Classification of Emotion using EMO-DB

Foundation of Deep Learning Project 2024-2025

by Chiara Genuardi and Giovanni Noè

Table of Contents

We used the Berlin Database of Emotional Speech, containing **535 wav** audio files of sentences spoken in **German** by professional **actors**.
Each sentence **expresses** one emotion:

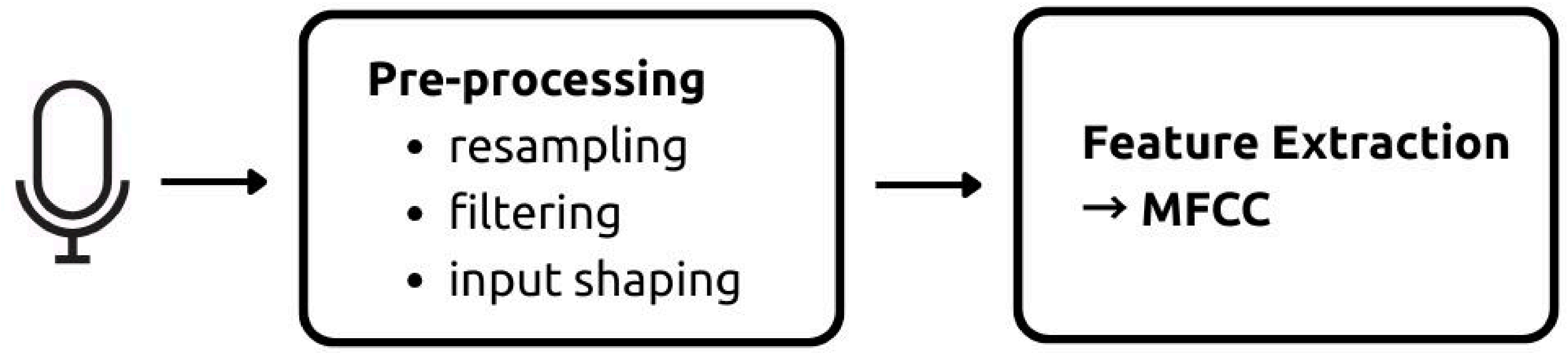
- anger
- boredom
- disgust
- happiness
- fear/anxiety
- sadness
- neutral version

None of these are present in the same numbers nor each speaker produced the same number of audios. This is taken care in LOSO and Data Augmentation sections.

Audio’s lengths span between 1.23 to 8.98 seconds.

Pre-processing steps:

- resampling → **16kHz**
- filtering → low pass **4kHz**
- input shape → **64000 samples ~ 4s**
  - padding (short) / trimming (long)

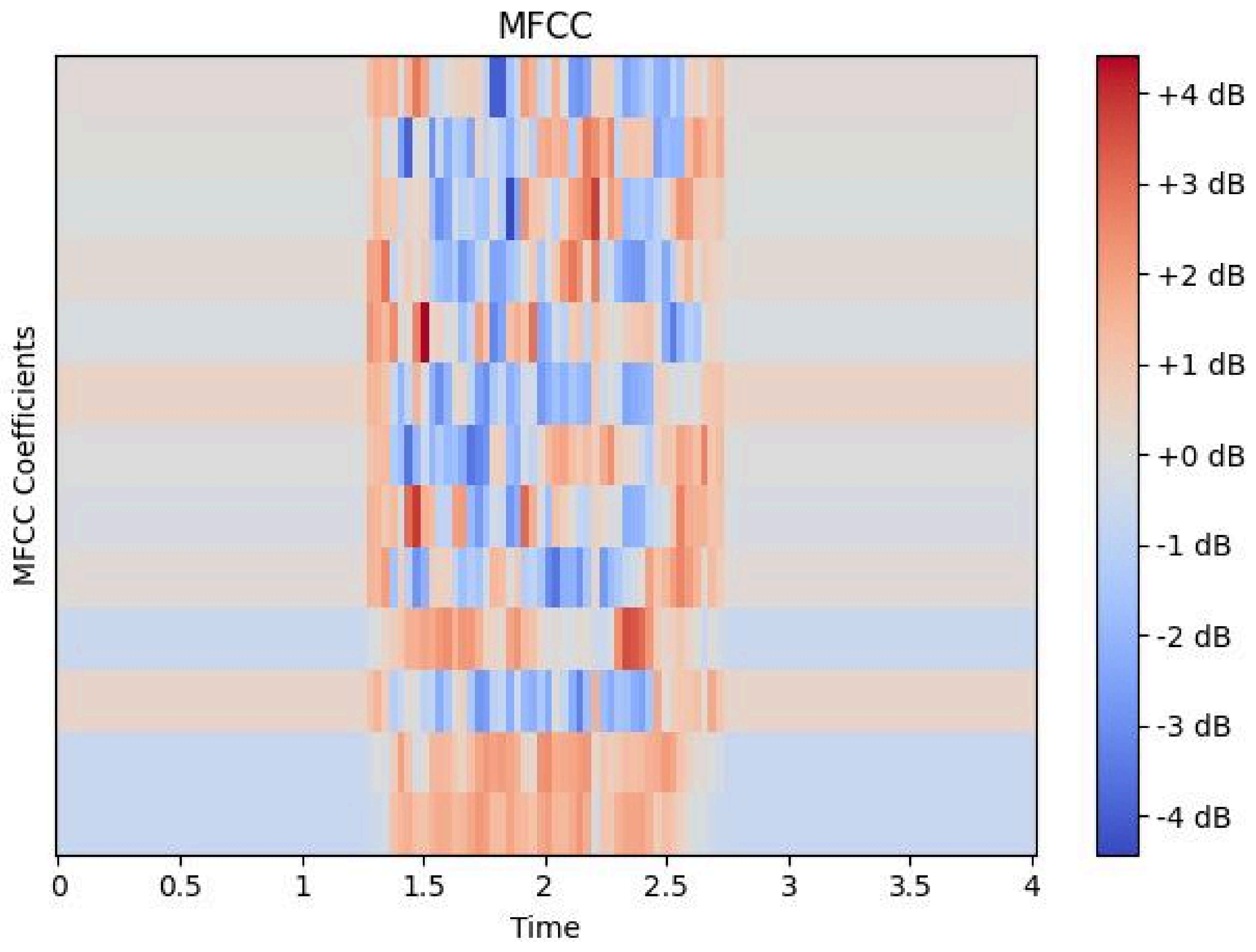


Mel-Frequency Cepstral Coefficients (**MFCC**) are widely used in audio recognition as they capture salient features of speech.

