PES
UNIVERSITY

Machine Learning Assignment

PROJECT REPORT

TEAM ID : 26

Speech Emotion Recognition

Name	SRN
Dinakar Emmanuel	PES2UG23CS178
Cheruku Manas Ram	PES2UG23CS147

Problem Statement

Using datasets such as RAVDESS or Emo-DB consisting of labeled emotional speech. Extract acoustic features like MFCC's, chroma, spectographic images etc from audio samples. Train models including CNN's, LSTM's, or hybrid architecture for multiclass emotional classification. Evaluate using accuracy and confusion matrices. Test data augmentation approaches like noise injection to improve robustness.

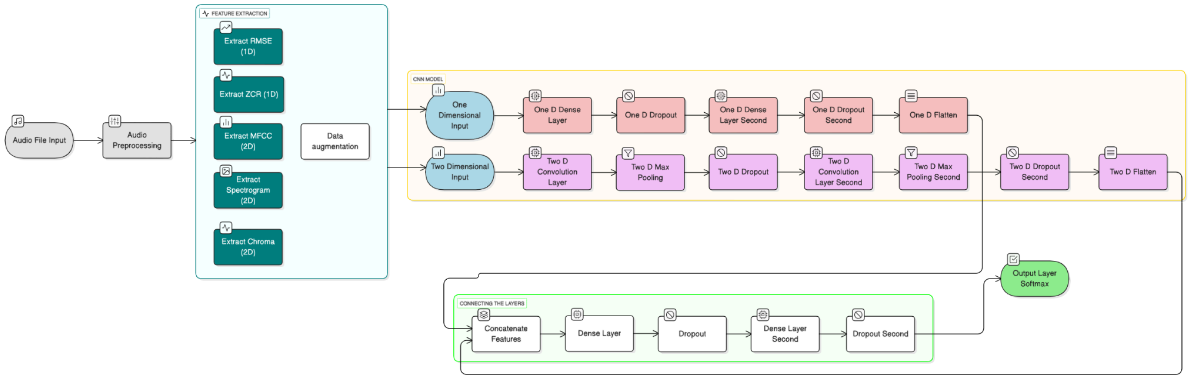
Objective / Aim

When given an audio file as input, the model is supposed to extract its features and classify it based on the emotion it conveys.

Dataset Details

- **Source:** Kaggle (RAVDESS audio dataset)
- **Size:** 1440 samples (24 actors (12 male, 12 female), 8 emotions)
- **Key Features:** MFCC, chroma, spectrogram, zero crossing rate, root mean square energy
- **Target Variable:** Emotion of speech

Architecture Diagram



Methodology

- We focused on audio-only speech. We split the datasets based on actors. Since there were 24 actors, we used 20 actors to train the model and 4 actors to test our trained our model.This ensured that the model was tested on it’s ability to recognise emotions from voices that it has never heard before.

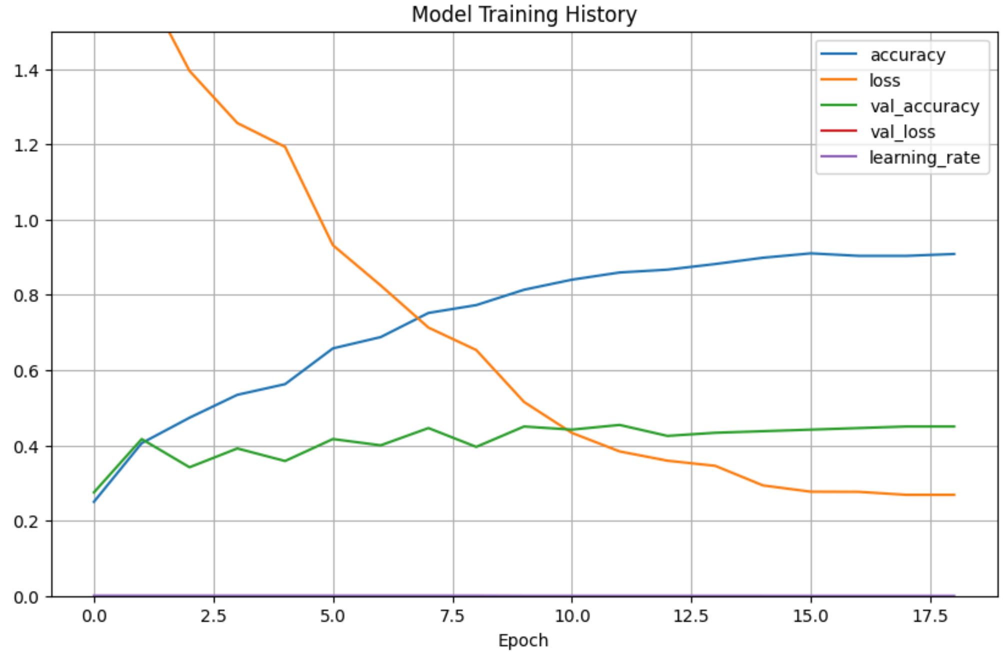
- A set of acoustic features were extracted from the dataset to train the model. We extracted **MFCC** to capture the timbre and vocal tract shape, **chroma** to represent the harmonic content (pitch), **ZCR** to measure the sharpness of the signal, **RMS Energy** to measure how loud the signal is and **Spectogram images.** We also experimented with more features like spectral rolloff and spectral centroid. The ZCR and RMSE were taken as 1-D features while the rest were represented as 2-D features.
- The features were extracted for short overlapping frames and then stacked. This would make multi-dimensional time-series for each audio file.
- Since the features were split into 1-D and 2-D features, the feature extraction was done separately in 2 branches. Once, the features were extracted, we concatenated both branches.
- We then used this to train our models (CNN, LSTM and hybrid architectures).
- We also augmented the data to make it more robust allowing for the features to be extracted better. We added low amplitude random gaussian noise. We shifted the pitch by a small random interval and we stretched the audio by speeding up or slowing down the audio.
- Using the features we extracted, we repeatedly trained our models, making small improvements at every iteration. We trained a 1D-CNN, 2D-CNN (Convolutional Neural Network) model using multiple blocks and layers to capture the local patterns in the sequences. We then trained a LSTM (Long Short-Term Memory) model, which is a recurrent neural network built by stacking layers to model the long-range temporal dependencies in the sequences. Finally, we trained a hybrid CNN-LSTM model which combined the architecture of the previous two models. We used CNN as our front end which identified local patterns and LSTM as our back end to process the sequences of the patterns to understand the temporal context.
- We normalised the features using StandardScaler (from scikit-learn) and fit it only on training data. Since the emotions are categorical, we one-hot encoded them. We used TensorFlow and Keras to build the models.

