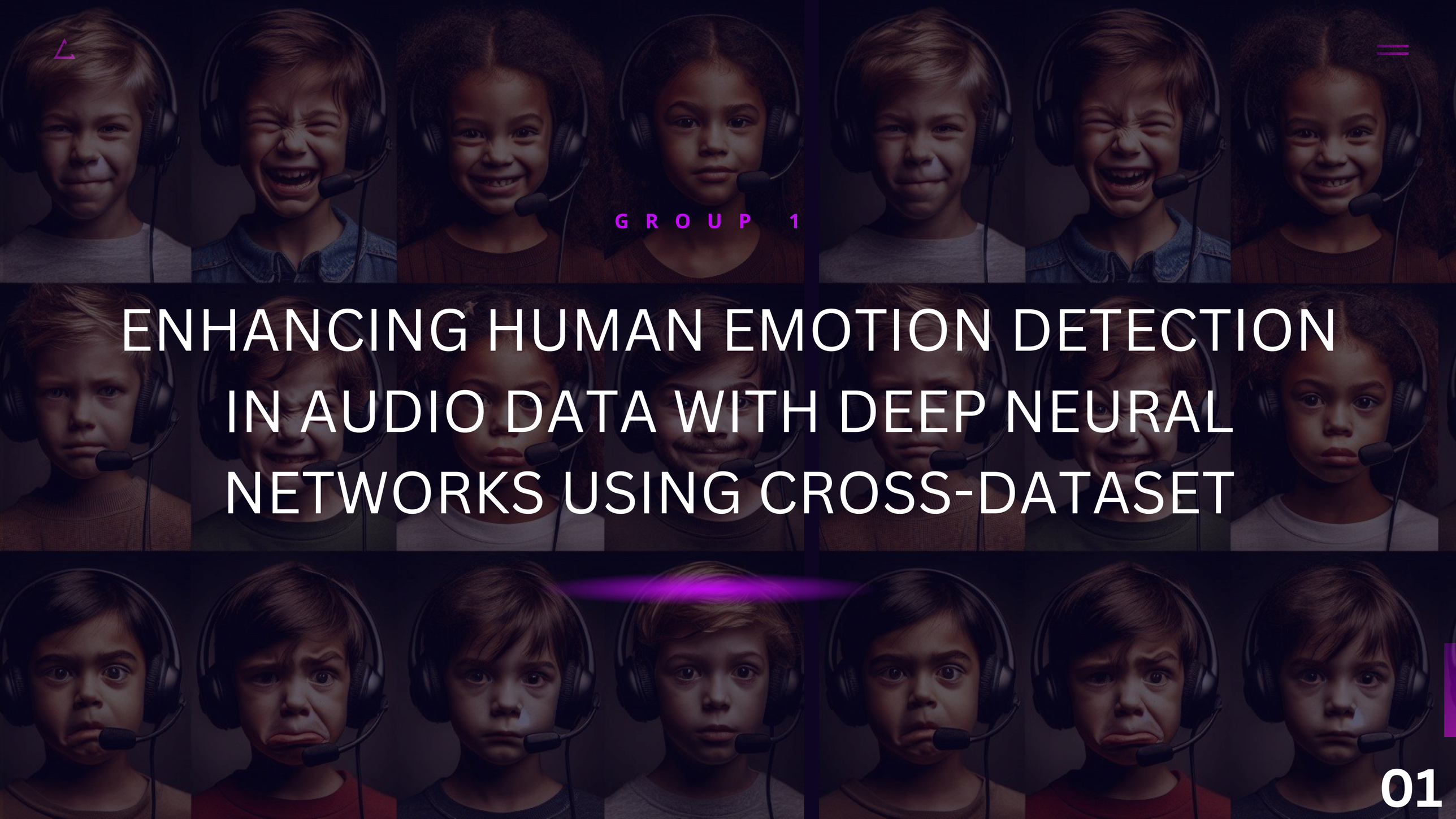


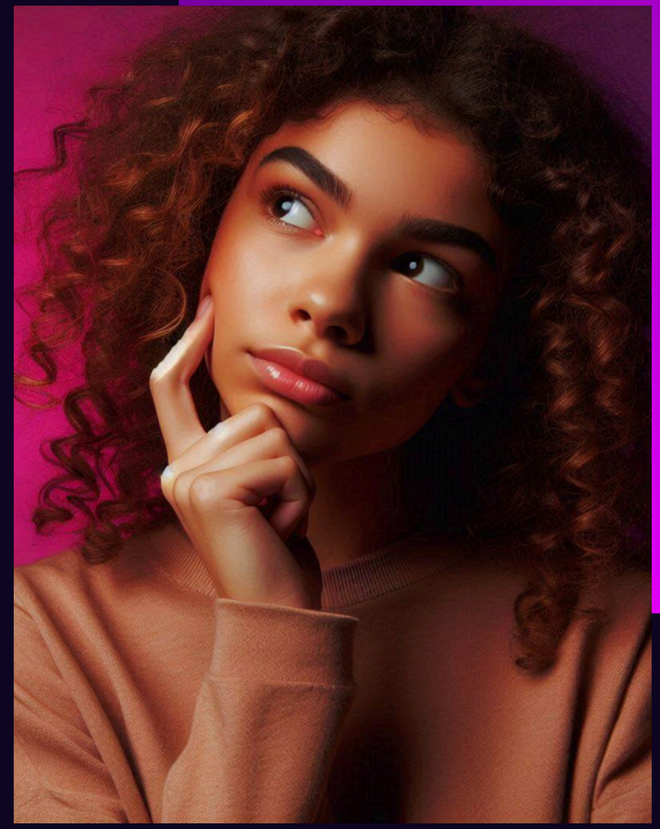
GROUP 1

ENHANCING HUMAN EMOTION DETECTION IN AUDIO DATA WITH DEEP NEURAL NETWORKS USING CROSS-DATASET

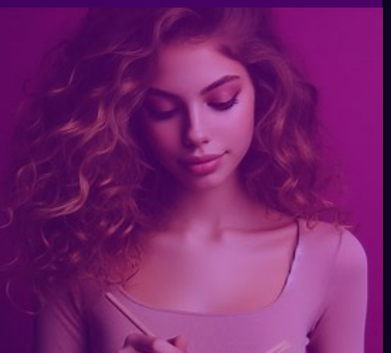
INTRODUCTION

- Interpreting spoken language is complex.
- Machines face significant challenges due to the myriad variables that influence sound waves during speech.
- Developing a model to detect emotions from speech using deep neural networks, specifically Convolutional Neural Networks (CNNs), and Mel-Frequency Cepstral Coefficients (MFCCs).
- Our model aims to accurately classify various emotions, enhancing interactions between humans and machines.

RESEARCH PROBLEM



- **Complexity of Human Emotions.**
- **Limited Progress in Audio-based Emotion Classification.**
- **Challenges with Existing Models:**
  - Difficulty in accurately classifying emotions across different datasets
  - Performance drops in real-world scenarios with background noise and speech variations
  - Need for more robust and generalizable models to handle diverse audio inputs



OBJECTIVES

RO1

To develop a Versatile Emotion Classification Model: Construct a neural network-based system that consistently identifies emotions from various audio datasets, enhancing its adaptability to different data characteristics.

R02

To enhance Model Generalization: Focus on improving the model's ability to maintain high accuracy when applied to new, unseen datasets, reducing performance discrepancies between different types of audio inputs.

R03

To implement Advanced Feature Extraction Techniques: Employ sophisticated signal processing methods to extract robust features from audio data that are effective across diverse datasets and help in accurately detecting emotions.

