Moodify

Speech Emotion-Based Music Recommendation System

PROJECT OVERVIEW

Problem Statement:

- Traditional music recommendation systems primarily rely on user behavior such as past interactions, listening history and search patterns to suggest songs.
- These systems lack emotional awareness and do not adapt to the how a user is feeling in real time.

Goal:

- To develop an AI-driven music recommendation system that analyzes speech to detect emotions and suggest music using speech processing and emotion recognition.
- Delivers personalized music recommendations that align with the user's current emotional state.

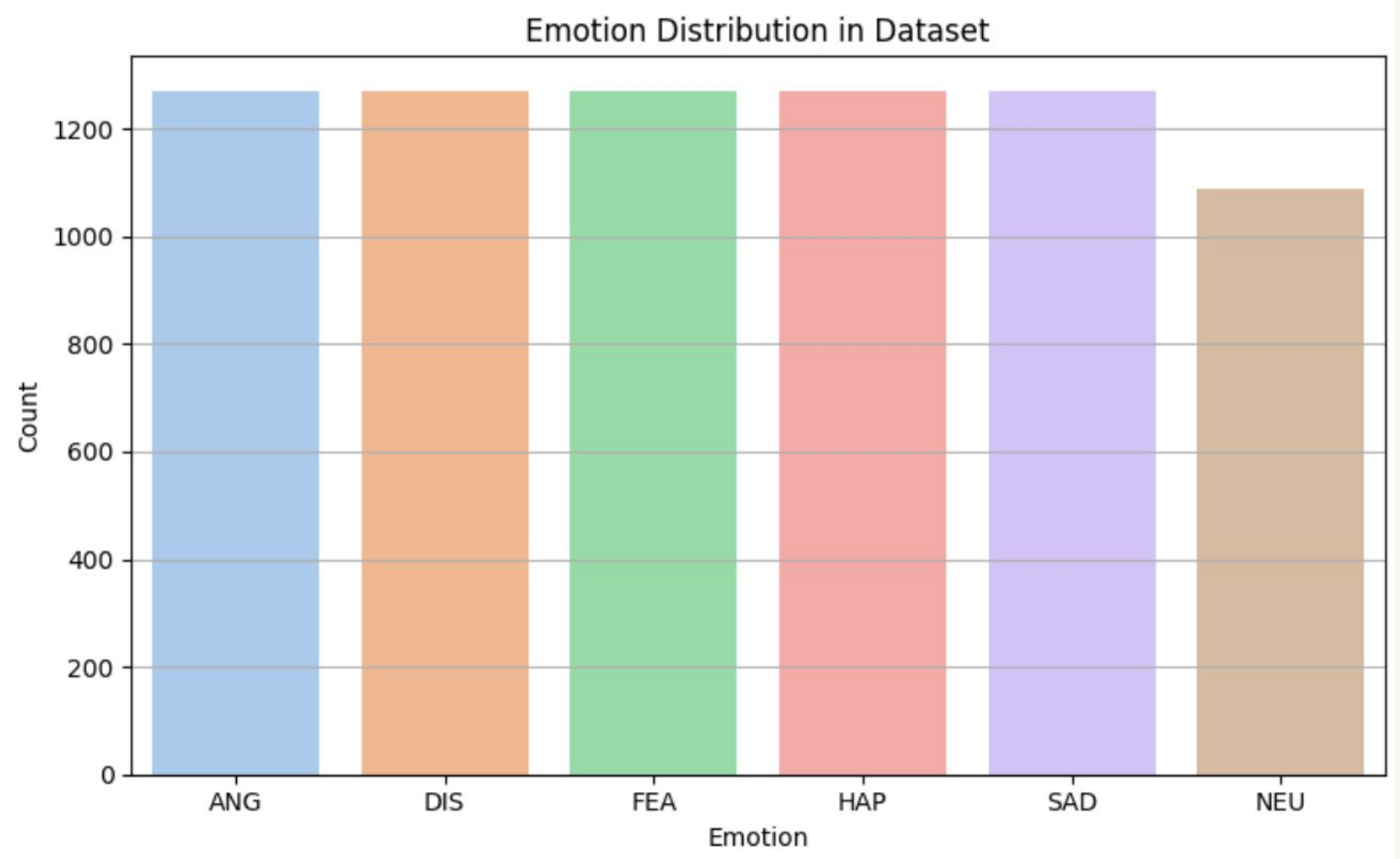
DATA SOURCE

CREMA-D(Crowd-sourced Emotional Multimodal Actors Dataset) is a publicly available dataset designed for speech emotion recognition.

