Agenda

Project Overview & Background	• General Idea • Background & Motivation • Related Work
Datasets Samples	• Datasets used • Sample Profiles • Sample of Augmentations
Approach	• Code Approach • Model Architecture
Training History Results & Confusion Matrix	• Training and Validation Accuracy Trends • Distribution of Model Prediction
Moving Forward	• Ideas for improvement and advancement

Overview

Designing a model to analyze the acoustic features of speech and extract emotional information from them. By using machine learning algorithms, the model can accurately identify different emotional states expressed in speech, such as anger, sadness, happiness, or excitement.

Improving social skills and emotional awareness in individuals with autism. Monitoring and treating mental health conditions. Enhancing customer service experiences. Improving education outcomes. Informing various fields such as marketing, politics, and human resources.

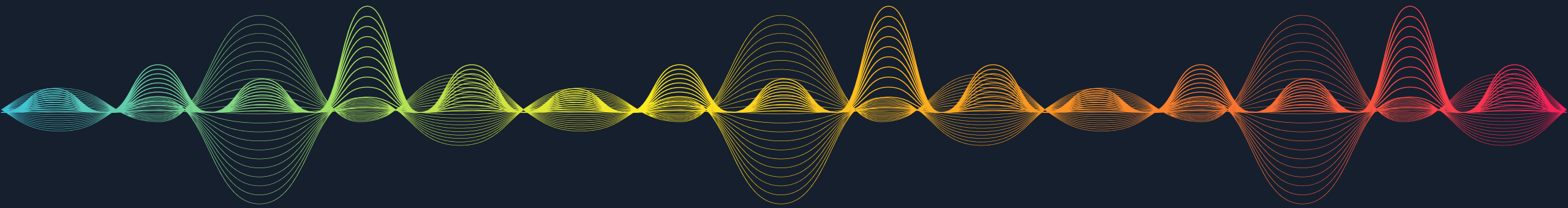
Related Work

Conventional models – Statistical & Machine Learning

- HMMs, GMMs, SVMs, ANNs

Modern models – Deep Learning

- CNNs, RNNs, LSTMs, Autoencoders
- Single vs. Multimodal
- CRNNs, DANNs



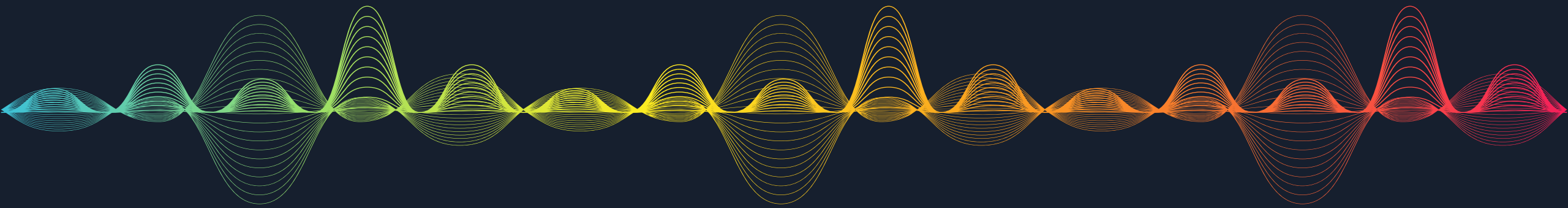
Related Work

Key challenges and limitations

- Lack of quality labeled speech emotion datasets
- Achieving high accuracy across all emotions
- Generalization of SER models