RELATED WORKS

Study	Dataset	Methods Used	Results	Comments
A real-time emotion recognition from speech using gradient boosting. (Iqbal et al,2019)	RAVDESS	Gradient Boosting, KNN, SVM	- RAVDESS (male): SVM & KNN: 100% accuracy in anger and neutral. Gradient Boosting better in happiness and sadness - RAVDESS (female): SVM: 100% in anger. Good overall except sadness. KNN: 87% in anger, 100% in neutral; Gradient Boosting poor in anger and neutral - Combined: SVM & KNN good in anger and neutral. KNN poor in happiness and sadness. SVM higher accuracy than gender-based datasets	- Performance varies by emotion and gender - SVM generally performs well except in sadness for females -Gradient Boosting outperforms in some emotions
Speech Emotion Recognition of Intelligent Virtual Companion for Solitudinarian (Alnahhas et al., 2022)	RAVDESS	Multilayer Perceptron (MLP) classifier, features: MFCC, Chroma, Mel Frequency Cepstrum, adaptive learning rate, 500 iterations	Achieved accuracy of 78.47%, better at detecting fearful, sad, calm, and angry emotions; weaker with happy and calm	Suggests further improvement through additional training datasets, transfer learning, and enhanced feature selection
Recognizing emotion from singing and speaking using shared models. (Zhang et al(2016)	RAVDESS	Simple model, single-task hierarchical model, multi-task hierarchical model (using Directed Acyclic Graph SVM (DAGSVM)[14])	multi-task hierarchical models for emotion recognition from both speech and song are more effective than single-task models when utilizing a common set of features.	- Focuses on both speech and song - Classifies only some of the available emotions

RESEARCH GAP

- Current Issue in Emotion Recognition from Audio:
  - Models excel on single datasets (e.g., RAVDESS).
  - Performance drops significantly with different datasets.

Critical Gap:

- Existing models lack flexibility and struggle to maintain accuracy across diverse datasets.