Key features:

- Contains 7,442 audio clips with acted emotional expressions.
- Covers six basic emotions - neutral, happy, sad, angry, fear, and disgust and four different emotion levels (Low, Medium, High, and Unspecified).
- Includes recordings from 91 professional actors (48 male, 43 female).
- Speakers are from various age groups and ethnic backgrounds including African-American, Asian, Caucasian, Hispanic, and Unspecified.
- Each audio clip is labeled with the emotion being acted and additional metadata.
- File Naming: Files are named in the form “ActorID_Statement_Emotion_Intensity.wav” e.g : 1001_DFA_ANG_XX.wav

DATASET ANALYSIS



REQUIREMENTS

- Python 3.11
- VS Code/Google Colab
- Libraries:
  - librosa
  - numpy
  - pandas
  - matplotlib,seaborn
  - scikit-learn
  - imblearn
  - tensorflow
  - spotipy
- UI - streamlit

WORKFLOW

- Phase 1 : Audio Processing & Feature Extraction
- Phase 2 :Speech Emotion Detection
- Phase 3 : Music Recommendation

Phase 1 : Audio Processing & Feature Extraction

1.Feature Extraction

a.Trimming Audio-remove silence

b.Zero Crossing Rate (ZCR) - measures signal noisiness

c.Root Mean Square Energy (RMSE) - Captures loudness/energy patterns

d.Spectral Features - Frequency distribution, learning intricate patterns in tone, intensity and rhythm within the audio signal.

e.MFCC (Mel-Frequency Cepstral Coefficients) - Identify emotion patterns.

f.Chroma Features – Pitch classes (happy/sad)

g.Mel Spectrogram - Perceive pitch and loudness

2.Label Encoding - Converts categorical emotion labels to numeric values

3.Feature Scaling – Standardizes/normalize feature values.

4. SMOTE - Synthetic Minority Over-sampling Technique. Fix imbalance in datasets

5. Audio Augmentation(Method 2) - Noise addition, pitch shifting, time stretching

Phase 2 : Speech Emotion Detection

METHODOLOGY

- Implemented 2 methods
- Used *Convolutional Neural Networks (CNNs)* model.
- Model Architecture
  - Two convolutional blocks each consisting of Convolutional layer, Batch normalization Max pooling layer and Dropout.
  - Flatten Layer to convert the 2D feature maps into a 1D array.
  - Fully Connected Layers: Dense layer, Dropout layer and Output layer.
- Compiled the model with the *Adam* optimizer, used categorical crossentropy as the loss function, suitable for multi-class classification tasks.

HYPERPARAMETER TUNING

- *Learning Rate:* Set to 0.0003 in the Adam optimizer for gradual, stable convergence.
- *Batch Size*: Fixed at 32 to balance memory use and training efficiency.
- *Epochs:* Defined as 100 for the number of training passes.
- *Convolution Filters*: Varying numbers (64, 128) capture features at different levels of complexity.
- *Kernel Size:* (3, 3) for detecting fine patterns in data.
- *Dropout Rate:* 0.5 in fully connected layers to prevent overfitting.
- L2 Regularization -Adds penalty to reduce overfitting
- *K-Fold Cross-Validation:* Set to 5 for robust performance evaluation(⅘ th data for training,⅕ th data for testing)
- *Callbacks:* Includes ModelCheckpoint to save the best model and EarlyStopping to halt training if validation loss remain unchanged.