Potential solutions / mitigations

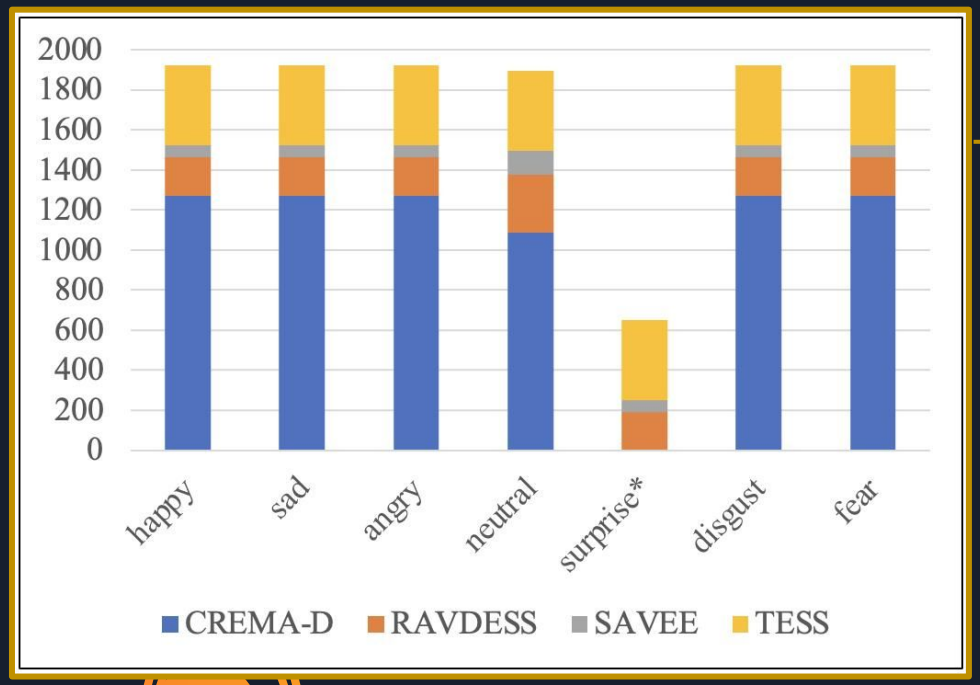
- Data augmentation, Transfer learning
- Semi-supervised models
- CRNNs, DANNs
- Multimodal



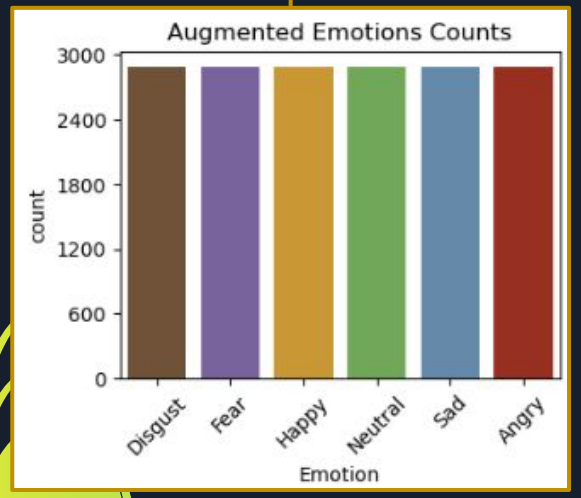
Datasets for Training

Crema – Tess – Ravdess – Savee

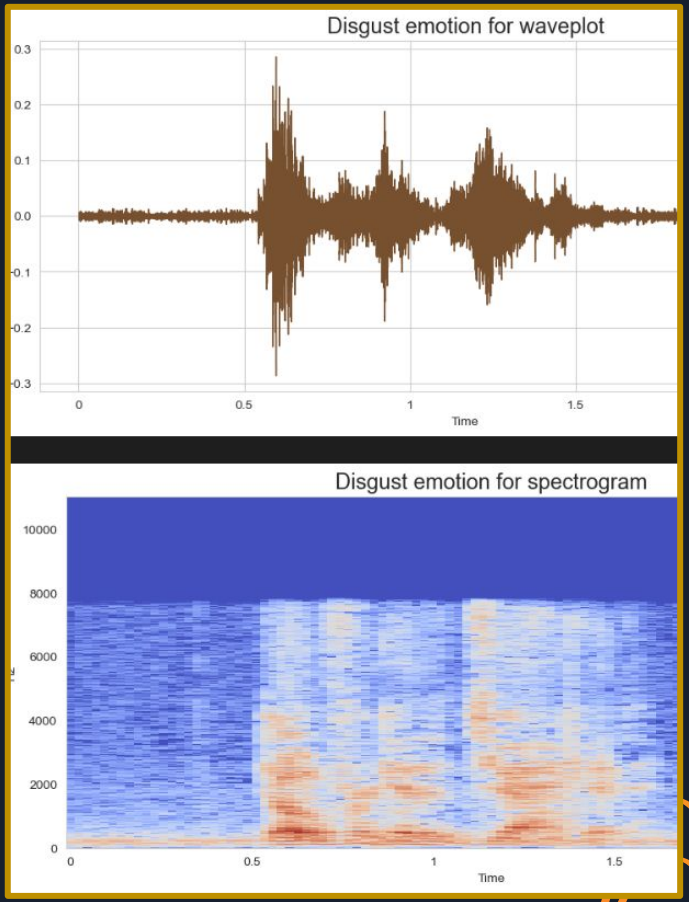
Well known emotion audio snippet datasets from Kaggle.com combined