Proposed Model

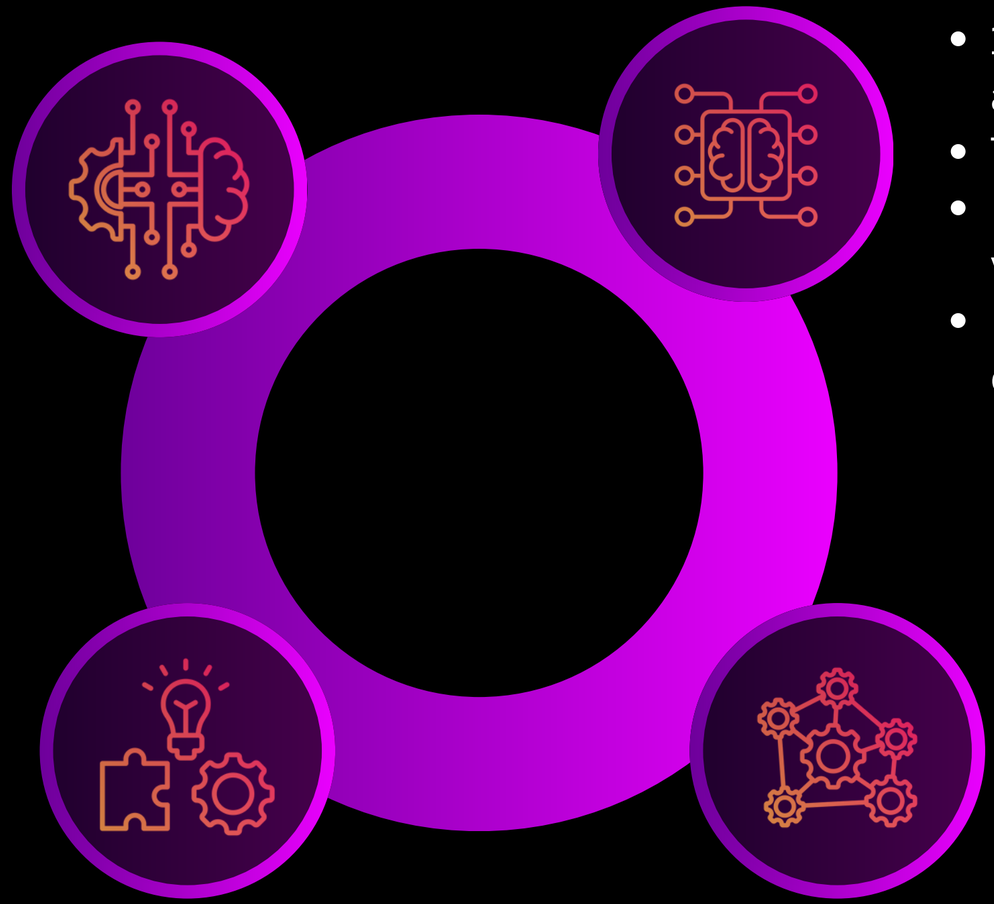
CLASSIFICATION OUTPUT

- Final Dense Layer
- Outputs 8 emotion classes
  - 0: 'neutral'
  - 1: 'calm'
  - 2: 'happy'
  - 3: 'sad'
  - 4: 'angry'
  - 5: 'fearful'
  - 6: 'disgust'
  - 7: 'surprised'
- Softmax Activation

FEATURE EXTRACTION WITH MFCC

Mel-frequency cepstral coefficients (MFCC):

- 40 features per audio file.
- Represents the amplitude spectrum in a compact form.



DEEP LEARNING ARCHITECTURE

- Input: 40 features per audio file.
- 1D CNN with ReLU activation, dropout, and max-pooling, Batch Normalization.
- Total layers: 45
- Residual Blocks concept to address the vanishing gradient problem
- Dense layer with softmax activation for classification.

AUDIO PREPROCESSING

- Frame Division
- Discrete Fourier Transformation (DFT)
- Logarithm of Amplitude Spectrum
- Mel-Frequency Scale Normalization

METHODOLOGY

1. **Analyze the Challenges of speech ecognitions and Literature Review**

2. **Data Collection**
used the TESS and RAVDESS datasets
combined them to create one dataset.
Resources

3. **Data Preprocessing**
Data Augmentation,
Label Encoding, Data Splitting,
Normalization, Data Transformation
(e.g., 03-01-**06**-01-02-01-12.wav).

4. **Feature Extraction**
Mel-frequency cepstral coefficients (MFCC) and Saving the features as a joblib X and Y files

5. **Model training and Testing**
33%-Testing
77%-Training

6. **Live Prediction**
Read resources to support your hypothesis.