- **13** coefficients
- **1024** FFT dimension, **512** hop lenght

  → **13 coefficients x 126 frame**

  → librosa.feature.mfcc

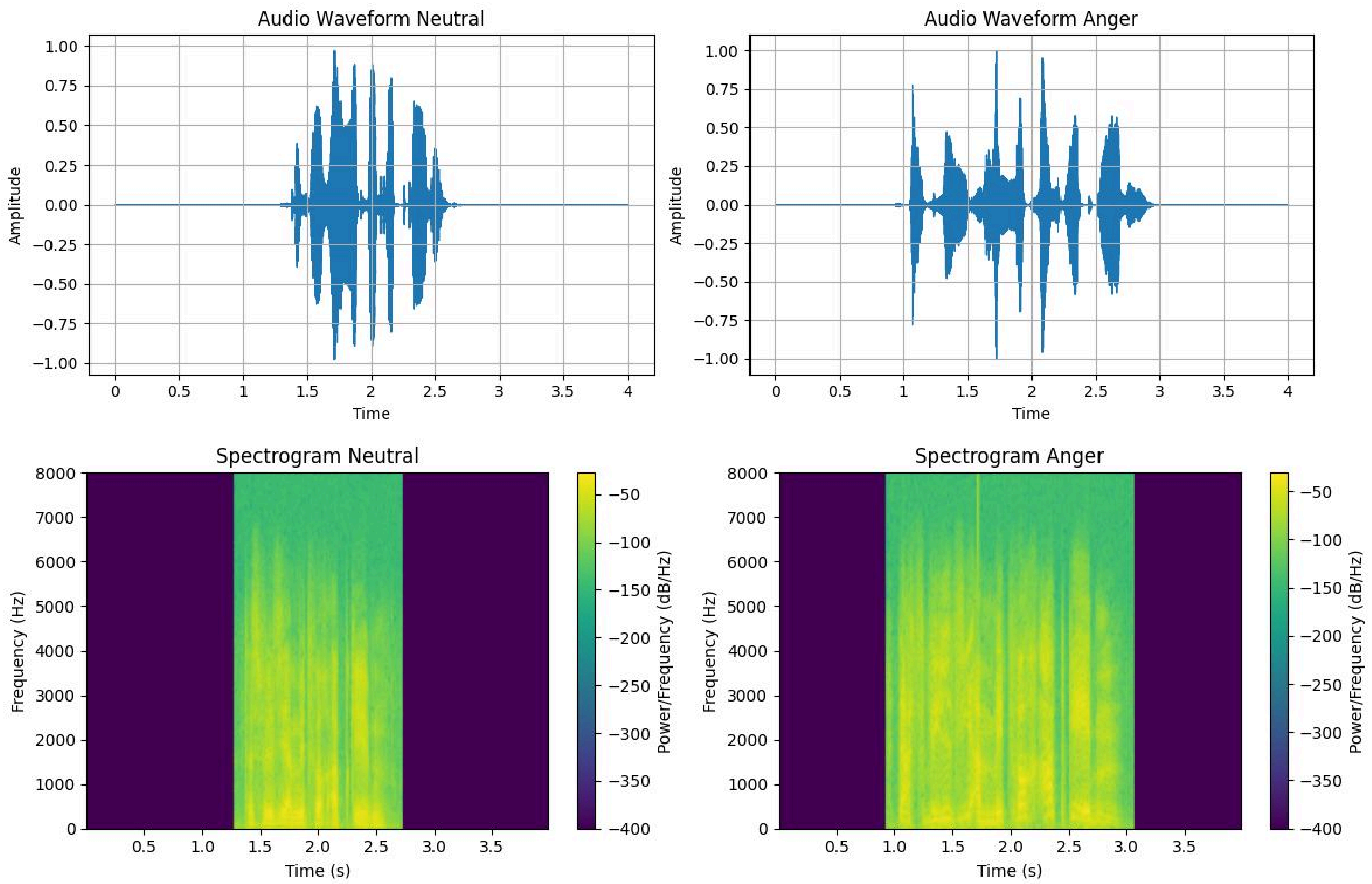


Low order coefficients → global properties
High order → finer details.

Basic exploratory analyses show noticeable differences for different emotions.

→ Audio waveform is shown as Amplitude vs Time.

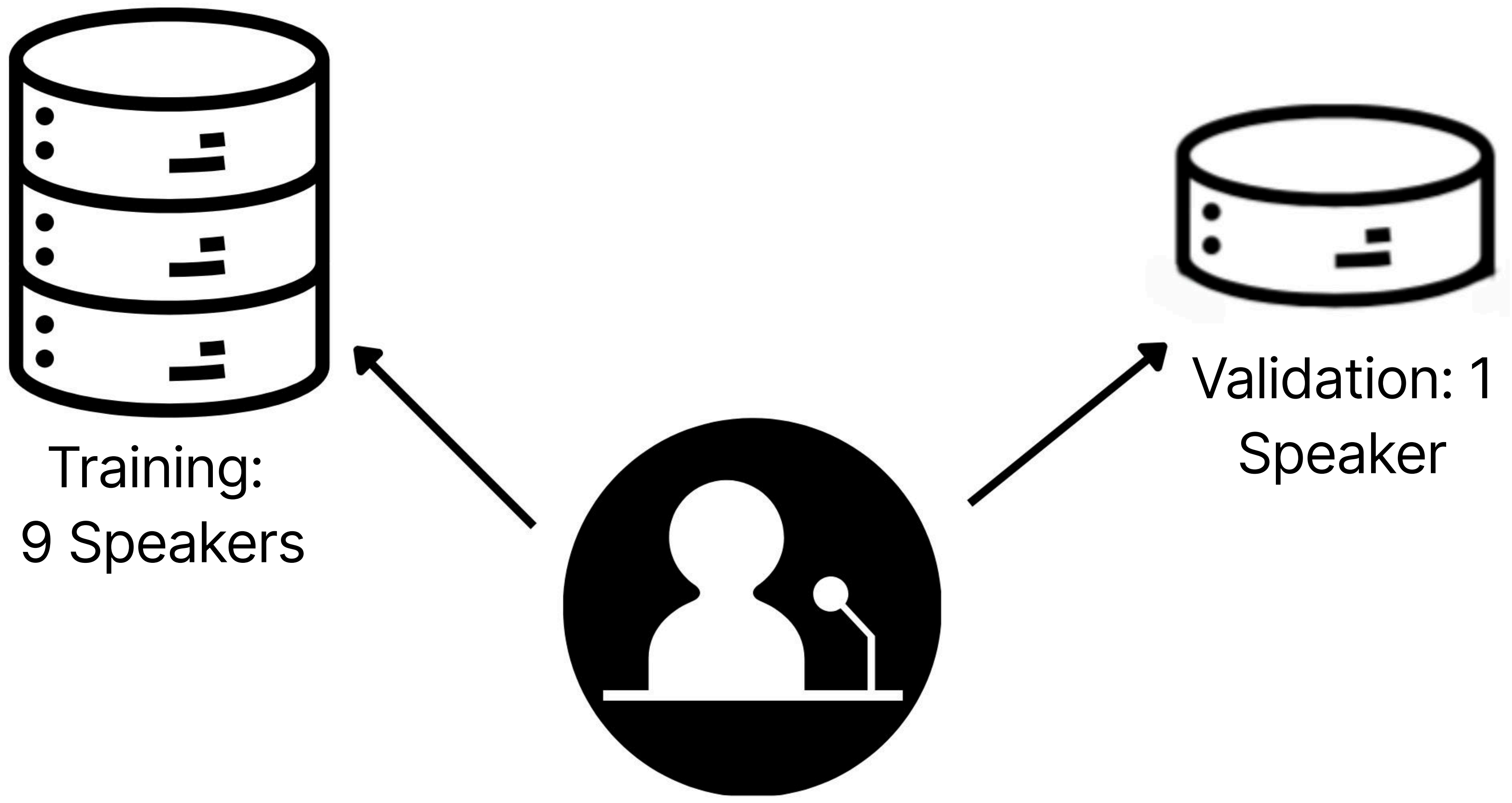
→ Spectrograms show the intensity of sound as a function of frequency and time.



