EMOTION CLASSIFICATION WITH SPEECH AUDIO

Student : Kerem Erciyes

University of Milano Bicocca | 2024

Supervised by: Asst. Prof. Flavio Piccoli
Mirko Paolo Barbato

ABSTRACT

In this project the task is to focuse on developing an advanced audio classification system utilizing deep learning techniques. The main task is to classify audio files into predefined categories. The system is built using a combination of Convolutional Neural Networks (CNNs), Time-Delay Neural Networks (TDNNs), Long Short-Term Memory (LSTM) networks, and an Attentive Statistical Pooling (ASP) mechanism to effectively capture both spatial and temporal features of audio data.

DATASET EXPLANATION

The name of the general dataset is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The audio durations at most 4 seconds.

OVERVIEW

Data Preparation

Audio features are extracted and standardized using Mel-Frequency Cepstral Coefficients (MFCCs). The dataset is split into training, validation, and test sets.

Models

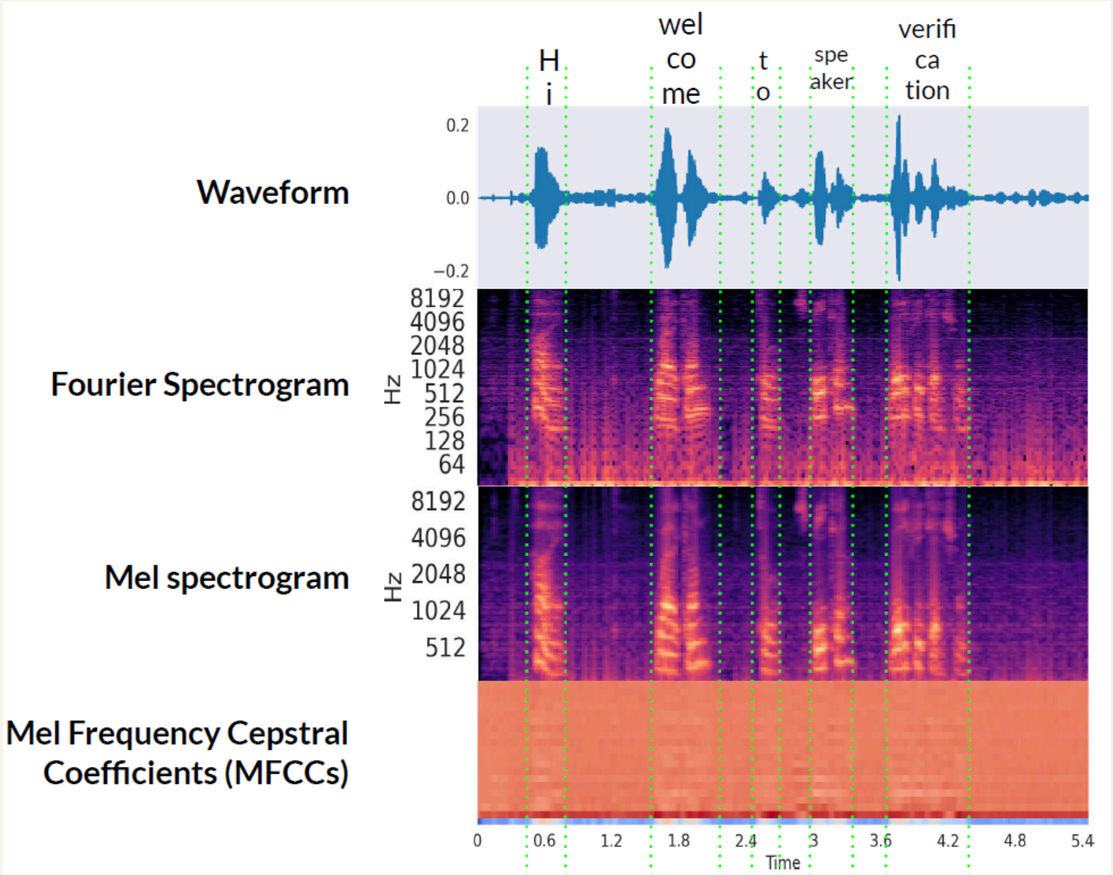
The model architecture combines CNNs, TDNNs, LSTM layers. This architecture is designed to leverage the strengths of each component in processing audio data.

