Speech Emotion Recognition: Problem Statement & Significance

The Challenge

- Speech Emotion Recognition (SER) is complex due to high variability in vocal expressions across different speakers.
- Overlapping acoustic features (e.g., pitch and energy) make it difficult to distinguish between emotions like "Fear" and "Happy."

Why It Matters

- **MentalHealth:** Detecting depression or anxiety markers in voice.
- **Human-Computer Interaction:** Enabling voice assistants to understand user sentiment, not just commands.
- **Market Demand:** The Affective Computing Market is expected to grow from USD 76.310 billion in 2025 to USD 192.189 billion in 2030, at a CAGR of 20.29%.

Literature Review & Research Gap

Classical

Relied on hand-craftedfeatures (MFCC, pitch) with SVMs; limited generalization.

2019

LSTM with attentionmechanisms (69%).

1

2

3

4

2017

CNN on spectrograms (64%).

2023

Pre-trained modelslikeWav2Vec2 (75%).

The Gap

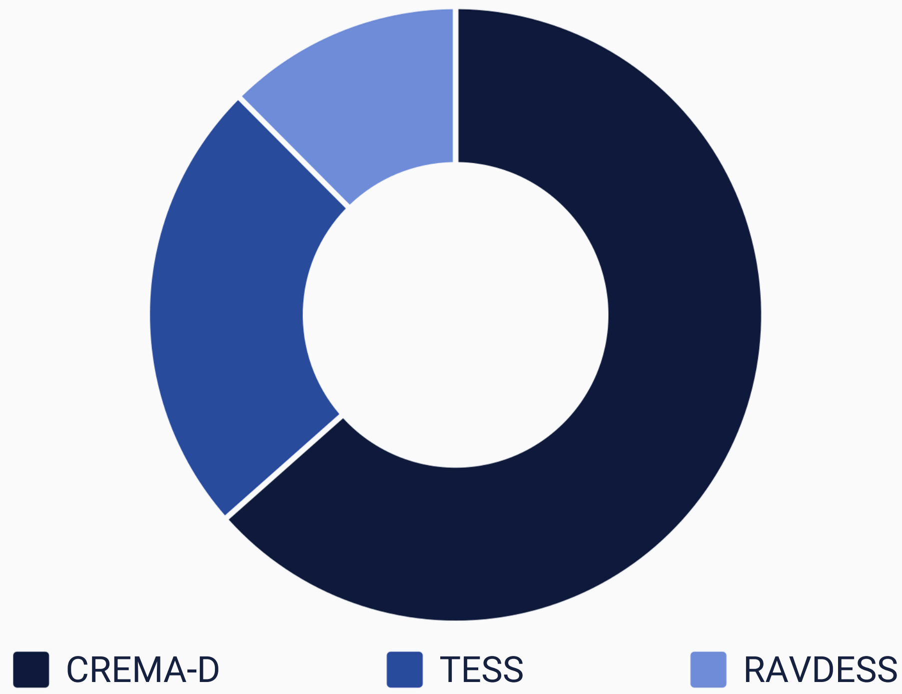
Mostresearch focuses on single architectures. Our approach uniquely fuses **Spatial (CNN)**, **Temporal (LSTM)**, and **Global (Transformer)** feature extraction to improve robustness.