Emo-DB is small and imbalanced, requiring strategies to enhance performance reliability and model generalization.

LOSO Cross-Validation

- Each speaker = one fold (10 folds total)
- Training: 9 speakers → Validation: 1 unseen speaker
- Same partitioning reused for all models



Data Augmentation

- Only on training sets
- Two strategies:
  - Noise addition (random values in MFCCs)
  - Time shifting (±0.8 sec)
- Balancing: ensure ≥150 samples per emotion
- Training sets expanded to 2-3 times the original size

Explored architectures:

- MLP (baseline)
- CNNs (1D, 2D)
- LSTM RNN
- Hybrid models (LSTM+CNN1D, LSTM+CNN2D)

Training setup:

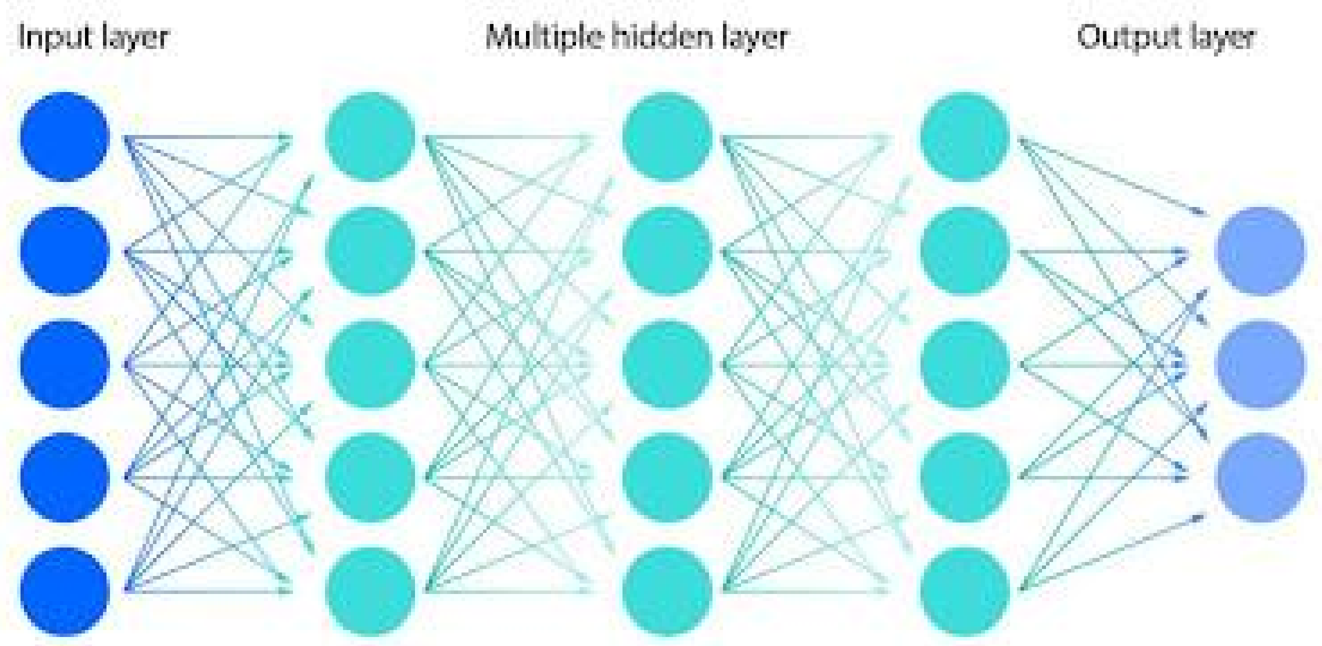
- Optimizers: RMSprop / Adam (only MLP)
- LR = 0.001, up to 25 epochs, early stopping

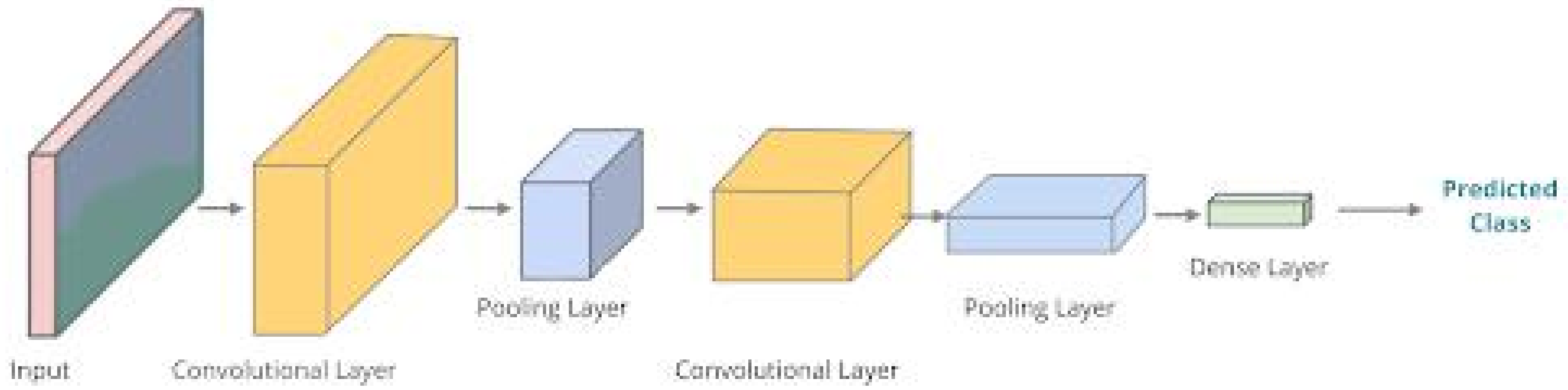
Evaluation metrics:

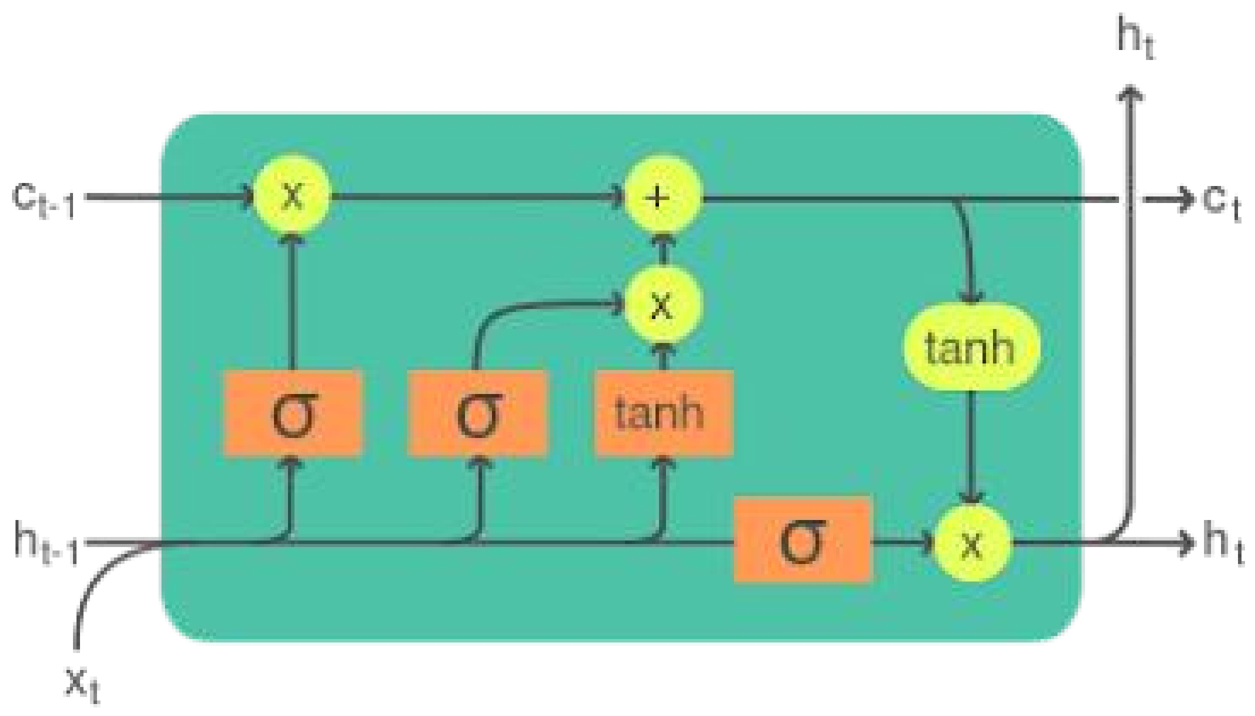
- Accuracy, Top-3 Accuracy, Precision, Recall, F1-score
- Confusion matrix, ROC, AUC

