Pattern Recognition

Speech Emotion Recognition

Team Members

Name	ID
Ahmed Ashraf	21010040
Ahmed Osama	21010037
Saifullah Mousaad	21010651

1D CNN

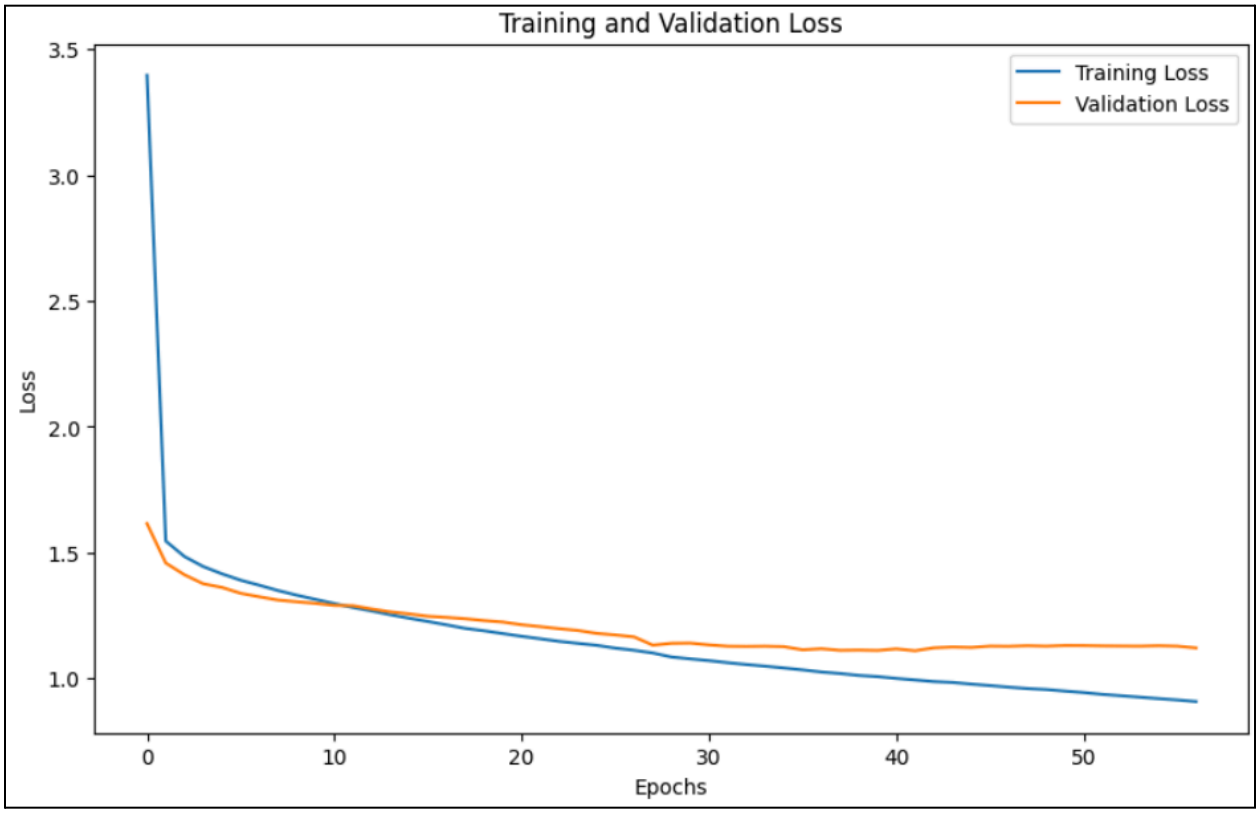
Approach

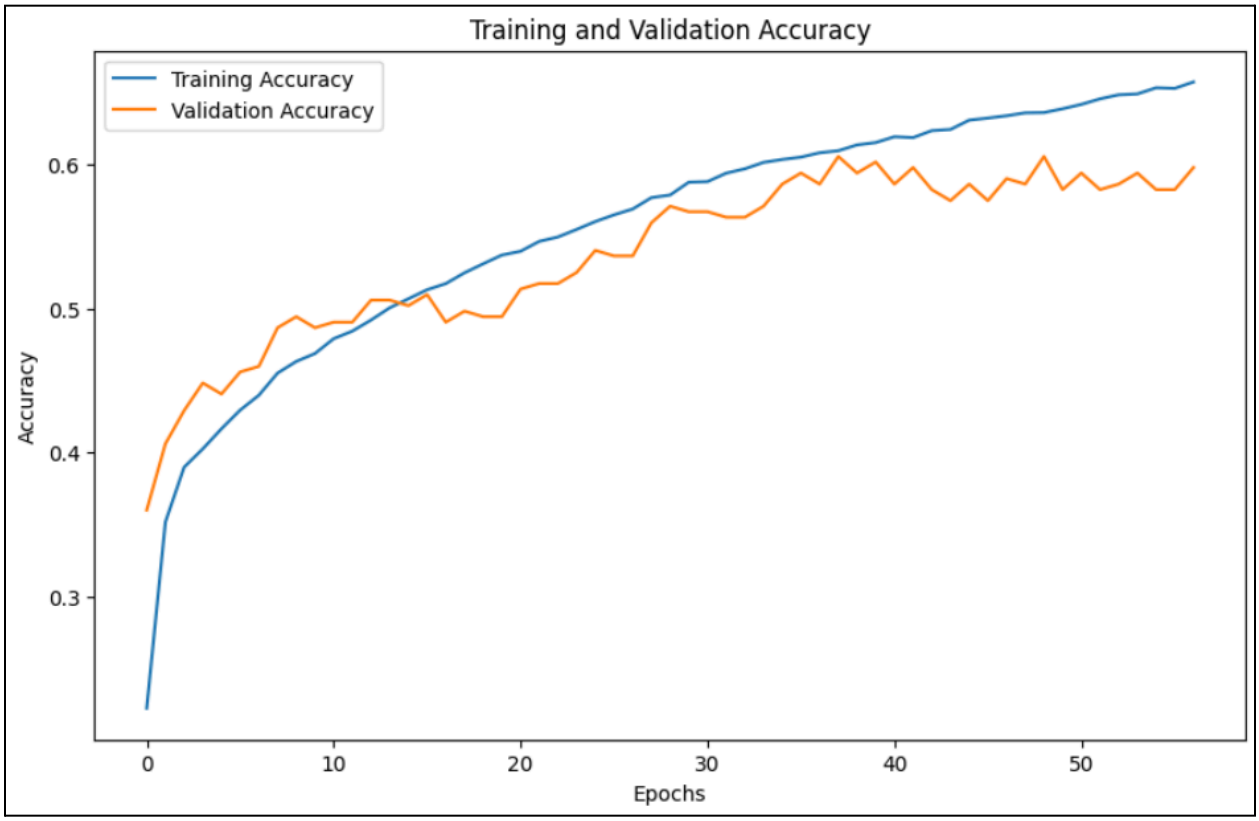
- Features used:
  - Zero Crossing Rate (ZCR).
  - Energy.
  - Mel-Frequency Cepstral Coefficients (MFCCs).
- Used 8-layers architecture.

Layer	Functionality
Input	Accepts 1D time-series features of shape (216, 15) (e.g., MFCC, ZCR, Energy).
Conv1D	Extracts local temporal patterns using 1D convolution with ReLU activation.
AveragePooling1D	Downsamples the time dimension by a factor of 2, keeping essential features.
Conv1D	Learns mid-level temporal representations.
AveragePooling1D	Further reduces time resolution, improving generalization.
Conv1D	Refines features with fewer filters, capturing high-level abstractions.
GlobalAveragePooling1D	Aggregates features across the entire time axis, reducing risk of overfitting.
Dense (ReLU)	Non-linear projection to a latent space.
Dense (Softmax)	Outputs probabilities for 6 emotion classes.