RESULTS : Accuracy (5-Fold Cross Validation)

Fold-wise Accuracy:

	Fold	Accuracy
0	Fold 1	0.5708
1	Fold 2	0.5659
2	Fold 3	0.5803
3	Fold 4	0.5626
4	Fold 5	0.5679

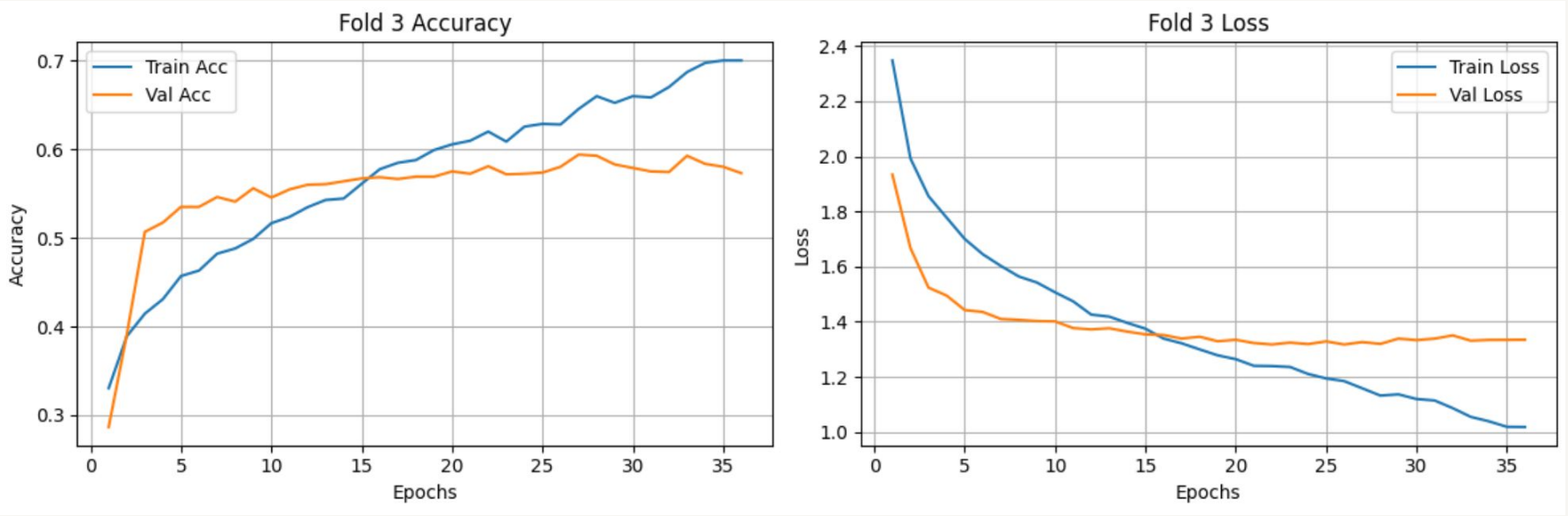
METHOD 1

Fold-wise Accuracy:

	Fold	Accuracy
0	Fold 1	0.6426
1	Fold 2	0.6345
2	Fold 3	0.6302
3	Fold 4	0.6352
4	Fold 5	0.6451