Results:

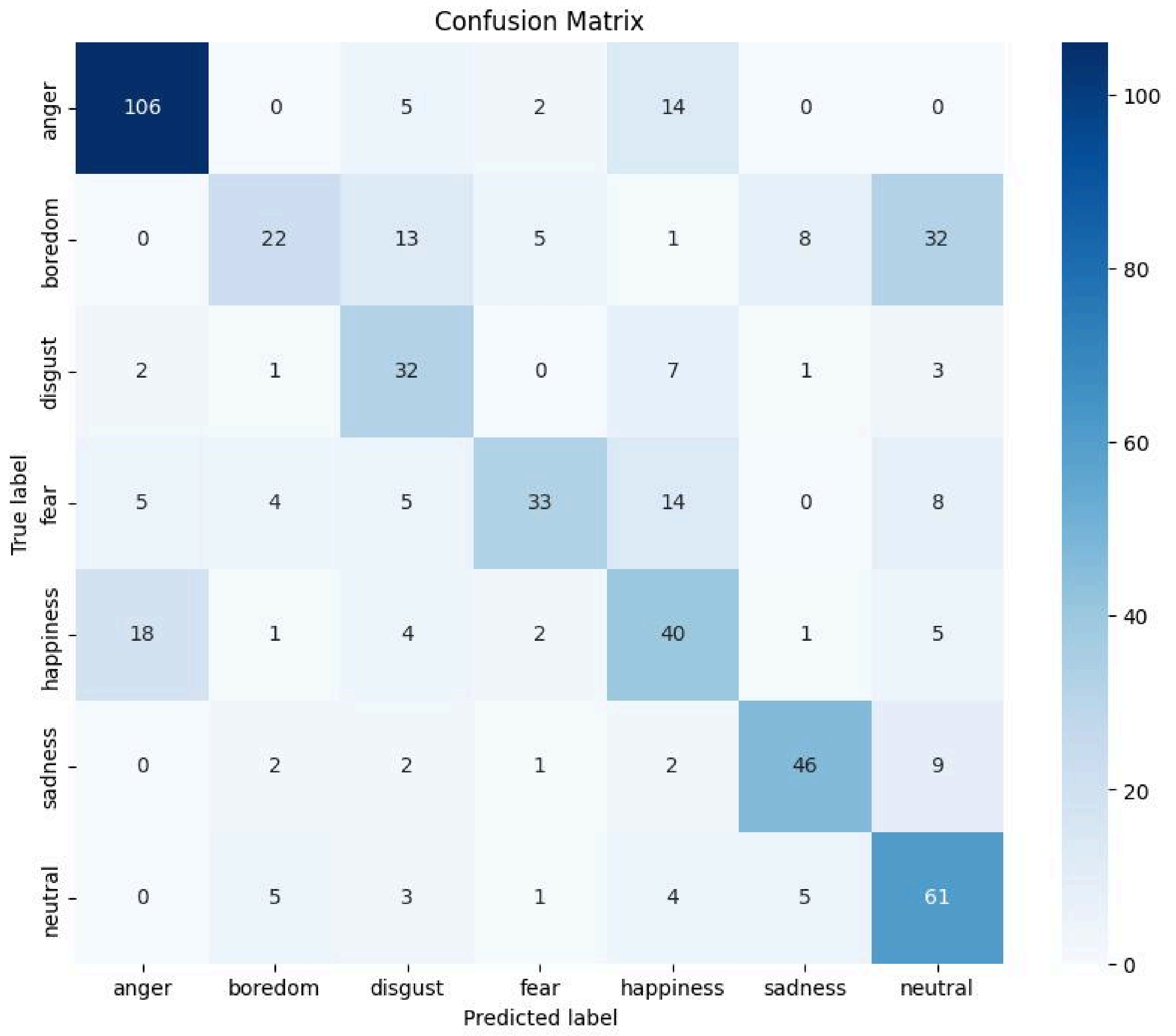
- Most models comparable to MLP
- Focus on baseline and best performing models







