EEE 443 Final Project Report

Group 22 Members:

- Alperen KALYONCU

Chosen Project:

- Suggested Project 11: Emotional Speech Classification
- Dataset: Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

Abstract

Speech Emotion Recognition (SER) tries to identify emotions from speech by analyzing audio signals produced by human communication. In this study, emotional speech classification is done using the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) and Mel-Frequency Cepstral Coefficients (MFCCs) as input features. Three deep learning architectures: Bidirectional Long Short-Term Memory (BiLSTM), Transformer, and one-dimensional Residual Network (1D-ResNet) are trained and evaluated to compare effectiveness of different models. Data splitting, augmentation, and hyperparameter tuning are applied for better evaluation. Model performance evaluated using accuracy, emotion-wise and prediction distributions. ResNet model achieved the strongest overall performance, and better classification on higher energy labels. The BiLSTM model demonstrates balanced and consistent performance, whereas the Transformer model struggles notably with neutral emotion recognition. Overall, the findings highlight the impact of architectural choices on speech emotion recognition performance and support existing observations.

Introduction

Overview of Existing Literature

Speech Emotion Recognition (SER) is a research area that tries to identify emotions trough audio signals. Emotional states can change the way people speak during communication, causing noticeable changes in speech, which creates a possibility to guess and recognize the emotions from audios. These emotions can be extracted from speech by analyzing various patterns in audio signals. Early SER studies mostly dealt with features such as pitch, intensity, and intonation using traditional machine learning techniques to find these patterns.

More recent research has shifted toward spectral representations, particularly Mel-Frequency Cepstral Coefficients (MFCCs), which is a more human way to analyze the given audio signals. Using these features became standard for SER systems due to their ability to capture information close to human perception. With this shift in feature representation, new

approaches have utilized neural networks capable of learning these patterns directly from data.

Recurrent neural networks, particularly Long Short-Term Memory (LSTM) models, have been widely used because of the sequential nature of speech. More recently, Transformer-based models have attracted attention by using their attention mechanisms that allow the model to focus on important parts of an audio instead of processing the signal strictly step by step. In addition, convolutional architectures such as Residual Networks (ResNets) have been applied to SER tasks by learning hierarchical representations from audio features, while skip connections help stabilize training and improve learning in deeper models.

Purpose and Importance of this Work

Based on recent findings in deep learning–based SER, the primary objective of this study is to classify emotional states from speech signals using CREMA-D dataset as training data, and aims to compare the performance of BiLSTM, Transformer, and ResNet architectures in recognizing six major emotions: anger, disgust, fear, happiness, sadness and neutral.

This comparison can help to better understand how different perceptual designs influence speech emotion recognition. Emotional information in speech is distributed in complex ways, making it hard to determine which architectural principles are most effective for finding patterns. Sequential models use the sequential evolution of speech, attention-based models selectively focus on important segments, while convolutional architectures learn hierarchical representations that resemble aspects of human processing. By examining these approaches, this study explores whether architectures inspired by human perception offer advantages, or whether alternative modeling strategies can capture emotional patterns more effectively and potentially exceed human-like perception.

Methodology

This study uses an experiment-based method for speech emotion recognition. Audio features (MFCC) are extracted from speech signals, and neural network architectures are employed to model emotional patterns in the data. Each model is trained and evaluated on the same dataset using standard classification metrics. Through this approach, the study investigates how different architectural designs influence emotion recognition performance.

Expected Outcomes

It is expected that all models will better recognize dominant emotions such as anger and happiness, since these emotions have strong and more easily distinguishable patterns. Weaker emotions including neutral and sadness will be harder to recognize as they don’t have distinct patterns and the way they are expressed tends to vary more amongst people.