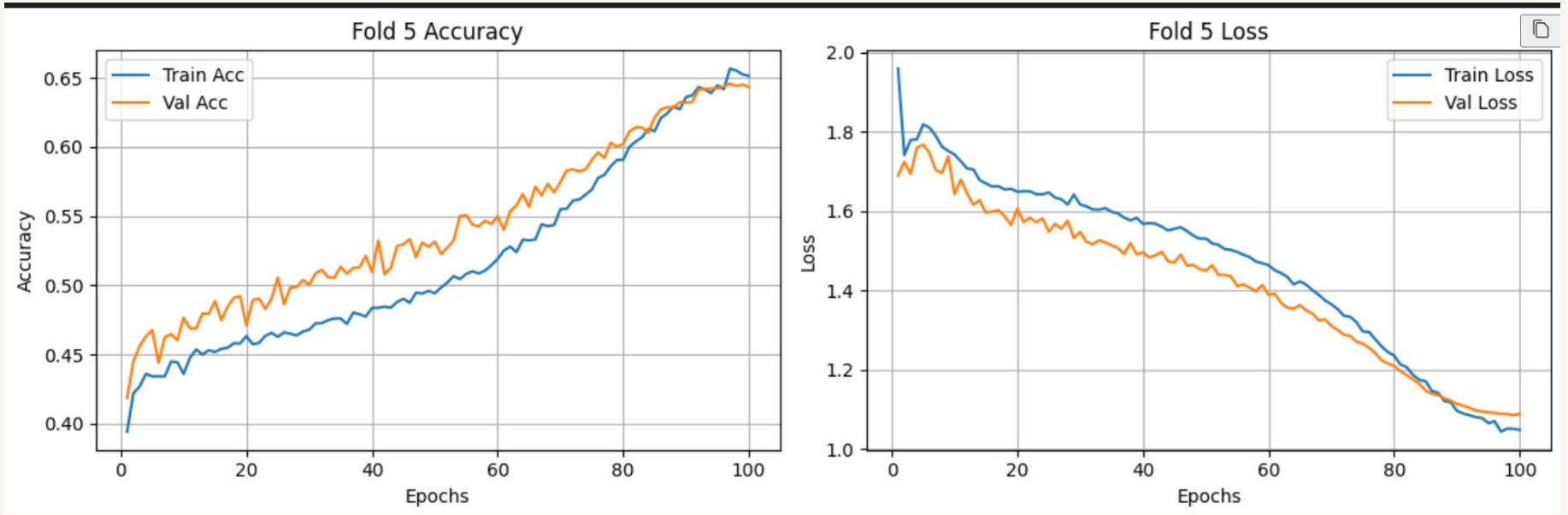
METHOD 2

RESULTS - Accuracy,Loss Plots



METHOD 1

METHOD 2



RESULTS - Classification Report

Classification Report:

	precision	recall	f1-score	support
ANG	0.81	0.86	0.83	1271
DIS	0.74	0.70	0.72	1271
FEA	0.82	0.58	0.68	1271
HAP	0.78	0.70	0.74	1271
NEU	0.70	0.88	0.78	1271
SAD	0.71	0.78	0.74	1271