RESEARCH OUTCOMES

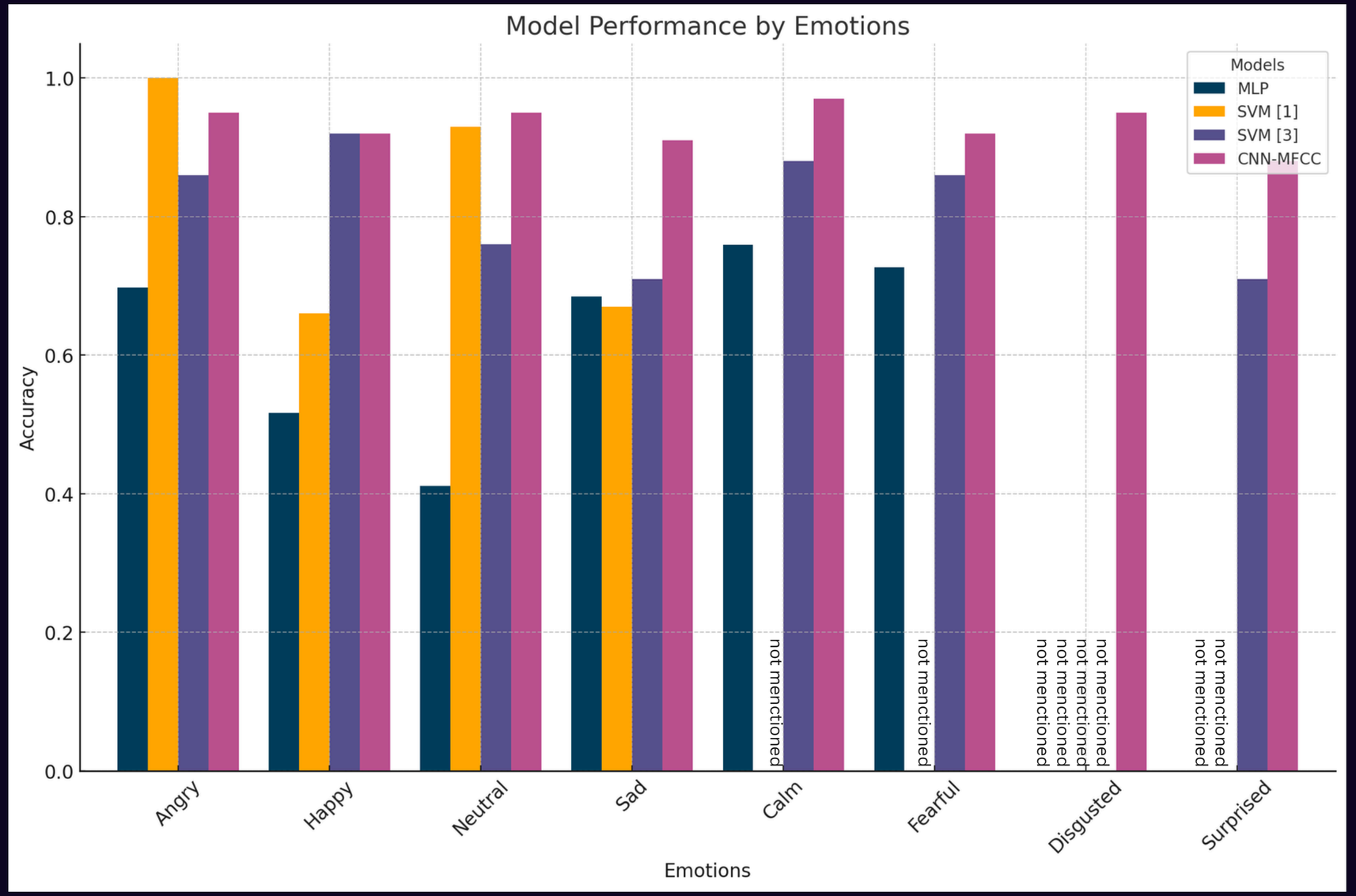
Model Performance

Emotion	Precision	Recall	F1-score	Support
0-Neutral	0.95	0.94	0.95	342
1-Calm	0.94	1.00	0.97	360
2-Happy	0.94	0.90	0.92	351
3-Sad	0.92	0.90	0.91	413
4-Angry	0.94	0.96	0.95	413
5-Fearful	0.91	0.94	0.92	387
6-Disgust	0.96	0.93	0.94	320
7-Suprised	0.95	0.94	0.95	320
Accuracy	-	-	0.94	2906
Macro avg	0.94	0.94	0.94	2906
Weighted avg	0.94	0.94	0.94	2906