Dataset Analysis & Distribution

DataSources

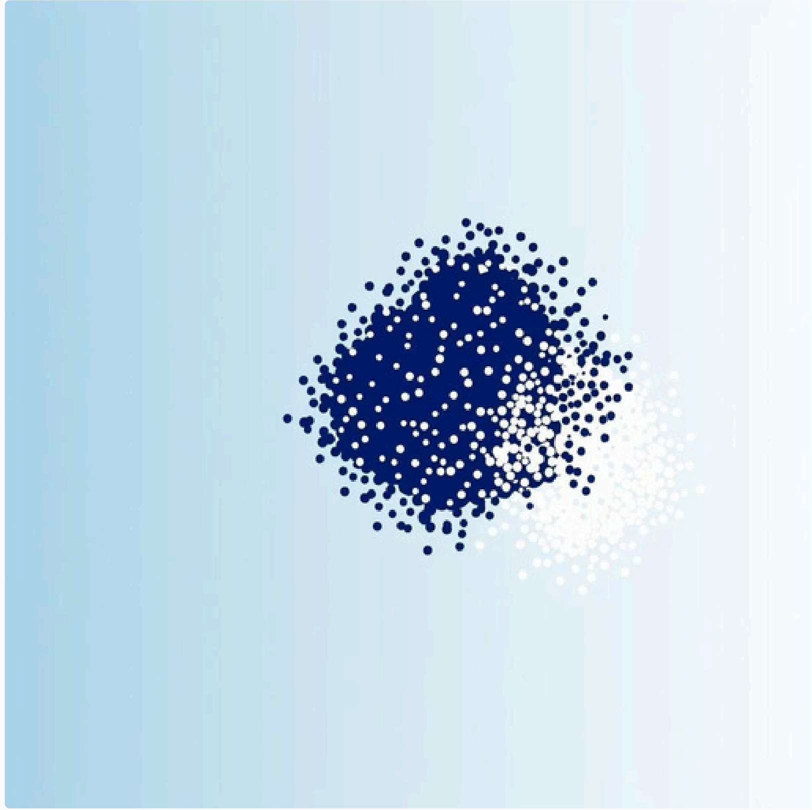
Combinedthreebenchmark speech datasets: **RAVDESS**, **TESS**, and **CREMA-D**.

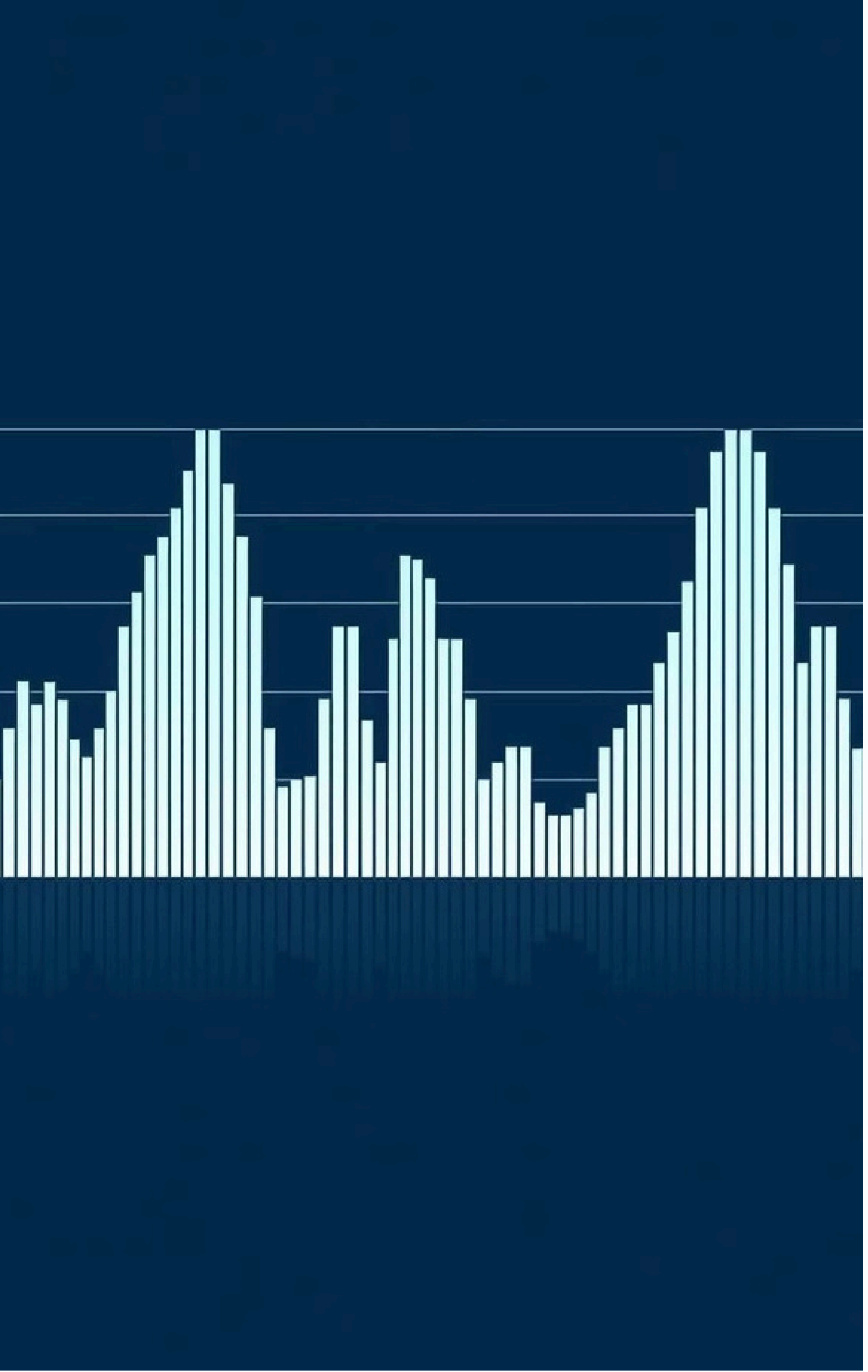
Total Volume: 11,682 audio samples.



