SPEECH SIGNAL PROCESSING

# Emotion Recognition using MFCC and Prasody Features extracted at Word, Syllable and Utterance Level

**Vedant Nipane (2021102040) and Himanshu Gupta (2022102002)**

04 April, 2025

## OVERVIEW

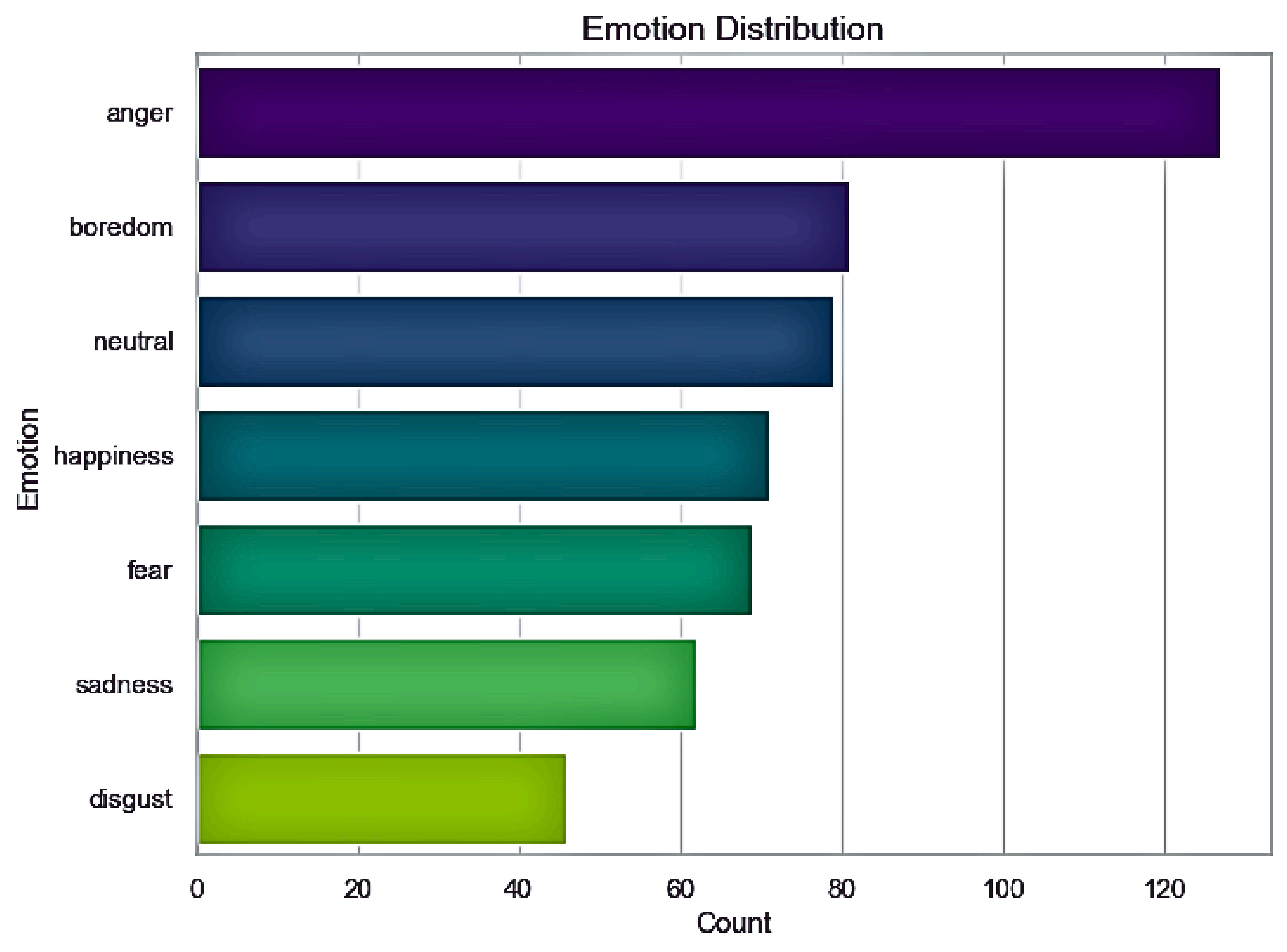
This emotion recognition system extracts Mel-Frequency Cepstral Coefficients (MFCCs) and prosodic features (pitch, energy, duration) at three linguistic levels to capture both spectral and temporal characteristics of emotional speech. We implement various machine learning classifiers to evaluate which combination of features and models yields the highest accuracy in emotion detection, with potential applications in human-computer interaction, mental health assessment, and customer service technologies

## OBJECTIVE AND GOALS

The objective of this project is to develop a machine learning system capable of accurately recognizing human emotions from speech signals using the EmoDB dataset. Our system aims to leverage multi-level acoustic features extracted at the utterance, word, and syllable level to classify speech into seven distinct emotional states: anger, boredom, anxiety, happiness, sadness, disgust, and neutral.

# DATASET OVERVIEW

- EmoDB: German emotional speech corpus with 535 utterances at 16kHz sampling rate
- Contains 7 emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral
- 10 different sentences spoken by professional actors expressing various emotions



# FEATURE EXTRACTION

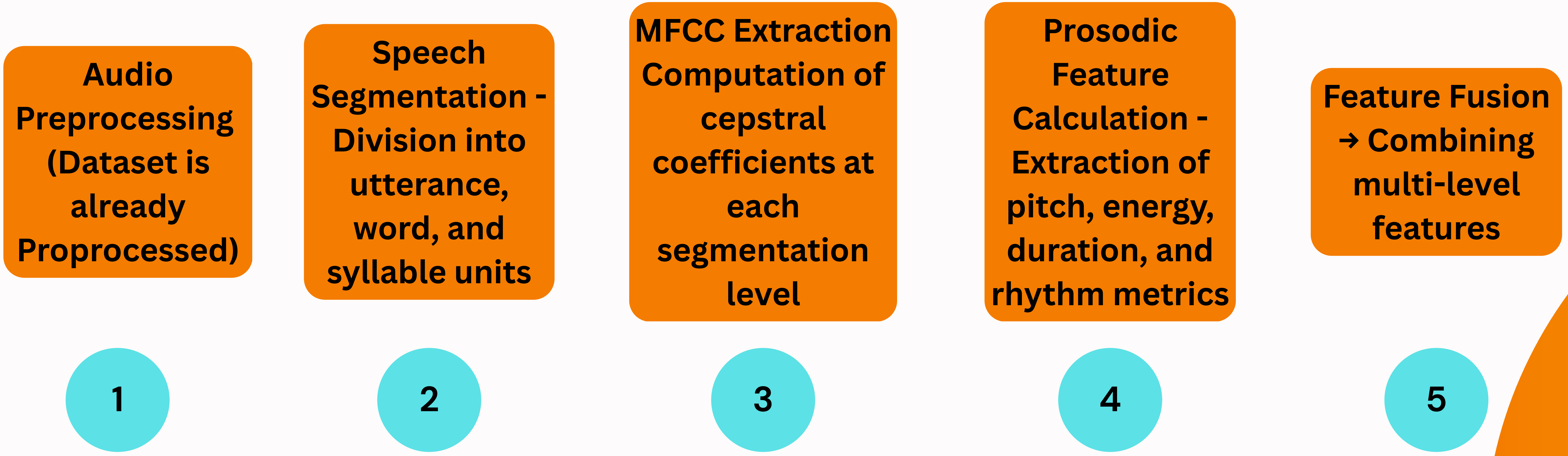
**MFCC features:**

- Mel-Frequency Cepstral Coefficients capture the spectral envelope of speech signals based on human auditory perception.
- MFCCs represent the short-term power spectrum, providing crucial information about vocal tract configuration.