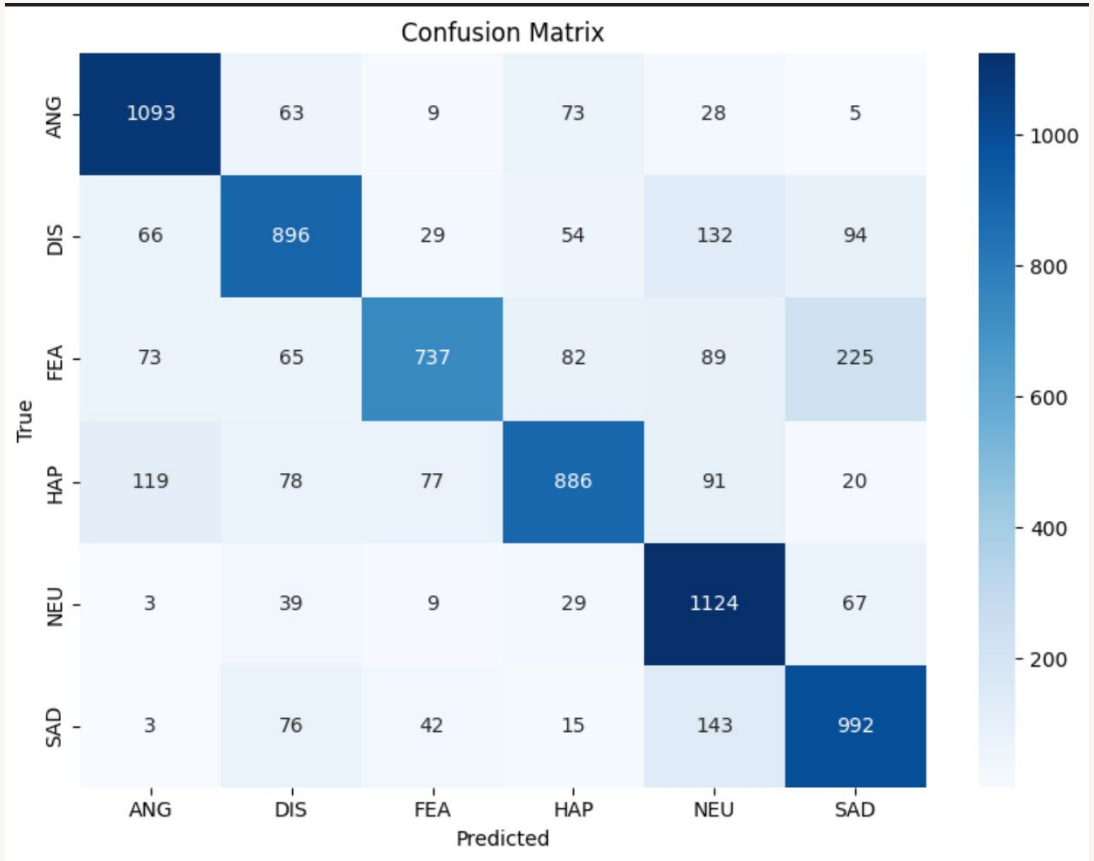
METHOD 1

METHOD 2

Classification Report:

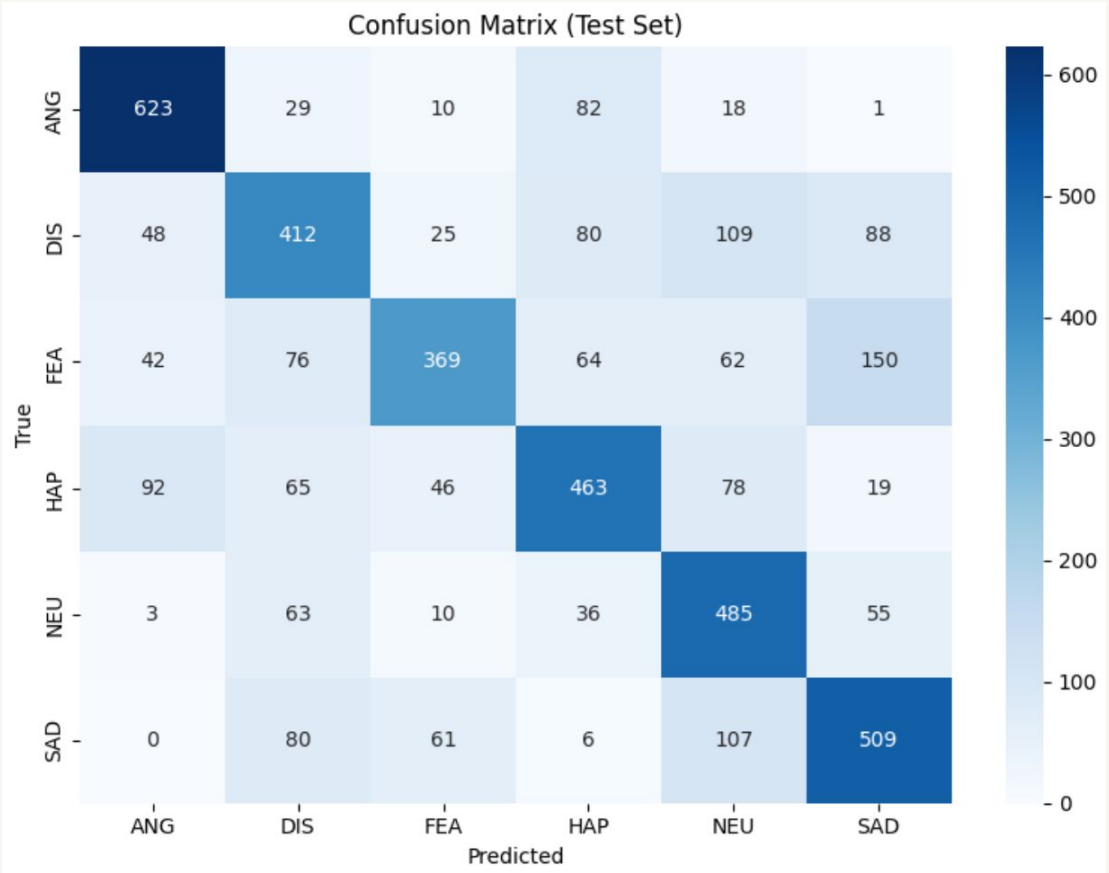
	precision	recall	f1-score	support
ANG	0.77	0.82	0.79	763
DIS	0.57	0.54	0.55	762
FEA	0.71	0.48	0.57	763
HAP	0.63	0.61	0.62	763
NEU	0.56	0.74	0.64	652
SAD	0.62	0.67	0.64	763