Multi-layer Perceptron - Baseline

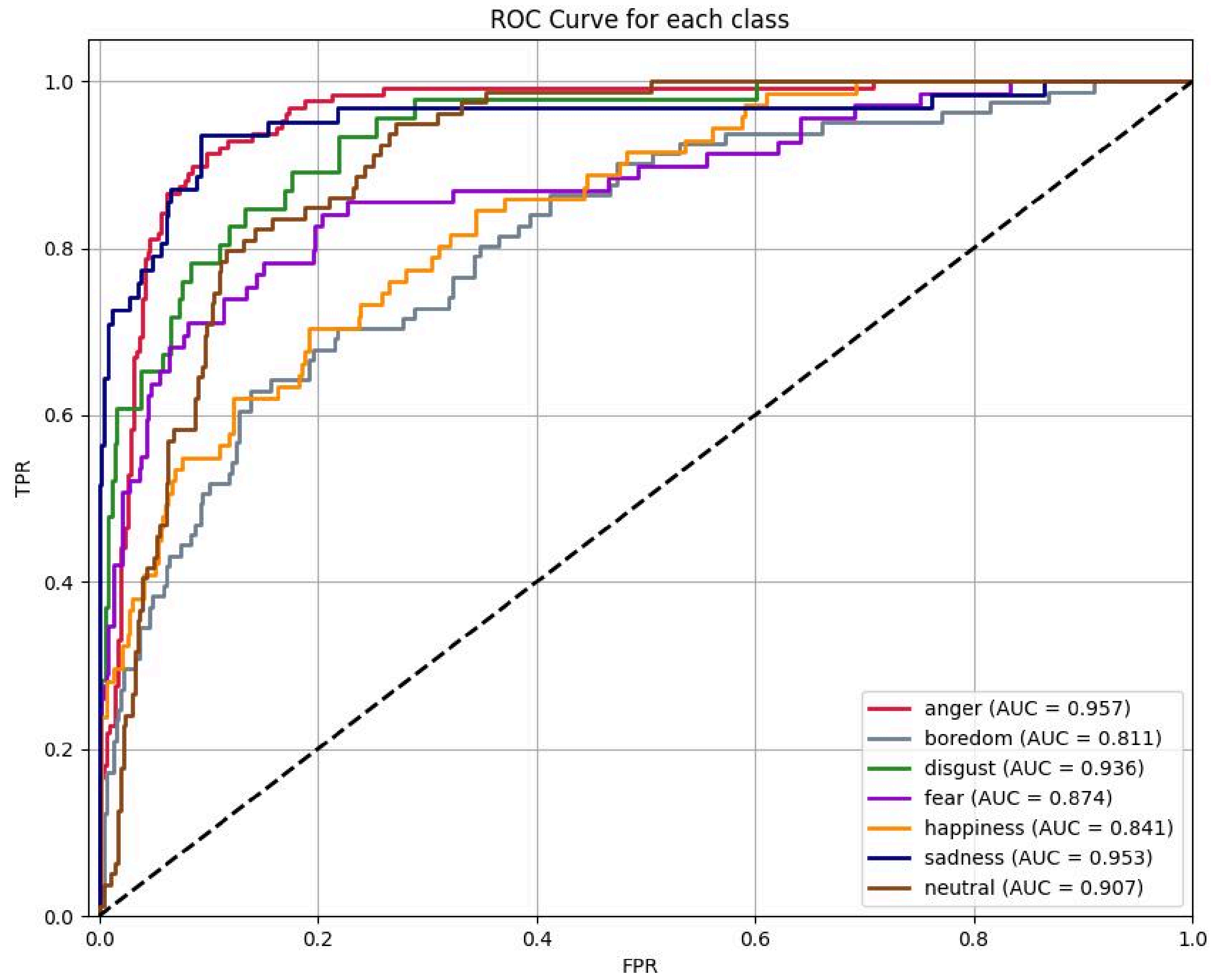


Architecture:

- Input: MFCCs reduced to 26 features
- 4 dense layers (256 → 128 → 64 → 32)
- ReLU activations, batch normalization
- Output: 7 neurons (emotions)

Results (LOSO CV):

- Avg. accuracy (per speaker): 0.50
- Overall accuracy: 0.66
- Top-3 accuracy: 0.85
- Weighted F1-score: 0.64



Findings:

- Best: Anger and Sadness (AUC = 0.957 and 0.953)
- Worst: Boredom (AUC = 0.811, often confused with neutral)
- Happiness frequently misclassified as anger

Architecture:

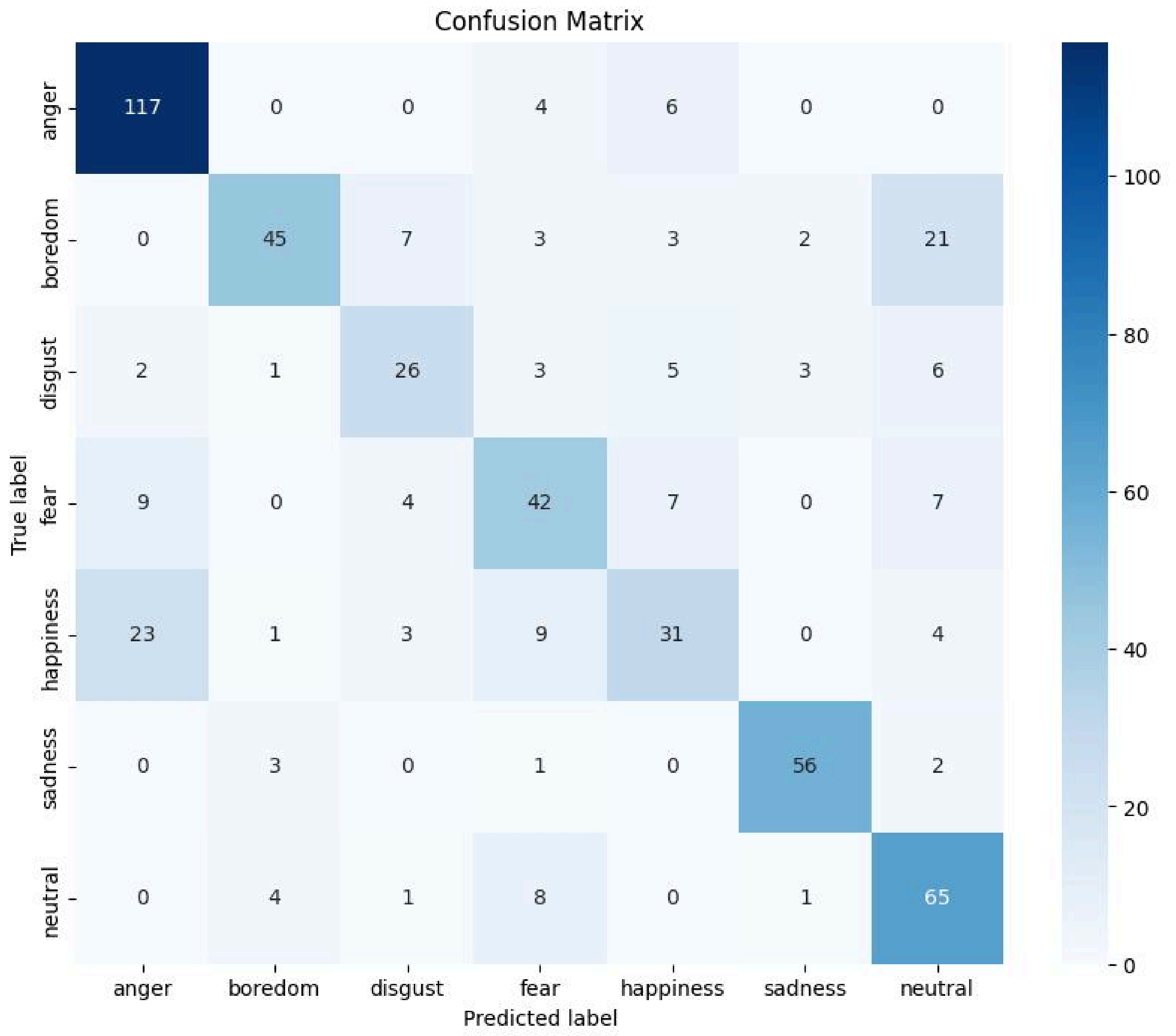
- Input: 126 frames x 13 coefficients x 1 channel
- 4 CNN2D layers (16 → 32 → 64 → 128)
- ReLU activations, batch normalization, max and global pooling
- dense layer (64) using ReLU
- Output: 7 neurons (emotions)

