

# speech emotion recognition

EmoDB Dataset | Custom CNNs

**Subject:** DPL302m

**Class:** AI1917

**Instructor:** Ho Le Minh Toan

**Group:** 2

# team members

Lai Le Dinh Duc | SE192137

Chau Le Quan | SE192913

Le Minh Khoi | SE192396

Nguyen Trung Kien| SE192209

Nguyen Minh Duc | SE192059

---

01

Introduction

---

04

Experiment

---

06

Conclusion

---

02

Approaches

---

05

Result

---

03

Method

table  
of  
content

# i. Introduction

## Speech Emotion Recognition

Speech Emotion Recognition (SER) is the process of analyzing and classifying human emotions from voice signals using audio processing and AI. It is crucial in areas like psychology, human-computer interaction, and customer service.



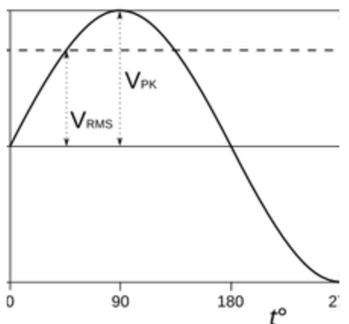
## Audio Data

In nature, audio exists as sound waves, they travel through air, liquids, or solids.

When recorded, these waves are stored as digital or electromagnetic signals. In digital form, sound is captured through sampling and quantization, with key parameters including: sampling rate, bit depth, and channels



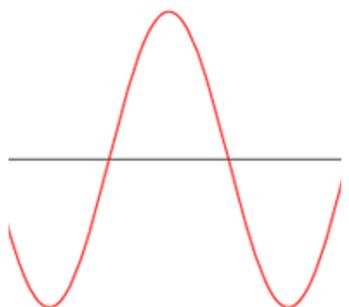
## Audio Features



*Time-Domain Features*

### RMS (Root Mean Square)

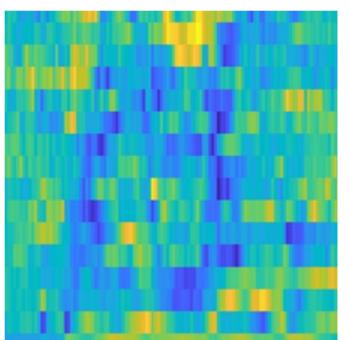
Represents the average energy level of the signal.



*Time-Domain Features*

### Zero-Crossing Rate (ZCR)

The rate at which the signal changes sign.



*Cepstral-Domain Features*

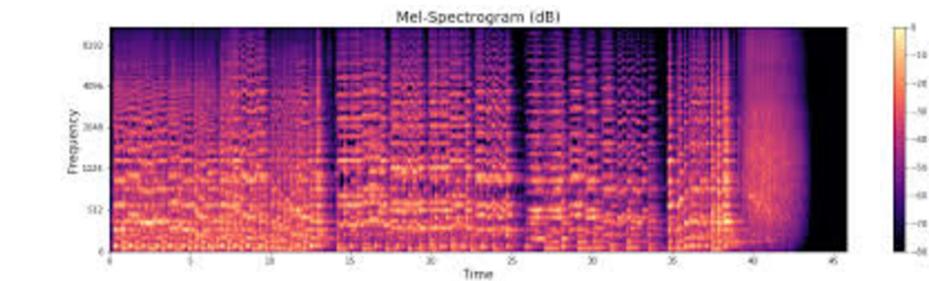
### MFCC (Mel-Frequency Cepstral Coefficients)

These coefficients mimic the way the human ear perceives sound.

## Time-Frequency Features

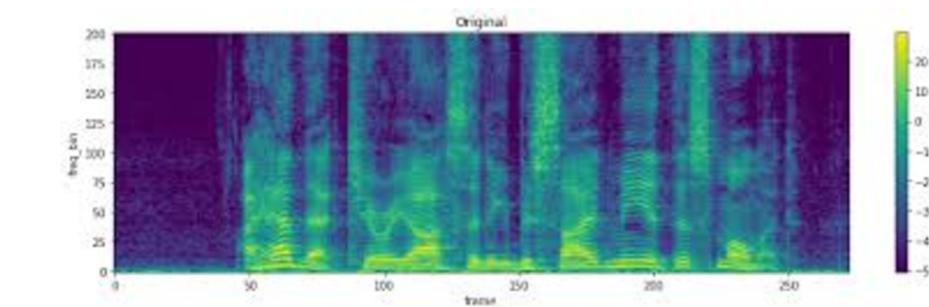
### Spectrogram

Indicate how the signal's spectral density evolves over time.