**Prosodic Features:**

- Prosodic features include pitch (FO), energy, duration, and rhythm patterns in speech.
- These features convey emotional information through variations in speaking rate, stress patterns, and intonation.

# FEATURE EXTRACTION PIPELINE



Audio Preprocessing (Dataset is already Proprocessed)

1

Speech Segmentation - Division into utterance, word, and syllable units

2

MFCC Extraction Computation of cepstral coefficients at each segmentation level

3

Prosodic Feature Calculation - Extraction of pitch, energy, duration, and rhythm metrics

4

Feature Fusion → Combining multi-level features

5

# FEATURE EXTRACTION: METHODOLOGY

MFCC Extraction:

1️⃣ Framing & Windowing – Divided the signal into short frames with a Hamming window to maintain continuity.

2️⃣ Fourier Transform – Converted each frame from the time domain to the frequency domain.

3️⃣ Mel-Scale Filtering – Applied a filter bank to emphasize frequencies perceived by human hearing.

4️⃣ Logarithm & DCT – Took the log of filter bank energies and applied Discrete Cosine Transform (DCT) to get MFCC coefficients.

5️⃣ Feature Aggregation – Extracted MFCCs at word, syllable, and utterance levels, adjusting feature lengths using averaging.

# FEATURE EXTRACTION: METHODOLOGY

1. **Word Level**
   - Extracted MFCC and Prosody features for each word in the utterance.
   - Features: 13 MFCCs + multiple prosody features (pitch, energy, duration, etc.).
   - Words vary in length, so features were standardized using averaging instead of zero-padding.
2. **Syllable Level**
   - Segmented words into syllables using phonetic properties.
   - Extracted MFCCs and Prosody per syllable.
   - Different utterances had different numbers of syllables, handled by averaging across syllables.

# FEATURE EXTRACTION: METHODOLOGY

**3. Utterance Level**

- ○ Extracted global-level statistics over the entire utterance.
- ○ Included mean, variance, skewness of MFCCs and prosody features.

📌 **Total Number of Features**

**After combining all levels:**

✅ **MFCC Features: 13 (per word, syllable, utterance)**

✅ **Prosody Features: Pitch, Energy, Duration, Jitter, Shimmer, etc.**

✅ **Final Feature Set: 408 features per utterance (speech file)**

# ML MODELS FOR EMOTION RECOGNITION

Till now, we tried classifying using 3 ML Models

**1. Support Vector Machines**

✅ Works well with high-dimensional feature sets like our 408 extracted features.

✅ Good generalization for supervised classification tasks.

**2. Gaussian Mixture Models**

✅ Probabilistic model suitable for speech processing.

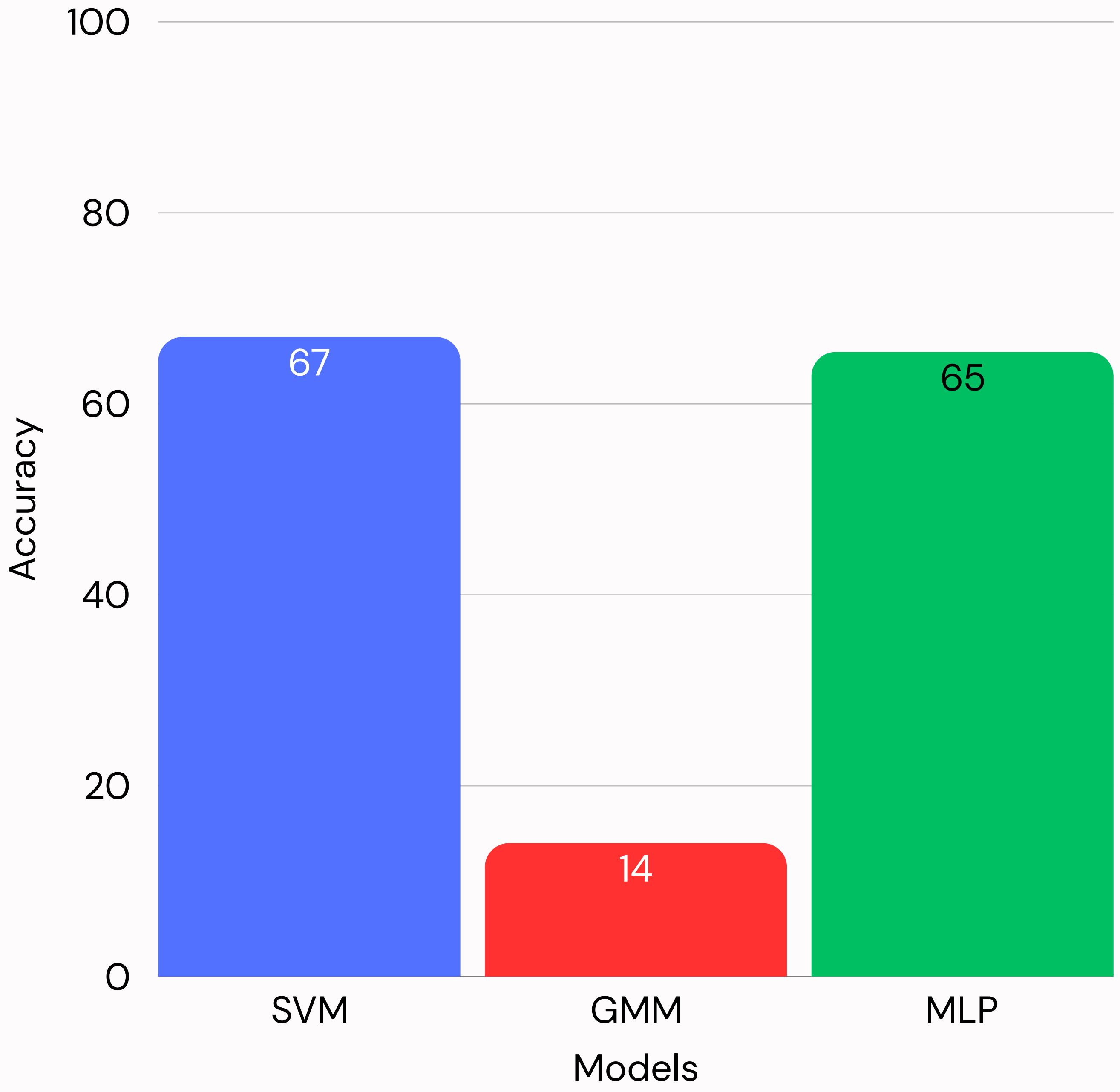
✅ Captures variation in feature distributions.

**3. Multi-Layer Perceptron (MLP)**

✅ Can model complex non-linear patterns in MFCC + Prosody features.

✅ Learns hierarchical representations better than SVM.