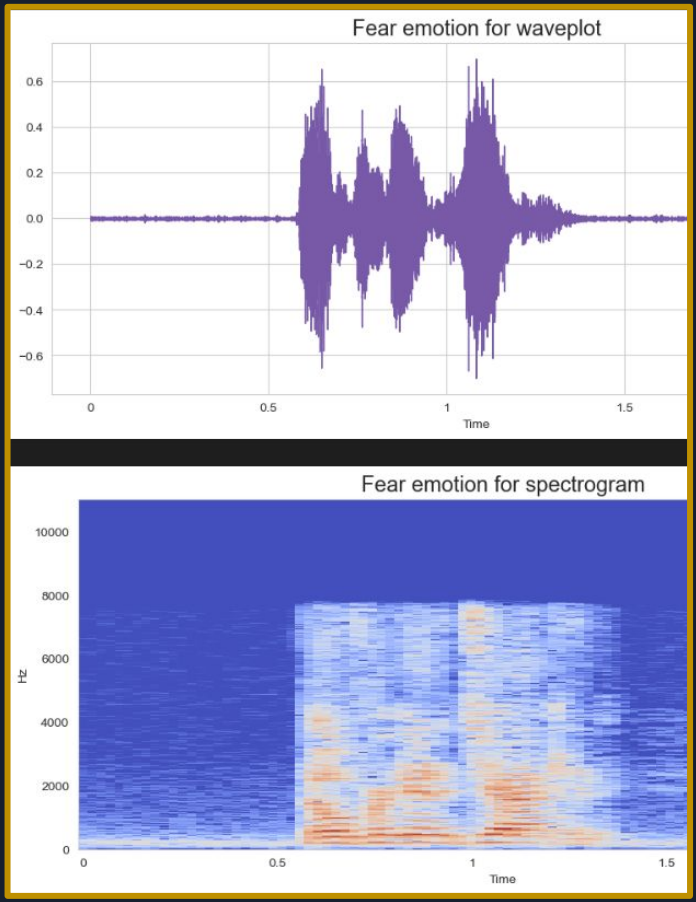


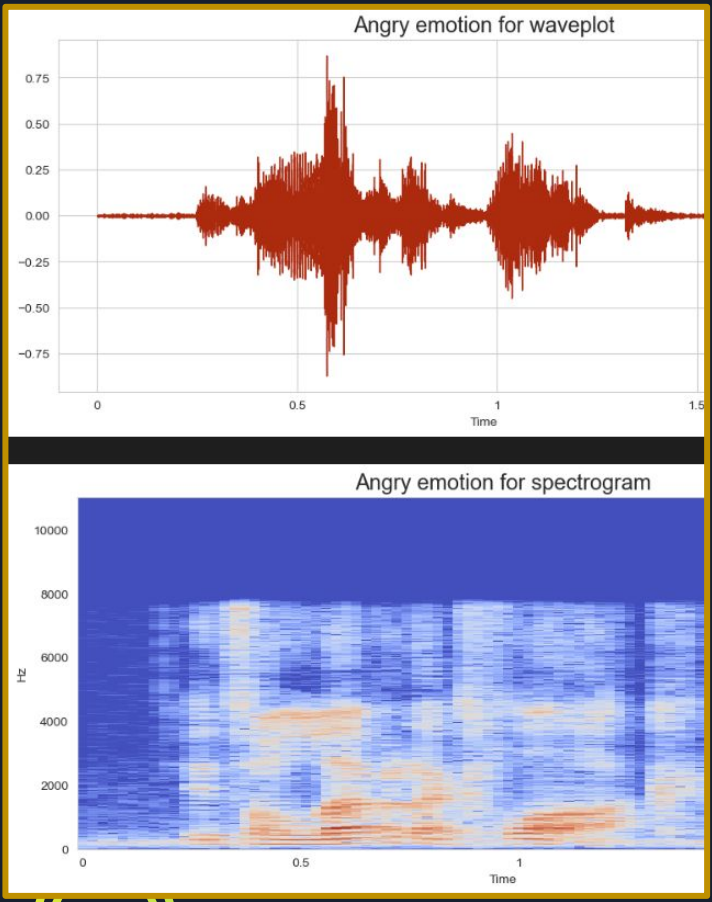
- Balanced through augmentation
- Surprise dropped due to significantly fewer points