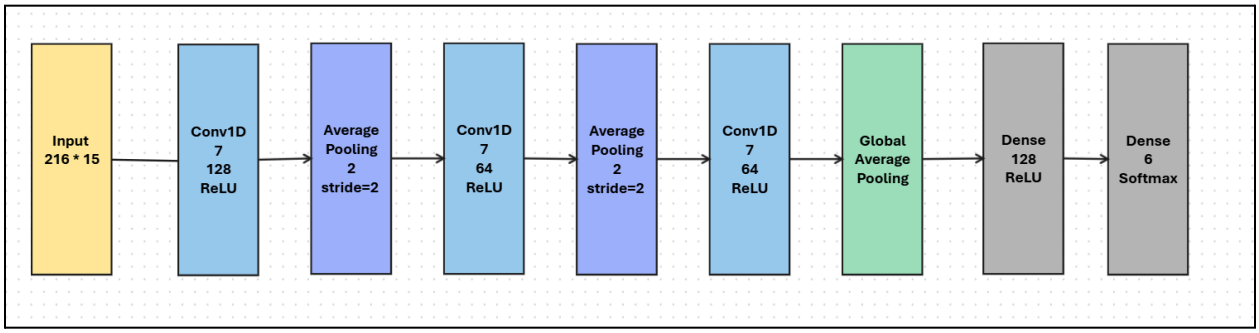
Results

- Automatic Tuning
  - Best Model: kernel size=7, filters=128
  - Best Accuracy: 0.5977 (Validation)
  - Test Accuracy: 0.5884 (58.84%)





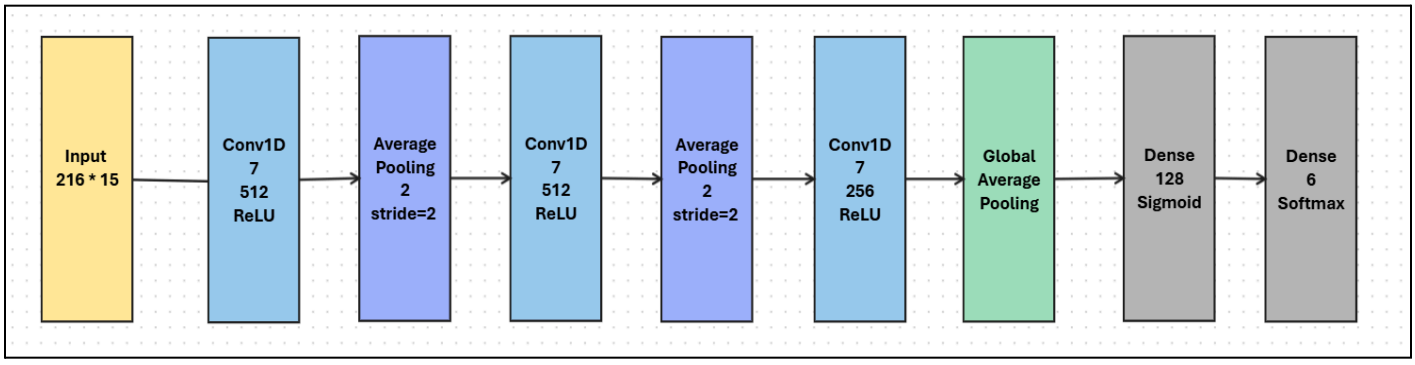
Full Architecture



- Manual Tuning (Trial and Error)