PIPELINE OF THE PROJECT

- Datapreparation
- Methodology
- Models
- Result
- Conclusion
- User Interface

DATA PREPARATION

1. **Waveform (amplitude x time)**
2. **Fourier Spectrogram (frequency x time)**
3. **Mel spectrogram (frequency x time)**



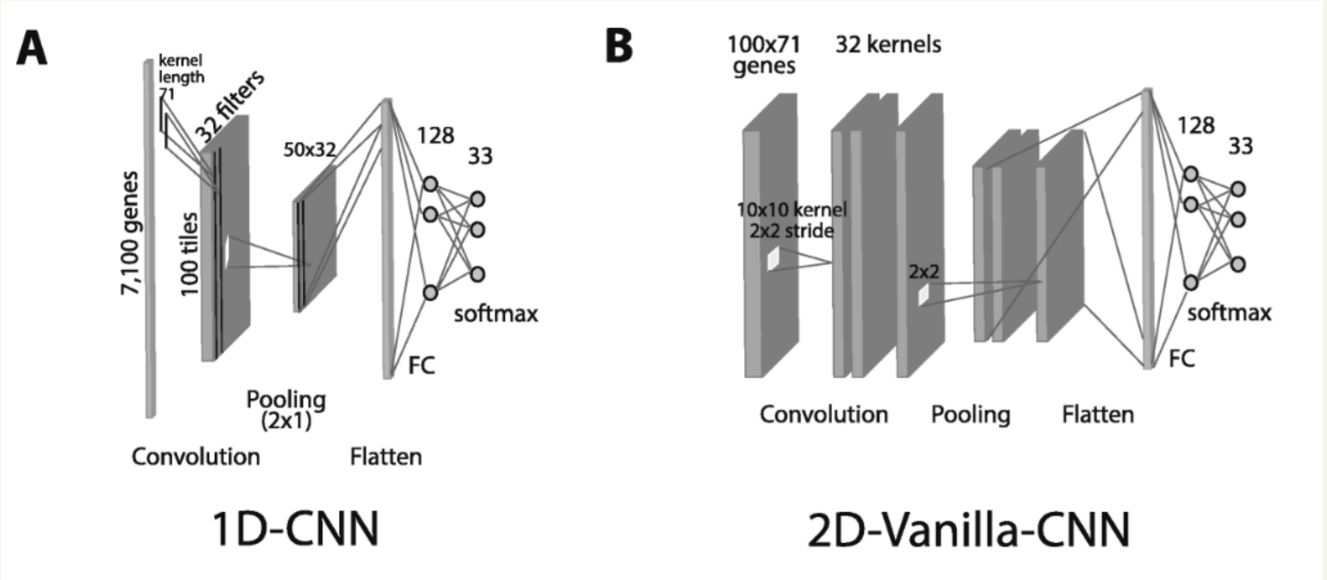
METHODOLOGY

1. Convolutional Neural Networks (CNNs): Applied for feature extraction
2. Time-Delay Neural Networks (TDNNs): Applied to model temporal context by capturing dependencies over varying time delays.
3. Long Short-Term Memory (LSTM) Networks: Employed to capture long-term dependencies.
4. Attentive Statistical Pooling (ASP): Used Temporal Average Pooling (TAP) method to aggregate frame-level embeddings into an utterance-level embedding.