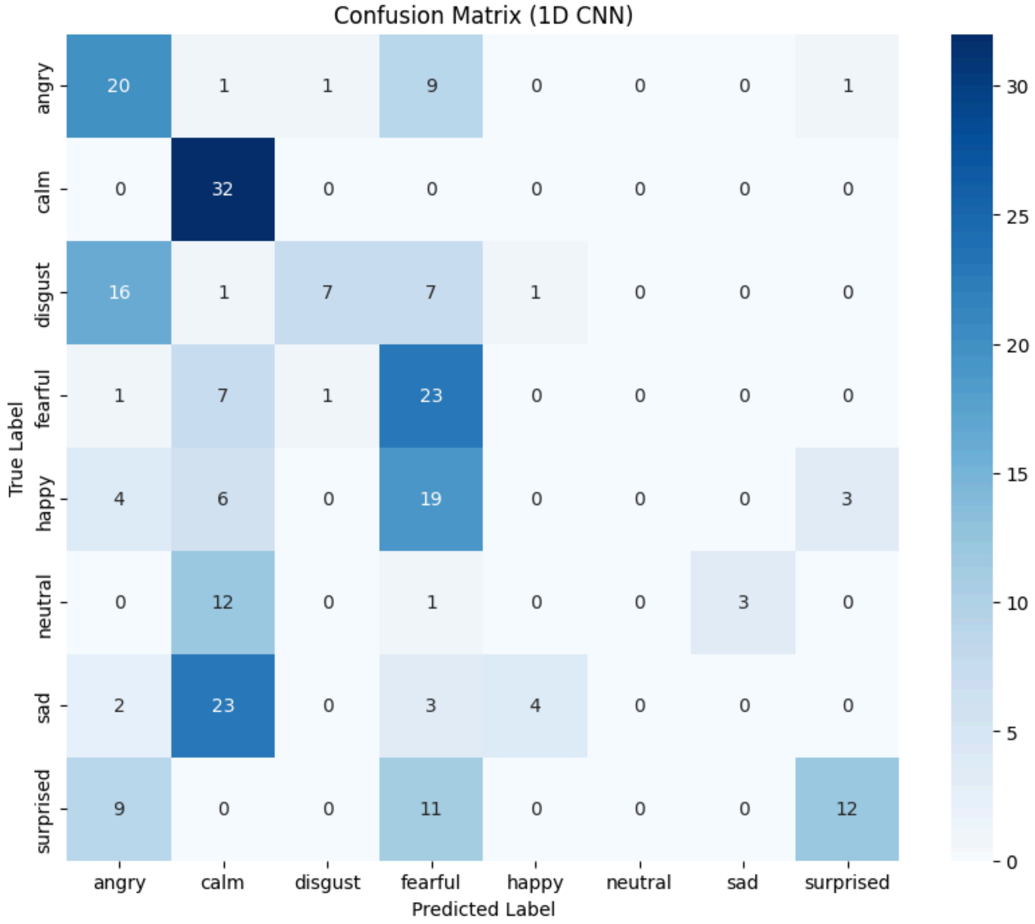
Results & Evaluation

Evaluation Metrics: Accuracy and Confusion Matrix.

CNN

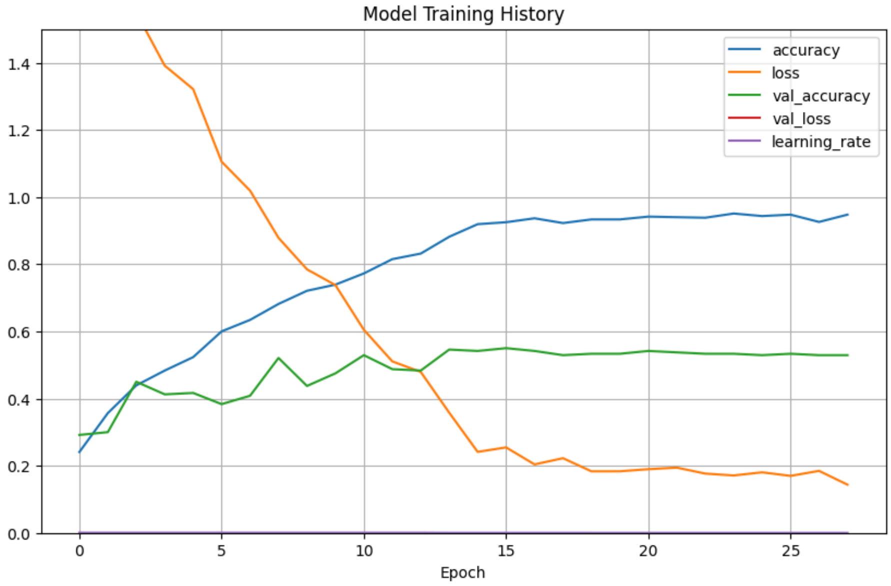
The first model without augmentation had an accuracy of 39.17%.

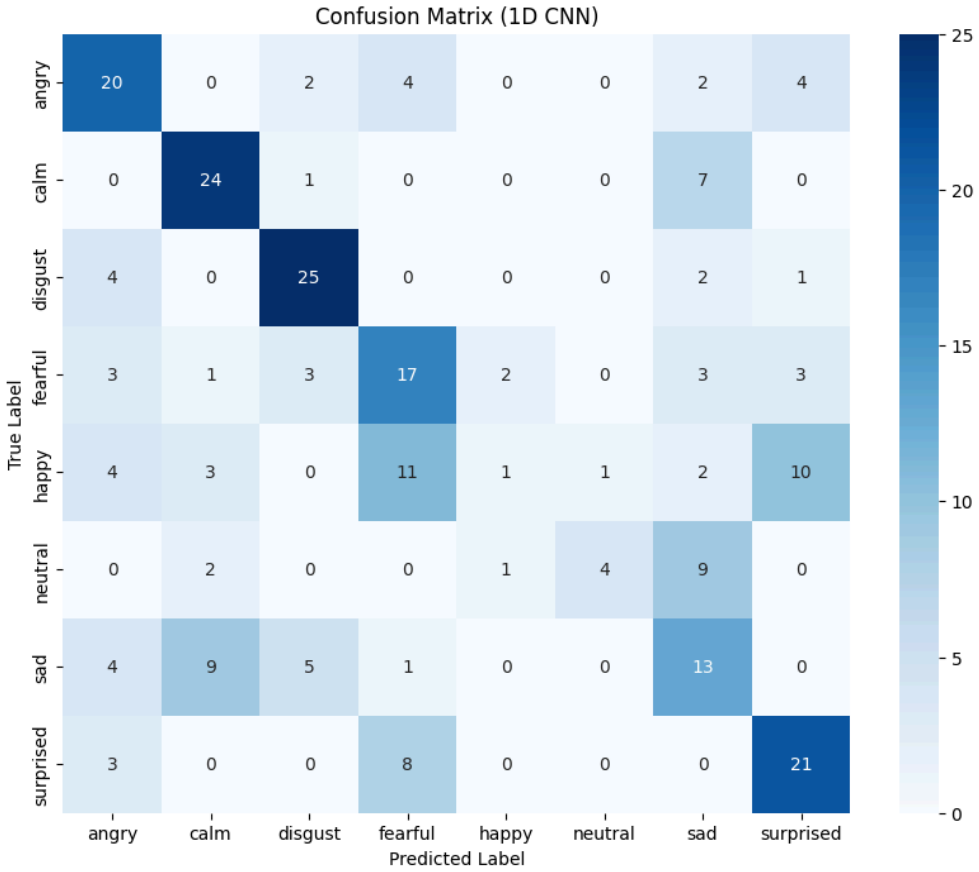




We can see from the confusion matrix that the model performed well in predicting labels like “calm” and “fearful”. But performed very poorly with labels like “happy” (mistaking for “fearful”), “neutral” (mistaking for “calm”) and “sad” (mistaking for “calm”) where it wasnt able to predict the emotion correctly at all.

After augmenting the model, we received an accuracy of 52.08%.





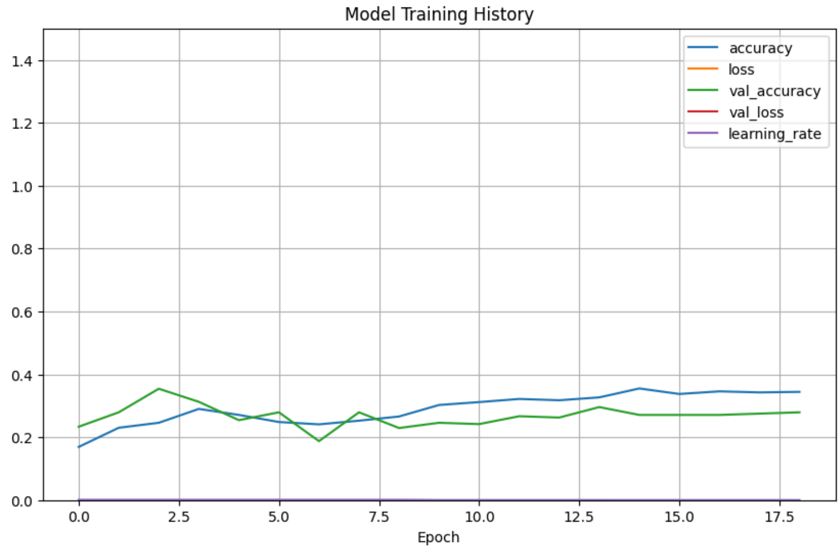
	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	20	0	2	4	0	0	2	4
calm	0	24	1	0	0	0	7	0
disgust	4	0	25	0	0	0	2	1
fearful	3	1	3	17	2	0	3	3
happy	4	3	0	11	1	1	2	10
neutral	0	2	0	0	1	4	9	0
sad	4	9	5	1	0	0	13	0
surprised	3	0	0	8	0	0	0	21