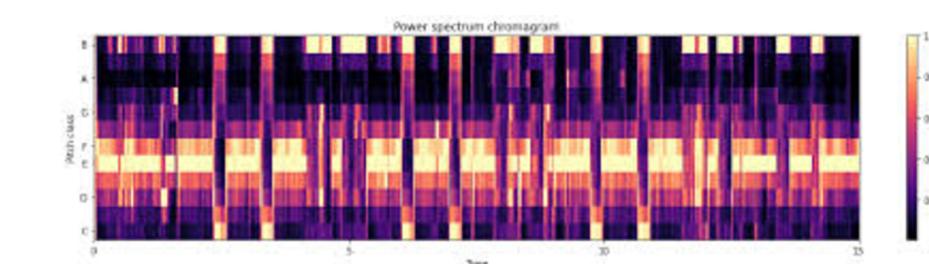
### Mel Spectrogram

A spectrogram with frequency scaled on the Mel scale.



### Chroma Features

Chroma features show energy across 12 pitch classes.



# ii. Approaches

## Data Processing Approaches

### Raw Audio Processing

In this approach, acoustic features are extracted and then fed into models. It's a well-established method, often enhanced with normalization and feature selection.

### Hybrid Processing

This approach feeds raw audio waveforms directly into deep learning models, enabling automatic feature learning. However, it often demands large datasets and significant computational power.

### Feature-Based Processing

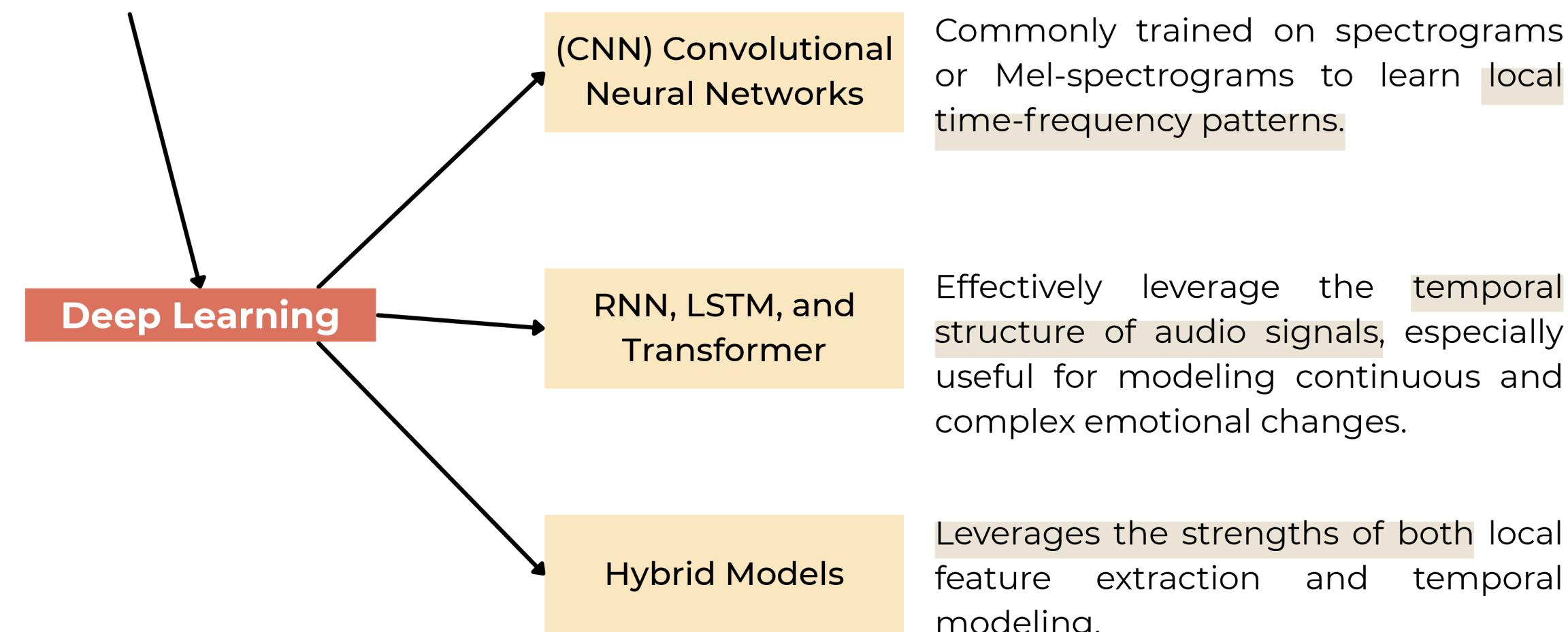
Modern systems often combine both approaches by using pre-trained models to extract rich features (embeddings) from raw audio, then fine-tuning them for emotion recognition.

## Traditional Machine Learning

Traditional ML rely on pre-extracted features and are well-suited for small to medium-sized datasets, but they require careful feature engineering and selection for optimal performance.

In our approach, we first focused on extracting 2D feature representations, which were used to train several simple CNN-based model variants.