RESEARCH OUTCOMES

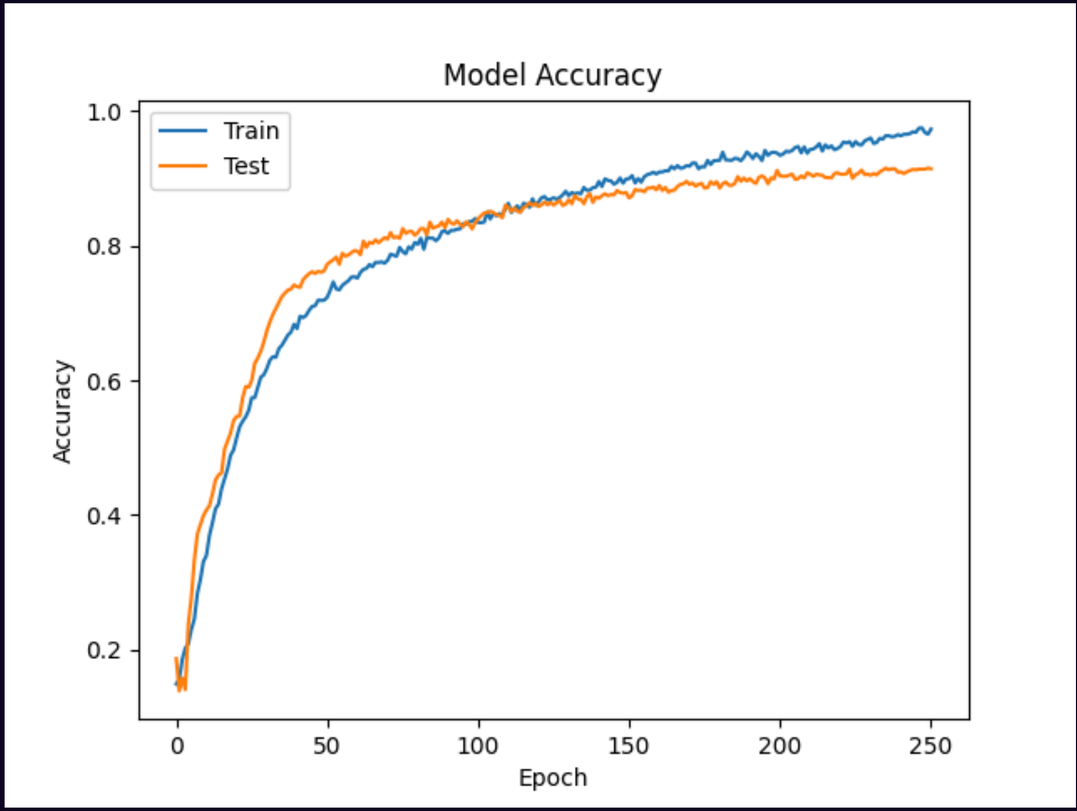
Class	MLP	(Iqbal et al,2019) SVM	(Zhang et al(2016) SVM	CNN-MFCC
Angry	0.698	1.0	0.86	0.95
Happy	0.517	0.66	0.92	0.92
Neutral	0.411	0.93	0.76	0.95
Sad	0.685	0.67	0.71	0.91
Calm	0.759	-	0.88	0.97
Fearful	0.727	-	0.86	0.92
Disgusted	-	-	-	0.95
Surprised	-	-	0.71	0.88