MODEL 1: CNN

Objective

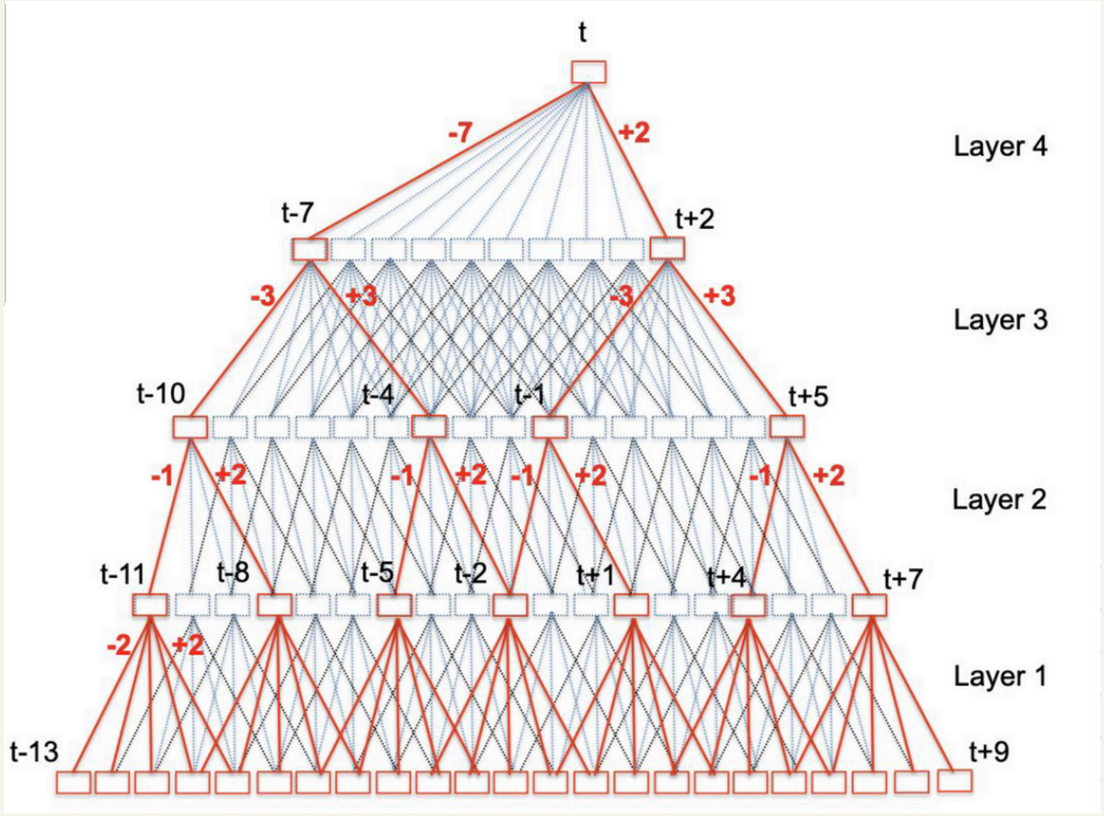
Used for initial feature extraction from audio data, capturing local patterns in the spectral representation (spatial dependencies).



MODEL 2: TDNN

Objective

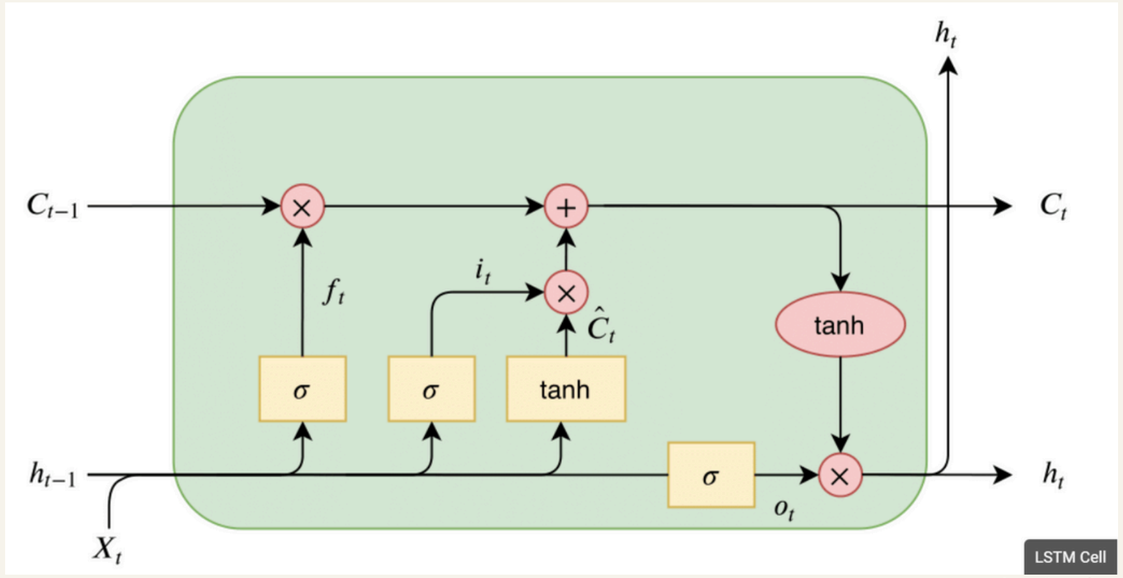
Capture temporal dependencies and context within the extracted features (short-term dependencies), which are crucial for understanding sequential audio features.