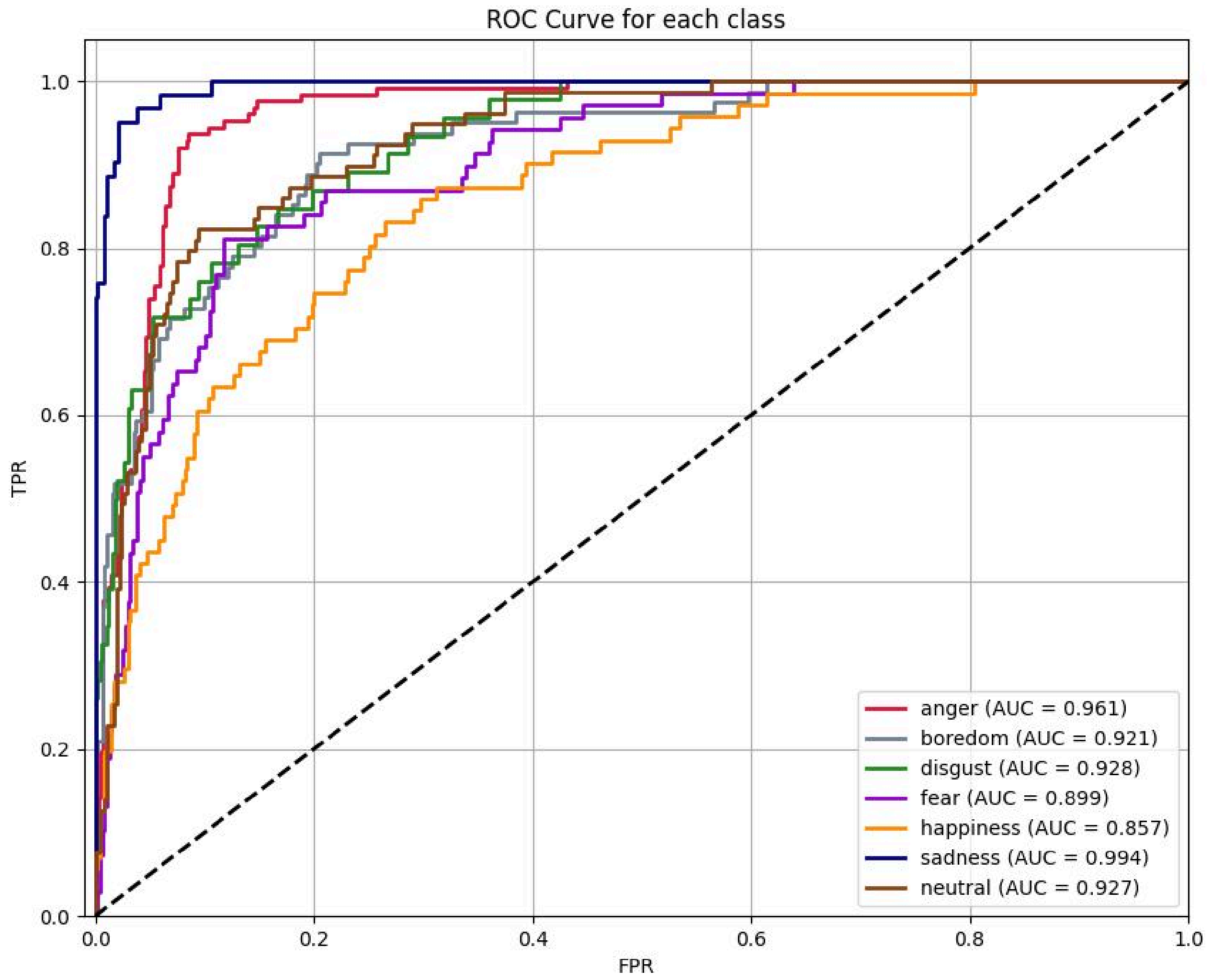
Results (LOSO CV):

- Avg. accuracy (per speaker): 0.63
- Overall accuracy: 0.71
- Top-3 accuracy: 0.93
- Weighted F1-score: 0.69

Findings:

- Best: Sadness (AUC = 0.994)
- Worst: Happiness (AUC = 0.857)





Convolutional Neural Network 2D + Long Short-Term Memory

Architecture:

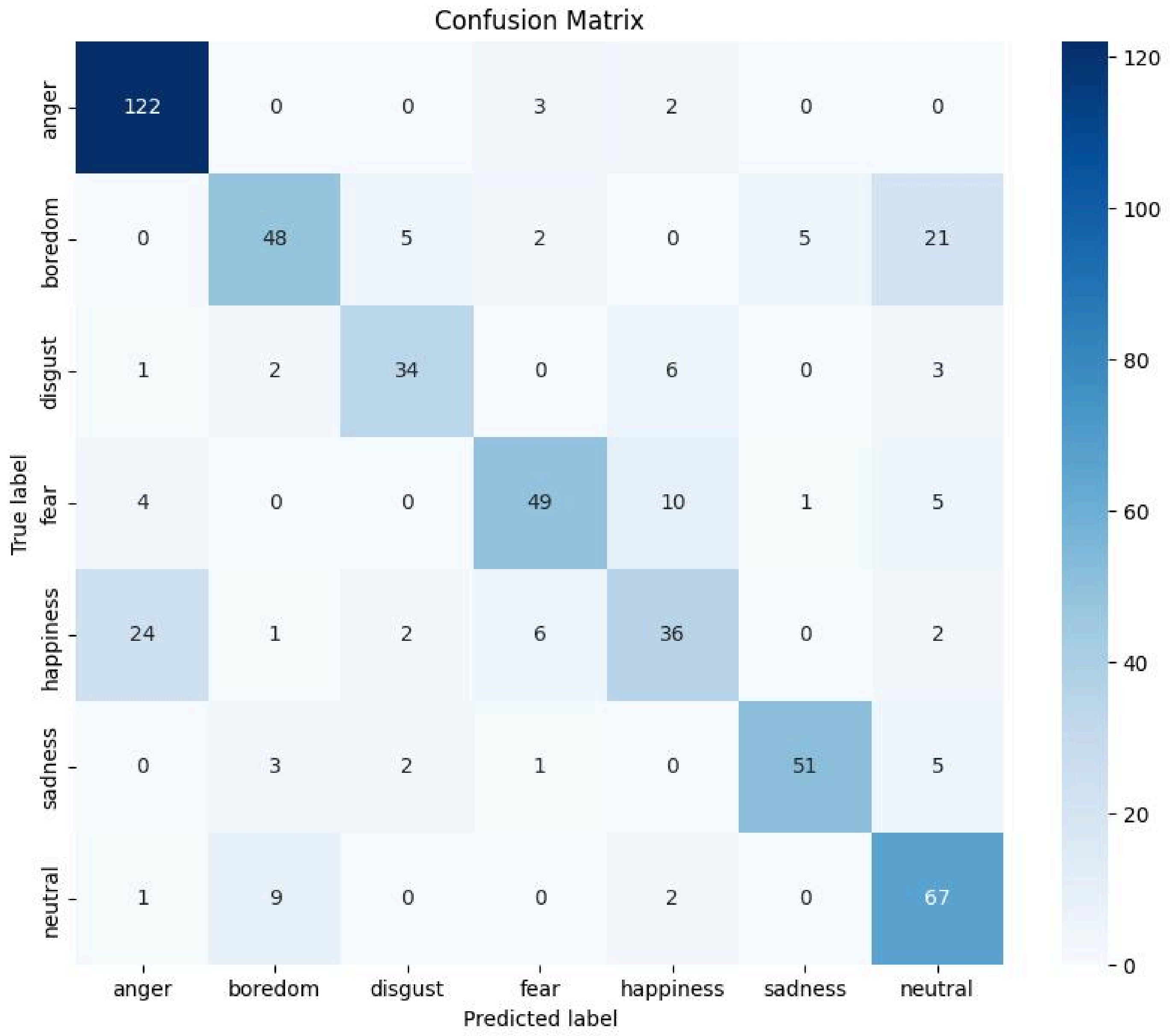
- Input: 126 frames x 13 coefficients x 1 channel
- 2 CNN2D layers (32 → 64)
- ReLU activations, batch normalization, max pooling
- 2 LSTM (128 → 64)
- Output: 7 neurons (emotions)