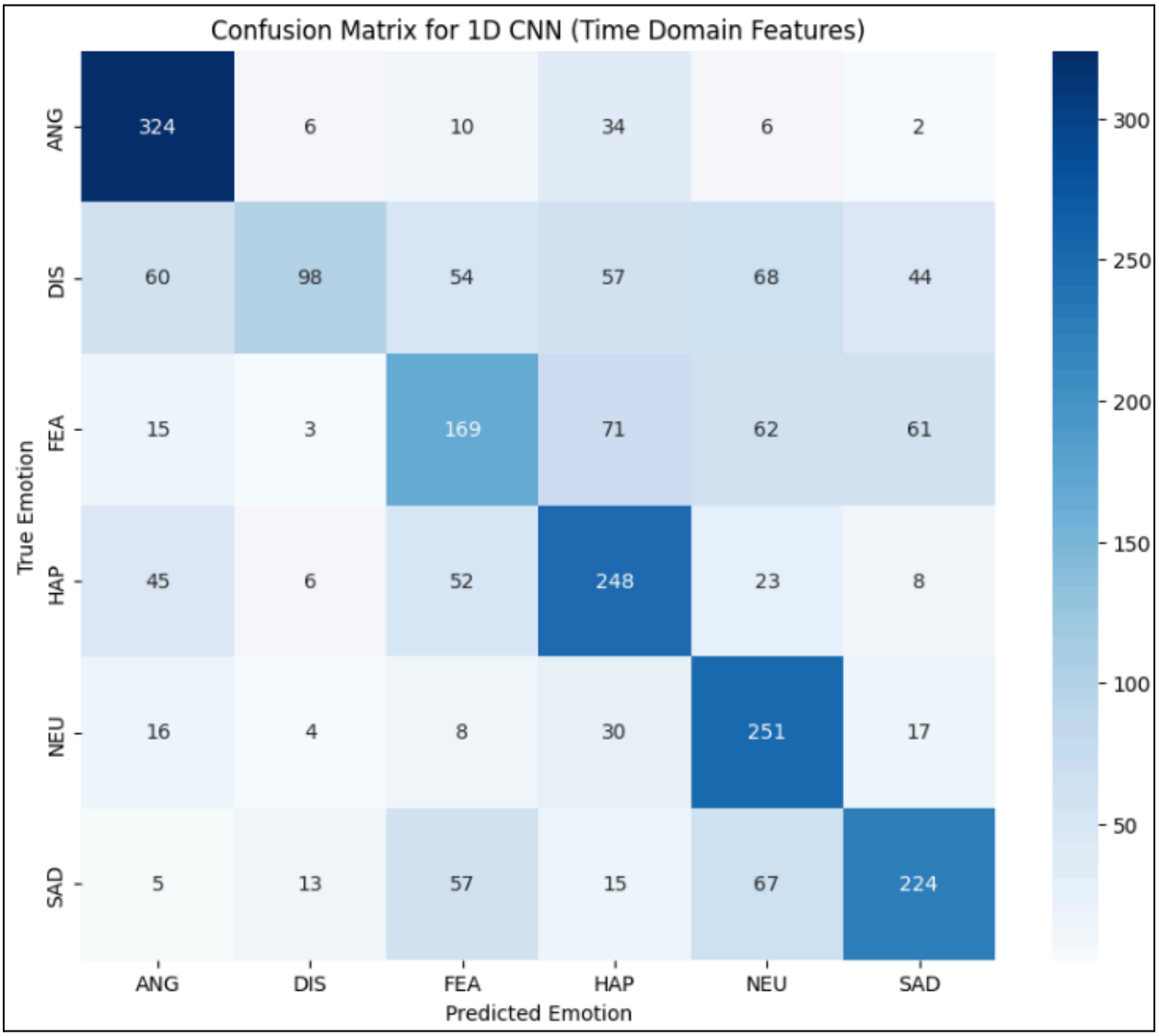
  We thought of some modifications that might introduce better accuracy based on the automatic tuning best results.

  - kernel size=7, filters=(512, 512, 256)
  - Validation Accuracy: 0.6245 (62.45%)
  - Test Accuracy: 0.6341 (63.41%)
  - Key Idea:
    - As there are 15 channels, we increased the filters to be 512 in both first and second conv layers and the third is halved as the channels became less due to average pooling.
    - Instead of using ReLU activation function in the dense layer before the softmax, we used sigmoid because it might highlight the differences between different emotions better than ReLU, as it amplifies the small values **(sigmoid =** $\frac{1}{1+e^{-x}}$**)**