MODEL 3: LSTM

Objective

Capture long-term dependencies and sequential information in the features. Enhancing the model's ability to remember previous information of the The audio sequences over extended periods.

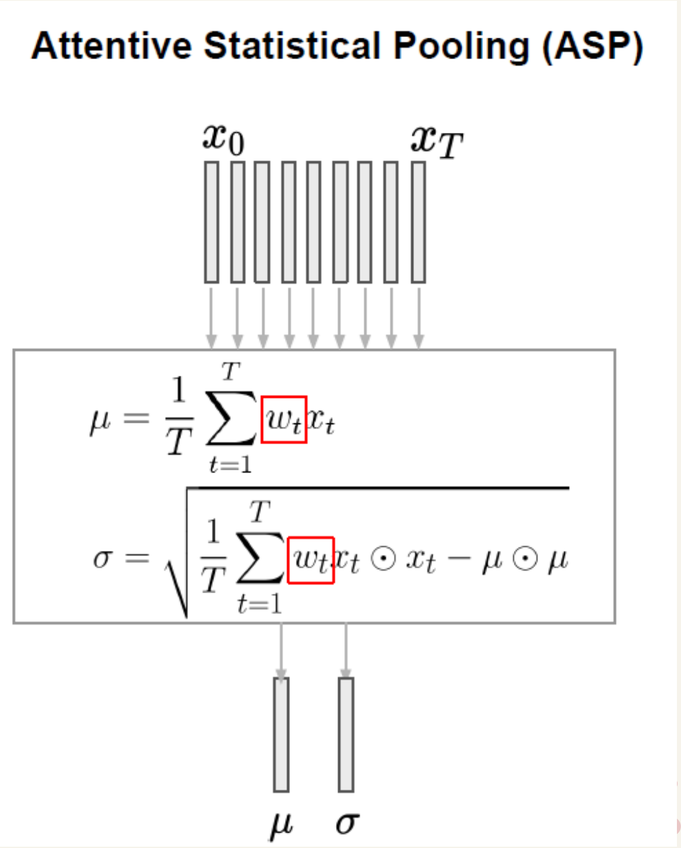


MODEL 4: TEMPORAL AGGREGATION

Attentive Statistical Pooling (ASP)

