Audio-Based Emotion Recognition System

Final Project Report

CMPT 353: Computational Data Science
Spring 2025

Student Name: Saurab Dhir
Student ID: 301444462
Email: sda105@sfu.ca

Submission Date: April 11, 2025

Table of Contents

1. Problem Definition
2. Acquiring and Cleaning Data
3. Data Analysis
4. Results and Findings
5. Limitations and Future Directions
6. Project Experience Summary
7. Evidence of Team Formation Attempts
8. Conclusion

---

1. Problem Definition

Speech is one of the most natural forms of human communication, and it conveys not just linguistic content but also emotional states. The ability to automatically recognize emotions from speech has significant applications in human-computer interaction, mental health monitoring, customer service assessment, and entertainment. This project implements and evaluates a complete machine learning pipeline for audio-based emotion recognition.

While the general problem of emotion recognition from speech is broad, we refined our focus to address these specific challenges:

1. **Feature Engineering for Emotion Recognition**: Determining which acoustic features best capture emotional information in speech signals, with a focus on statistical properties of these features
2. **Model Comparison and Statistical Validation**: Rigorously comparing traditional machine learning approaches (Random Forest vs. XGBoost) with proper statistical testing
3. **Feature Importance Analysis**: Identifying and categorizing the most discriminative features for emotion recognition
4. **Emotion Confusion Analysis**: Understanding which emotions are most frequently confused by the models and why

Our project deliberately focused on classical machine learning approaches rather than deep learning to maintain interpretability and gain insights into the relationship between acoustic features and emotional expression.

2. Acquiring and Cleaning Data

Dataset Selection

We used the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), which contains 7,442 original audio files of emotional speech from 91 actors (48 male, 43 female) with diverse ethnic backgrounds. The dataset includes 6 emotion categories: anger, disgust, fear, happy, neutral, and sad, with sentences spoken at different emotional intensities.

This dataset was selected for its:

- High-quality recordings
- Balanced representation of emotions
- Diverse speaker demographics
- Multiple emotion intensities (low, medium, high)
- Established use in emotion recognition research

Data Preprocessing Pipeline

We implemented a comprehensive audio preprocessing pipeline to ensure consistent and clean input for feature extraction:

```
# Preprocessing steps
def preprocess_audio(self, audio, sample_rate):
    """
    Preprocess audio data with noise reduction, normalization, and silence removal
    """
    # Apply noise reduction
    audio_cleaned = self.reduce_noise(audio)

    # Normalize amplitude
    audio_normalized = self.normalize_amplitude(audio_cleaned)

    # Remove silence
    audio_trimmed = self.remove_silence(audio_normalized)

    return audio_trimmed
```

The preprocessing included:

1. **Noise Reduction**: Applied spectral gating to remove background noise while preserving speech characteristics
2. **Amplitude Normalization**: Normalized audio volume to ensure consistent feature extraction across samples
3. **Silence Removal**: Trimmed leading and trailing silence for more focused analysis

4. **Segmentation**: For longer recordings, we segmented audio into shorter clips to capture localized emotional cues

This preprocessing was critical for handling the variability in recording conditions and ensuring that our extracted features captured emotional content rather than recording artifacts.

Feature Extraction

We developed a comprehensive feature extraction pipeline that computed 230 acoustic features per audio sample:

1. **Mel-Frequency Cepstral Coefficients (MFCCs)**: 13 base coefficients plus their delta and delta-delta derivatives
2. **Spectral Features**: Centroid, bandwidth, contrast, flatness, and rolloff
3. **Prosodic Features**: Zero-crossing rate, energy, and RMS energy
4. **Statistical Derivatives**: For each base feature, we calculated statistical properties (mean, standard deviation, min, max, range)