Sample Audio Profiles

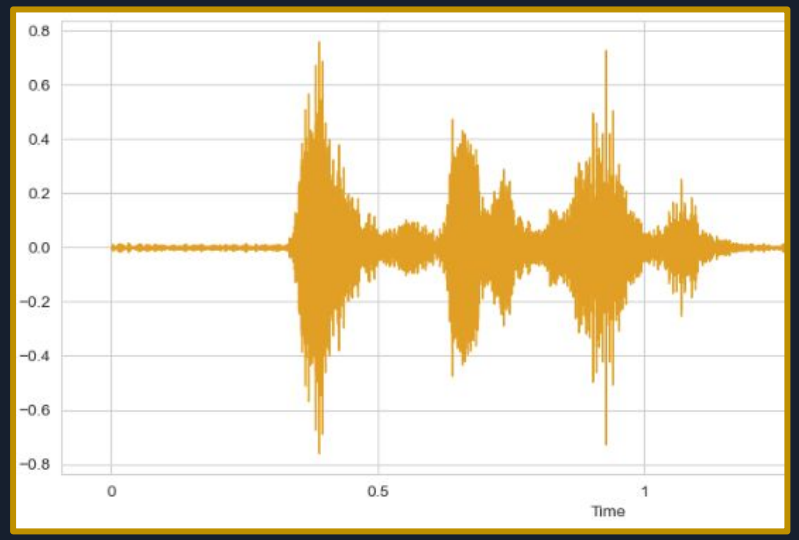




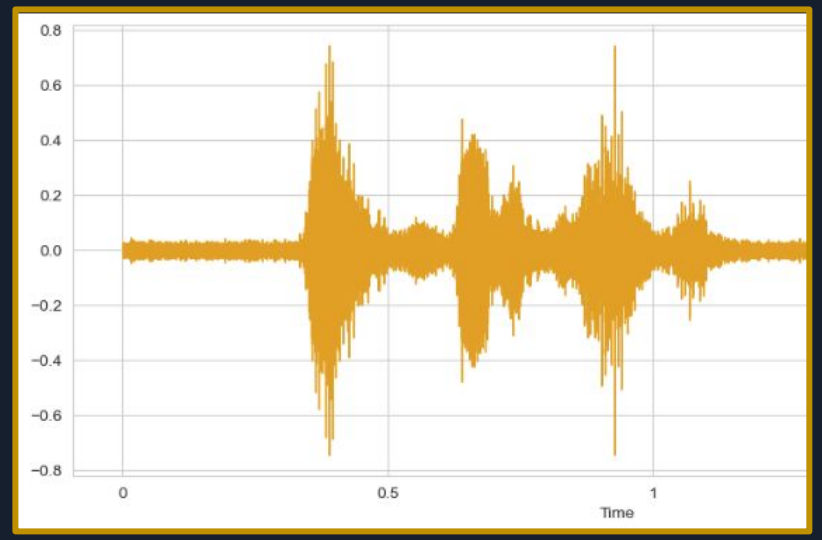


Sample Data Augmentations

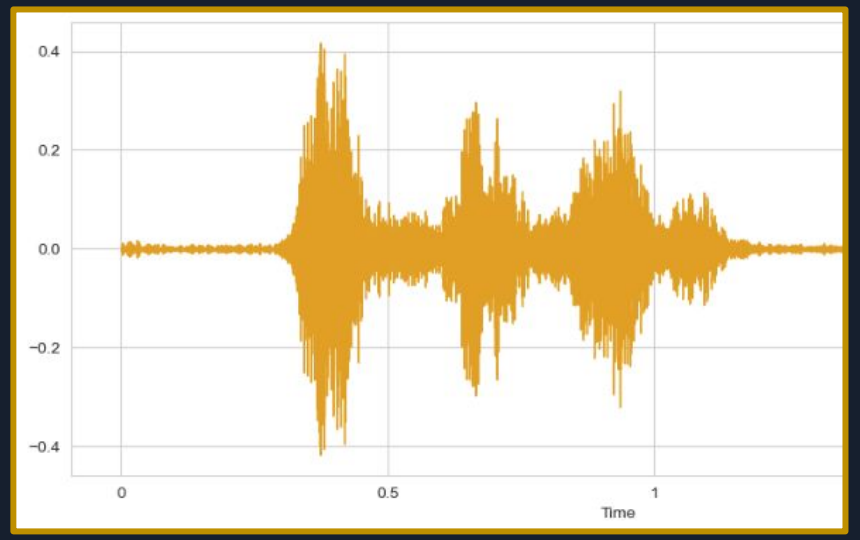
Original Sample



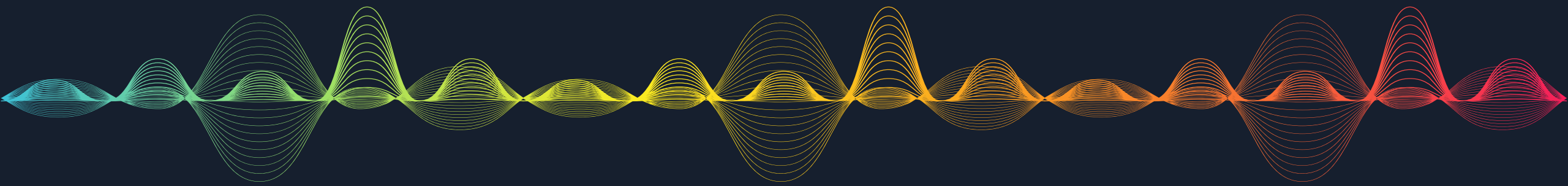
Noise Augmentation



Pitch Augmentation



Subtle Effects on the Original Sample decreasing Overfitting



Approach

Starting Points

Audio Processing Librosa, pytorch, sklearn, tqdm

Machine Learning Pytorch Library for Neural Networks

Data Crema, Savee, Ravdess, & Tess Data Sets

PreProcessing

Data Augmentation Stretch, Pitch, Shift and Noise data.
Balance data with augmentations

Feature Extraction Extract Audio Features. Zero Crossing Rate, Root Mean Squared Error, MFCC

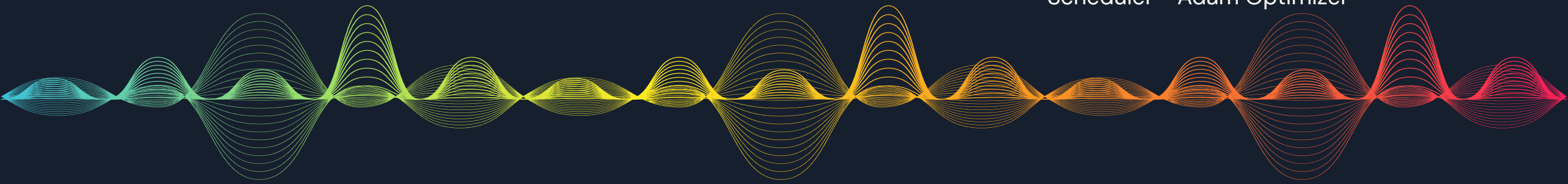
Processing One Hot Encoding on Labels
Standard Scaling Data