# RESULTS



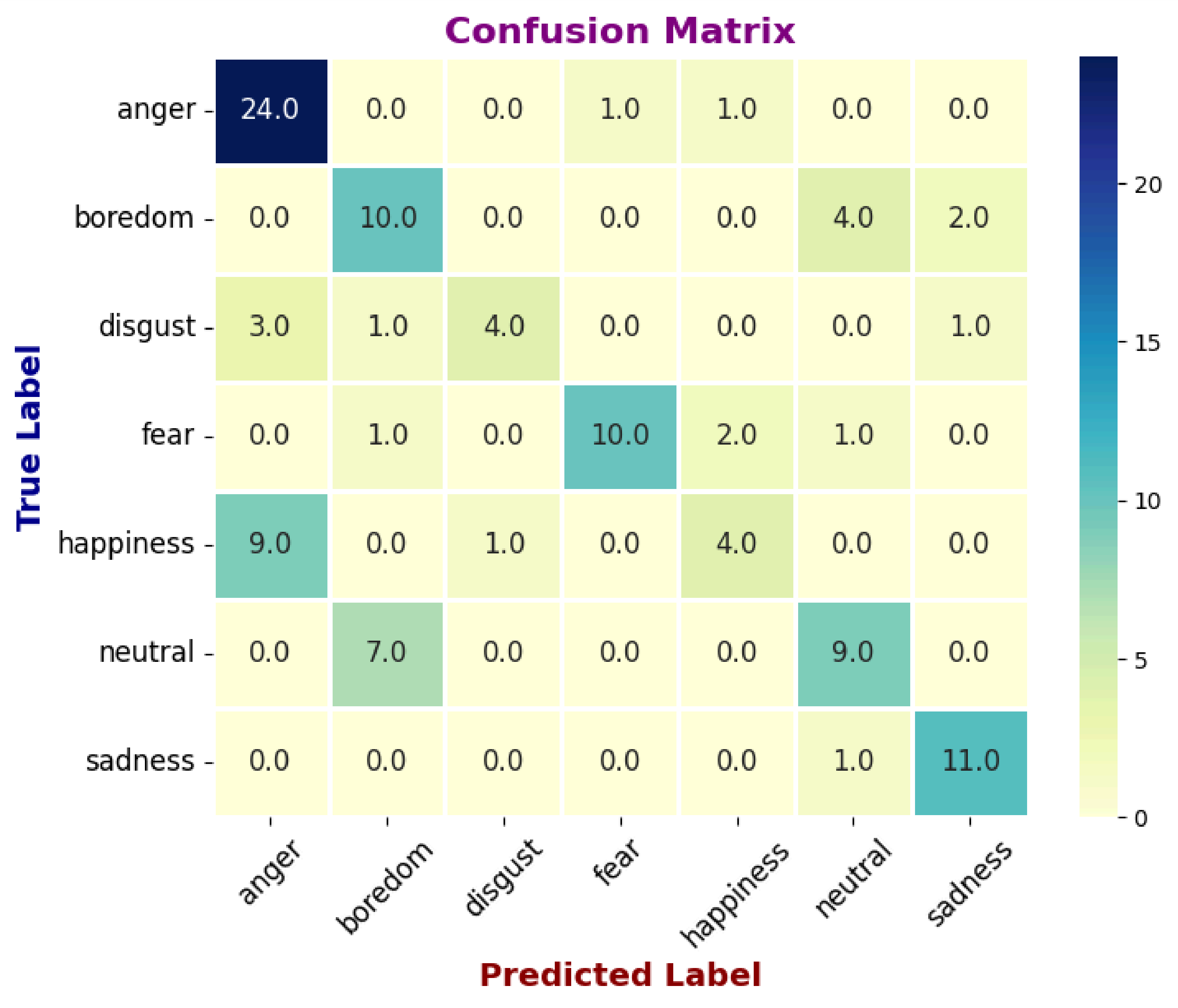
**Accuracy on test set**
**(20 % of whole dataset)**

✅ SVM (67%) performed best so far.

✅ MLP (65.4%) is close but may improve with fine-tuning.

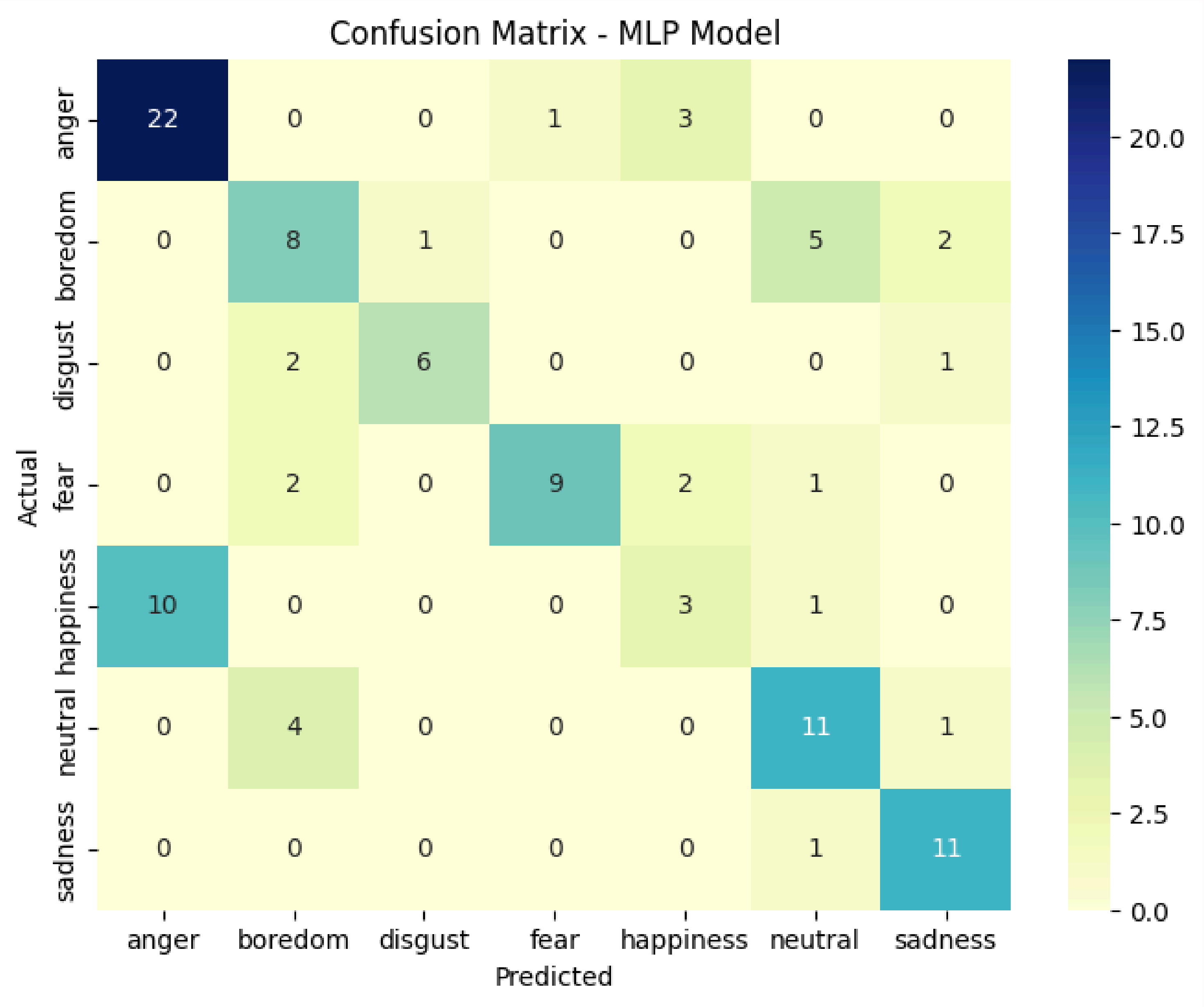
❌ GMM was the worst (14%) and dropped.