Class Imbalance Challenge

Emotionslike"Surprise" are underrepresented (5.1%) compared to "Happy" or "Angry" (15.9% each), requiring careful handling during training.





Feature Engineering for Speech

Input Representation

Rawaudiowaveforms(16kHz sampling rate, 3-second fixed duration).

Feature Extraction Strategy

128-dimensionalMel-spectrograms: Captures the frequency spectrum over time, ideal for CNN processing.

Preprocessing: Amplitude-to-dB conversion, normalization, and padding/truncation.

Data Augmentation

Applied**NoiseInjection**,**Time Shifting**, and **Speed Perturbation** to simulate different recording conditions and improve model generalization.

Neural Network Architectures

1. CNN

Role: Extracts local spectral patterns from Mel-spectrograms.

Performance: 62.01% accuracy.

2. Bi-LSTM

Role: Modelstemporal dependencies in speech sequences (Forward + Backward context).

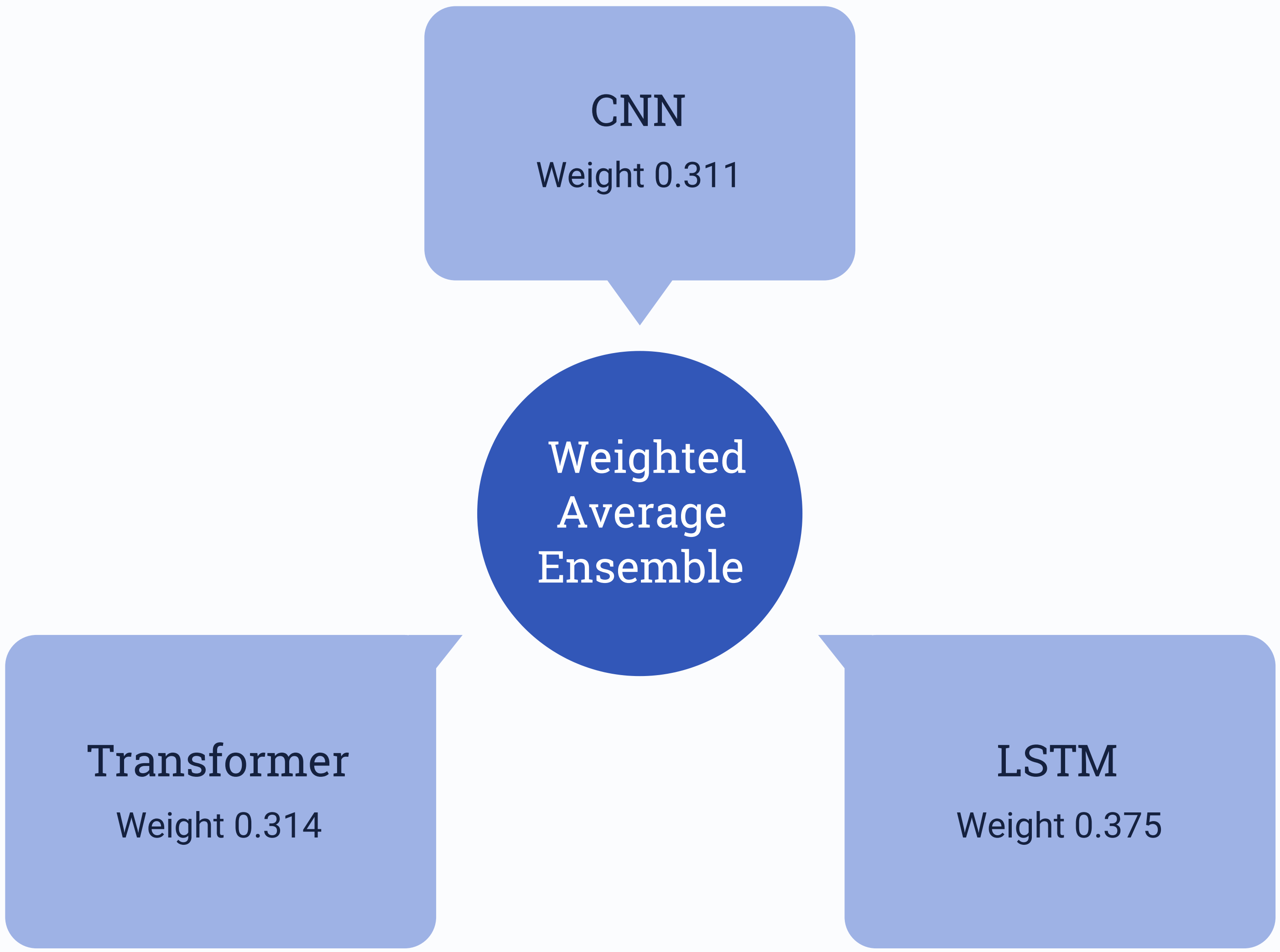
Performance: 67.66% accuracy (Best Individual Model).

3. Transformer

Role: Captures global relationships using self-attention mechanisms.

Performance: 45.07% accuracy (Limited by dataset size).

Ensemble Architecture Design



The Fusion Strategy

A **Weighted Average Ensemble** combines the outputs of all three models.

$$Output = 0.311 \cdot CNN + 0.375 \cdot LSTM + 0.314 \cdot Transformer$$

Key Insight

The systemlearned to weigh the **LSTM highest (37.5%)** because temporal dynamics are most critical for speech emotion.

Total Parameters: 3.3 Million.

Result: 66.74% Test Accuracy, outperforming the CNN and Transformer individually.

Training Strategy & GPU Optimization

Optimization Techniques

- **Hardware:** Dual T4 GPUs.
- **Mixed Precision (FP16):** Reduced memory usage and increased speed.
- **DataParallelism:** Distributed batch processing across GPUs.

Hyp erparameters

BatchSize:256 | Optimizer: Adam | LR Scheduling: Cosine Annealing.