RESULTS - Confusion Matrix



METHOD 1

METHOD 2



RESULTS

- *Accuracy : Method 2 give better accuracy than Method 1*
- *Plots :*
  - *Training and Validation Accuracy Graphs:* The training and validation accuracy curves converge and stabilize, showing no overfitting.
  - *Training and Validation Loss Graphs:* Loss decreases quickly in the early epochs and stabilizes, indicating effective learning.
- *Classification Report : Precision,Recall and F1-score is higher for method 1*
- *Confusion Matrix : Method 1 has less misclassifications*
  - Method 1 shows strong overall performance with high true positive rates, especially for NEU and ANG, but some confusion exists between FEA and SAD.
  - Method 2 has slight performance drop with higher misclassification for SAD and FEA, indicating the model struggles more on unseen data, particularly distinguishing between similar emotional tones.

Phase 3 : Music Recommendation

1. Each emotion is linked to a suitable music genre through a `mood_map` dictionary

```
mood_map = {
    'HAP': 'pop',
    'SAD': 'classical',
    'ANG': 'rock',
    'NEU': 'chill',
    'FEA': 'metal',
    'DIS': 'emo'
}
```

2. Connect to Spotify API - Set Up Spotify Developer Credentials and Authenticate using '***spotipy***'
3. Search for Tracks by Genre
4. Display Top 5 Recommended Songs (Using Content based filtering by Spotify)

Outputs

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 23ms/step/usr/local/lib/python3.1
  warnings.warn(
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 49ms/step
Prediction Probabilities:
ANG: 28.92%
DIS: 2.65%
FEA: 27.45%
HAP: 36.24%
NEU: 1.31%
SAD: 3.43%

Top tracks for HAP mood:
BIRDS OF A FEATHER by Billie Eilish
Die With A Smile by Lady Gaga
Sailor Song by Gigi Perez
Good Luck, Babe! by Chappell Roan
No One Noticed by The Marías


Audio File: 1059_WSI_HAP_XX.wav
Actual Emotion:    HAP
Predicted Emotion: HAP
```

```
Analyzing /content/drive/MyDrive/CIS 579- AI Project/AudioWAV/1060_TAI_SAD_X
/usr/local/lib/python3.11/dist-packages/librosa/core/spectrum.py:266: UserWa
  warnings.warn(
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 61ms/step
Prediction Probabilities:
ANG: 0.03%
DIS: 5.94%
FEA: 16.45%
HAP: 0.40%
NEU: 0.25%
SAD: 76.92%

Top tracks for SAD mood:
Experience by Ludovico Einaudi
Adieux by Ludovico Einaudi
Cello Suite No. 1 in G Major, BWV 1007: I. Prélude by Johann Sebastian Bach
Suite bergamasque, L. 75: III. Clair de lune by Claude Debussy
Gymnopédie No. 1 by Erik Satie