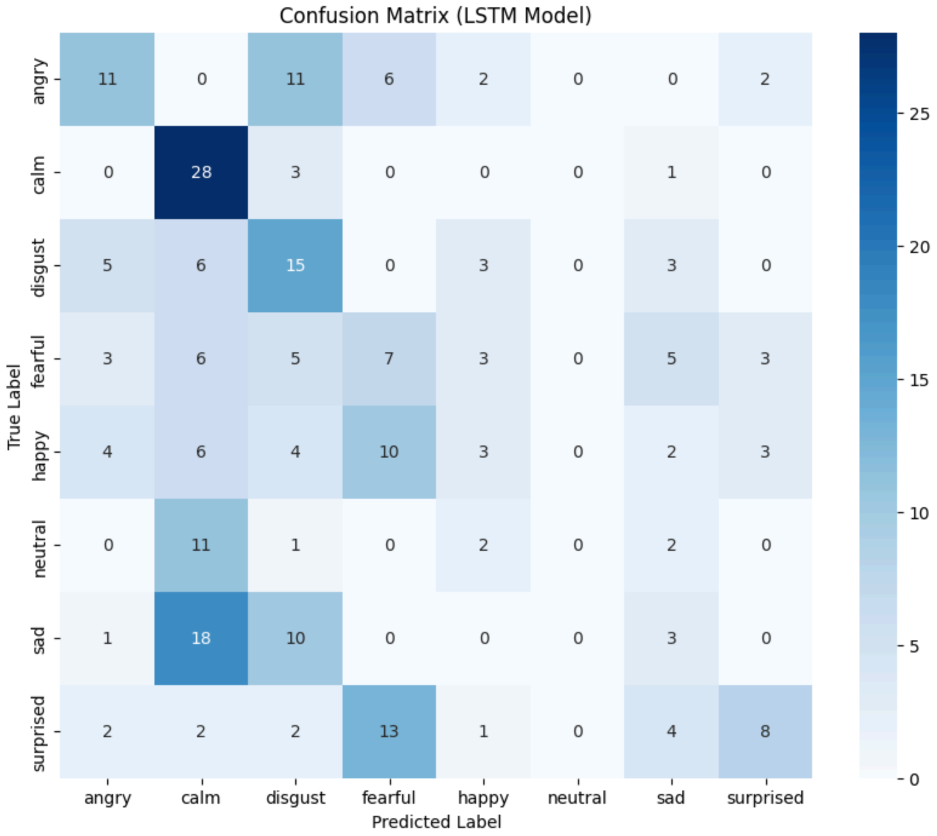
We can now see from the confusion matrix that the model is now predicting better, especially with emotion like “sad” where it is able to predict correctly 40%

of the time, unlike previously where it wasnt able to predict the emotion at all. We can also notice the predictions of emotion such as “disgust” and “suprised” get better.

LSTM

The first model without augmentation achieved an accuracy of 31.25%.

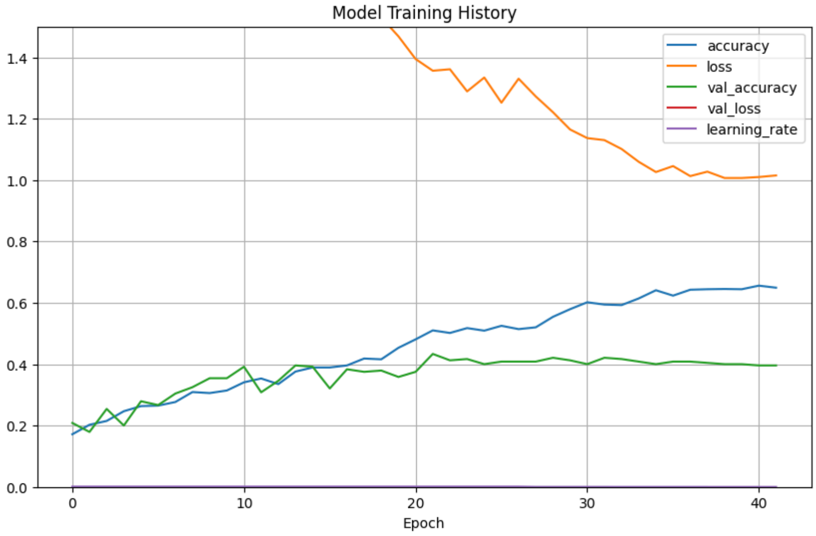


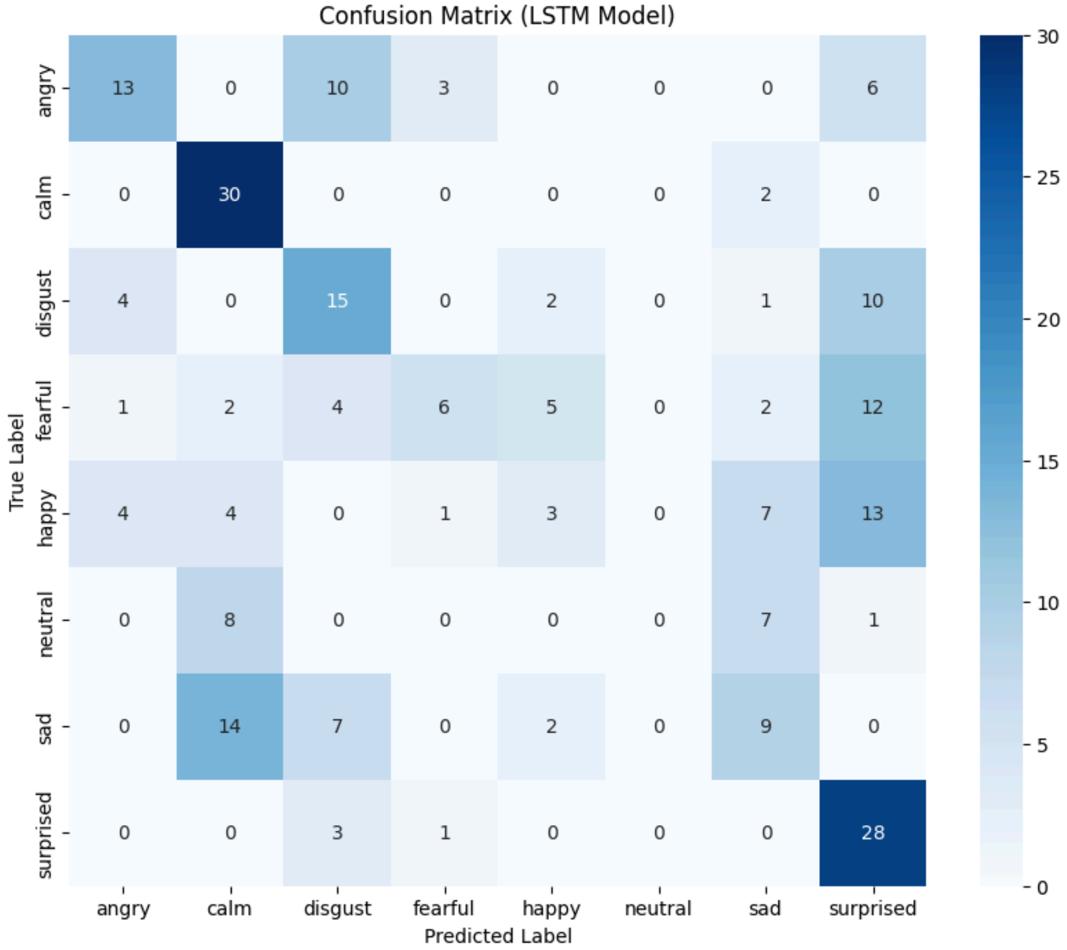


True \ Predicted	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	11	0	11	6	2	0	0	2
calm	0	28	3	0	0	0	1	0
disgust	5	6	15	0	3	0	3	0
fearful	3	6	5	7	3	0	5	3
happy	4	6	4	10	3	0	2	3
neutral	0	11	1	0	2	0	2	0
sad	1	18	10	0	0	0	3	0
surprised	2	2	2	13	1	0	4	8

We can see from the confusion matrix that even LSTM is best at predicting “calm”, with “neutral” being its worst emotion to predict.

After augmentation, the accuracy jumps to 43.33%.