Objective

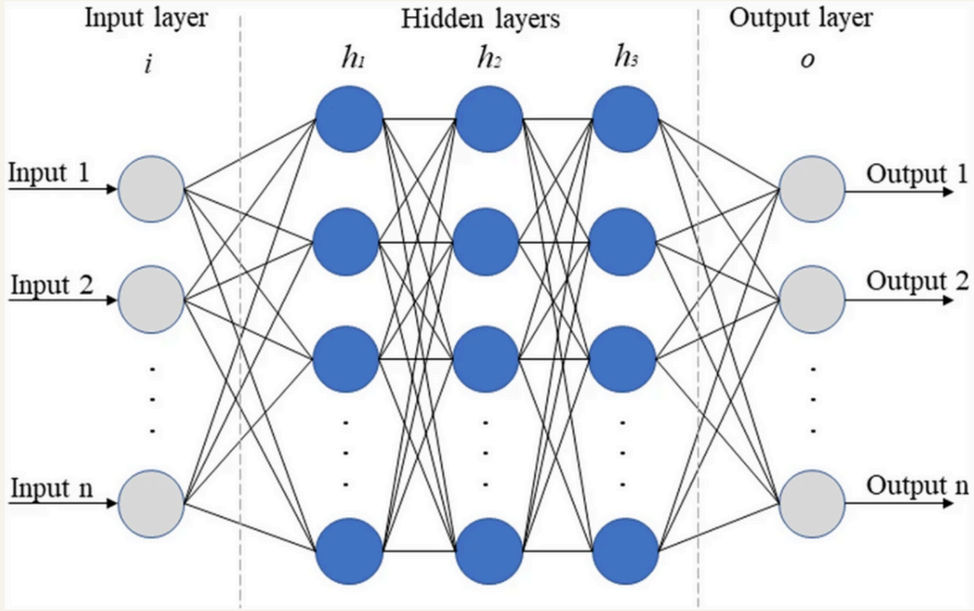
Incorporated to focus on the most relevant parts of the sequence, combining learned attention weights with statistical measures (mean and variance) for robust feature aggregation.



MODEL 5: MULTI LAYER PERCEPTRON

Objective

In order make the classification, whole model's output connected to fully connected neural network with 2 fully connected layer and 8 outputs (number of classes) structure. And neurons vote for final class decision.



COMBINED MODEL

Implementation steps:

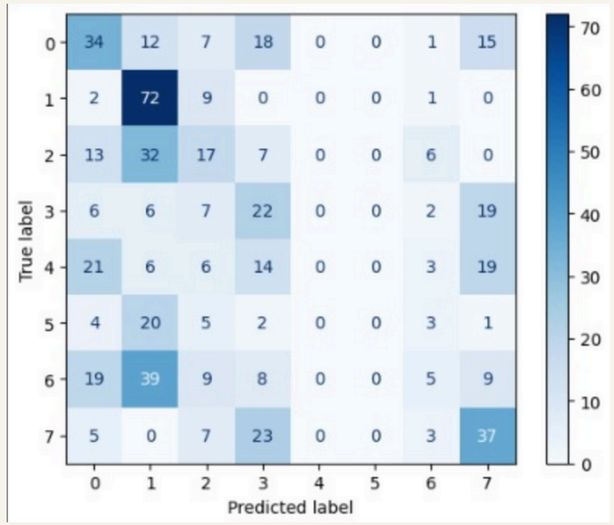
1. Each convolutional layer captures local patterns in the MFCCs, such as edges and textures, through convolution operations.
2. TDNNs apply dilated convolutions with different context sizes and dilation rates, allowing them to capture features over varying time scales. Each TDNN layer focuses on different temporal patterns and dependencies in the audio features.
3. The LSTM layer processes the sequence of features, capturing dependencies over long time periods and maintaining information across time steps. Generates a sequence of features that incorporate long-term temporal dependencies in the audio signal.
4. Attention weights are computed using the Attention mechanism, which assigns different importance to different time steps in the LSTM output. This process captures the central tendency and variability of the most important features, effectively summarizing the sequence at the end for fully connected layers.