Train Test Split Train 80% / Test 10% / Validate 10%

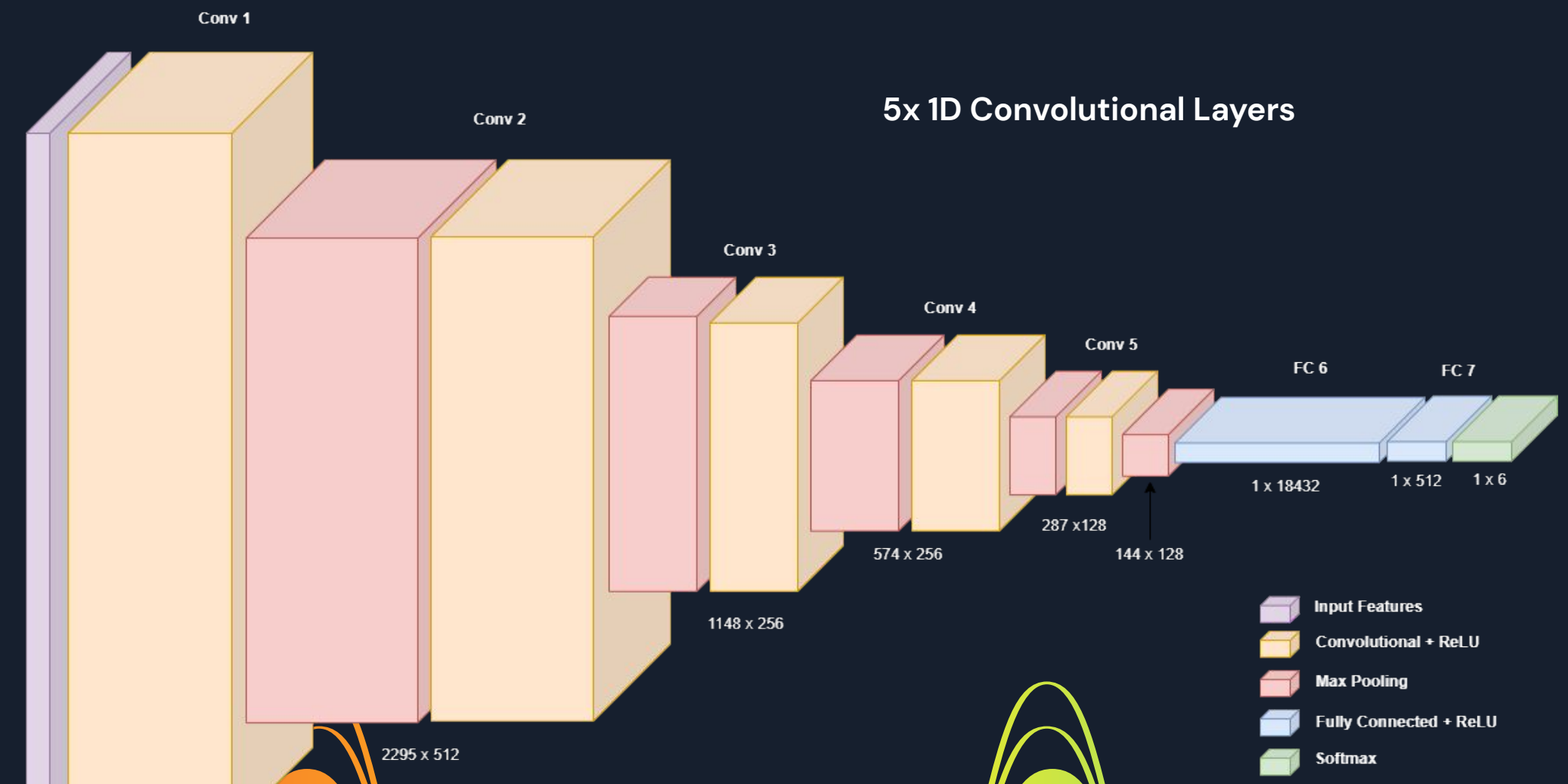
Training

Model 5 Layer CNN, with batch normalization and max pooling

Training Early Stopping - Patience of 10
Criterion - Cross Entropy Loss
Scheduler - Adam Optimizer