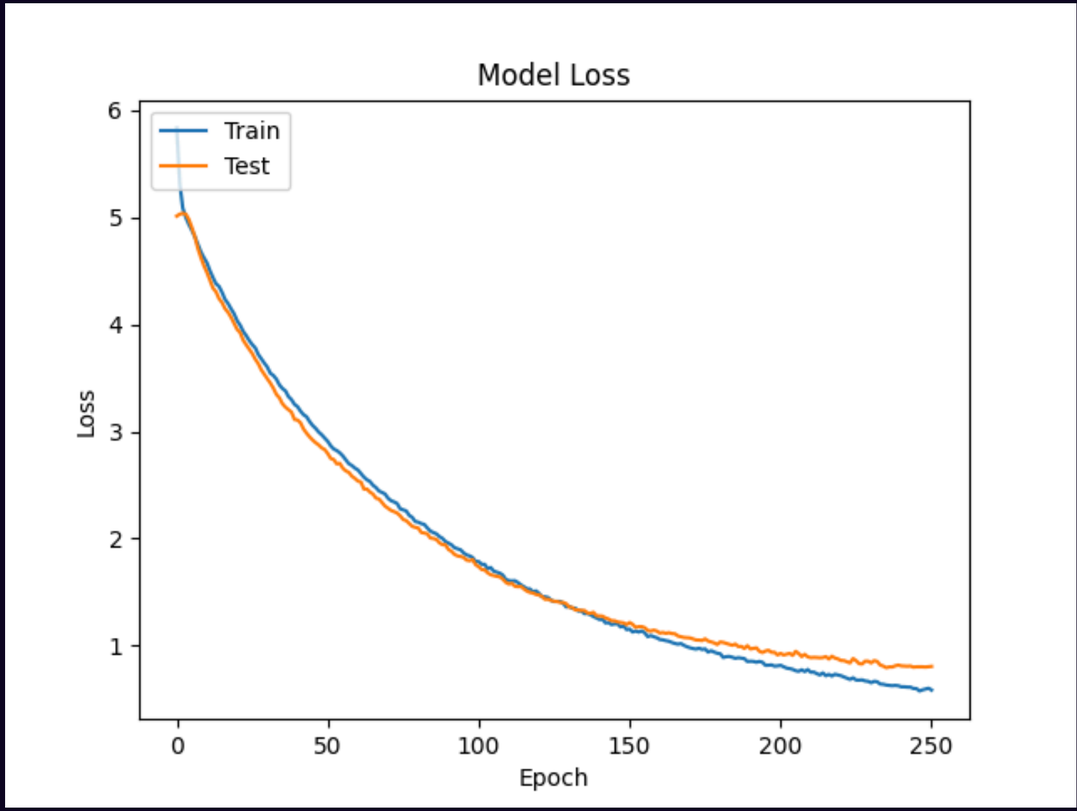
RESEARCH OUTCOMES



RESEARCH OUTCOMES

Fine Tuning of the Model body text





CONCLUTION

A model that combines different datasets for emotion recognition from audio achieves high performance across various new datasets, ensuring accurate predictions in different situations.

RO1 To develop a Versatile Emotion Classification Model

RO2 To enhance Model Generalization

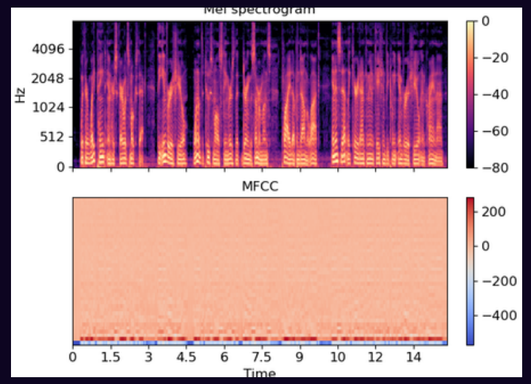
RO3 To implement Advanced Feature Extraction Techniques

RAVDASS +TESS

```
1/1 [==============================] -
Prediction is: sad
```

```
              accuracy                           0.94      2906
             macro avg       0.94      0.94      0.94      2906
          weighted avg       0.94      0.94      0.94      2906

[[322  11   1   5   0   1   1   1]
 [  0 360   0   0   0   0   0   0]
 [  0   3 315   2  10  12   2   7]
 [ 14   6   5 370   1  15   0   2]
 [  0   2   2   3 396   2   6   2]
 [  0   0   3  16   5 362   0   1]
 [  1   0   3   4   8   4 298   2]
 [  1   0   5   2   2   3   5 302]]
```



MEET OUR TEAM

Minuli Kannangara - IM/2019/032

Shiran SuriyaPathiraja - IM/2019/068

Sandushi Weraduwa - IM/2019/112

THANK YOU

Q & A

Resources

DATASET

RAVDESS

- RAVDESS Dataset: The Ryerson Audio-Visual Database of Emotional Speech and Songs.
- It contains 7,356 files from 24 actors (12 female, 12 male), vocalizing two lexically-matched statements in North American English (speech + songs)
- Emotions included are neutral, calm, happy, sad, angry, fearful, disgust, and surprise.
- Has 192 audio files to each emotion
- file naming convention:
  - (e.g., 03-01-**06**-01-02-01-12.wav).

TESS

- The Toronto Emotional Speech Set (TESS) includes recordings of 200 target words spoken in the carrier phrase "Say the word _".
- There are 2800 recordings per actor.
- The dataset includes recordings from two actresses, aged 26 and 64, simulating seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.
- Has 400 audio files for each emotion.

Combined Dataset

- Each has 400+192 = 592 files (except calm)
- calm has only 192 files from RAVDESS
  - technique : Data Augmentation (Duplicate 192*2 = 384)-> 384+192=576