# CONFUSION MATRIX



| True Label \ Predicted Label | anger | boredom | disgust | fear | happiness | neutral | sadness |
|---|---|---|---|---|---|---|---|
| anger | 24.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 |
| boredom | 0.0 | 10.0 | 0.0 | 0.0 | 0.0 | 4.0 | 2.0 |
| disgust | 3.0 | 1.0 | 4.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| fear | 0.0 | 1.0 | 0.0 | 10.0 | 2.0 | 1.0 | 0.0 |
| happiness | 9.0 | 0.0 | 1.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| neutral | 0.0 | 7.0 | 0.0 | 0.0 | 0.0 | 9.0 | 0.0 |
| sadness | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 11.0 |

SVM



| Actual \ Predicted | anger | boredom | disgust | fear | happiness | neutral | sadness |
|---|---|---|---|---|---|---|---|
| anger | 22 | 0 | 0 | 1 | 3 | 0 | 0 |
| boredom | 0 | 8 | 1 | 0 | 0 | 5 | 2 |
| disgust | 0 | 2 | 6 | 0 | 0 | 0 | 1 |
| fear | 0 | 2 | 0 | 9 | 2 | 1 | 0 |
| happiness | 10 | 0 | 0 | 0 | 3 | 1 | 0 |
| neutral | 0 | 4 | 0 | 0 | 0 | 11 | 1 |
| sadness | 0 | 0 | 0 | 0 | 0 | 1 | 11 |

MLP

# NEXT STEPS

Next Steps for Emotion Recognition

- **Feature Engineering & Selection**
  - Use PCA to reduce dimensionality and identify key features.
  - Analyze feature importance for better classification.
- **Model Optimization & Exploration**
  - Fine-tune SVM & MLP for better accuracy.
  - Test LSTM, BiLSTM, and CNN for improved sequence modelling.
  - Explore ensemble models for enhanced performance.
- Benchmarking & Literature Review (Optional - Based on feedback)
  - Compare results with research on datasets like IEMOCAP (Interactive Emotional Dyadic Motion Capture) & RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song).
  - Identify improvements for real-world applications.

# RESOURCES REFERRED

1️⃣ "Speech Emotion Recognition: Features, Classification Models, and Databases"
Zheng, Z., Zhang, Y., & Yu, H. (2019)

- Overview of key speech features (MFCC, prosody, spectral) and ML models used in emotion recognition.

2️⃣ "Extraction and Representation of Prosodic Features for Language and Speaker Recognition"
Leena Mary & B. Yegnanarayana, IIT Madras

- Discusses prosodic feature extraction techniques and their role in speech analysis.

3️⃣ "Deep Learning for Audio-Based Emotion Recognition: A Review"
Trigeorgis, G., Ringeval, F., & Schuller, B. (2018)

- Explores CNNs, RNNs, and hybrid deep learning models for speech emotion recognition.

SPEECH SIGNAL PROCESSING

# End Evaluation Presentation

# RECAP

- Dataset Exploration
- Feature Extraction:
- Extracted features MFCC and Prosody features at
  - Word-Level Features
  - Syllable-Level Features
  - Utterance-Level Features
- Combined all features into a single feature set per file
- Used mean-based aggregation to standardize feature lengths
- Combined Machine Learning Models for classifying emotions
  - Support  Vector Machines
  - Gaussian Mixture Models
  - Multi-Level Perceptron
- Previous Results;
  - SVM - 67%
  - MLP - 65.4%
  - GMM - 14%

# LIMITATIONS THAT WERE IDENTIFIED

- Combined features made it difficult to assess individual contribution of MFCC or prosody
- Averaging removed valuable temporal and structural cues (especially in prosody)
- No dedicated models per feature level — lacked granular analysis

# POST MID-EVALUATION IMPROVEMENTS

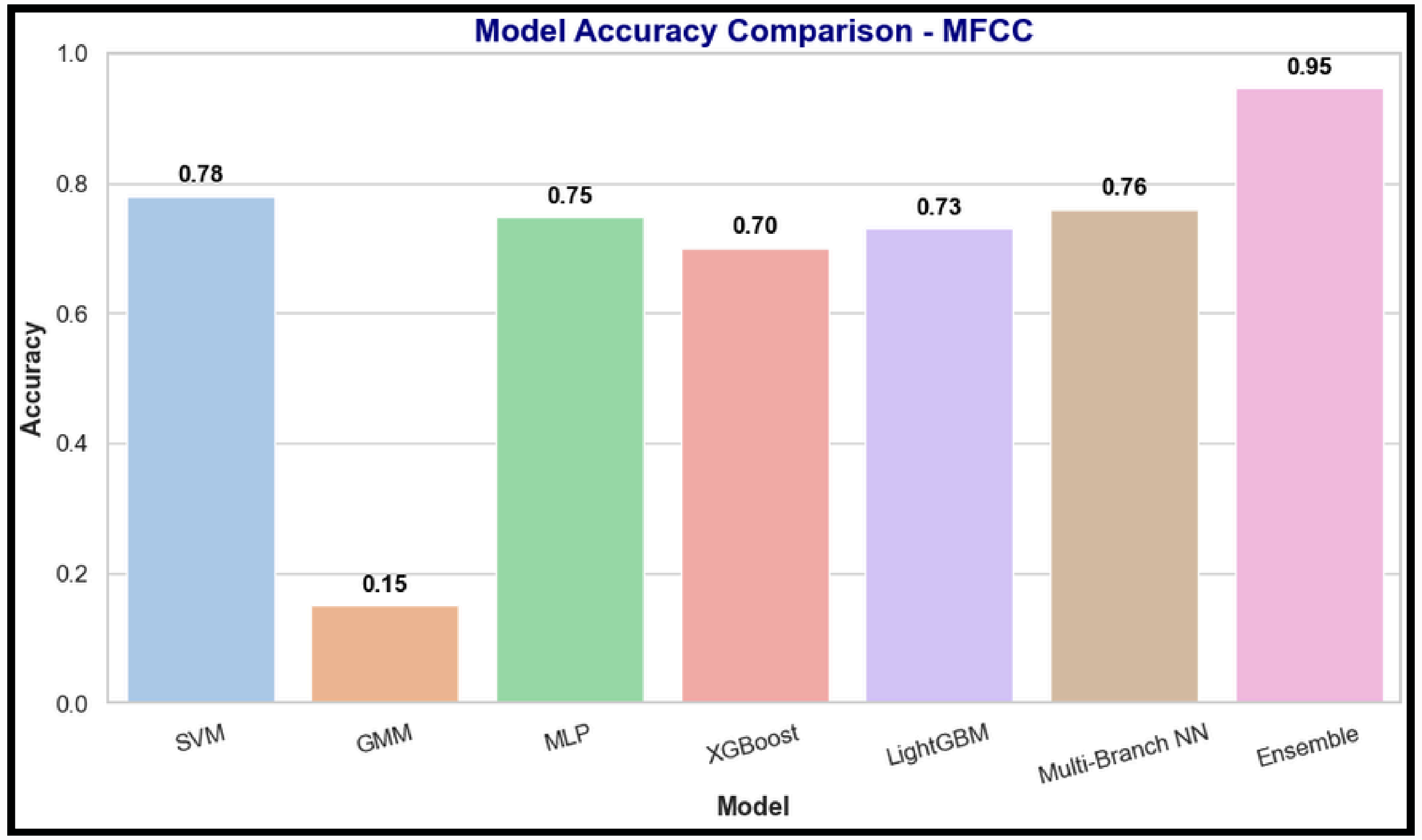
- **Feature Pipeline changes**
  - Extracted MFCCs only at utterance level (not repeated across all levels)
  - Created 4 separate feature sets: mfcc_features.csv, prosody_word.csv, prosody_syllable.csv, prosody_utterance.csv
  - Removed zero-padding and used mean-based resampling for consistency
- **Model Re-Architecture**
  - Trained separate models for each feature type using:
    - SVM, GMM, MLP, XGBoost, LightGBM
  - Implemented Ensemble Voting:
    - Within each feature level (SVM + MLP + XGBoost + LightGBM)
    - Across feature levels (final ensemble)
- **Deep Learning Extension**
  - Developed Multi-Branch Neural Networks for each feature set
  - Prepared groundwork for future fusion into a single multi-input model
- **Spectrogram Analysis of different emotions (audio)**
  - By analyzing pitch and spectrogram patterns across emotion, gender, and age using Praat, we found that variations in pitch height, contour, and spectral brightness effectively distinguish emotional states and speaker traits.