## Machine Learning Approaches



We then extended this by generating 3D feature representations, allowing us to design and experiment with custom 3D Convolutional Neural Network (3D CNN) architectures for enhanced performance.

# iii. Methodology

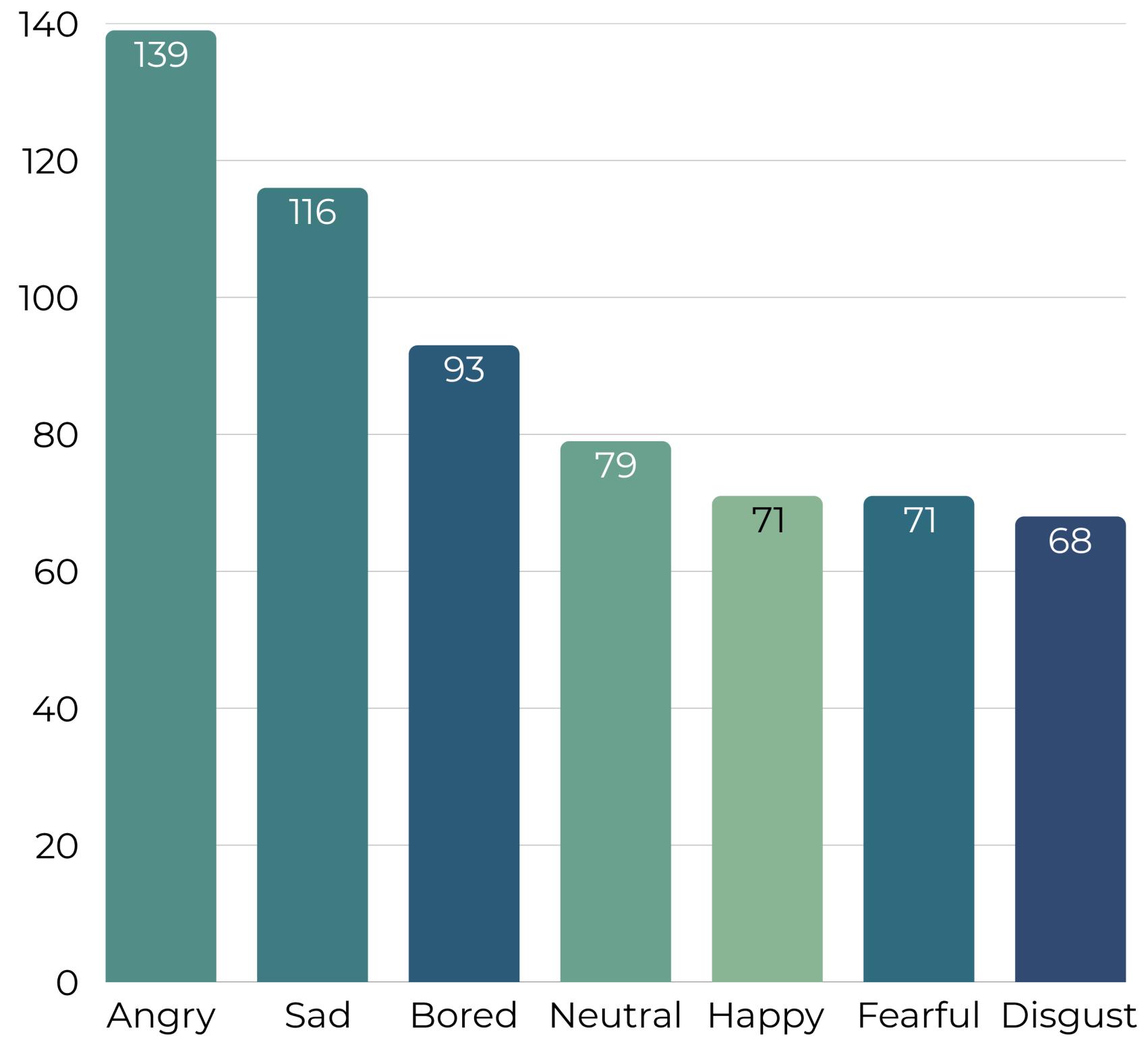
## Dataset

Name: EmoDB (Berlin Database of Emotional Speech)