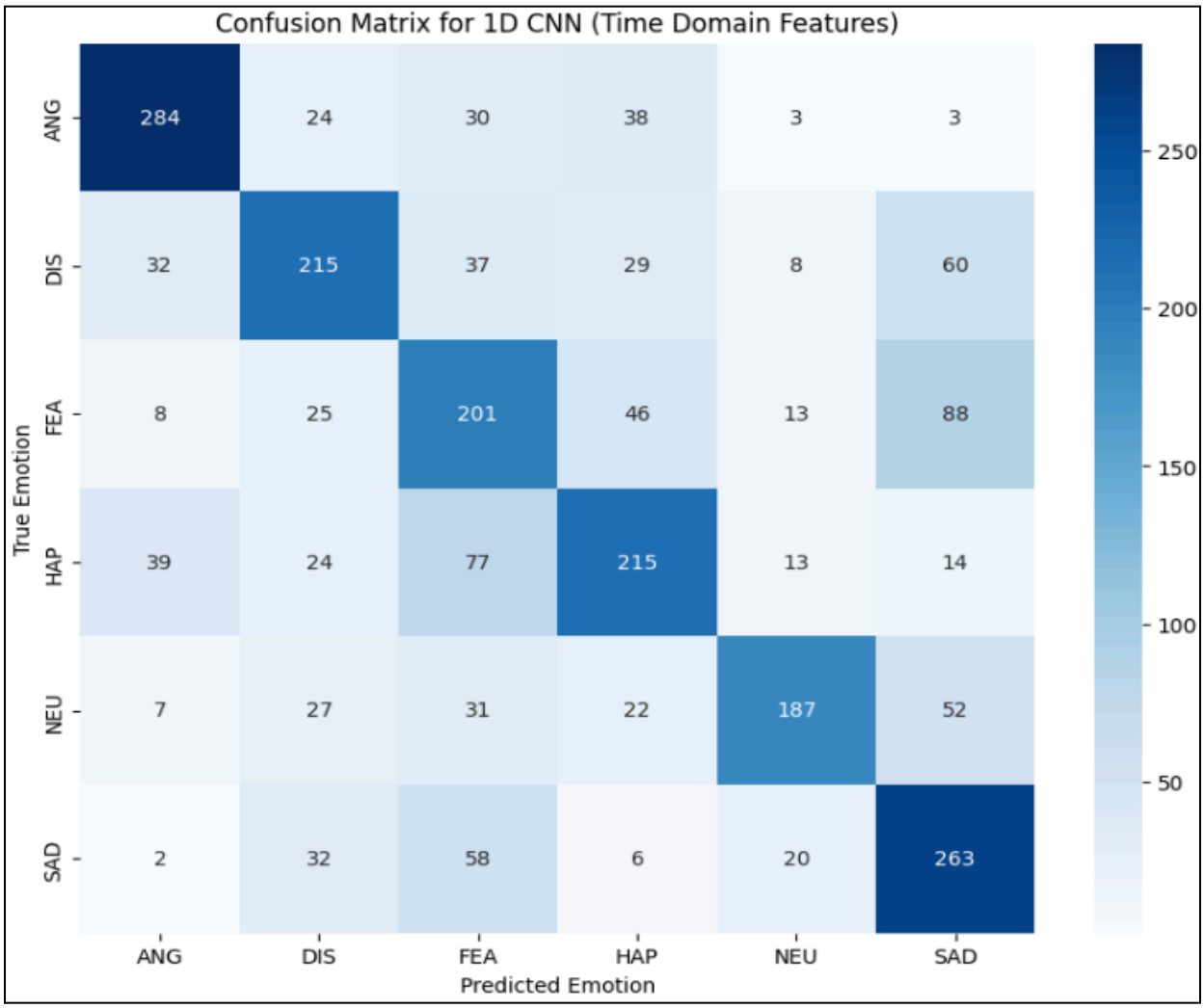
Analysis

- Automatic Tuning



- The issue is concentrated on FEAR and DISGUST emotions, especially DISGUST, where it can be predicted to be any of the 6 emotions. This indicates that the characteristics of this voice tone gather from all other tones. Hence, it's misleading for the model to correctly predict it.

- Manual Tuning



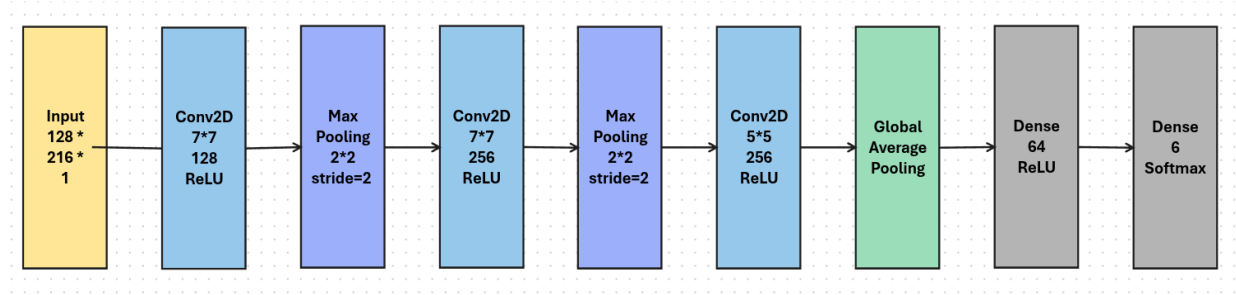
- With Modifications applied, DISGUST became more easily detected. Though similarities between voice tones are still present like the closeness between FEAR and SAD, both have quiet tones causing a relatively good number of samples to be misclassified.