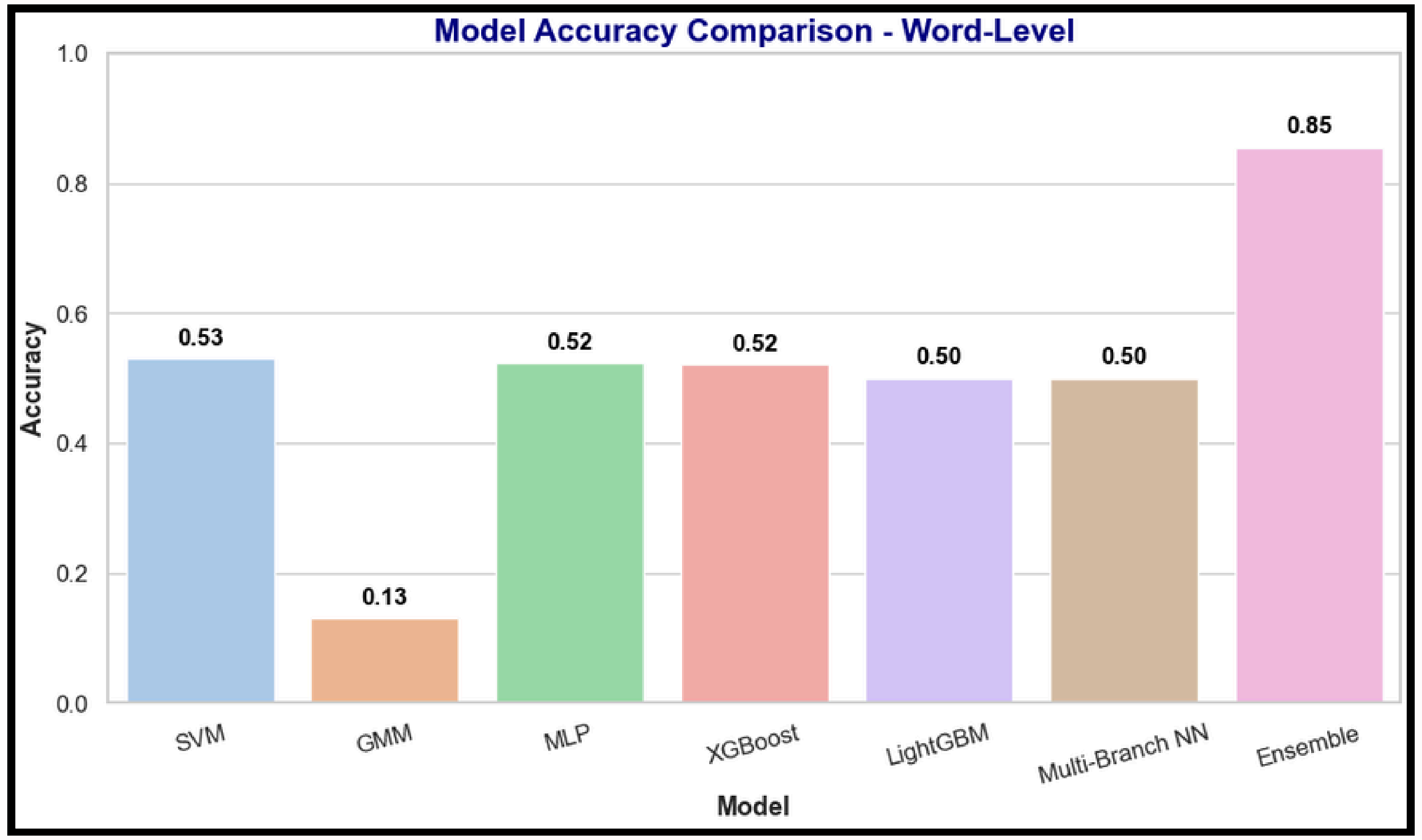
# MODEL PIPELINE IMPLEMENTED

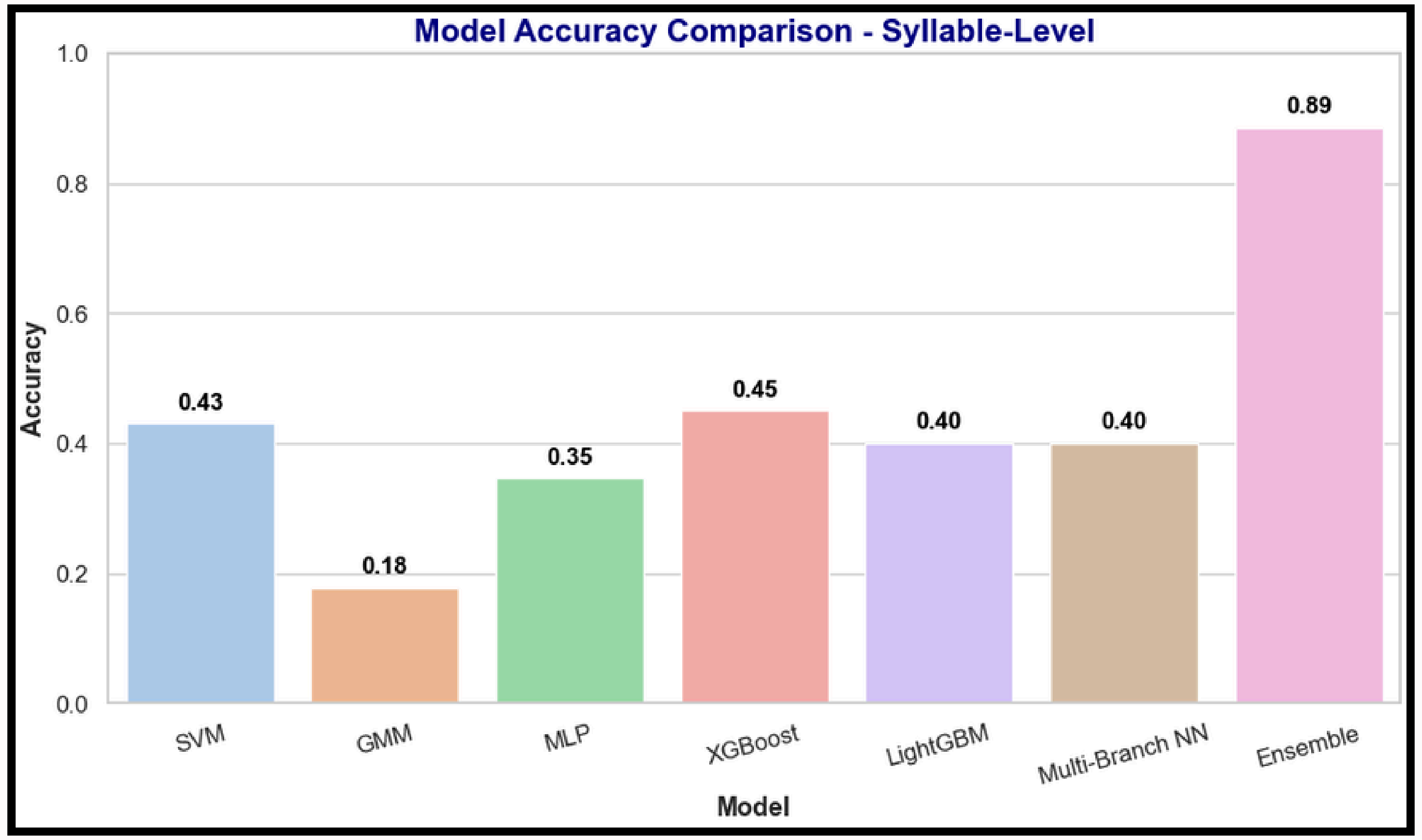
- **Machine Learning Models**
  - Support Vector Machine (SVM): Tuned with grid search; best performance on MFCC
  - Gaussian Mixture Model (GMM): Baseline probabilistic model; underperformed overall
  - Multi-Layer Perceptron (MLP): Fully connected feedforward network with dropout
  - XGBoost & LightGBM: Gradient boosting models for structured/tabular features
- **Deep Learning Models**