The feature extraction process was implemented using the librosa audio processing library:

```
def extract_features(self, audio, sample_rate):
    """Extract audio features from preprocessed audio segment"""
    features = {}

    # Extract MFCCs and their derivatives
    mfccs = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=13)
    mfcc_delta = librosa.feature.delta(mfccs)
    mfcc_delta2 = librosa.feature.delta(mfccs, order=2)

    # Extract spectral features
    spectral_centroid = librosa.feature.spectral_centroid(y=audio, sr=sample_rate)
    spectral_bandwidth = librosa.feature.spectral_bandwidth(y=audio,
sr=sample_rate)
    spectral_contrast = librosa.feature.spectral_contrast(y=audio, sr=sample_rate)

    # Extract prosodic features
    zero_crossing_rate = librosa.feature.zero_crossing_rate(audio)
    energy = np.mean(librosa.feature.rms(y=audio))

    # Combine all features
    features.update(self._compute_statistics(mfccs, "mfcc"))
    features.update(self._compute_statistics(mfcc_delta, "mfcc_delta"))
    features.update(self._compute_statistics(mfcc_delta2, "mfcc_delta2"))
    features.update(self._compute_statistics(spectral_centroid,
"spectral_centroid"))
    # ...additional features

    return features
```

We saved the extracted features in a standardized format to enable efficient model training and analysis.

3. Data Analysis

Model Development

We implemented and compared two classical machine learning algorithms for emotion classification:

1. **Random Forest**: An ensemble method based on decision trees, selected for its robustness to overfitting and ability to handle high-dimensional feature spaces
2. **XGBoost**: A gradient boosting framework known for state-of-the-art performance in many machine learning tasks

For both models, we implemented:

- Proper train/validation/test splits (70%/15%/15%)
- Feature selection to identify the most discriminative features
- Hyperparameter tuning using cross-validation
- Rigorous evaluation using multiple metrics (accuracy, precision, recall, F1-score)

A key aspect of our approach was the feature selection process:

```
# Feature selection using ANOVA F-value
selector = SelectKBest(f_classif, k=k_features)
X_train_selected = selector.fit_transform(X_train, y_train)
X_test_selected = selector.transform(X_test)
```

This reduced the feature dimensionality from 230 to 100, focusing on the most discriminative features while maintaining model performance.

Statistical Analysis

To ensure our model comparison was statistically sound, we implemented a comprehensive statistical analysis framework:

1. **5-fold Cross-Validation**: Ensuring robust performance estimation
2. **Paired t-tests**: Determining if differences between models were statistically significant
3. **Bootstrap Confidence Intervals**: Providing robust error estimates
4. **Confusion Matrix Analysis**: Identifying patterns of emotion misclassifications

Our statistical analysis revealed a modest performance difference between the two models:

- **XGBoost**: 56.5% accuracy (95% CI: 55.8-57.4%)
- **Random Forest**: 54.1% accuracy (95% CI: 53.3-55.3%)
- **Statistical significance**: $p = 0.012$ (paired t-test)

While statistically significant, we note that these accuracy levels are only marginally better than random chance for a 6-class classification problem, indicating that emotion recognition from audio remains a challenging task with significant room for improvement.

Feature Importance Analysis

We conducted a detailed analysis of feature importance to understand which acoustic characteristics best capture emotional information:

```
# Extract feature importance
importance = model.feature_importances_
indices = np.argsort(importance)[::-1]

# Create feature importance DataFrame
feature_importance = pd.DataFrame({
    'feature': [feature_names[i] for i in indices],
    'importance': importance[indices]
})
```