Performance Gains

80%+

GPU Utilization

Increased from 18%

2.8x

Training Speedup

Reduced from 37s to 19s/epoch

Results - Overall Performance Metrics

Accuracy

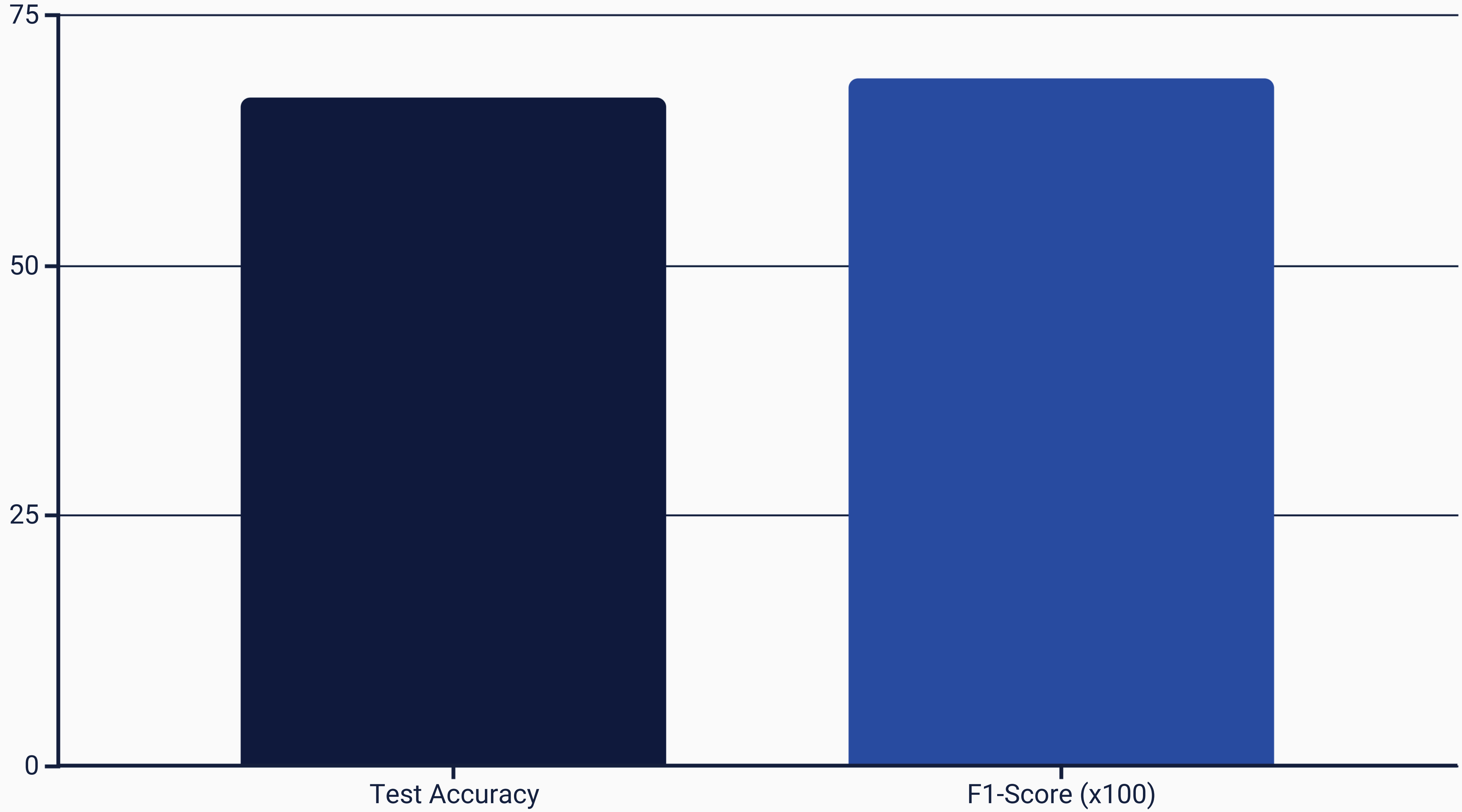
TestAccuracy: 66.74% (Within industry benchmark of 65-75%).

Macro F1-Score: 0.687 (Good balance across classes).

Efficiency Metrics

Training Time:6.3 minutes for 20 epochs.

Inference Speed: 23ms per sample, making the system capable of **real-time speech processing**.