  - Multi-Branch Neural Network (MBNN)

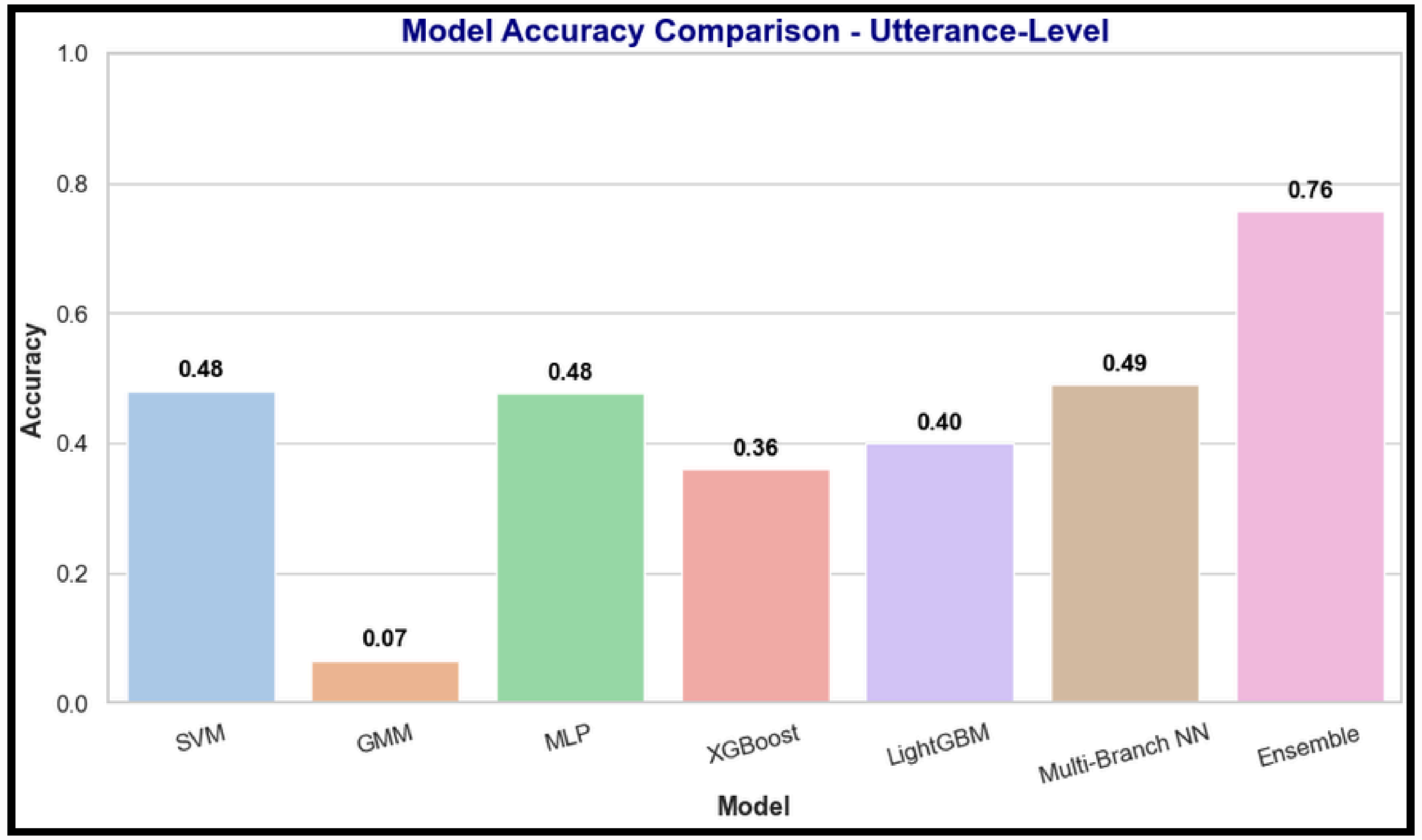
  One model trained independently on each feature type
    - MFCC Branch
    - Word-Level Prosody Branch
    - Syllable-Level Prosody Branch
    - Utterance-Level Prosody Branch
- **Ensemble Methods (Gave Highest Accuracy)**
  - Feature-Level Ensemble Voting: Combined multiple models (SVM, MLP, XGBoost, LightGBM) trained on the same feature set
  - Global Ensemble Voting: Combined predictions from MFCC, word, syllable, and utterance ensemble models