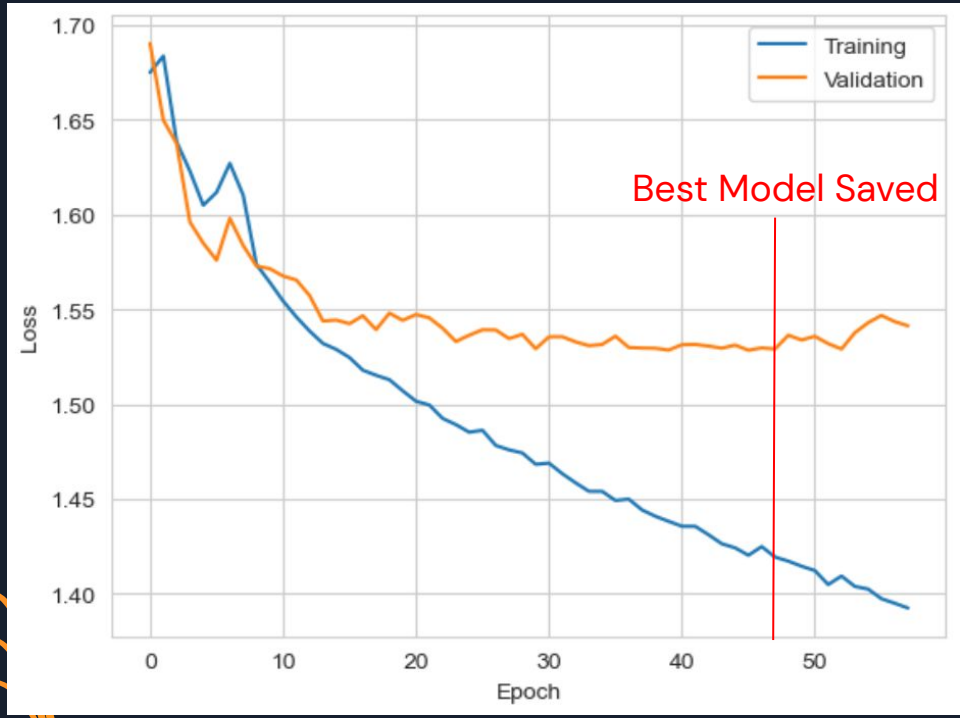
Model Architecture



Training History

Model predictive power: 47.9% on 6 classifications

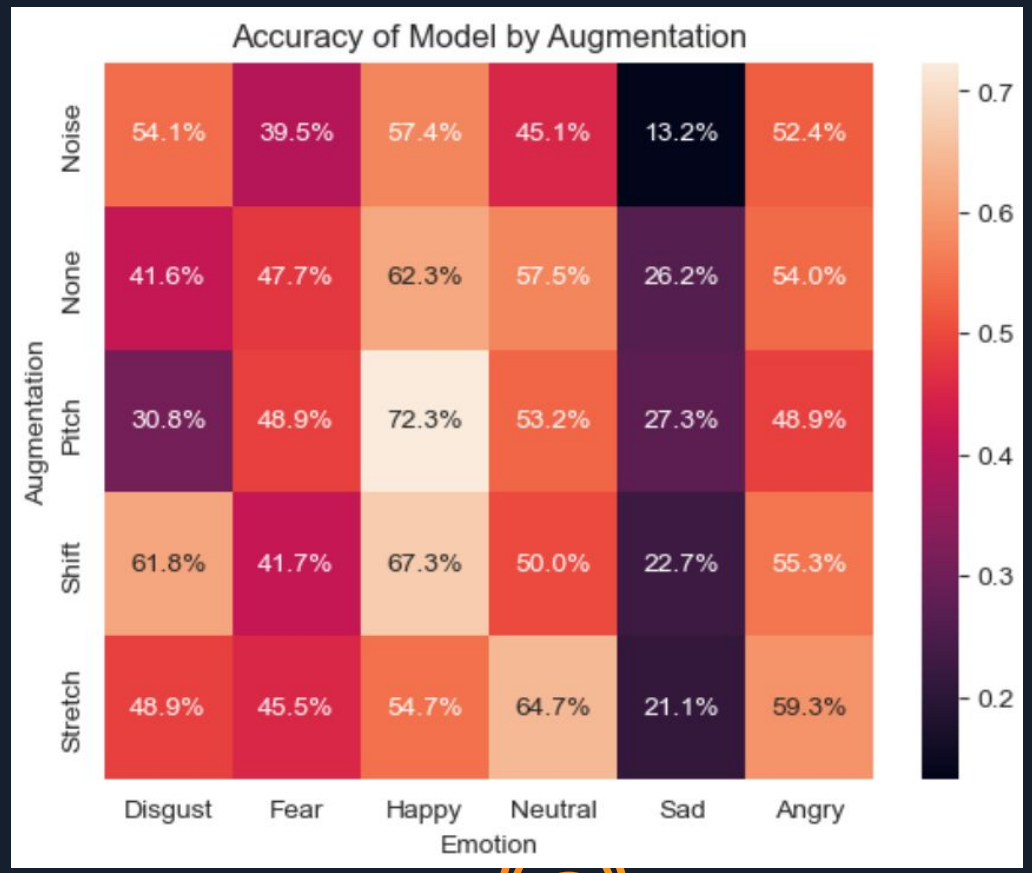
Cross Entropy Loss

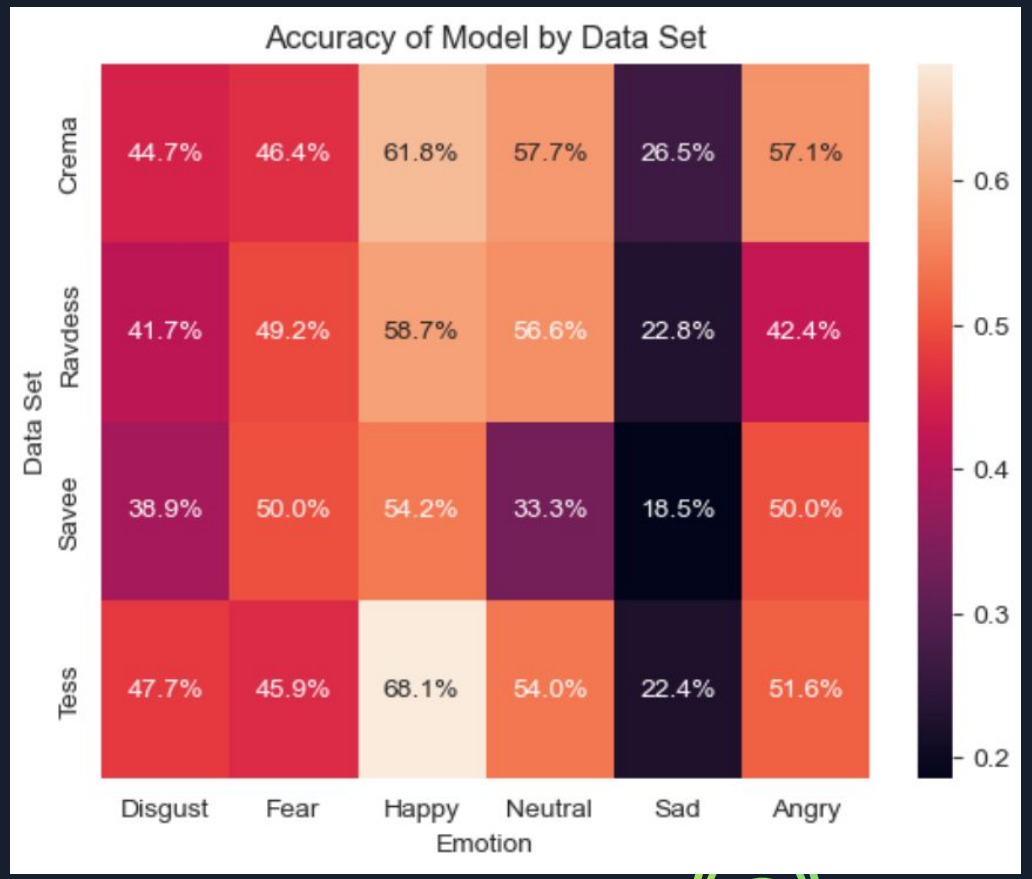


Confusion Matrix Total Set

Actual vs Predicted Emotions

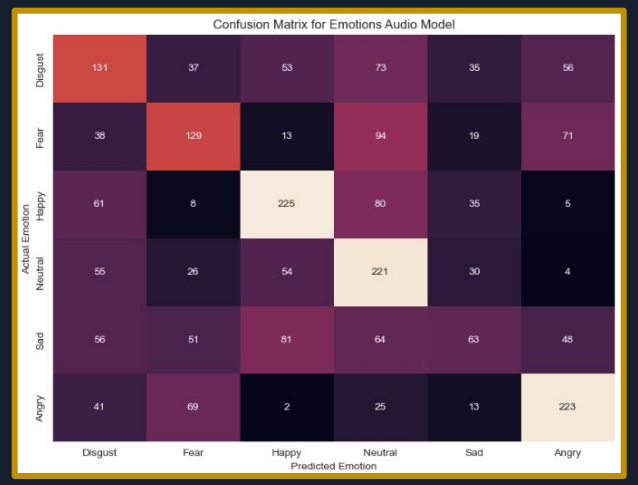
Evaluation By Augmentation and Data Set