Number of Samples: 535 utterances

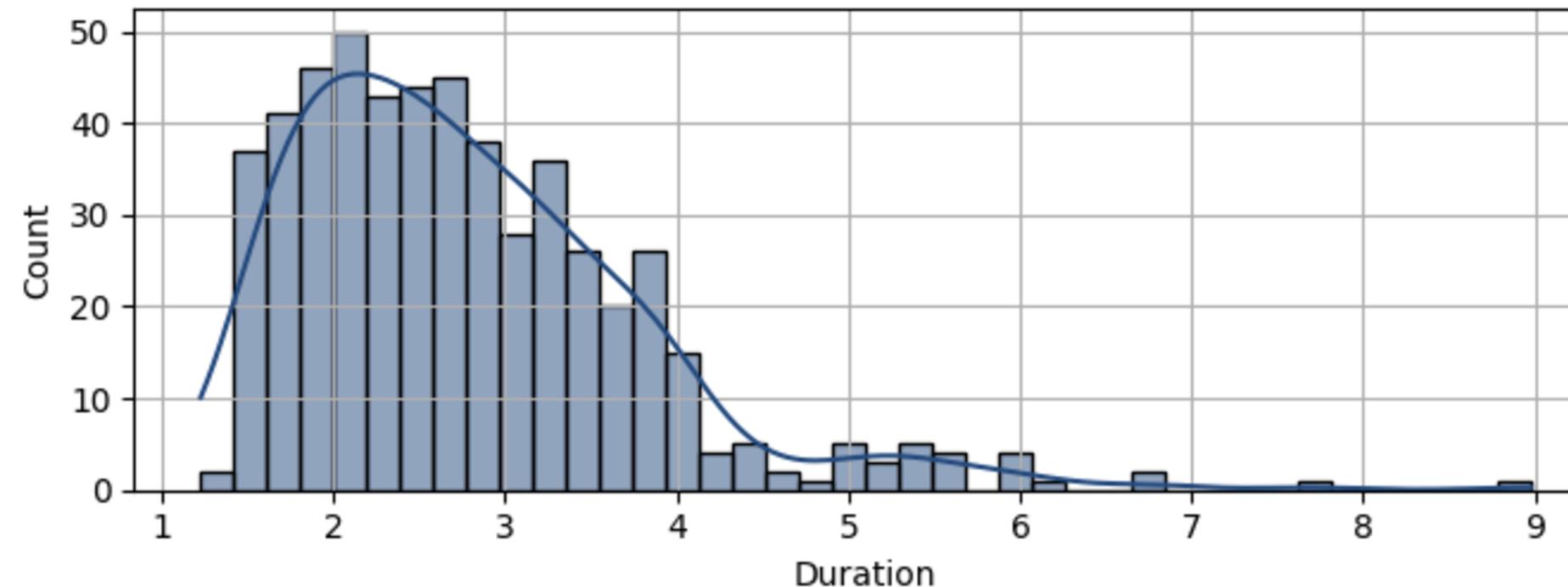
Language: German

Speakers: 10 professional actors (5 male, 5 female)



## Dataset

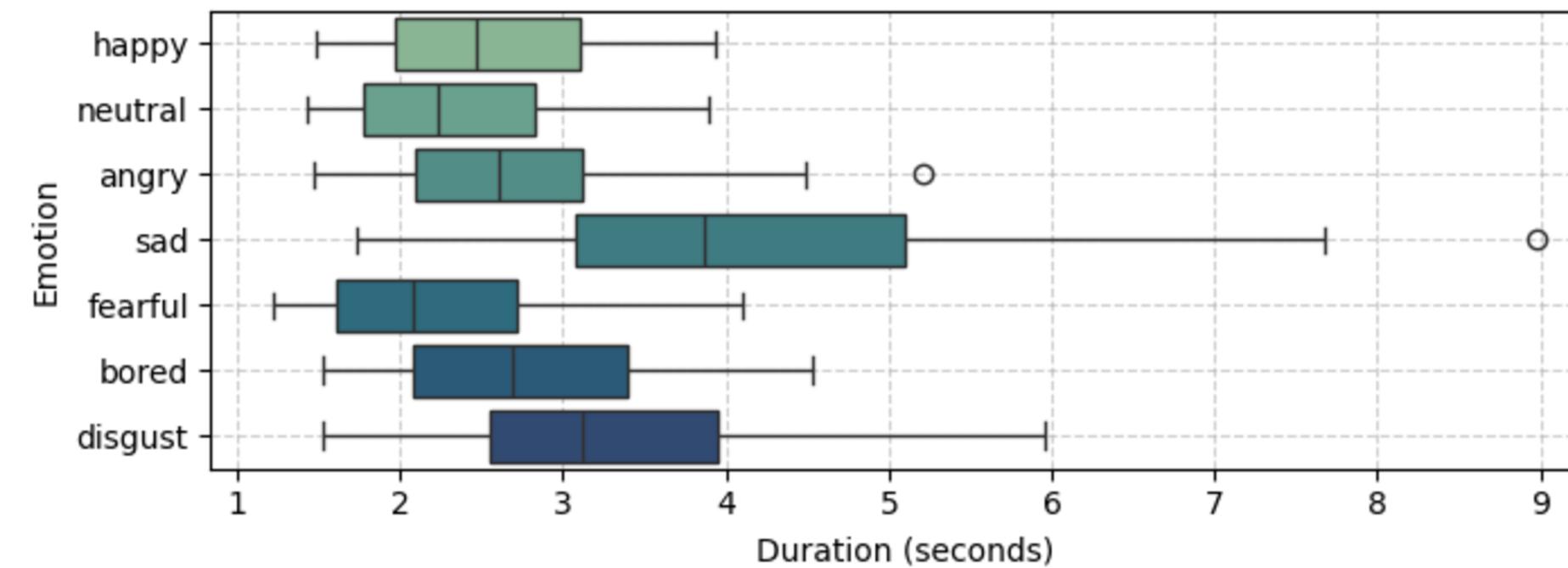
Most recordings are between 1.5 to 3.5s, peaking around 2s. The distribution is right-skewed, with a few longer recordings reaching up to 9 seconds.