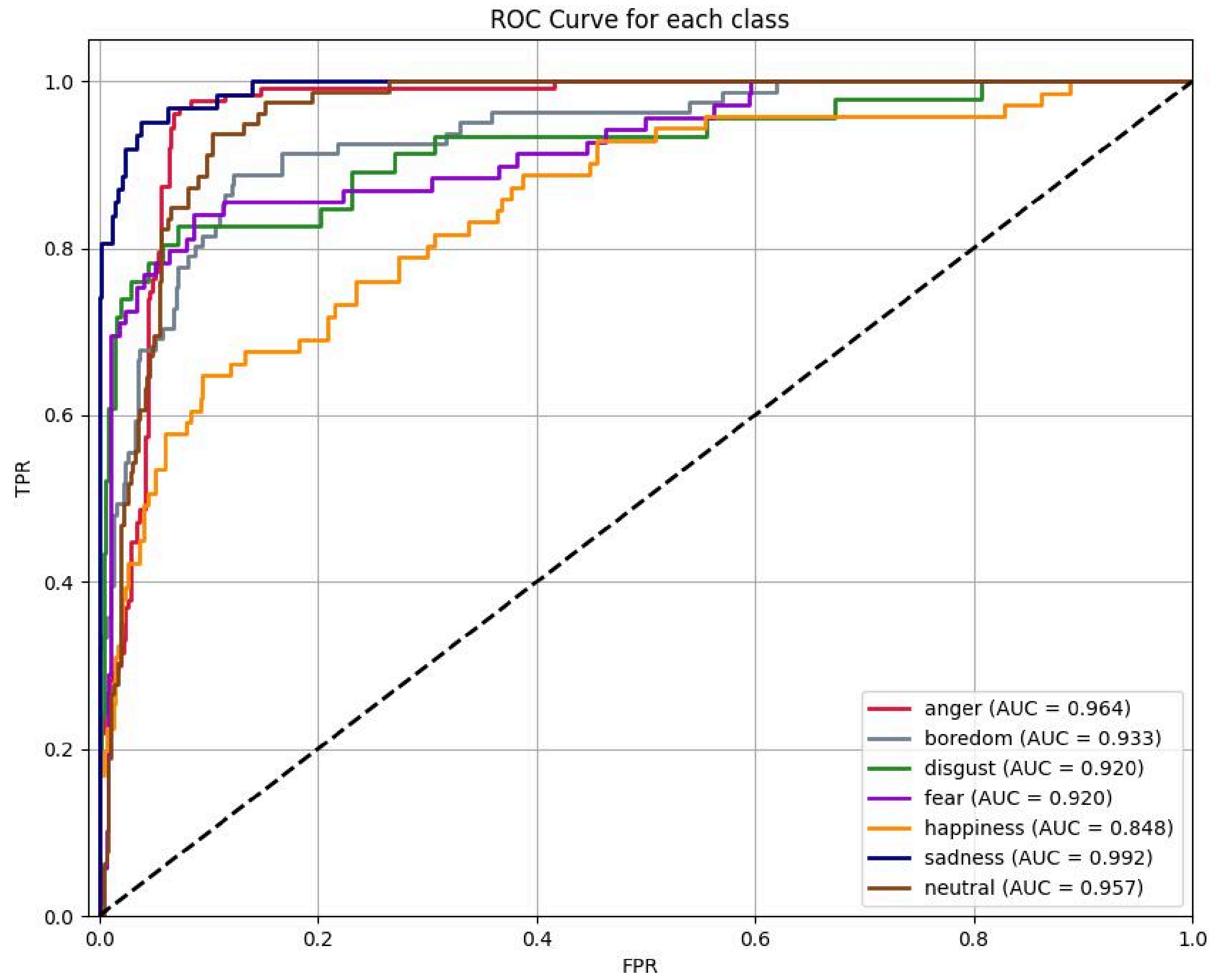
Results (LOSO CV):

- Avg. accuracy (per speaker): 0.65
- Overall accuracy: 0.76
- Top-3 accuracy: 0.95
- Weighted F1-score: 0.75

Findings:

- Best: Sadness (AUC = 0.992)
- Worst: Happiness (AUC = 0.848)





Evaluation setup:

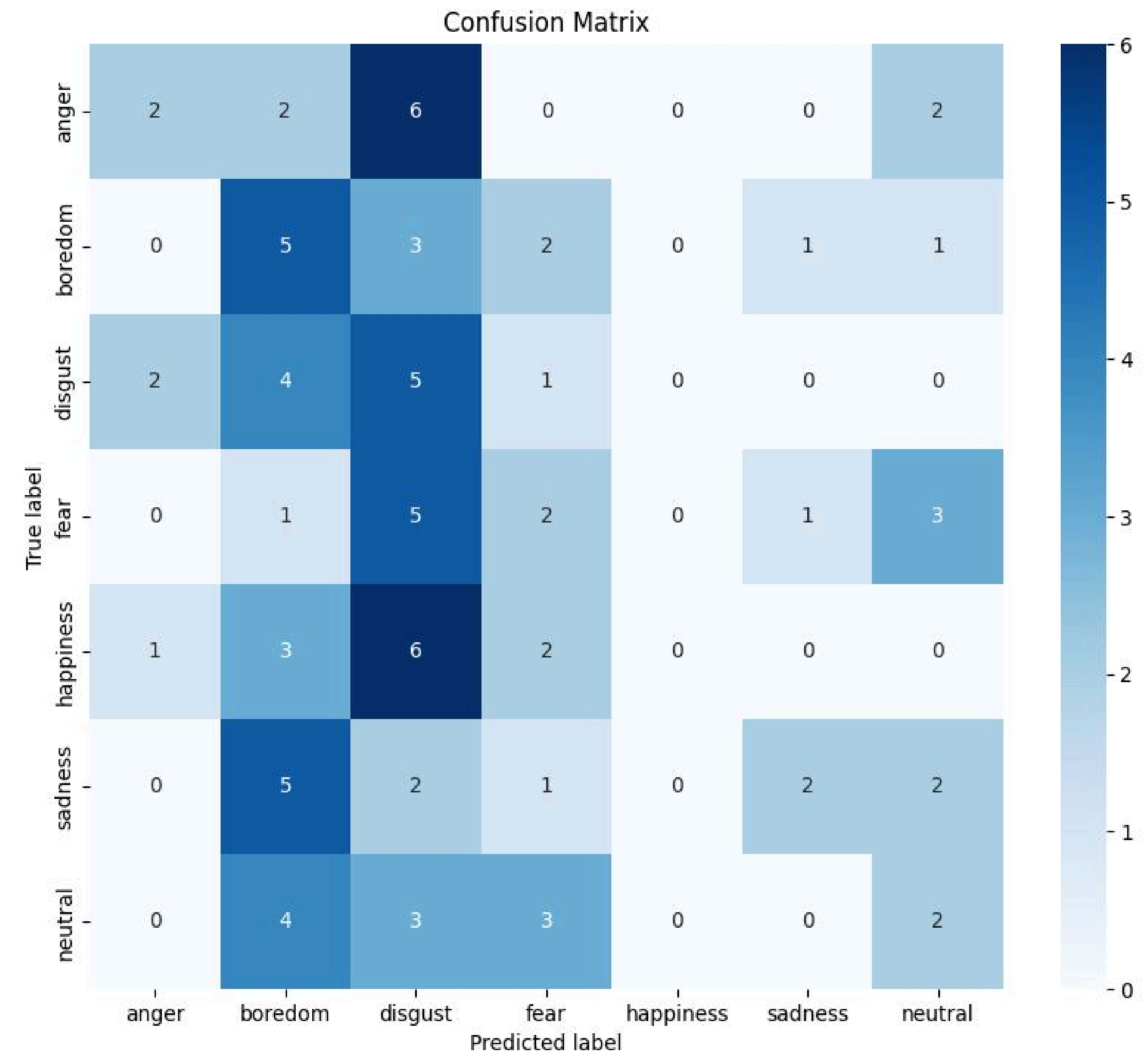
- New test set created (not in EMO-DB)
  - 84 audio samples (12 per emotion)
  - 3 Italian speakers + 1 German speaker
  - Same preprocessing pipeline (+ standardization)
- CNN2D + LSTM Trained on full EMO-DB + augmentation