- However, with an increasing number of filters, the model could extract better features and perform better than before.

2D CNN

Approach

- Worked on Mel Spectrogram DB.
- Used 8-layers architecture after trying multiple architectures (increasing number of dense layers, decreasing convolutional layers, different activation functions and adding max, average pooling).



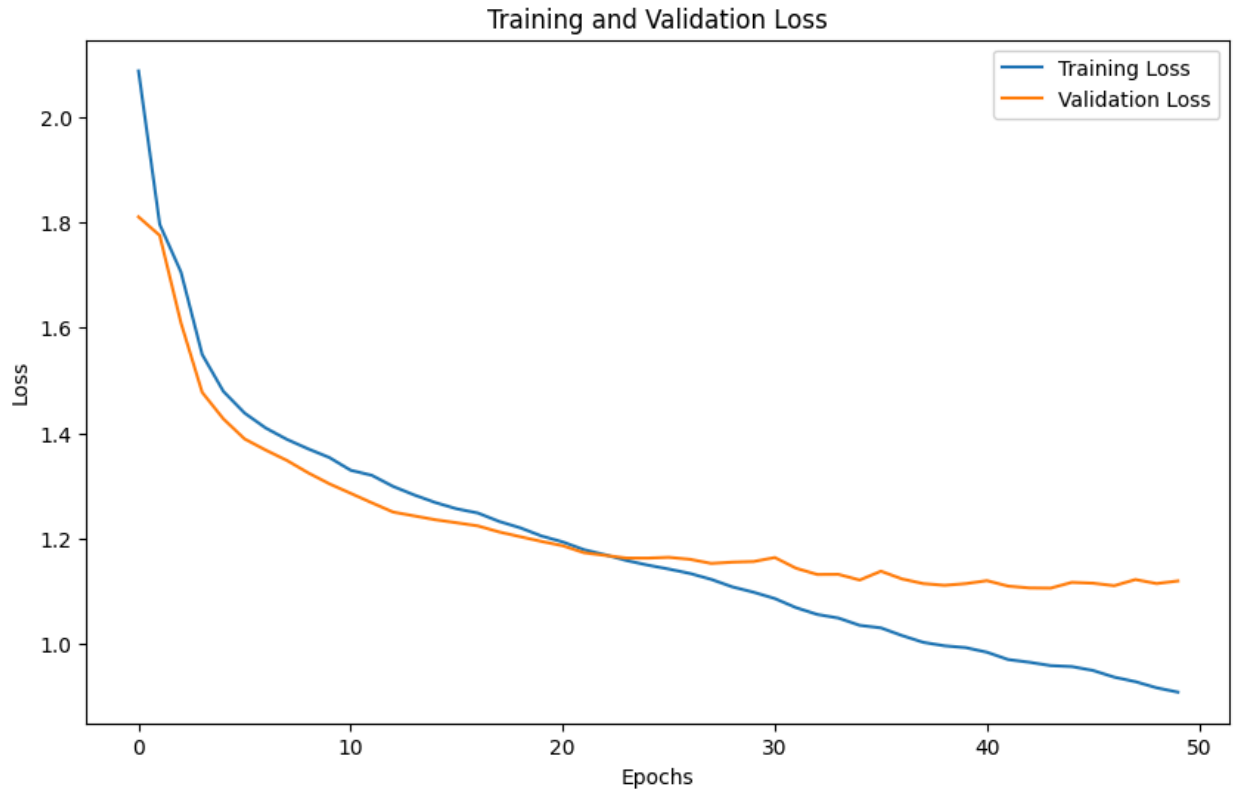
Layer	Functionality
Input	Accepts Mel spectrograms of shape (128, 216, 1).
Conv2D	Extracts low-level features with ReLU activation.
MaxPooling2D	Reduces spatial dimensions by 2×, retaining dominant features.
Conv2D	Captures mid-level features.
MaxPooling2D	Further downsampling.
Conv2D	Captures high-level features.
GlobalAveragePooling2D	Averages all spatial locations, reducing risk of overfitting.
Dense	Non-linear projection to a latent space.
Dense	Outputs probabilities for 6 emotion classes.