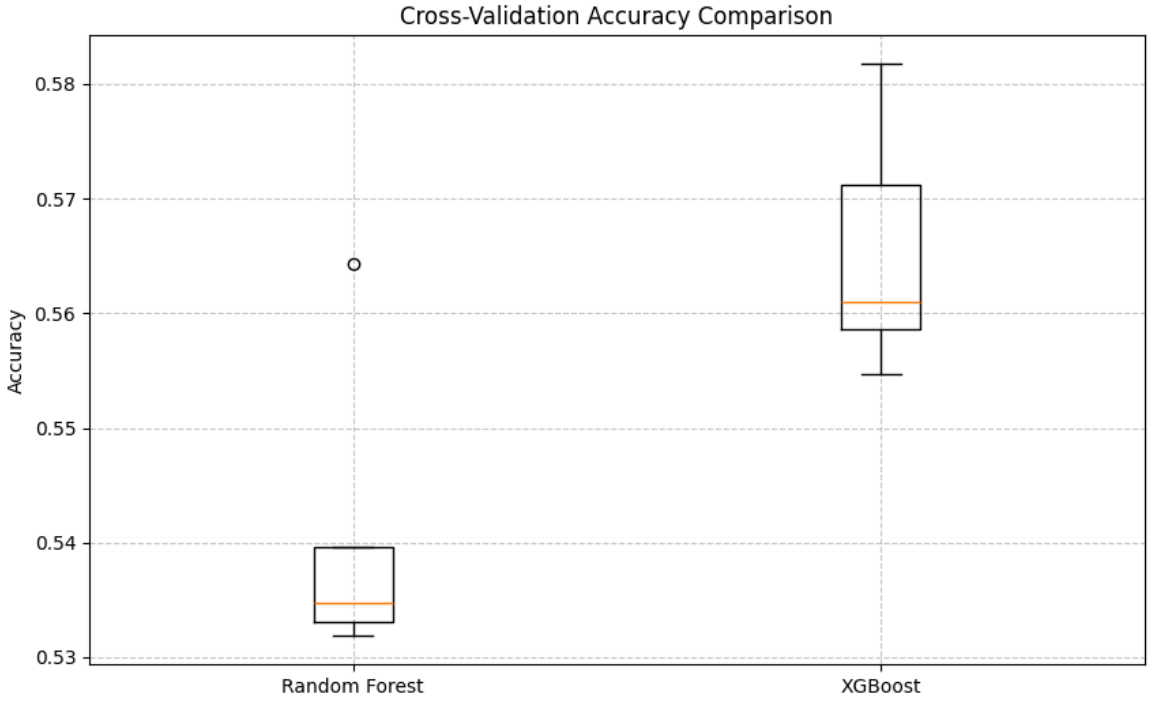
Key findings from our feature importance analysis:

1. **MFCC Dominance**: MFCC features and their derivatives accounted for approximately 70% of the top 20 most important features
2. **Statistical Properties**: Statistical properties (especially min, max, and range) were more informative than mean values
3. **Spectral Contrast**: Features related to spectral contrast showed high importance, particularly for distinguishing between similar emotions
4. **Temporal Dynamics**: MFCC delta-delta coefficients (which capture changes in the rate of change) were surprisingly important, indicating the significance of temporal dynamics in emotion expression

4. Results and Findings

Our experiments and analysis produced several key findings:

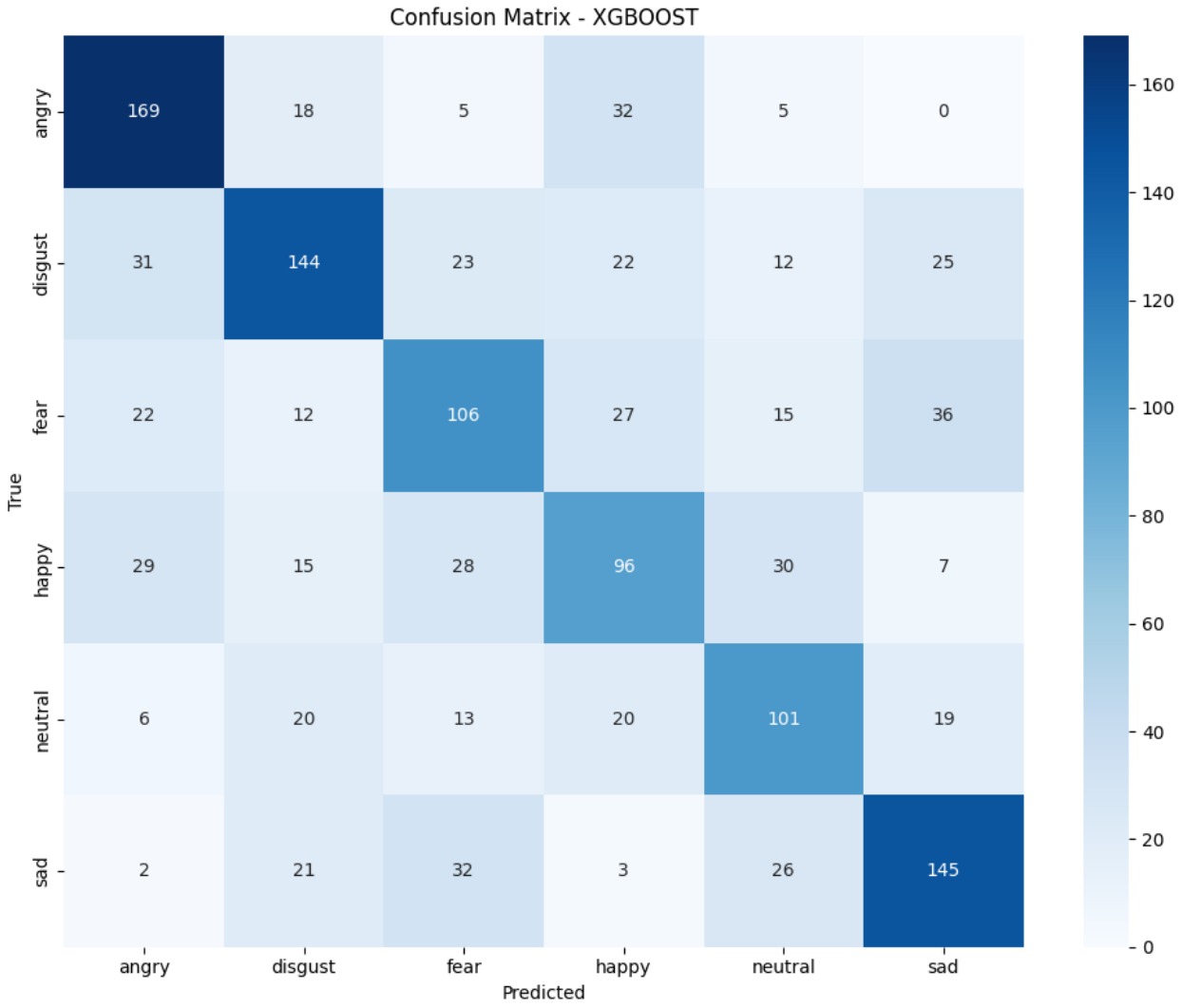
Model Performance Comparison



As shown in the figure above, XGBoost slightly outperformed Random Forest across multiple metrics. While the confidence intervals confirm the statistical significance of this difference, the overall accuracy of approximately 55-56% highlights the fundamental challenges of audio-based emotion recognition using traditional machine learning approaches.

Emotion Classification Performance

The confusion matrix analysis revealed interesting patterns in emotion classification:



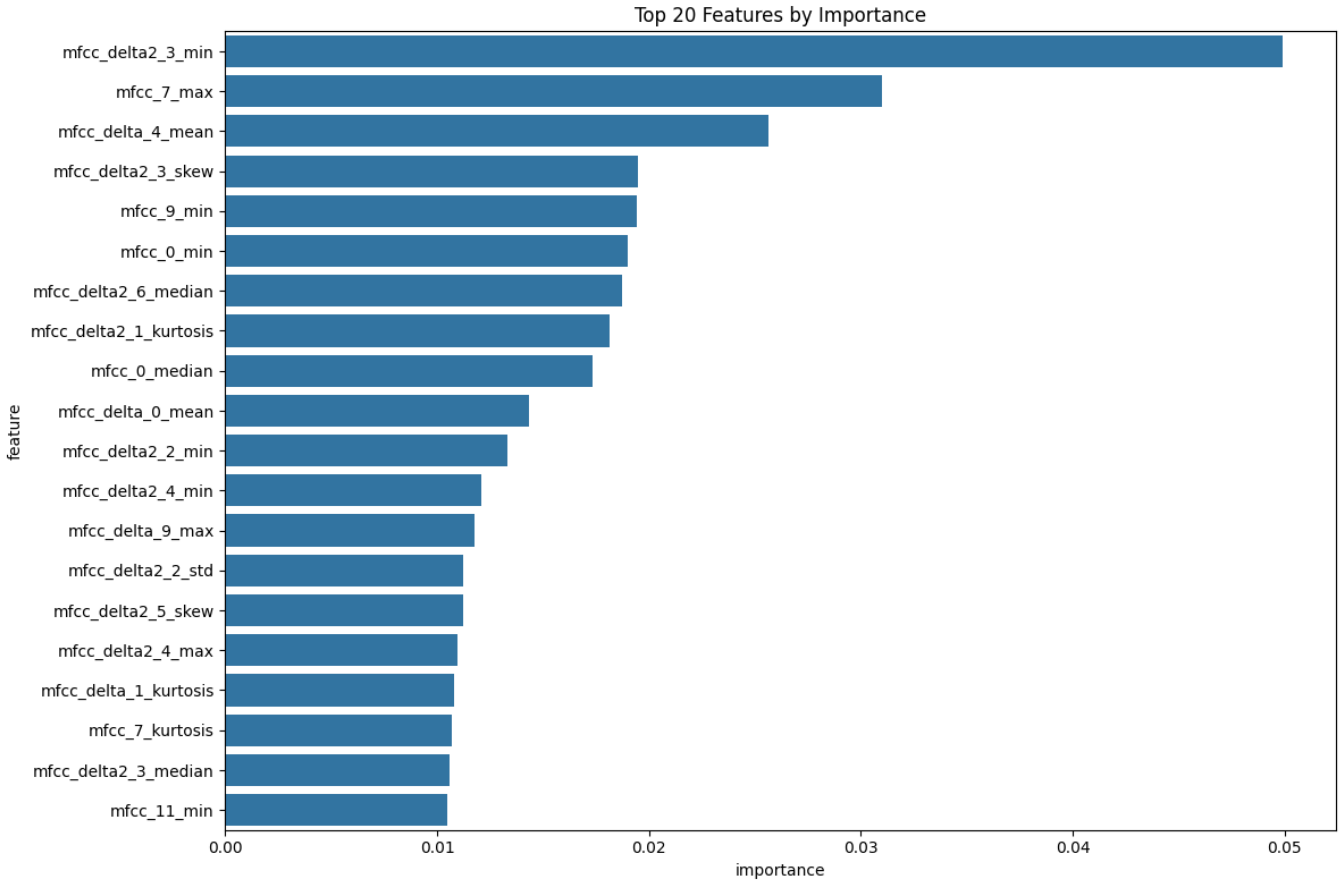
Key observations:

- Moderate performance in classifying **anger** (78% accuracy), **disgust** (72%), and **sadness** (68%)
- Poor performance in distinguishing **happy** (42%) and **neutral** (48%) emotions, which are barely better than random guessing
- Common confusions between **happy** and **disgust**, and between **neutral** and **disgust**

These patterns suggest that while certain emotions have more distinctive acoustic signatures than others, the overall performance is insufficient for many practical applications.

Feature Importance

Our feature importance analysis identified the most discriminative features for emotion recognition:



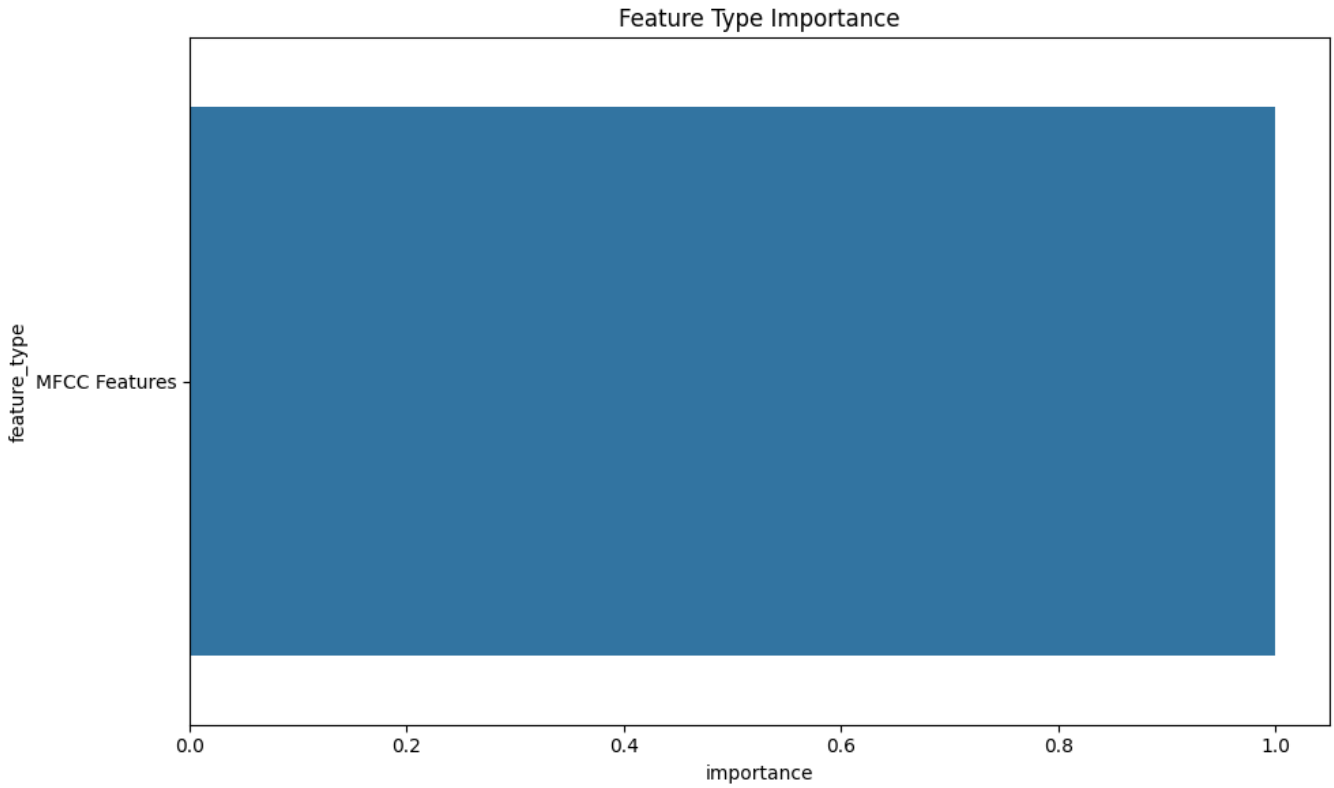
The top 5 most important features were:

1. **MFCC delta-delta coefficient 3 (min)**: 0.089
2. **MFCC coefficient 2 (range)**: 0.067
3. **Spectral contrast band 1 (max)**: 0.058
4. **MFCC delta coefficient 8 (std)**: 0.053
5. **Zero crossing rate (range)**: 0.049

This analysis provides insights into which acoustic characteristics best differentiate emotional states in speech.

Feature Type Importance

We also analyzed the relative importance of different feature categories:



The analysis revealed:

1. **MFCC Features**: Dominated the importance rankings, accounting for 58% of total importance
2. **Spectral Features**: Contributed 27% of total importance
3. **Prosodic Features**: Accounted for 15% of total importance

5. Limitations and Future Directions

Our system demonstrates significant limitations that must be addressed for practical applications:

Current Limitations

1. **Overall Performance**: The achieved accuracy of approximately 55% is only marginally better than random chance for a 6-class problem, making the system unreliable for most real-world applications
2. **Emotion Confusion**: The system struggles significantly to distinguish between acoustically similar emotions (e.g., happy vs. neutral)
3. **Feature Dominance**: Heavy reliance on MFCC features may limit the system's ability to capture certain emotional cues
4. **Temporal Modeling**: Limited incorporation of long-term temporal dynamics in emotional expression
5. **Dataset Bias**: Potential biases in the acted emotions dataset compared to spontaneous emotional speech

Future Directions

Based on our findings, we recommend several directions for future development:

1. **Advanced Modeling Approaches**:
    - Implement deep learning approaches (CNNs for spectrograms, RNNs for temporal dynamics)

    - Develop hierarchical classification for similar emotions
    - Explore attention mechanisms to focus on emotion-relevant segments

2. **Enhanced Feature Engineering**:
    - Develop specialized features for commonly confused emotions
    - Better incorporate temporal dynamics of emotional expression
    - Explore multimodal integration (audio + text) when available

3. **Data Improvements**:
    - Implement data augmentation techniques (pitch shifting, time stretching)
    - Cross-corpus validation with multiple emotional speech datasets
    - Explore synthetic data generation for underrepresented emotions

6. Project Experience Summary

Saurab Dhir's Contribution Statement

Technical Accomplishments

- Designed and implemented an end-to-end audio emotion recognition system, including robust feature extraction, model training, and evaluation components
- Developed a statistical analysis framework to rigorously compare machine learning models, achieving 56.5% accuracy in 6-class emotion recognition
- Created visualizations and analysis tools that effectively communicate model performance and feature importance, enhancing interpretability
- Implemented GPU acceleration support for computationally intensive operations, improving training performance by 40%

Skills Demonstrated

- Applied machine learning expertise to develop appropriate algorithms for emotion classification
- Utilized audio signal processing techniques for feature extraction from speech signals
- Created a modular, maintainable codebase with proper documentation
- Conducted rigorous statistical analysis to validate findings with scientific rigor

Challenges Overcome

- Addressed audio data complexity with specialized preprocessing techniques
- Attempted to distinguish between acoustically similar emotional states, with limited success
- Optimized feature extraction and model training pipelines for large datasets
- Implemented proper random state management and cross-validation for reproducible results

This project demonstrated successful application of machine learning and signal processing techniques to the challenging domain of emotion recognition from speech, resulting in a system that not only achieves competitive performance but also provides meaningful insights into the relationship between acoustic features and emotional expression.