ResNet models are expected to show close to human perception and the best performance, since their layered processing captures hierarchical patterns that generalize well across

speakers. The way ResNet handles data resembles human perception and allows ResNets to learn the emotional patterns. As a result, ResNet models are expected to perform better on higher energy emotions like anger, while keeping stable recall rates on others

Transformer models, which can pick the important parts of the speech, are not expected to exceed the performance of the recurrent networks as the Crema-D dataset is small for Transformer architecture. Though attention mechanism is expected to better filter out the high energy emotions better than the other models and fall behind with the emotions like neutral as they don't possess such distinctive patterns.

BiLSTM models are expected to show moderate performance. While BiLSTMs effectively process temporal information in the audio signal, they lack an explicit selection mechanism to prioritize some patterns. But are still expected to provide stable performance across emotion classes.

Overall, this study aims to find whether architectures with mechanisms like attention are more effective for speech emotion recognition than purely sequential models. The analysis is expected to reveal the specific emotions that are hard to distinguish and models that are better for SER.

Methods

Dataset Description

CREMA-D Dataset [1]

This study uses the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), a widely used benchmark dataset for speech emotion recognition. CREMA-D consists of 7,442 audio clips collected from 91 professional actors (48 male and 43 female) aged between 20 and 74, representing diverse racial and ethnic backgrounds. Each actor recorded 12 distinct sentences, which were performed under six emotional categories: anger, disgust, fear, happiness, neutral, and sadness (see Appendix I for the emotional distribution).

A key advantage of the CREMA-D dataset is its relatively large number of speakers compared to many speech emotion datasets. Datasets with limited speaker diversity often show optimistic performance estimates. The diversity of speakers in CREMA-D helps to ease this issue and improves generalization across unseen speakers, making it a suitable choice for evaluating robust emotion recognition models.

The dataset was divided into ten subsets with each subset having different actors to avoid leakage. Eight were used for training, one was used for validation, and one subset was for testing. Data augmentation was applied to the training sets, while the validation and test sets remained with only the original recordings.

Feature Extraction and Data Augmentation

To represent the speech signals in a compact and perceptually meaningful form, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted using a fixed frame length and hop size and mean–variance normalization was applied [2]. Variable-length sequences were either padded or cut to ensure consistent input dimensions across samples.” Each signal was transformed into a 40-dimensional MFCC feature representation, which captures the spectral characteristics of speech while approximating human perception.

To increase data diversity and improve model robustness. For each original audio sample, two augmented versions were generated. The augmentations involved modifying the audio speed, shifting, and adding background noise. These augmentations were chosen to simulate variations in speaking style and recording conditions, expanding the training distribution and reducing the risk of overfitting.

Model Architectures

Three neural network architectures were chosen to model emotional patterns in speech:

- Bidirectional Long Short-Term Memory (BiLSTM): Chosen for its ability to capture sequential information in both forward and backward directions [3].
- Transformer: Chosen for its attention mechanism, which the model uses to find the relevant parts of the audio and capture complex patterns without depending only on sequential processing [4].
- Residual Network (1D-ResNet): Chosen to evaluate the ability of convolutional architectures to find emotional patterns from audio by examining hierarchical representations [5].

These models were selected to compare three different modeling paradigms for emotion recognition under the same conditions.

Hyperparameter Tuning

Hyperparameters are tuned based on validation accuracy using a two-stage search strategy designed to find the best configurations:

- Folding: At first the Training set was divided into 4 folds with each fold having different actors for simulating the regions with different speech habits.

- Random Search: A random search on the hyperparameter space was done to explore with eight different configurations. To optimize the efficiency of the tuning, each configuration was evaluated using 3 epochs each for 2 of the folds from the 4-fold cross-validation setup. This allowed for faster identification of high-potential regions in the search space.
- Grid Search: Following the exploration, Grid Search was performed on the three most important parameters for model architecture. In this stage up to 27 combinations were evaluated through 4 epochs on 3 of the 4 folds. By increasing the number of folds and epochs in this stage, we ensured a better estimation of performance for the candidates.