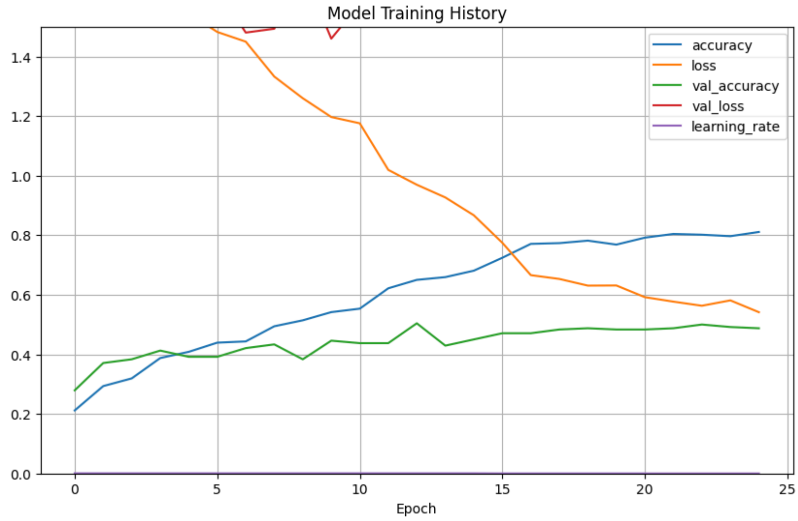


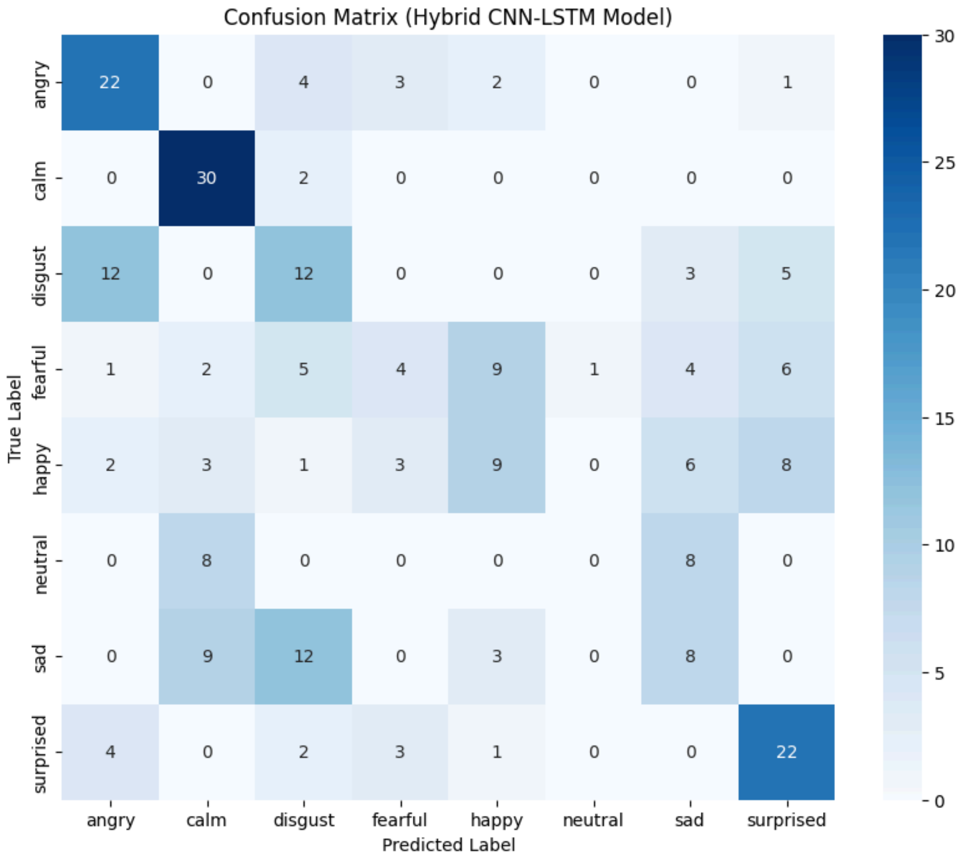


We can see that although the accuracy has increased by about 12%, it is still not able to predict “neutral”. But this model has gotten better at predicting emotions such as “calm”, “sad” and especially the emotion of “surprise” which it is now able to predict almost 88% of the time as opposed to before augmentation where it was only able to predict “surprise” 25% of the time.

Hybrid CNN-LSTM Model

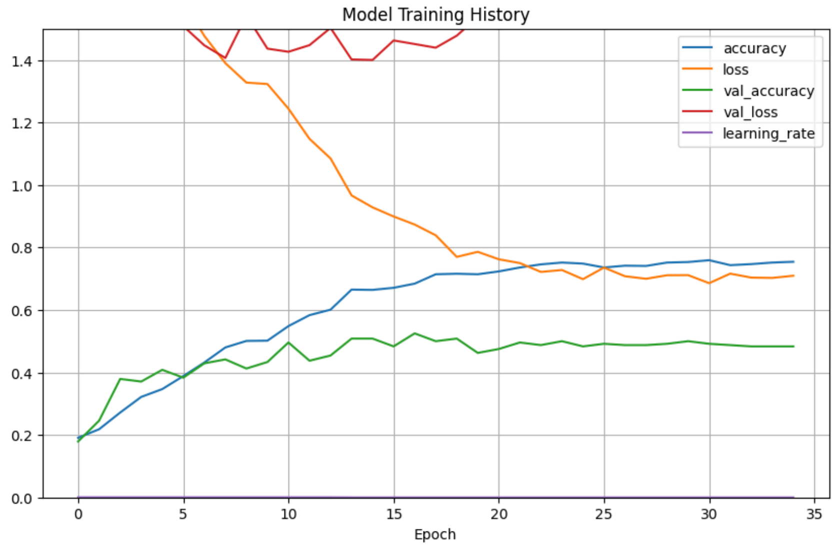
Before augmentation, it achieved an accuracy of 44.58%

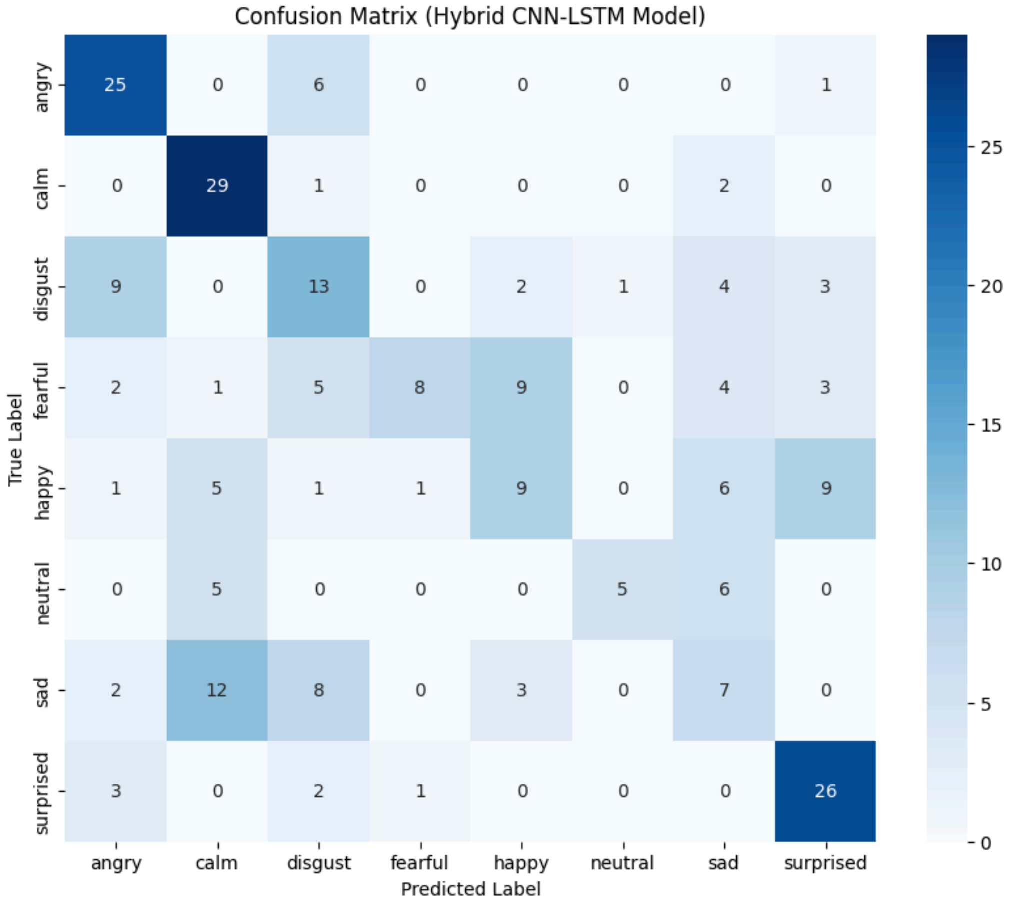




We can see from the confusion matrix that as expected, even the hybrid model is predicting "calm" almost perfectly with neutral being it's worst prediction never being able to predict it correctly.