HYPERPARAMETER TUNING

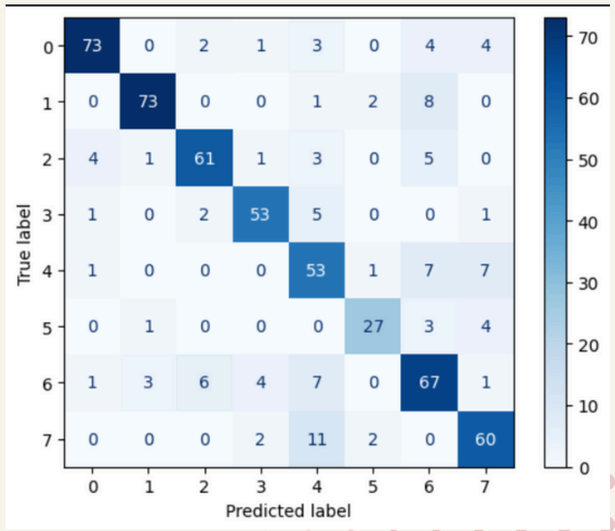
Epoch 1/10, Loss: 2.0642919540405273
Epoch 2/10, Loss: 2.046820640563965
Epoch 3/10, Loss: 2.0282816886901855
Epoch 4/10, Loss: 2.1179161071777344
Epoch 5/10, Loss: 1.9386979341506958
Epoch 6/10, Loss: 1.9509611129760742
Epoch 7/10, Loss: 1.907228946685791
Epoch 8/10, Loss: 1.8198908567428589
Epoch 9/10, Loss: 1.6874719858169556
Epoch 10/10, Loss: 1.9100912809371948

-Learning Rate
-Batch Normalization
-Weight Initialization
-Epoch number increased

Test Accuracy: 32.47%



-Added conv layers
-Kernel sizes changed
-Early stopping added



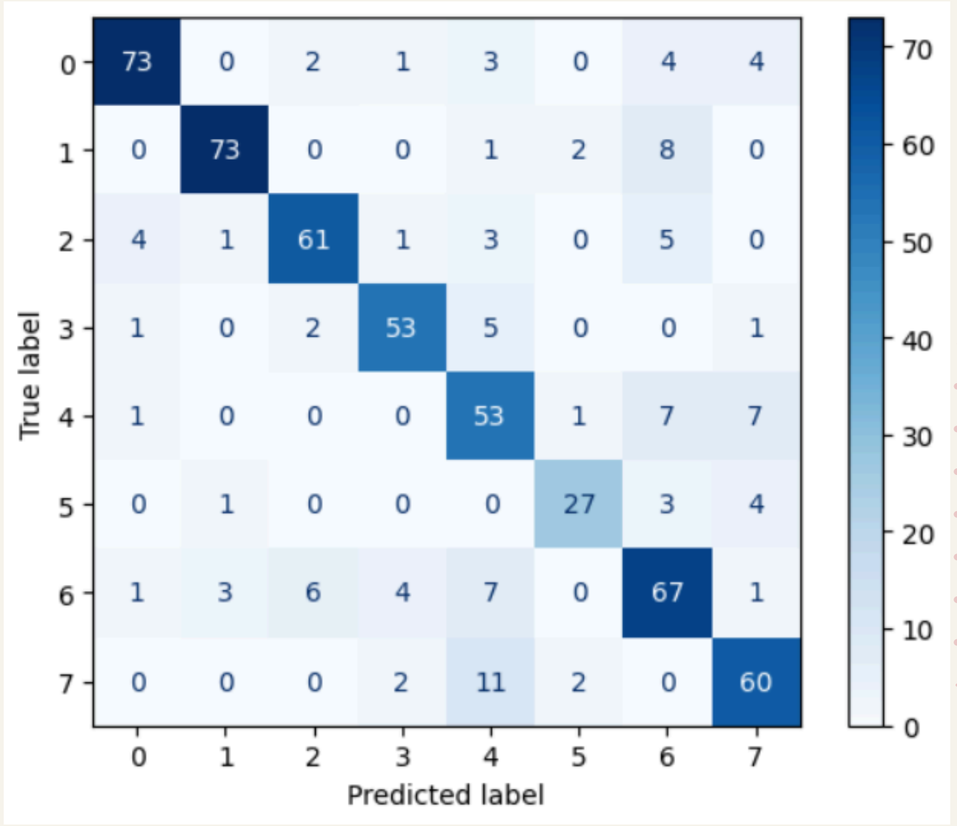
RESULT

Epoch 57/80, Loss: 0.00643

Val Loss: I.4726

Early stopping triggered.