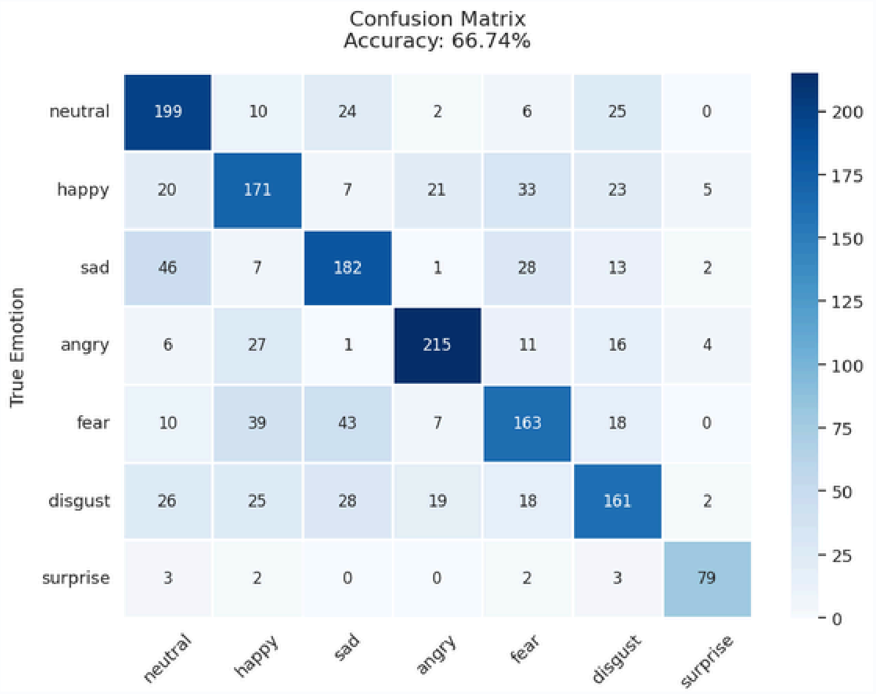
Results - Per-Class Analysis & Confusion

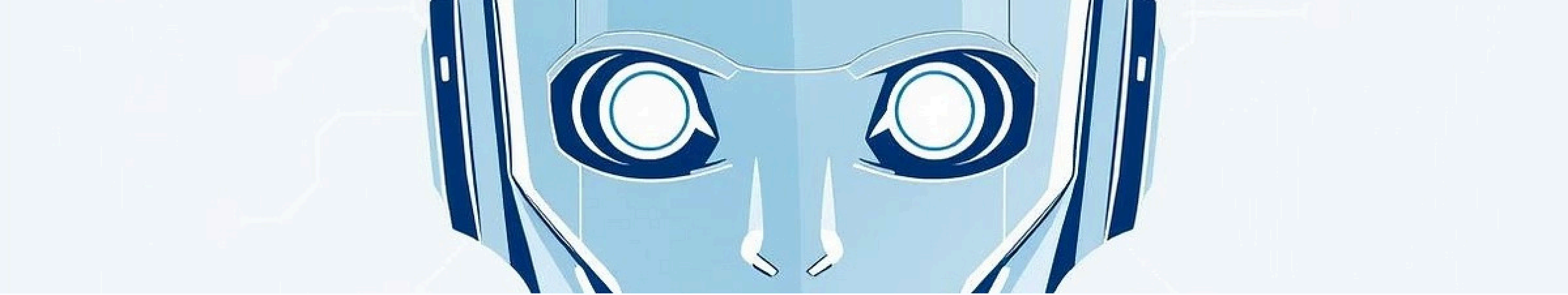
Best Performing Emotions

- **Surprise:** 88.8%(Distinct acoustic signature).
- **Angry:** 76.8%.

ChallengingEmotions (Confusions)

- **Disgust:** 57.7% (Lowest accuracy due to similarity to other negative emotions).
- **Sad vs. Neutral:** Confused due to similar low arousal characteristics.
- **Fear vs. Happy:** Confused due to overlapping high-pitch features.





Ablation Studies & Future Work

Ablation Study (Architecture Impact)

- Removing **LSTM**causedthe largestdrop (-3.54%), proving it is the critical component.
- Removing the Transformer had negligible impact (+0.36%), suggesting it requires more data.

Future Directions

- **Short-term:** Use SpecAugment and weighted loss functions to handle class imbalance.
- **Long-term:** Fine-tune pre-trained speech models (Wav2Vec2, HuBERT) for higher accuracy and cross-lingual support.