# RESULTS


Model Accuracy Comparison - MFCC


Model Accuracy Comparison - Word-Level


Model Accuracy Comparison - Syllable-Level


Model Accuracy Comparison - Utterance-Level

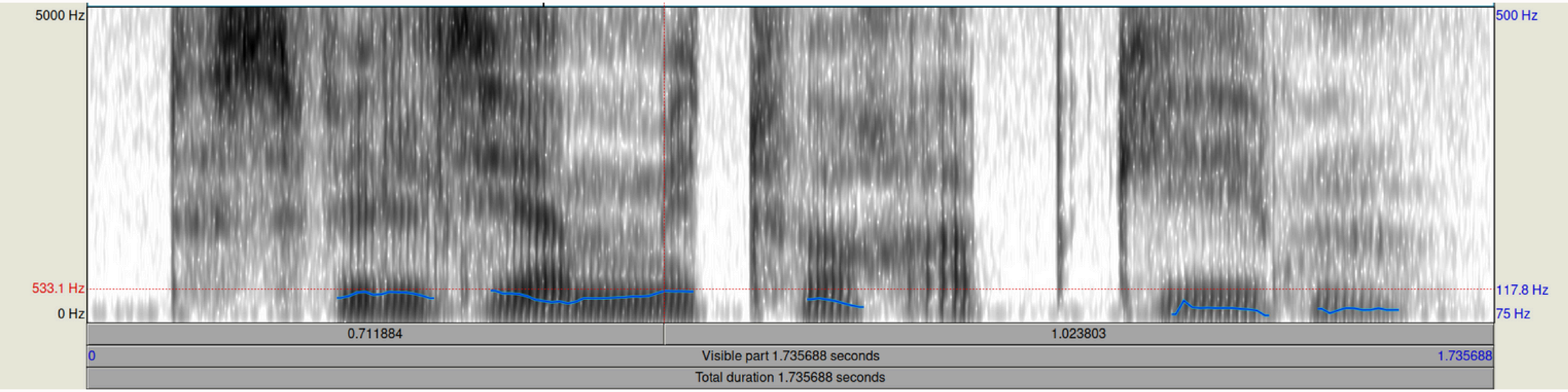
# MODEL-WISE COMPARISON (ACROSS ALL FEATURE TYPES)

- Ensemble Voting consistently achieved the highest accuracy
  - Combines outputs of multiple models, improving generalization and reducing variance
- SVM performed strongly on MFCC and structured prosodic data
  - Effective in high-dimensional spaces and handles non-linear boundaries well with kernels
- MLP showed stable results, particularly for MFCC and utterance-level prosody
  - Learns complex non-linear relationships, though sensitive to data quality and feature representation
- XGBoost and LightGBM performed well on prosody features
  - Handle heterogeneous tabular data and capture non-linear feature interactions efficiently
- GMM underperformed on all feature types
  - Assumes Gaussian distribution and struggles with high-dimensional, overlapping emotion classes
- Multi-Branch Neural Networks showed consistent performance across all features
  - Learn abstract patterns per feature type but without inter-feature synergy when used independently

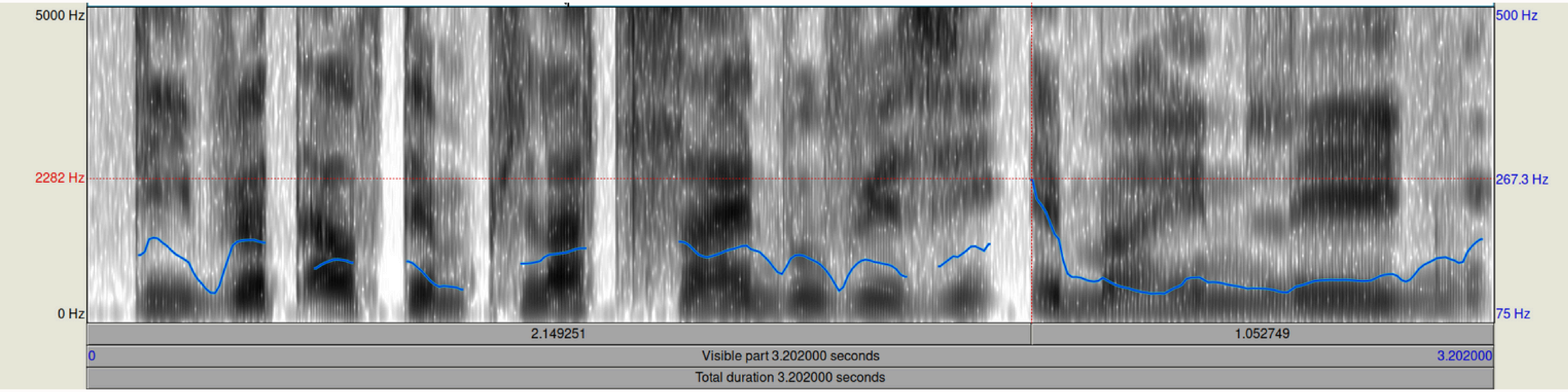
# FEATURE-WISE COMPARISON (ACROSS ALL MODELS)

- **MFCC** Features provided the ***best overall performance*** across most models
  - Capture spectral and phonetic characteristics directly tied to emotional expression
- **Word-Level Prosody** gave moderate individual performance, improved significantly with ensemble
  - Captures stress and emphasis patterns reflective of emotion at the word level
- **Syllable-Level Prosody showed limited standalone performance**, but improved through ensemble learning
  - Encodes rhythm and micro-intonation, which are more subtle and context-dependent
- **Utterance-Level Prosody was the least effective across models**
  - Aggregated features **lack granularity and fail to capture fine emotional transitions**
- Ensembles of feature-specific models performed best when features were complementary
  - Combining different perspectives allowed better emotion classification than individual models

# SPECTROGRAM AND PITCH ANALYSIS

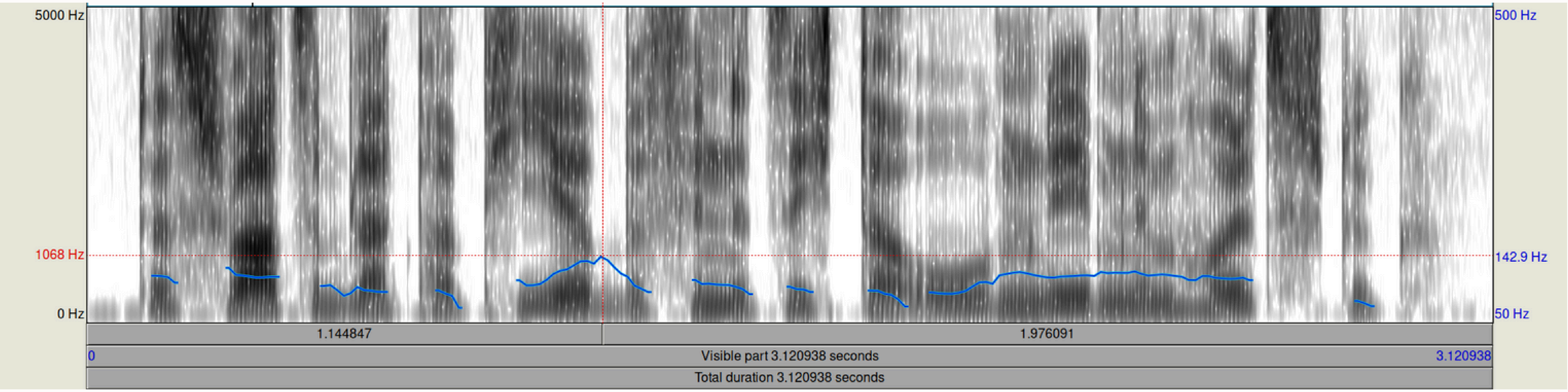


Emotion: **Sad**
Gender : **Male**
Pitch: **Low and Flat**
Spectrogram: **Dim, slow pitch movement**