After augmentation, it gives an accuracy of 50.83%.

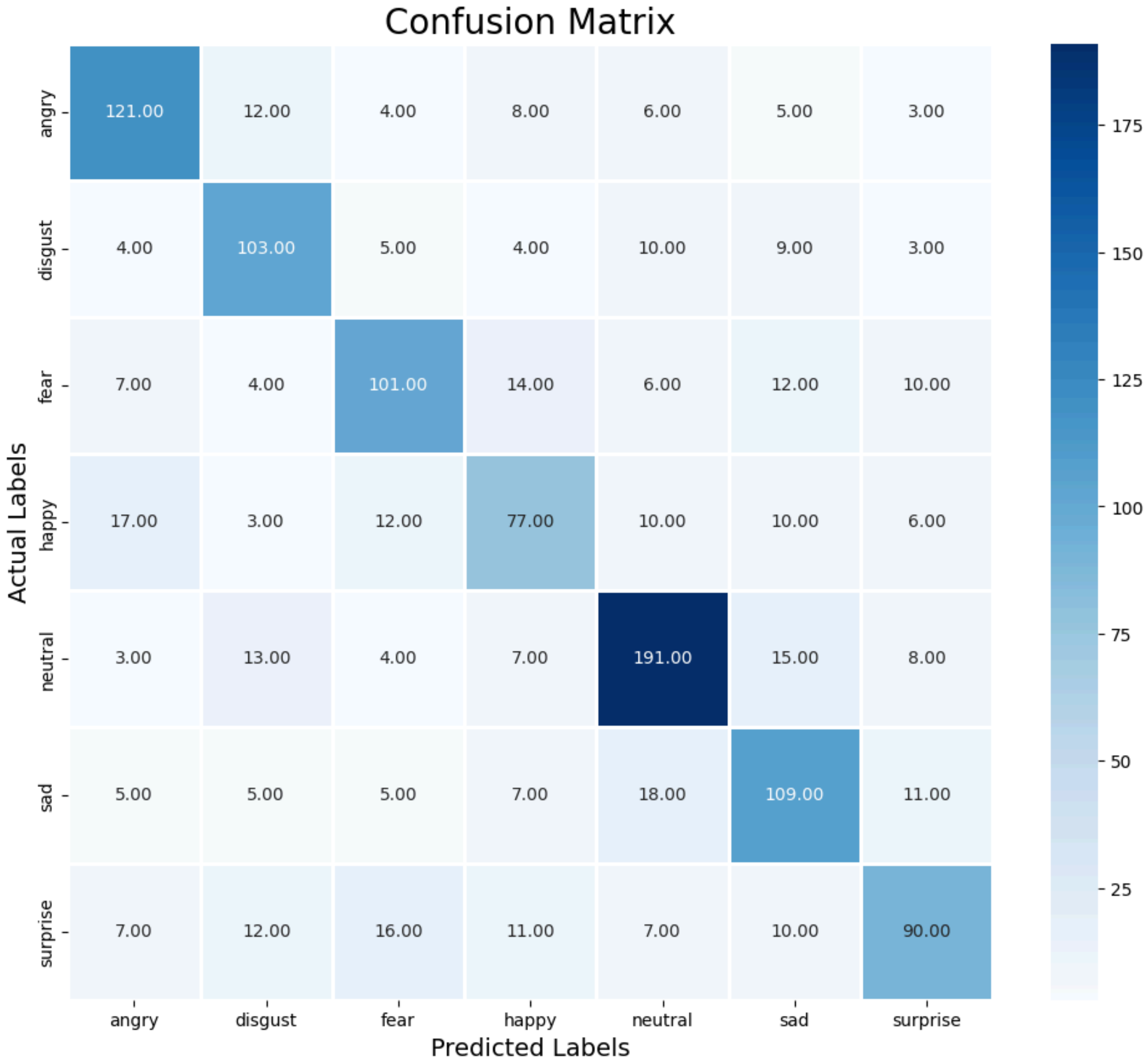


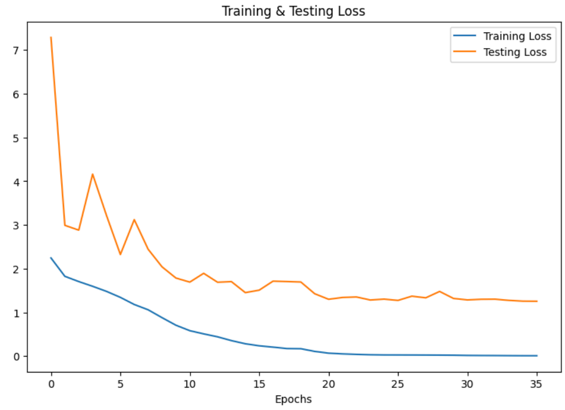


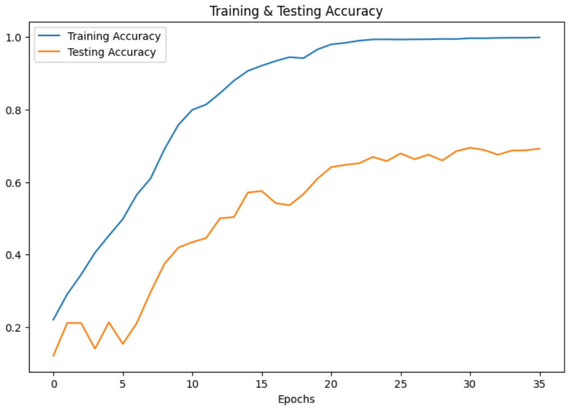
We can see from the confusion matrix that it is now able to predict "neutral", which seems to be the hardest emotion to predict since all the models struggle to predict it, sometimes (16% of the time). We can also see that emotions like “angry” and “fearful” are now getting predicted slightly better.

These models were trained on MFCC, ZCR, RMSE and Chroma features. We also trained models by replacing Chroma with spectrogram features.