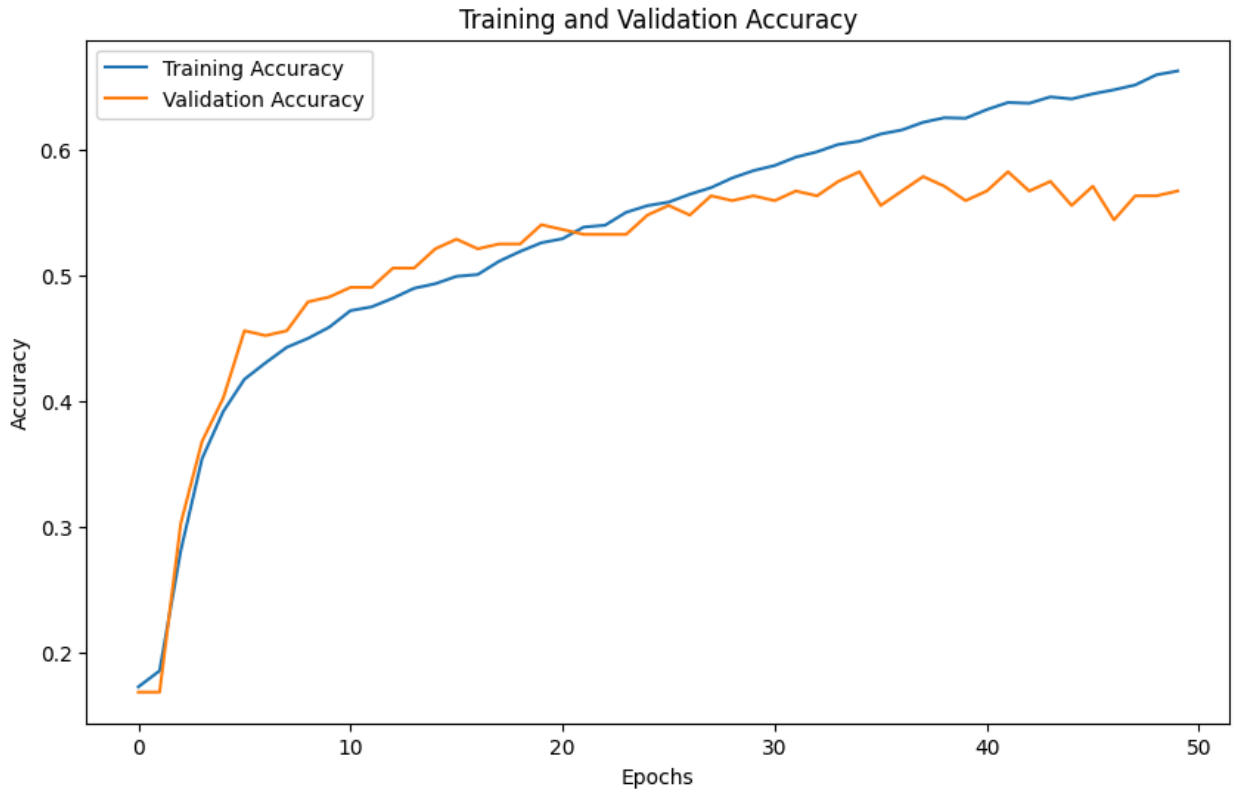
- Applied hyperparameter tuning on two hyperparameters with fixed number of strides (2) and fixed number of epochs (50):
  - **Number of filters:** 32, 64, 128
  - **Kernel size:** (3, 3), (5, 5), (7, 7)