Audio File: 1060_TAI_SAD_XX.wav
Actual Emotion:    SAD
Predicted Emotion: SAD
```

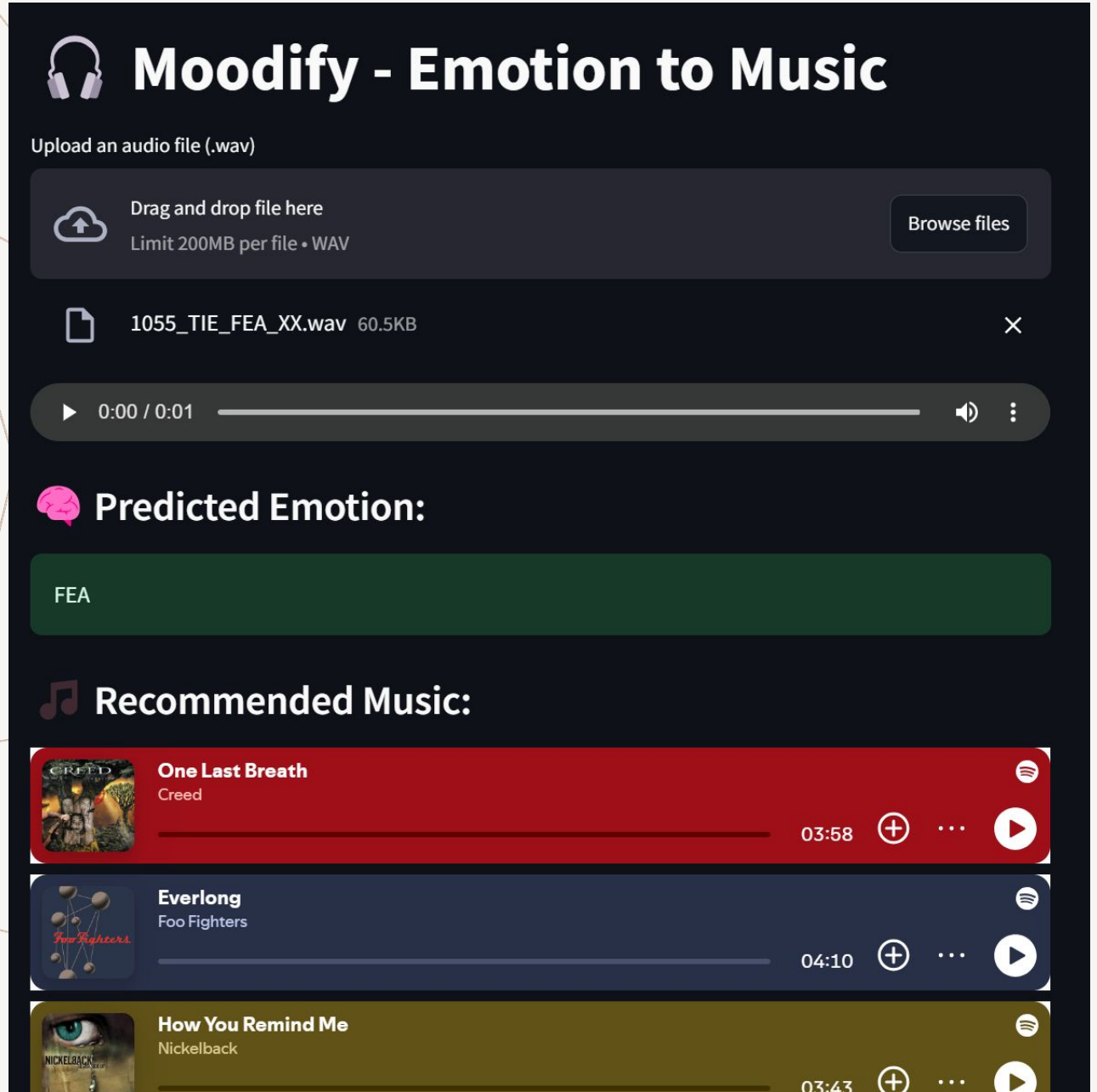
```
Analyzing /content/drive/MyDrive/CIS 579- AI Project/Audio
/usr/local/lib/python3.11/dist-packages/librosa/core/spect
  warnings.warn(
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 164ms/step
Prediction Probabilities:
ANG: 96.74%
DIS: 0.62%
FEA: 0.35%
HAP: 2.27%
NEU: 0.02%
SAD: 0.01%

Top tracks for ANG mood:
505 by Arctic Monkeys
Sweater Weather by The Neighbourhood
Dreams - 2004 Remaster by Fleetwood Mac
Mr. Brightside by The Killers
Iris by The Goo Goo Dolls


Audio File: 1002_IEO_ANG_HI.wav
Actual Emotion:    ANG
Predicted Emotion: ANG
```

```
Analyzing /content/drive/MyDrive/CIS 579- AI Proje
/usr/local/lib/python3.11/dist-packages/librosa/co
  warnings.warn(
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 67ms/step
Prediction Probabilities:
ANG: 0.01%
DIS: 0.69%
FEA: 0.01%
HAP: 0.21%
NEU: 98.92%
SAD: 0.15%