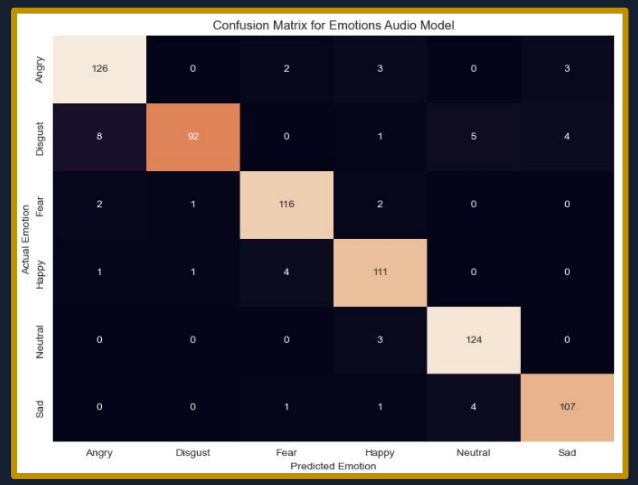


Datasets Individually

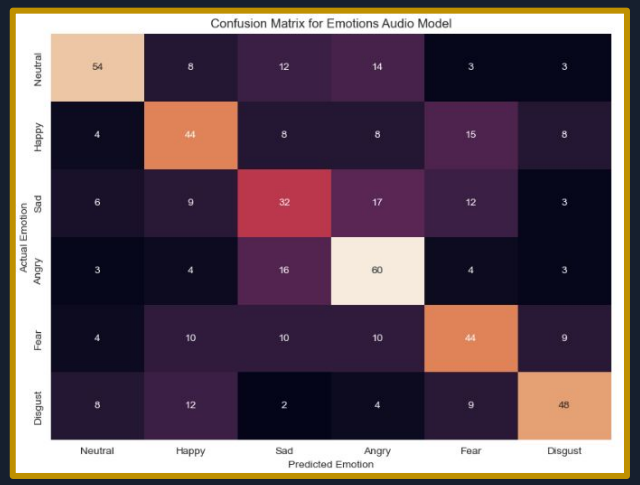
Crema
Accuracy 43.3%



Tess
Accuracy 93.6%



Ravdess
Accuracy 54.2%



Savee
Accuracy 46.8%

Moving Forward

Area of Improvements:

- Investigate difficulties in the “Sad” emotion- theorized to be a stylistic emotion which is more different between people-includes facial and body expression and is therefore difficult to capture with just sound as opposed to someone being “Fearful” or “Angry” which tends to be more distinct sounding
- Never enough data! Need more samples to include a ‘Surprised” emotion to expand model (IEMO-CAP Dataset)
- Evaluating Different model architectures and audio features. Fine tune the model to improve the performance