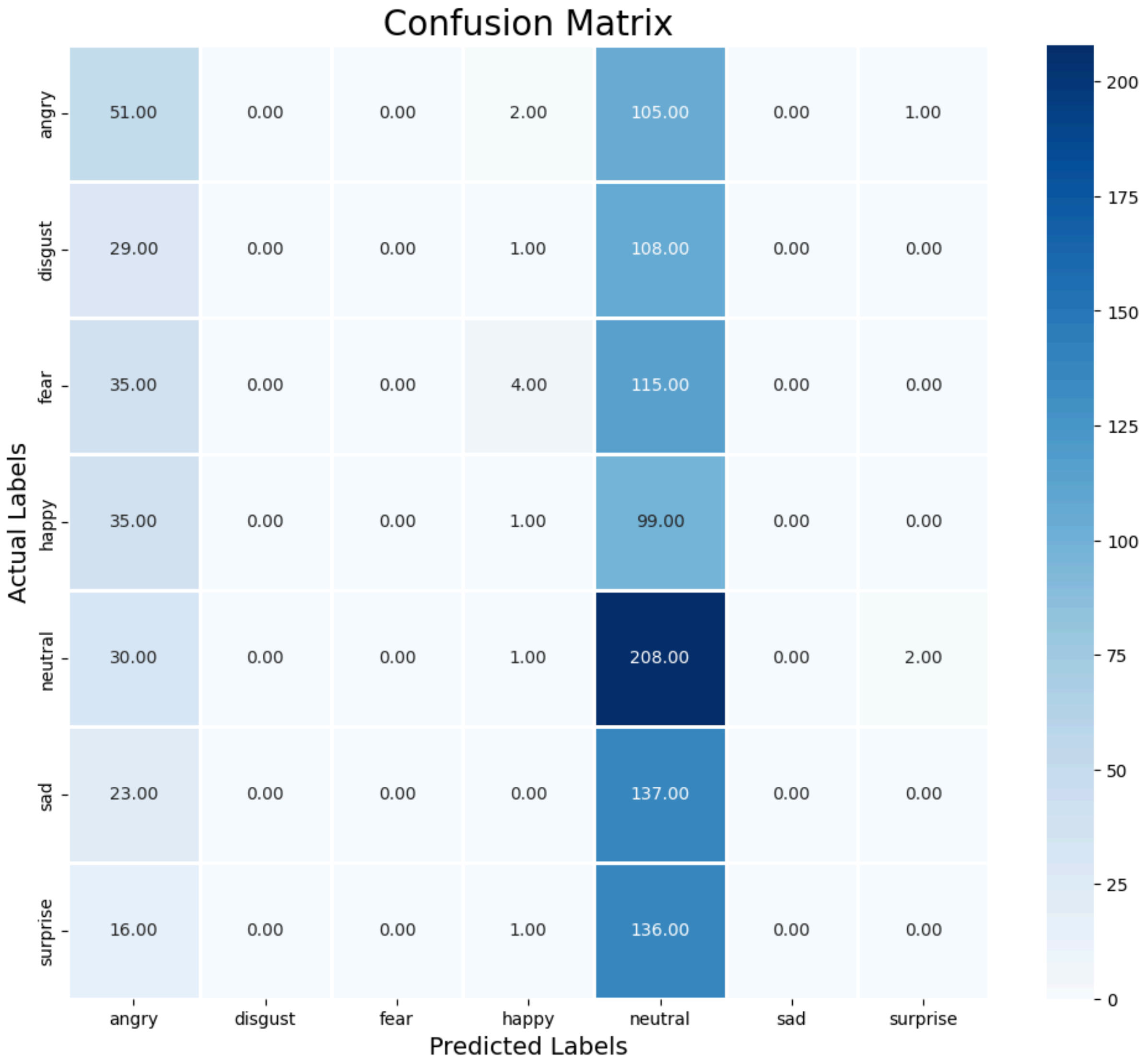
CNN



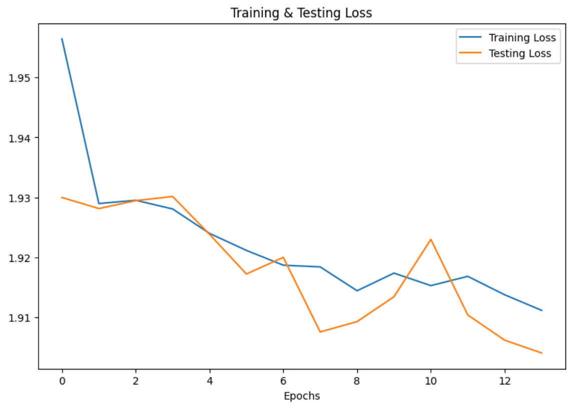

Training & Testing Loss
Training Loss
Testing Loss
0
1
2
3
4
5
6
7
0
5
10
15
20
25
30
35
Epochs

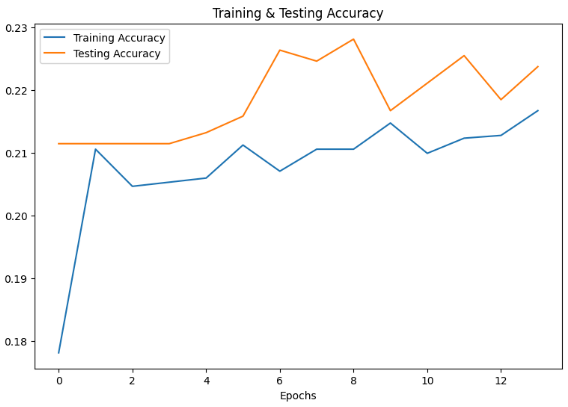

Training & Testing Accuracy
Training Accuracy
Testing Accuracy
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
Epochs

LSTM

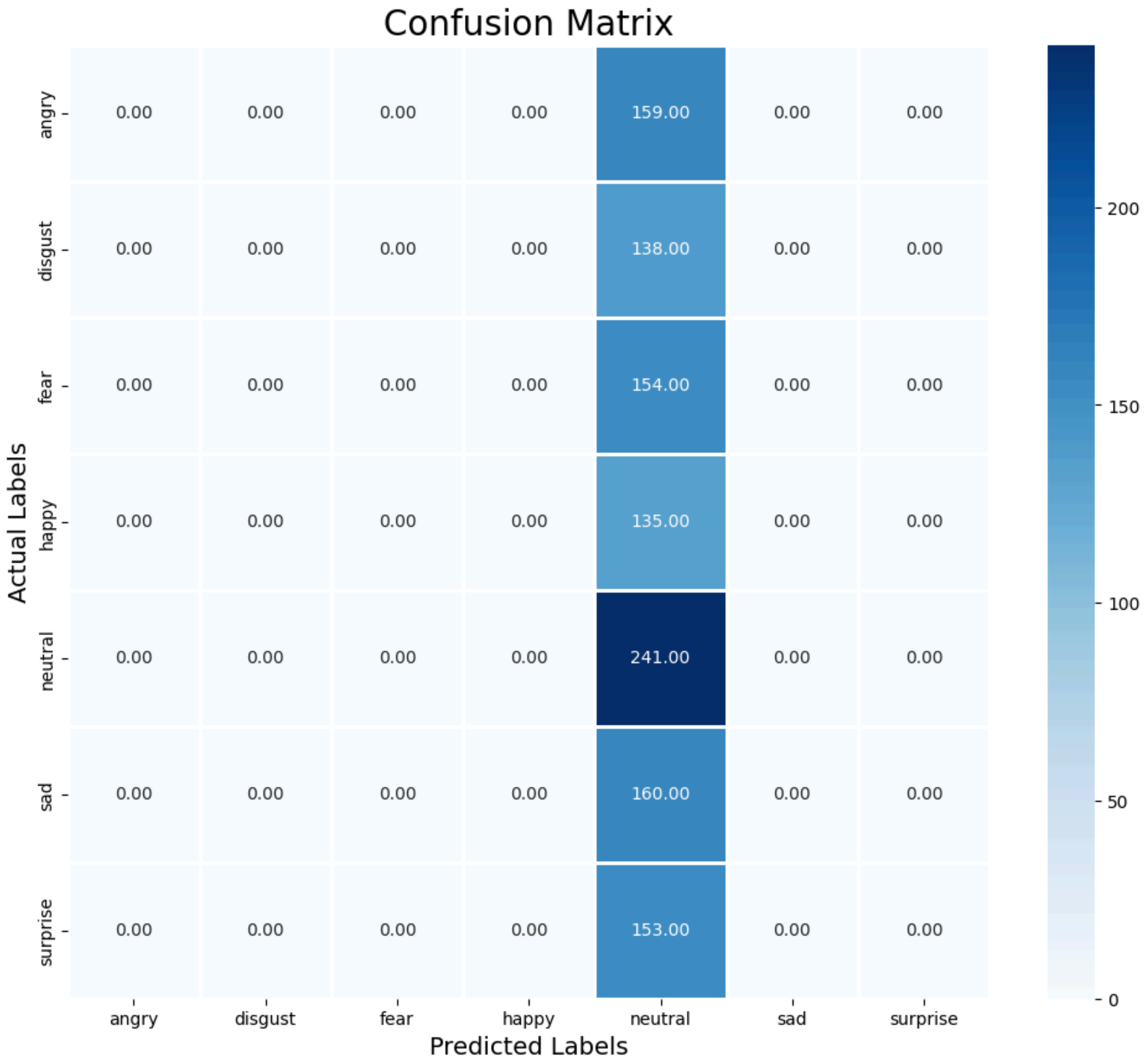


	angry	disgust	fear	happy	neutral	sad	surprise
angry	51.00	0.00	0.00	2.00	105.00	0.00	1.00
disgust	29.00	0.00	0.00	1.00	108.00	0.00	0.00
fear	35.00	0.00	0.00	4.00	115.00	0.00	0.00
happy	35.00	0.00	0.00	1.00	99.00	0.00	0.00
neutral	30.00	0.00	0.00	1.00	208.00	0.00	2.00
sad	23.00	0.00	0.00	0.00	137.00	0.00	0.00
surprise	16.00	0.00	0.00	1.00	136.00	0.00	0.00