Sad utterances are generally longer on average, with several outliers

Neutral, happy, and angry samples tend to be shorter and more consistent, mostly ranging between 1.5s and 3.5s.

Disgust exhibits the greatest variation among the shorter-duration emotions, including a few noticeably longer recordings.



## Feature Extraction

### Resampling & Mono Conversion

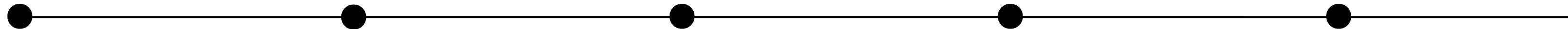
All .wav audio files were loaded and resampled to a 44.1 kHz sampling rate.

### MFCC Feature Extraction

Each segment was converted into a (128, 128) MFCC matrix.

### Final Dataset Dimensions

The final feature tensor had shape (1234, 128, 128, 3), representing 1234 segments from 535 clips



### Segmentation with Overlapping Windows

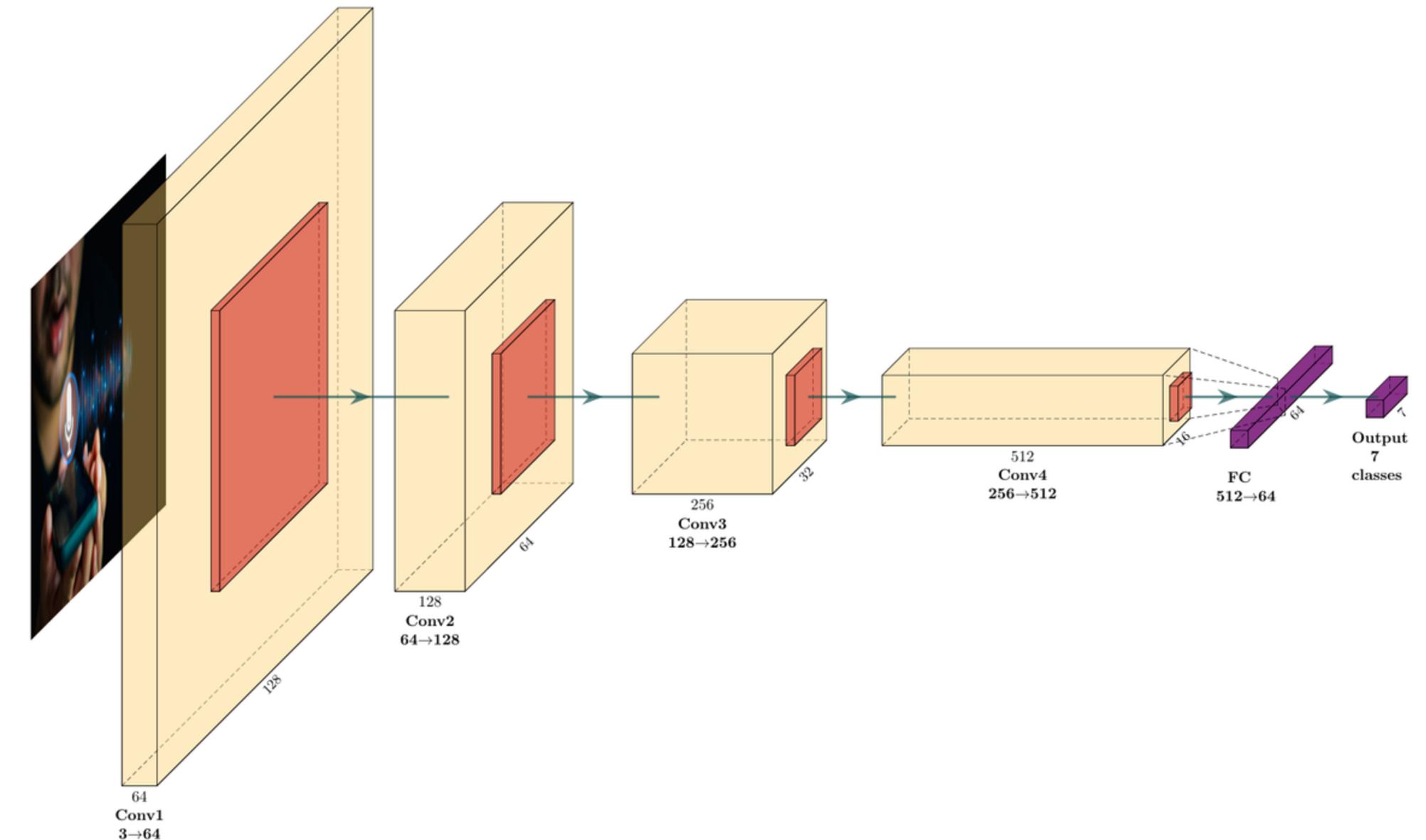
Each audio clip was divided into smaller overlapping segments using a sliding window approach.