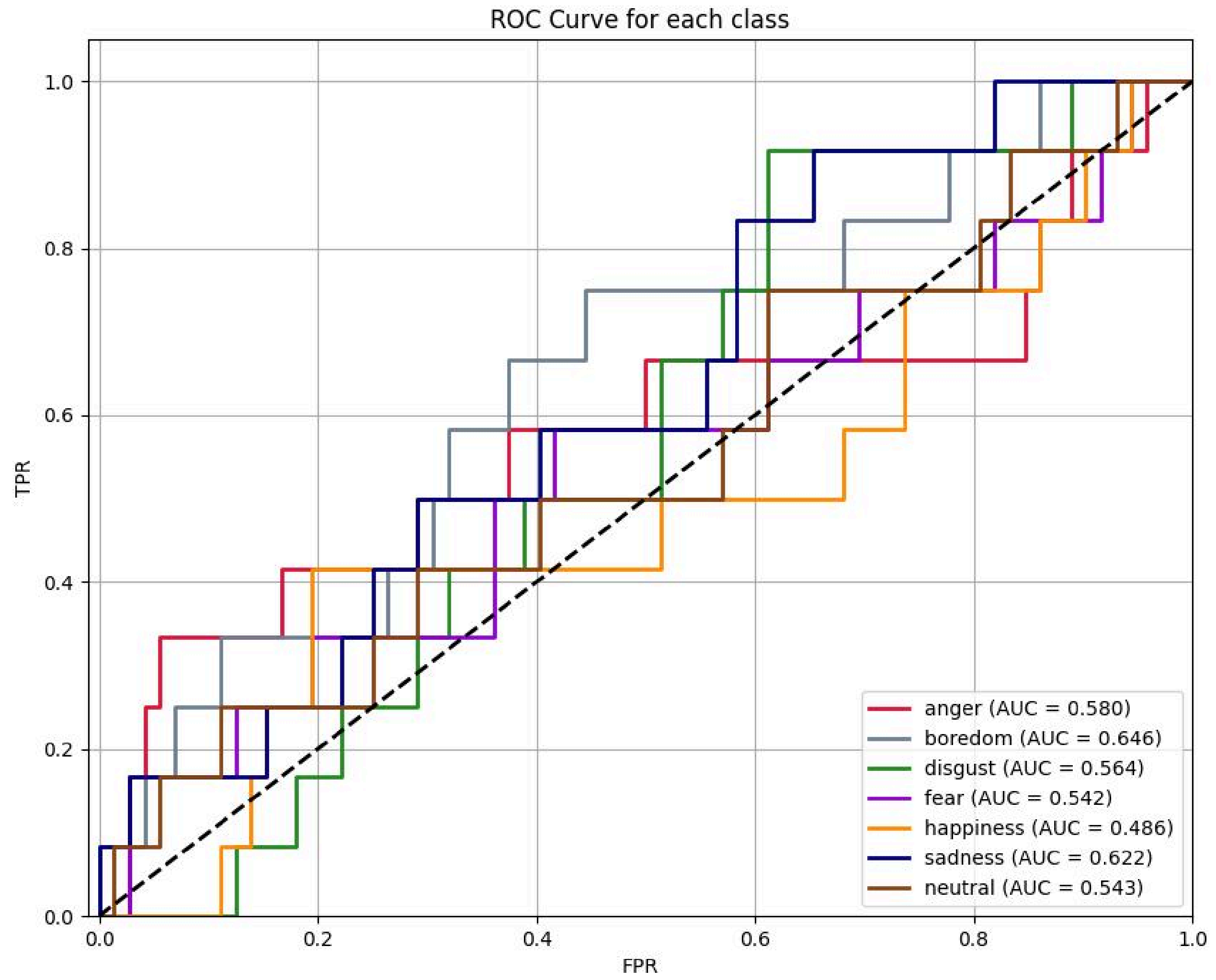
Results:

- Accuracy: 0.24
- Top-3 accuracy: 0.56
- Strong bias: boredom and disgust → over-predicted
- No samples classified as happiness

Key issues:

- Language gap (German vs Italian expression)
- Recording conditions (home setups vs lab studio)
- Non-professional acting (less consistent emotion portrayal)





- **Goal:** build deep learning models for emotion recognition from audio (EMO-DB)
- **Pipeline:** preprocessing → MFCC extraction → augmentation → LOSO CV → model training & testing
- **Models tested:** MLP (baseline), CNN1D, CNN2D, LSTM, CNN1D + LSTM, CNN2D + LSTM
- **Best architecture:** CNN2D + LSTM
  - Validation accuracy: 0.77
  - Top-3 accuracy: 0.94
  - AUC range: 0.880 to 0.987
- **Key issue:** poor generalization to real-world data
  - Test set (own recordings) → accuracy dropped to 0.24
  - Sensitive to language, recording environment, acting quality
- **Takeaway:** promising results within EMO-DB, but limited robustness in uncontrolled conditions

Subramanian, R. R., Sireesha, Y., Reddy, Y. S. P. K., Bindamrutha, T., & Harika, M. (2021). Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms. IEEE Xplore. https://ieeexplore.ieee.org/document/9675492

Doshi, K. (2021, February 12). Audio Deep Learning Made Simple – State-of-the-Art Techniques. https://ketanhdoshi.github.io/Audio-Intro/

GeeksforGeeks. (2025, July 23). Mel-frequency Cepstral Coefficients (MFCC) for Speech Recognition. https://www.geeksforgeeks.org/nlp/mel-frequency-cepstral-coefficients-mfcc-for-speech-recognition/

GeeksforGeeks. (2025, July 23). Preprocessing the audio dataset. https://www.geeksforgeeks.org/data-analysis/preprocessing-the-audio-dataset/

Ma, E. (2019, June 1). Data Augmentation for Audio. Medium. https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6

Thanks for Your Attention

by Chiara Genuardi and Giovanni Noè