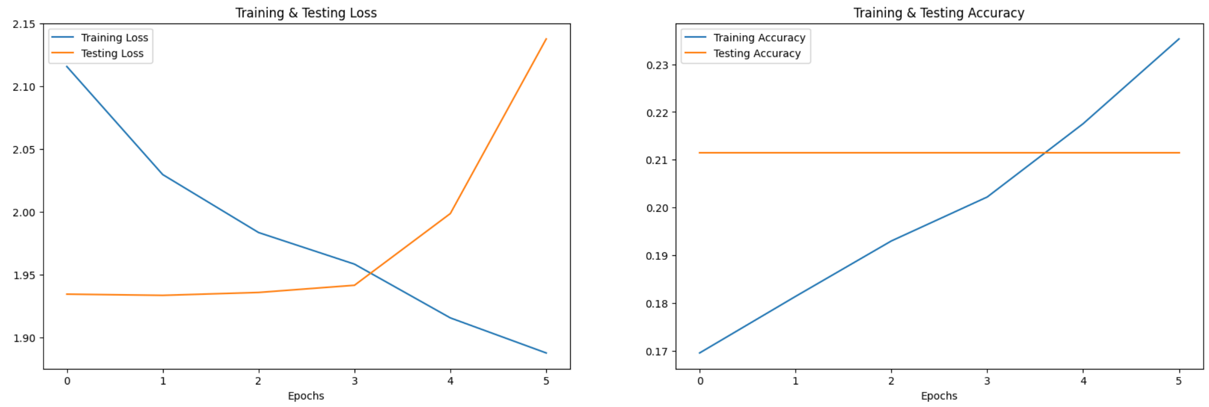




CNN-LSTM Hybrid

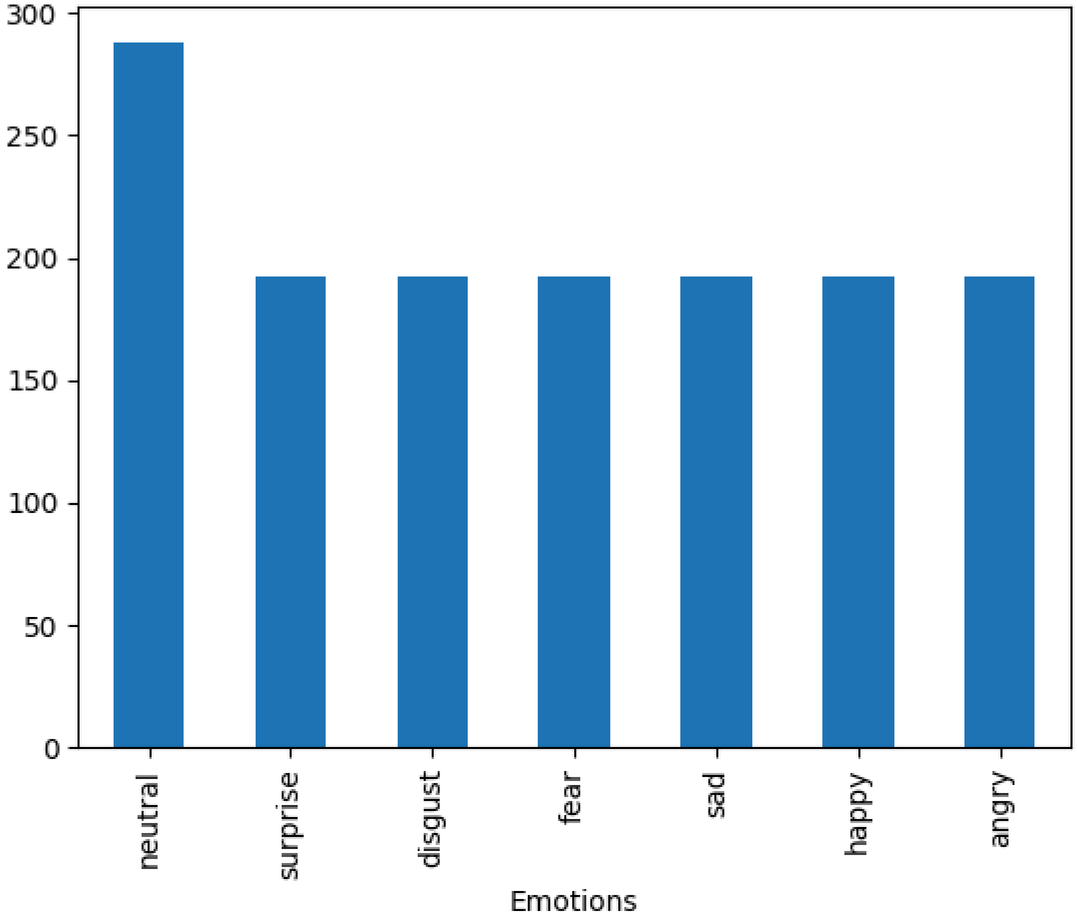


	angry	disgust	fear	happy	neutral	sad	surprise
angry	0.00	0.00	0.00	0.00	159.00	0.00	0.00
disgust	0.00	0.00	0.00	0.00	138.00	0.00	0.00
fear	0.00	0.00	0.00	0.00	154.00	0.00	0.00
happy	0.00	0.00	0.00	0.00	135.00	0.00	0.00
neutral	0.00	0.00	0.00	0.00	241.00	0.00	0.00
sad	0.00	0.00	0.00	0.00	160.00	0.00	0.00
surprise	0.00	0.00	0.00	0.00	153.00	0.00	0.00



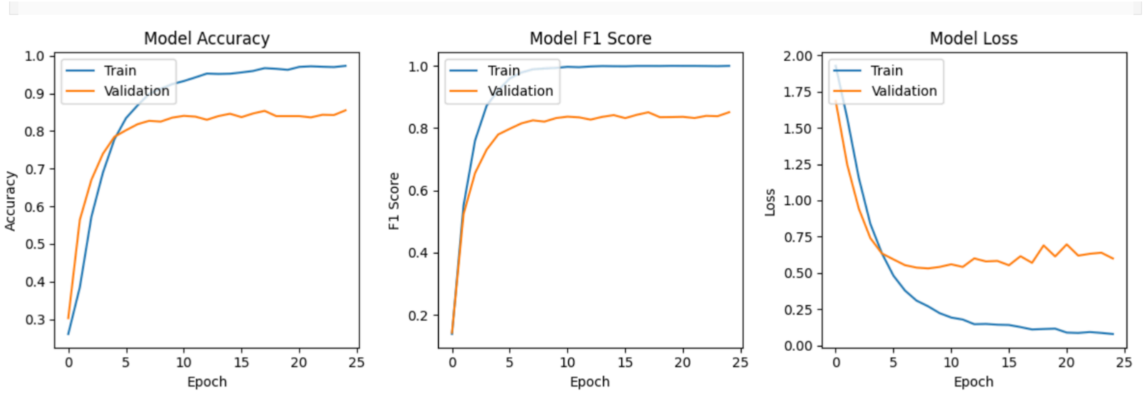
The CNN model had the highest accuracy of 69.47%

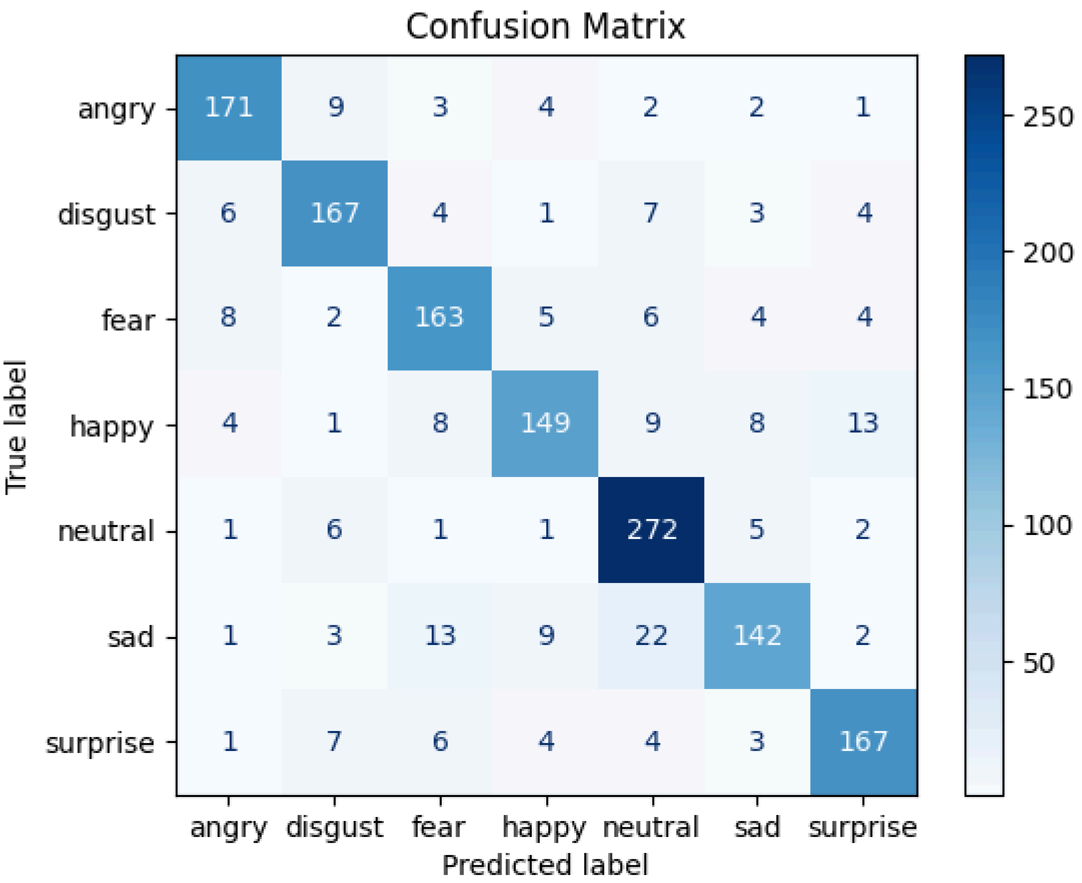
The LSTM and hybrid models suffered (22% and 21% accuracy respectively) to properly classify the test set. It can be observed that most of the predictions made by the LSTM and hybrid models are ‘neutral’. The RAVDESS dataset has 288 audio files labeled as ‘neutral’ and 192 files for each of the other emotions.