### Delta and Delta-Delta Features

To capture temporal dynamics, first- and second-order MFCC deltas were computed and stacked, forming a (128, 128, 3) input per segment.

## Model Architecture

A custom Convolutional Neural Network was implemented using PyTorch to classify emotional speech based on MFCC feature maps. The model was designed to process 3-channel spectrogram like inputs of shape (128, 128, 3) for each windowed segment.



# iv. Experiment

01

## Hardware Resources

Model was trained on Kaggle using an NVIDIA Tesla P100 GPU with 16 GB VRAM, an Intel Xeon CPU at 2.30 GHz (2 cores), and 13 GB of RAM.

02

## Training Configuration

**Batch Size:** 4  
**Epochs:** 200  
**Val Ratio:** 0.15  
**Test Ratio:** 0.15  
**Epochs:** 200  
**Loss Function:**  
CrossEntropyLoss  
**Optimizer:** Adam (learning rate=1e-4)  
**Scheduler:**  
ReduceLROnPlateau (patience=5, factor=0.5)

03

## Logging Organization

During training phase, loss, hyperparameters, accuracy, F1 score, confusion matrix, and model artifacts were tracked with Weights & Biases.

The best model was selected based on the highest validation F1 score and saved for evaluation.

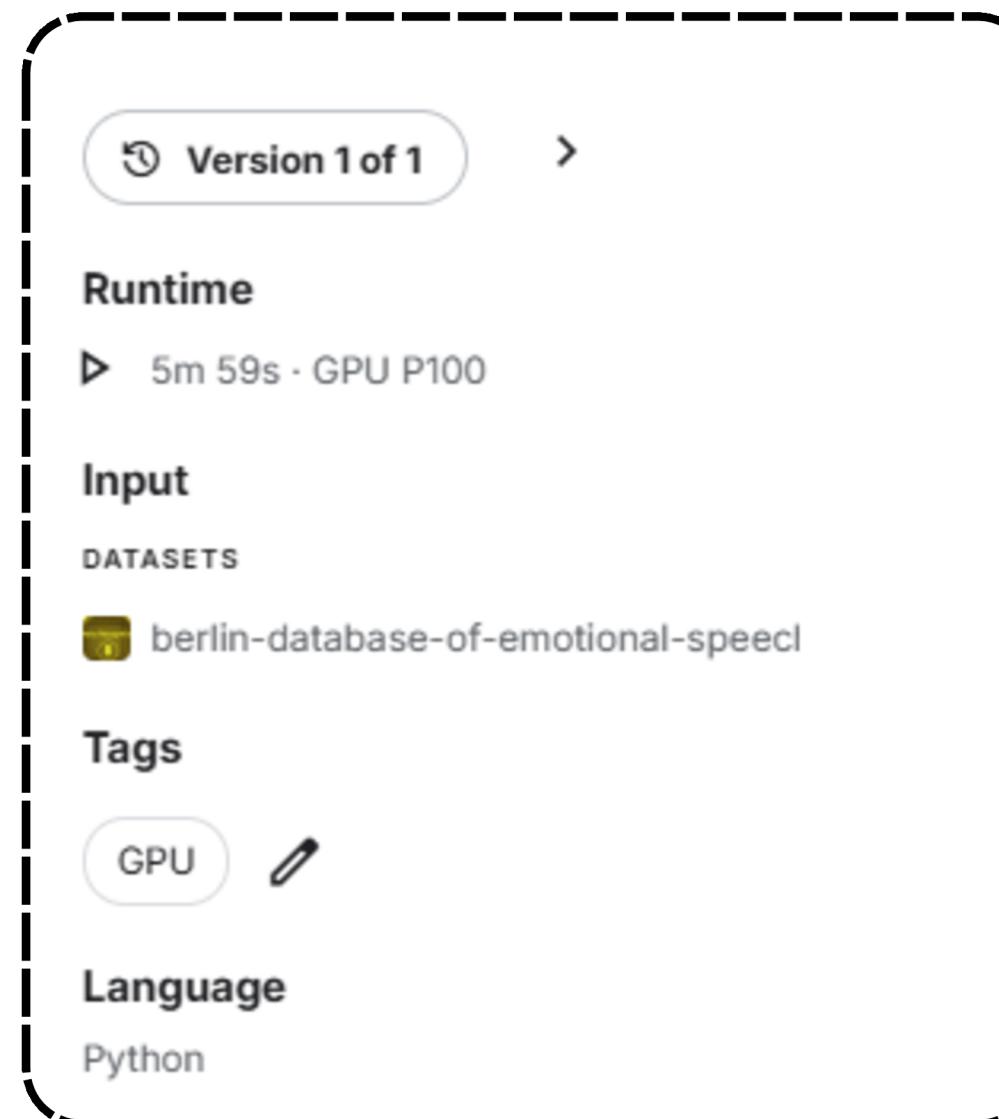
04

## Evaluation Metrics

The EmoDB dataset has imbalanced classes, making accuracy biased toward the dominant class. Instead, F1-macro, which averages F1 scores across all classes, provides a more balanced evaluation.