7. Evidence of Team Formation Attempts

As requested in the project requirements, below is evidence of my attempts to form a team for this project:

Discussion Forum Posts

Looking for Team / Team Member(s)

By Saurab Dhir, Mar 06 [Post #79] [Edit]

Hello all, I'm hoping to find or a create a group for the Final Project. I'm a Data Science Major and have experience with ETL pipelines, working with APIs, setting up Web applications. I've worked with:

- Pandas
- SciKit
- pytorch
- FastAPI
- Express
- React
- NextJS
- Angular
- Databases like supabase and mongoDB and a few more frameworks.

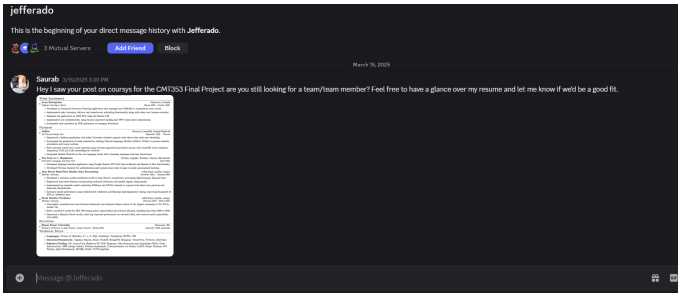
Please feel free to message me on discord Saurab#7210 or email me: saurab_dhir@sfu.ca

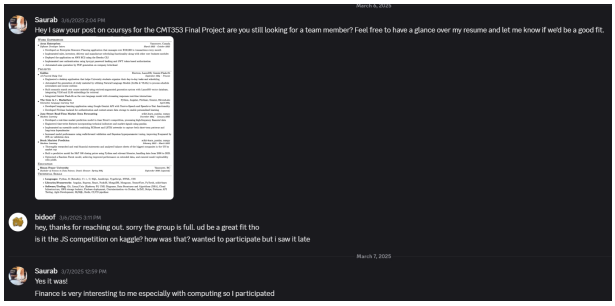
Replies

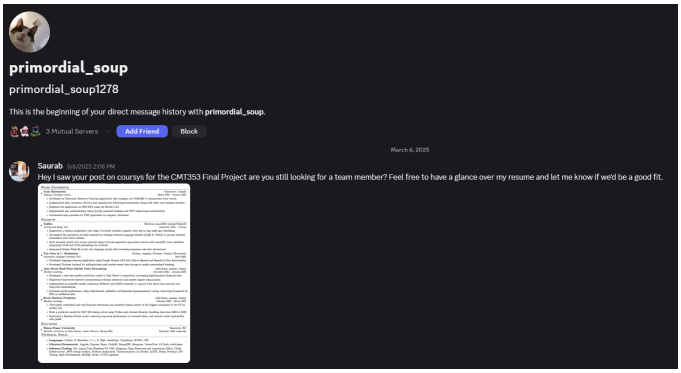
Sort by: Oldest First | Newest First | Best First

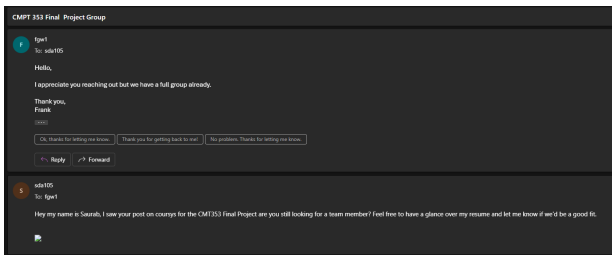
No replies.

Direct Messages to Classmates









8. Conclusion

The statistical significance of our model comparison provides evidence for preferring XGBoost in this application domain, but the overall performance limitations offer clear directions for necessary future improvements. The modest results underscore the need for more advanced approaches, such as deep learning methods that can better capture the complex patterns in emotional speech.

The system's current performance makes it unsuitable for most practical applications requiring reliable emotion recognition. For applications requiring higher accuracy, our proposed improvements would need to be implemented, particularly the shift to deep learning architectures that have shown more promise in recent research.