This imbalance could have been the factor that caused the LSTM models to underperform.

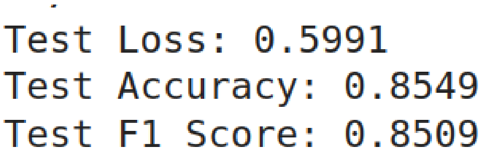
After multiple iterations of improving the model, we were able to train a model with a test accuracy of over 85%. This model used the CNN architecture that combines the features extracted from both 1D and 2D.





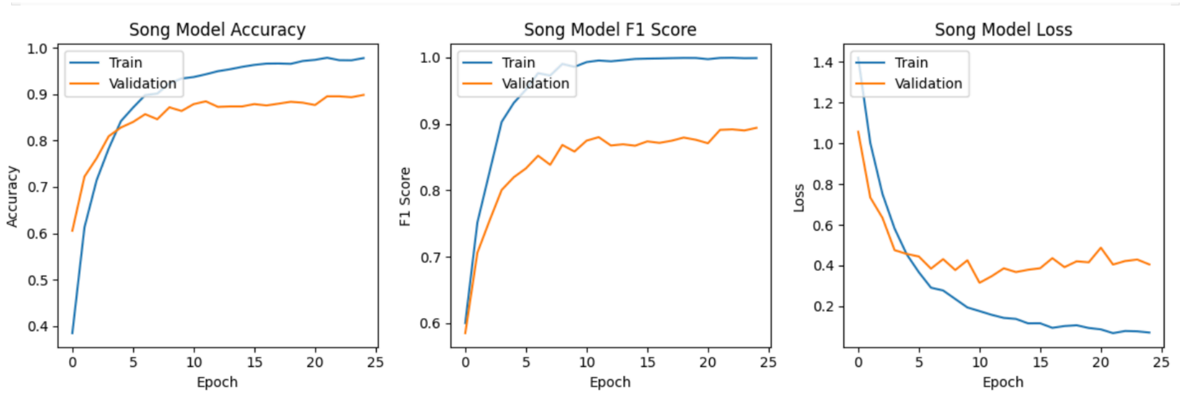
Confusion Matrix

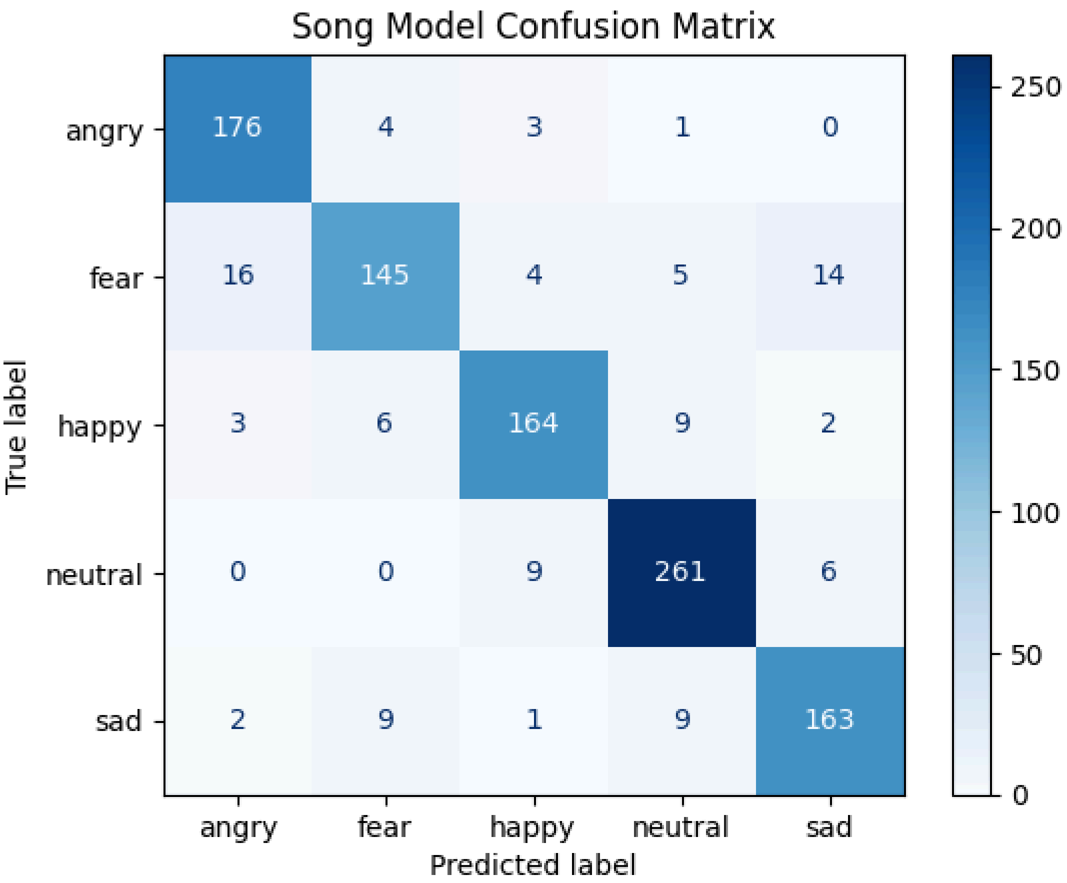
True label \ Predicted label	angry	disgust	fear	happy	neutral	sad	surprise
angry	171	9	3	4	2	2	1
disgust	6	167	4	1	7	3	4
fear	8	2	163	5	6	4	4
happy	4	1	8	149	9	8	13
neutral	1	6	1	1	272	5	2
sad	1	3	13	9	22	142	2
surprise	1	7	6	4	4	3	167

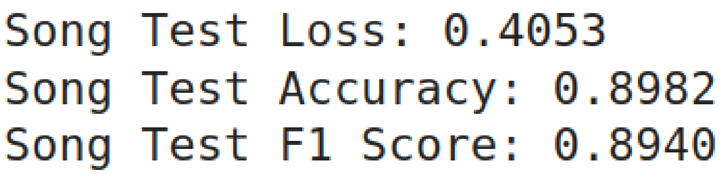


```
Test Loss: 0.5991
Test Accuracy: 0.8549
Test F1 Score: 0.8509
```

We experimented with the song dataset as well where the dialogue is the same, but the actors “sing” the dialogue.







```
Song Test Loss: 0.4053
Song Test Accuracy: 0.8982
Song Test F1 Score: 0.8940
```

Conclusion

The objective was to train and evaluate different models with different features extracted and see how each model is able to predict emotions. Through an iterative process of feature extraction and data augmentation, we were able to improve the performance of each of the models, with the multi input CNN model which used the zcr and rmse (1-d) and mfcc, spectogram and chroma (2-d) features having the highest accuracy (> 85%). We have learnt about what each feature extracted is and how augmenting the data can change the way features are extracted from these raw audio files. We have learnt about the process of working with audio data. We have learnt that model development is an iterative process, at each step we look at metrics to try to improve the model by making changes to which features are extracted and how data needs to be augmented for better feature extraction for different types of models and model training.