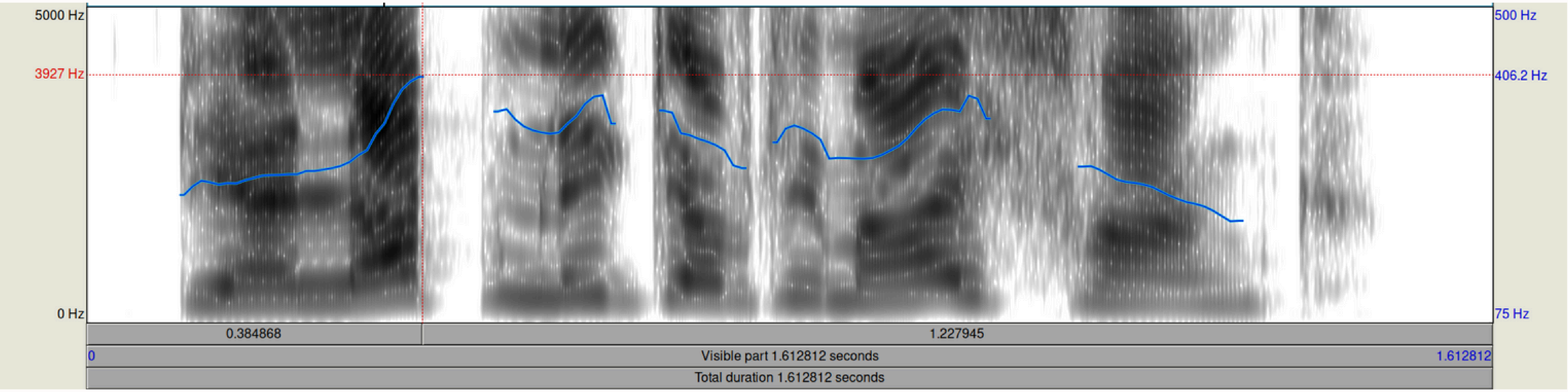


Emotion: **Disgust**
Gender : **Male**
Pitch: **Irregular**
Spectrogram: **Patchy brightness, rough transitions**

# SPECTROGRAM AND PITCH ANALYSIS



Emotion: **Neutral**
Gender : **Male**
Pitch: **Mid-low**
Spectrogram: **Even tone, smooth transitions, low variation**



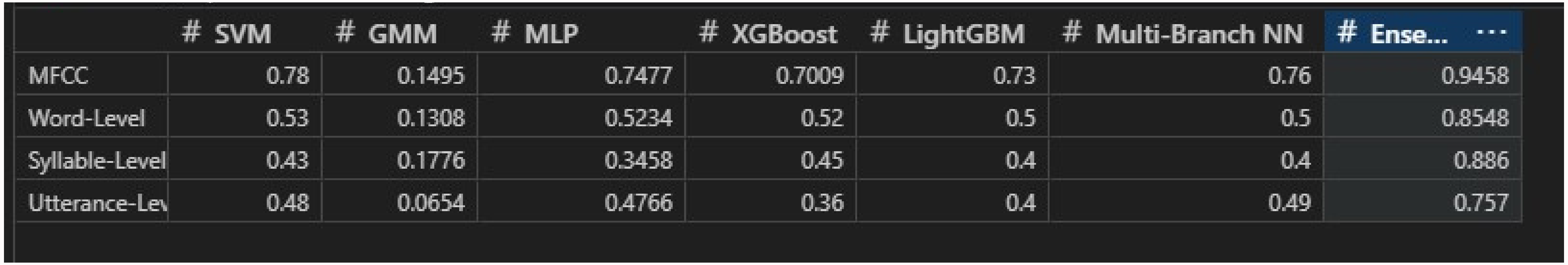
Emotion: **Angry**
Gender : **Female**
Pitch: **High**
Spectrogram: **Brighter, more pitch variation**

# SPECTROGRAM ANALYSIS SUMMARY

| *Emotion* | *Gender* | *Pitch Characteristics* | *Spectrogram Appearance* |
|---|---|---|---|
| **Angry** | **Male** | High | Bright, dense, sharp edges, forceful transitions |
| **Angry** | **Female** | Very high | Brighter, more pitch variation, sharper changes |
| **Boredom** | **Male** | Low, monotone | Dim, flat, low variation |
| **Boredom** | **Female** | Low-moderate | Slightly higher pitch, flat patterns |
| **Disgust** | **Male** | Irregular | Patchy brightness, rough transitions |
| **Disgust** | **Female** | Irregular, higher | Uneven spectrogram, more pitch fluctuation |
| **Fear/Anxiety** | **Male** | Shaky, high-moderate | Jittery bursts, irregular pitch patterns |
| **Fear/Anxiety** | **Female** | Very high & shaky | Tremble-like patterns, sharper intensity variations |
| **Happy** | **Male** | Varied, high-moderate | Rhythmic, melodic, bright |
| **Happy** | **Female** | High & sing-songy | Very melodic, fast rising/falling pitch |
| **Sad** | **Male** | Low & flat | Dim, slow pitch movement |
| **Sad** | **Female** | Low-moderate, flatter | Minimal pitch movement |
| **Neutral** | **Male** | Mid-low | Smooth, even tone, low variation |
| **Neutral** | **Female** | Mid | Slightly higher pitch, consistent transitions |

# KEY TAKEAWAYS

- MFCC features were the most consistent and effective across all models
- Prosody features (pitch and energy) at word and syllable levels revealed valuable rhythmic and stress-based emotional cues
  - Highlighted the role of local variations in speech in conveying emotions
- Utterance-level prosody, being coarse-grained, failed to capture finer emotional transitions
  - Confirmed that emotional signals are often embedded in short-term speech patterns
- Feature granularity matters: emotion is more detectable at word/syllable scale than at sentence scale
- Statistical analysis (Friedman Test) confirmed that feature type significantly affects model performance
  - P-value: 0.0047 — performance variation across feature sets is not by chance



|  | # SVM | # GMM | # MLP | # XGBoost | # LightGBM | # Multi-Branch NN | # Ense... ··· |
|---|---|---|---|---|---|---|---|
| MFCC | 0.78 | 0.1495 | 0.7477 | 0.7009 | 0.73 | 0.76 | 0.9458 |
| Word-Level | 0.53 | 0.1308 | 0.5234 | 0.52 | 0.5 | 0.5 | 0.8548 |
| Syllable-Level | 0.43 | 0.1776 | 0.3458 | 0.45 | 0.4 | 0.4 | 0.886 |
| Utterance-Lev | 0.48 | 0.0654 | 0.4766 | 0.36 | 0.4 | 0.49 | 0.757 |

# RESOURCES REFERRED

1️⃣ "Speech Emotion Recognition: Features, Classification Models, and Databases"
Zheng, Z., Zhang, Y., & Yu, H. (2019)

- Overview of key speech features (MFCC, prosody, spectral) and ML models used in emotion recognition.

2️⃣ "Extraction and Representation of Prosodic Features for Language and Speaker Recognition"
Leena Mary & B. Yegnanarayana, IIT Madras

- Discusses prosodic feature extraction techniques and their role in speech analysis.

3️⃣ "Deep Learning for Audio-Based Emotion Recognition: A Review"
Trigeorgis, G., Ringeval, F., & Schuller, B. (2018)

- Explores CNNs, RNNs, and hybrid deep learning models for speech emotion recognition.

4️⃣ "Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus" Syed Abbas Ali, Anas Khan, & Nazia Bashir (NED University)

- Investigates how pitch-based prosodic features influence the accuracy of various classifiers in speech emotion recognition.

# Thanks!