$$\text{Macro-F1} = \frac{1}{L} \sum_{i=1}^L F1_i$$

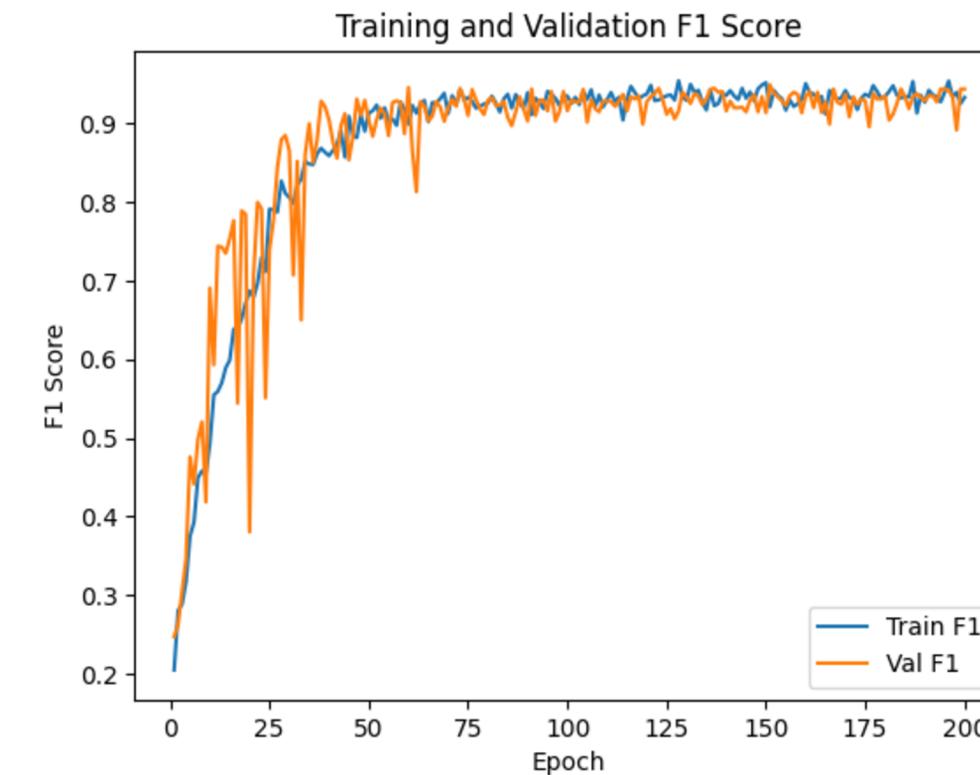
# v. Results



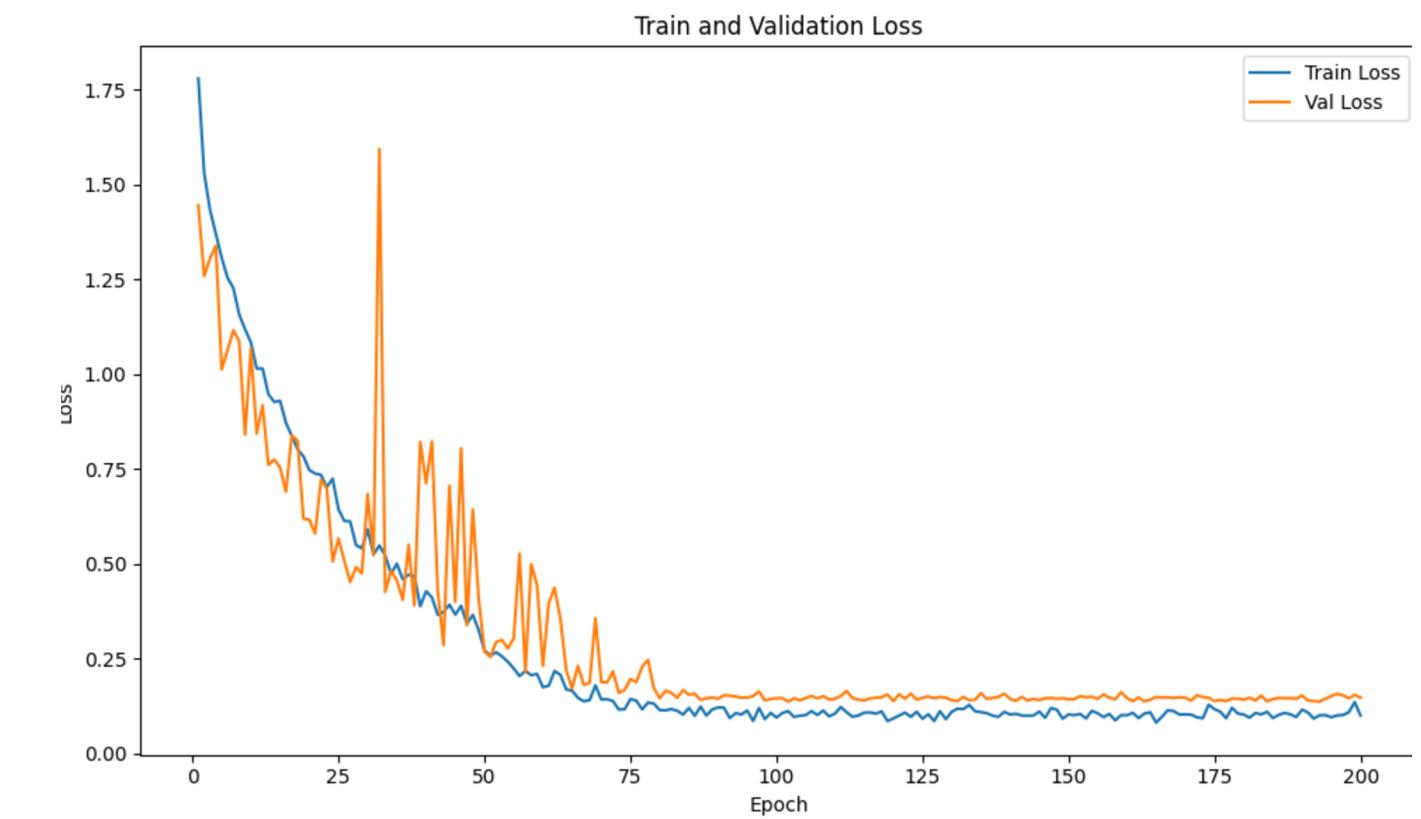
## Overview

The training required approximately 6 minutes ( 358.6 seconds) to complete 200 epochs. The best F1 score achieved on the validation set was 96%.

## Metrics Visualization



The training F1 score increases steadily to about 0.94. The validation F1 score fluctuates early but quickly stabilizes at comparable levels, indicating good generalization and minimal overfitting.



The training loss (blue curve) drops quickly in the initial epochs and converges below 0.1, indicating effective objective minimization.

The validation loss (orange curve) shows early fluctuations with several spikes but gradually stabilizes around 0.15, suggesting improved generalization over time.

## Classification Report

<b>Class</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
fear	0.94	0.88	0.91	17
disgust	0.95	1.00	0.98	21
happy	0.91	0.95	0.93	21
angry	0.97	0.95	0.96	40
sad	1.00	1.00	1.00	38
neutral	1.00	0.95	0.98	21
boredom	0.97	1.00	0.98	28
<b>Accuracy</b>			0.97	186
<b>Macro avg</b>	0.96	0.96	0.96	186
<b>Weighted avg</b>	0.97	0.97	0.97	186

- **Per-class F1-scores:** Most classes achieve  $F1 > 0.90$ , showing good separation between emotion categories.
- **Macro average F1-score:** 0.96 → demonstrates the model's balanced performance across all classes.
- **Perfect scores:** Classes 4, 5, and 6 achieve near-perfect or perfect precision/recall/F1
- **Overall:** Macro F1 of 0.96 → very high on EmoDB dataset, proving that the training pipeline is effective.

# vi. Conclusion

In this study, we developed a custom CNN architecture for speech emotion recognition using the EmoDB dataset. Our CNN achieved a macro-F1 score of 0.97 on the validation set and 0.96 on the test set, highlighting its effectiveness in capturing emotional cues from speech.

Future work will focus on adapting SER to Vietnamese and experimenting with hybrid CNN–attention architectures for enhanced real-time performance.

# Thank You.

LOOKING FORWARD TO YOUR FEEDBACK AND QUESTIONS.