Results

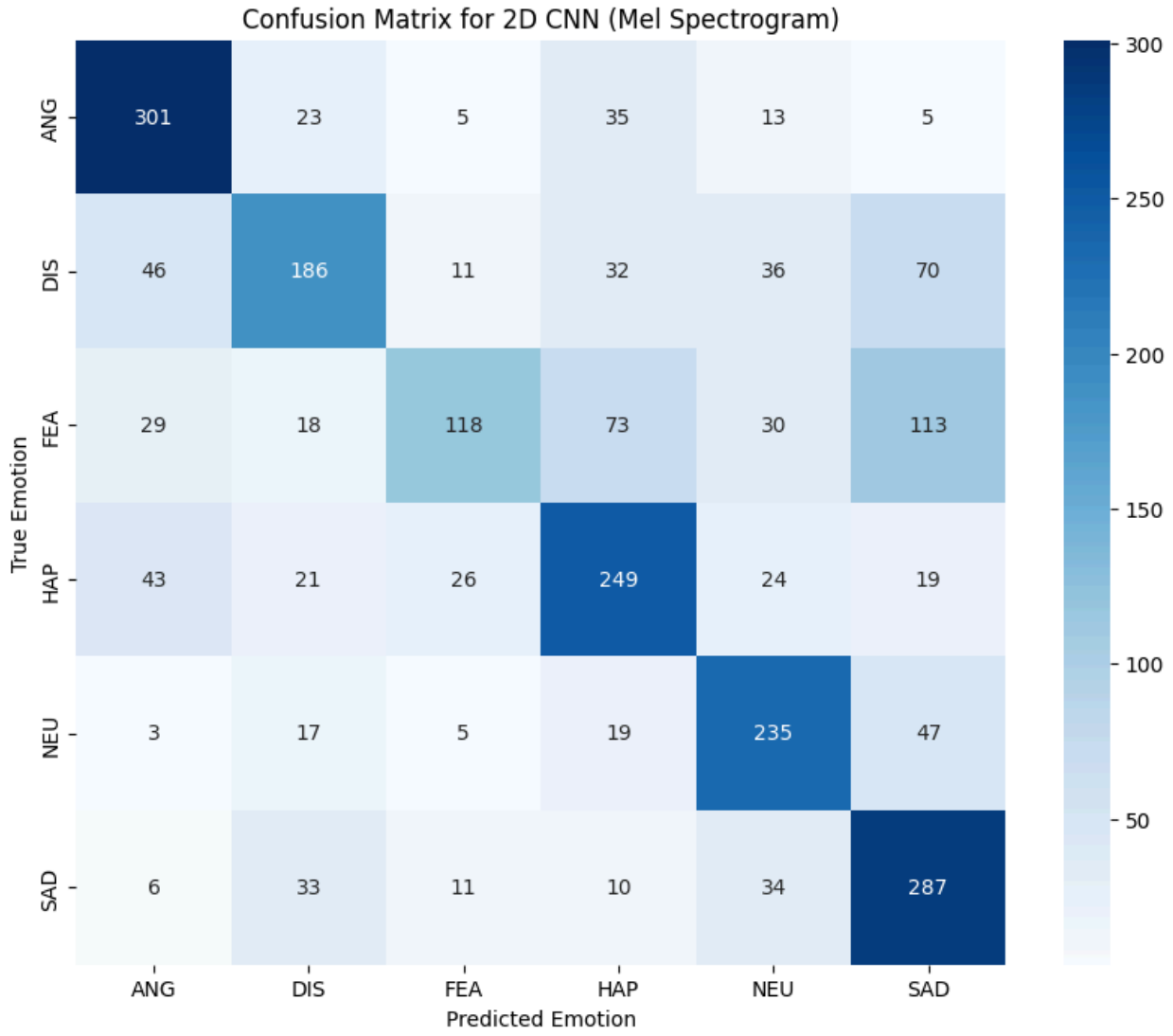
After Hyperparameter Tuning:

- Best Model: kernel size=(7, 7), filters=128
- Best Accuracy: 0.5670498013496399
- Test Accuracy: 0.6162 (61.62%)
- f1-score: 0.6039295817257896




Training and Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy
Epochs
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
50

Analysis



- Most confusing pair: True FEA predicted as SAD (normalized error: 0.2966)

```
Classification Report:
              precision    recall  f1-score   support

         ANG       0.70      0.79      0.74       382
         DIS       0.62      0.49      0.55       381
         FEA       0.67      0.31      0.42       381
         HAP       0.60      0.65      0.62       382
         NEU       0.63      0.72      0.67       326
         SAD       0.53      0.75      0.62       381

    accuracy                           0.62      2233
   macro avg       0.63      0.62      0.61      2233
weighted avg       0.63      0.62      0.60      2233
```