Top tracks for NEU mood:
Another Love by Tom Odell
Embrace It by Ndotz
Soft Spot by keshi
4 Morant (Better Luck Next Time) by Com Truise
Another Love by Tom Odell


Audio File: 1088_IWW_NEU_XX.wav
Actual Emotion:    NEU
Predicted Emotion: NEU
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 69ms/step
Prediction Probabilities:
ANG: 16.69%
DIS: 47.68%
FEA: 1.53%
HAP: 11.19%
NEU: 22.29%
SAD: 0.62%

Top tracks for DIS mood:
The Middle by Jimmy Eat World
[redacted]
Misery Business by Paramore
Hard Times by Paramore
Falling Down - Bonus Track by Lil Peep


Audio File: 1089_IWW_DIS_XX.wav
Actual Emotion:    DIS
Predicted Emotion: DIS
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 38ms/step
/usr/local/lib/python3.11/dist-packages/libros
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/libros
  warnings.warn(
Prediction Probabilities:
ANG: 26.79%
DIS: 11.31%
FEA: 50.52%
HAP: 10.95%
NEU: 0.12%
SAD: 0.31%

Top tracks for FEA mood:
One Last Breath by Creed
Everlong by Foo Fighters
How You Remind Me by Nickelback
Can't Stop by Red Hot Chili Peppers
Numb by Linkin Park

File: 1042_IEO_FEA_HI.wav
Actual Emotion:    FEA
Predicted Emotion: FEA
```

UI


Moodify - Emotion to Music
Upload an audio file (.wav)
Drag and drop file here
Limit 200MB per file • WAV
Browse files
1055_TIE_FEA_XX.wav 60.5KB
0:00 / 0:01
Predicted Emotion:
FEA
Recommended Music:
One Last Breath
Creed
03:58
Everlong
Foo Fighters
04:10
How You Remind Me
Nickelback
03:43

CHALLENGES FACED

- *Limited Model Accuracy* - Accuracy capped around 65%, possibly due to overlapping emotional features and limited feature representation.
- *Class Imbalance* - Emotions like Neutral had fewer samples, leading to biased predictions.
- *High Computational Load* - Voice data and deep learning models required long training times and high resources.
- *Lack of Diverse Data* - Limited variation in speakers and accents reduced generalization
- *Complex Feature Extraction* - Required significant time to research and implement effective audio features.

LESSONS LEARNT

- **Data Balance is Crucial** – Handling class imbalance with techniques like SMOTE significantly improves model accuracy for underrepresented emotions.
- **Quality Feature Extraction** – Properly tuned audio features (e.g., MFCCs, chroma) are essential for capturing emotional cues effectively.
- **Model Development** - Cross-validation helps prevent overfitting on emotionally skewed data.
- **Ethical Considerations -** Avoid reinforcing emotional biases or triggering unwanted emotional states.

FUTURE SCOPE

- Better mood map
  - Use Multiple Genres per Emotion
  - Search using emotion + genre keywords
- Multimodal Emotion Recognition - Combine voice with facial expressions or text (transcripts) for better accuracy and richer context.
- Use of pre trained audio models like Wav2Vec 2.0, YAMNet, or Whisper to extract high-level audio embeddings instead of hand-crafted features.
- Expand dataset diversity - Collect or use open-source datasets with more diverse speakers across various demographics and recording conditions.

REFERENCES

[1] Picard, R. W. (1997). Affective Computing. MIT Press. A foundational book introducing emotion-aware computing systems.

[2] Verma, G. K., & Tiwary, U. S. (2014). Multimodal emotion recognition using facial, speech and textual features. Multimedia Tools and Applications, 76(1), 4403–4425. https://doi.org/10.1007/s11042-016-3819-0

[3] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

[4] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C & Wilson, K. (2017). CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 131-135). IEEE. https://doi.org/10.1109/ICASSP.2017.7952132

[5] Spotify for Developers. (n.d.). Web API Reference. Retrieved April 21, 2025, from https://developer.spotify.com/documentation/web-api