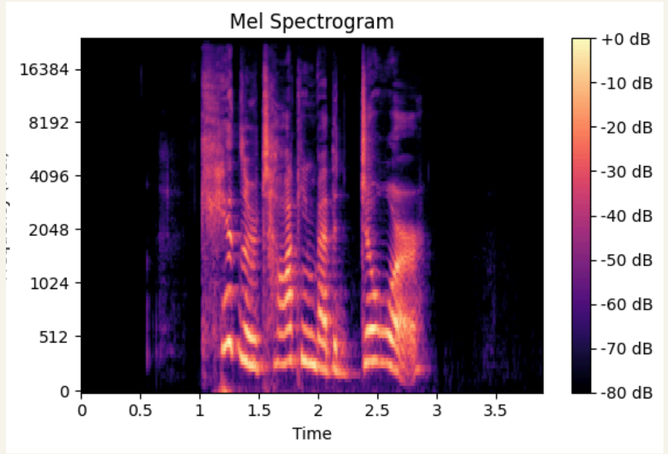
Test Accuracy: 8I.08%

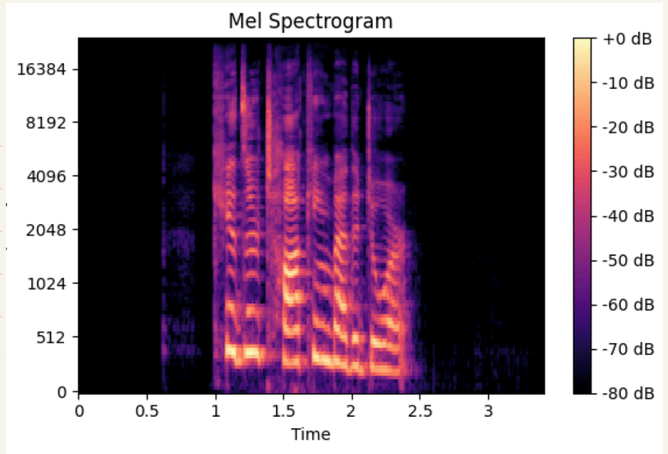


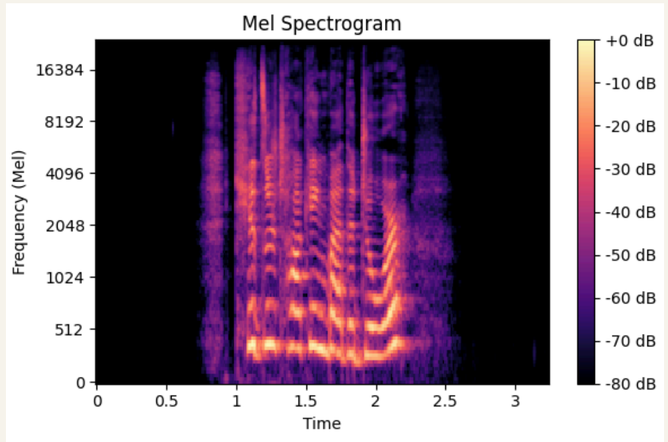
11

CONCLUSION

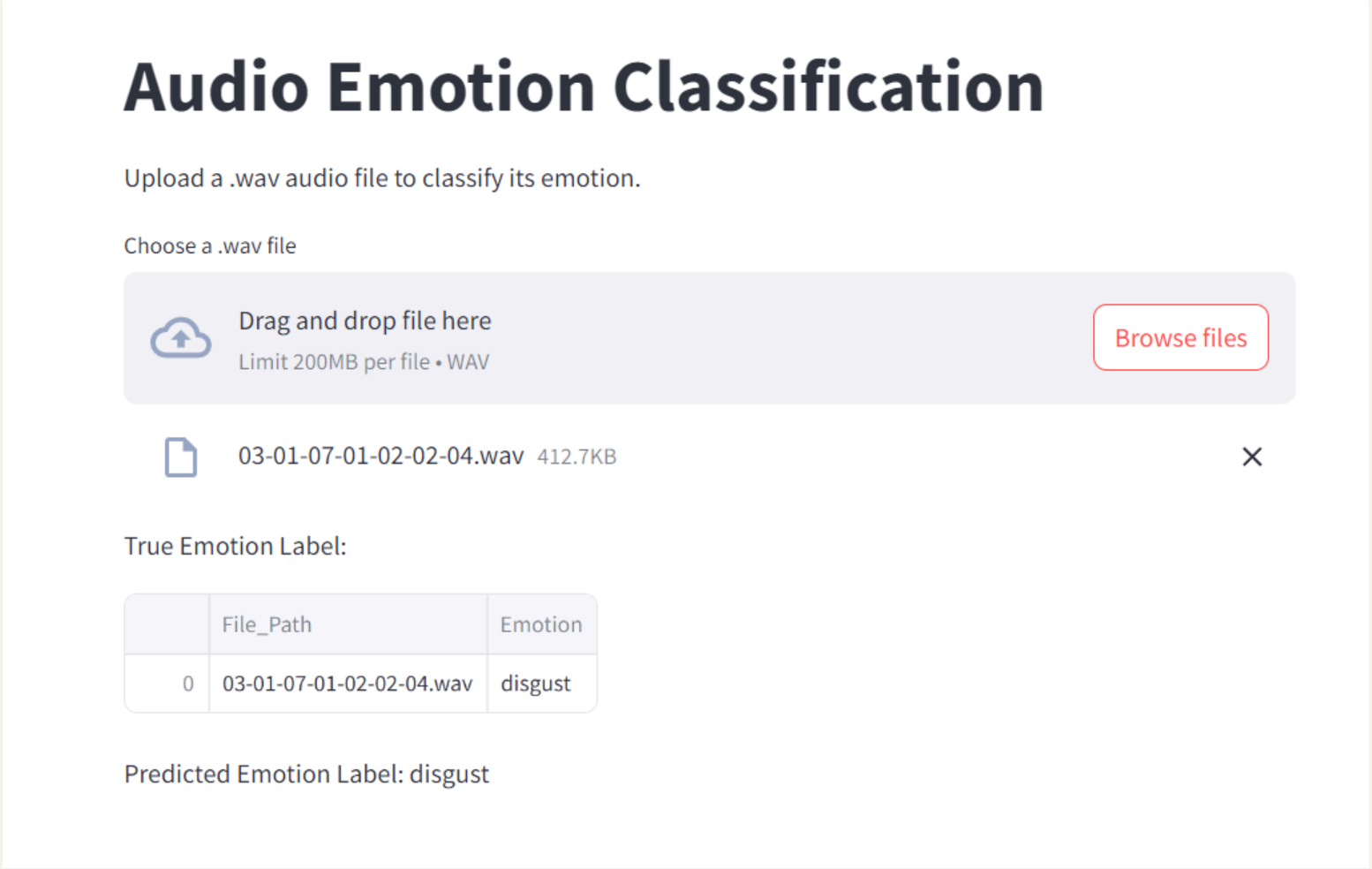
Overall, the result is good, but there is an issue with label 5. The model predicts all actual label 5 instances correctly, meaning the true positive rate is very high. However, due to the fixed random seed, the data shuffling is always done in the same way, causing fewer label 5 instances to appear in the validation data. Therefore, training and evaluating the model with a balanced dataset, along with balanced validation and test sets, would yield much better values. The model's biggest error rate is predicting label 4 for data that is actually label 7. The first figure is label 7, the second figure is label 4, and the third figure is label 6.







USER INTERFACE



REFERENCE

Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

Piccoli, F., & Barbato, M. P. (2024). Advanced Computational Techniques for Big Imaging and Signal Data: Audio [PowerPoint slides]. University of Milano-Bicocca.

Okabe, K., Koshinaka, T. and Shinoda, K., 2018
arXiv preprint arXiv:1803.10963

University of Milano Bicocca | 2024

THANK YOU

Presented By : Kerem Erciyes