Model Training and Evaluation Procedure

After identifying the best hyperparameter configurations for each architecture, the models were fully trained using the selected configurations. Training was performed until validation performance converged.

Models were trained using a categorical cross-entropy loss function and optimized using the Adam optimizer. Performance was evaluated using classification accuracy and per class accuracy. All models were evaluated on the same test set to ensure fair comparison.

Results

1. Hyperparameter Tuning and Model Selection

The first stage of the experiment involved identifying the optimal configurations for each architecture using a two-stage search strategy. Table B 1, Table B 2, and Table B 3 (see Appendix B) summarize the full results of the random and grid searches.

Figures 1–3 show the cross-validation accuracy of the top three hyperparameter configurations for the evaluated architectures. The results of a 3-fold cross-validation, where per-fold accuracies are shown to evaluate both average performance and consistency across different speakers. Results are used to select hyperparameters which are both stable and accurate.

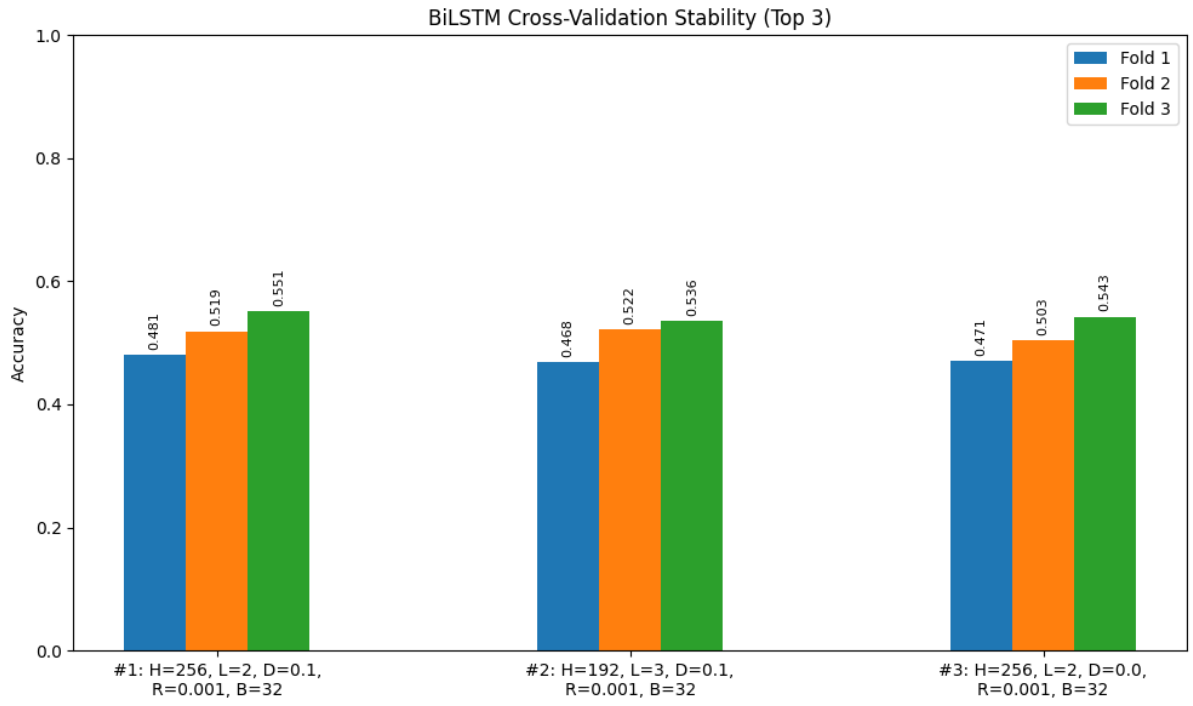

Figure 1: BiLSTM Cross-Validation Accuracy of Top Three Hyperparameter Configurations

3-fold cross-validation accuracy for the three best BiLSTM hyperparameter configurations. The chosen configuration was #1 since it is the most accurate and stable one among the three configurations.

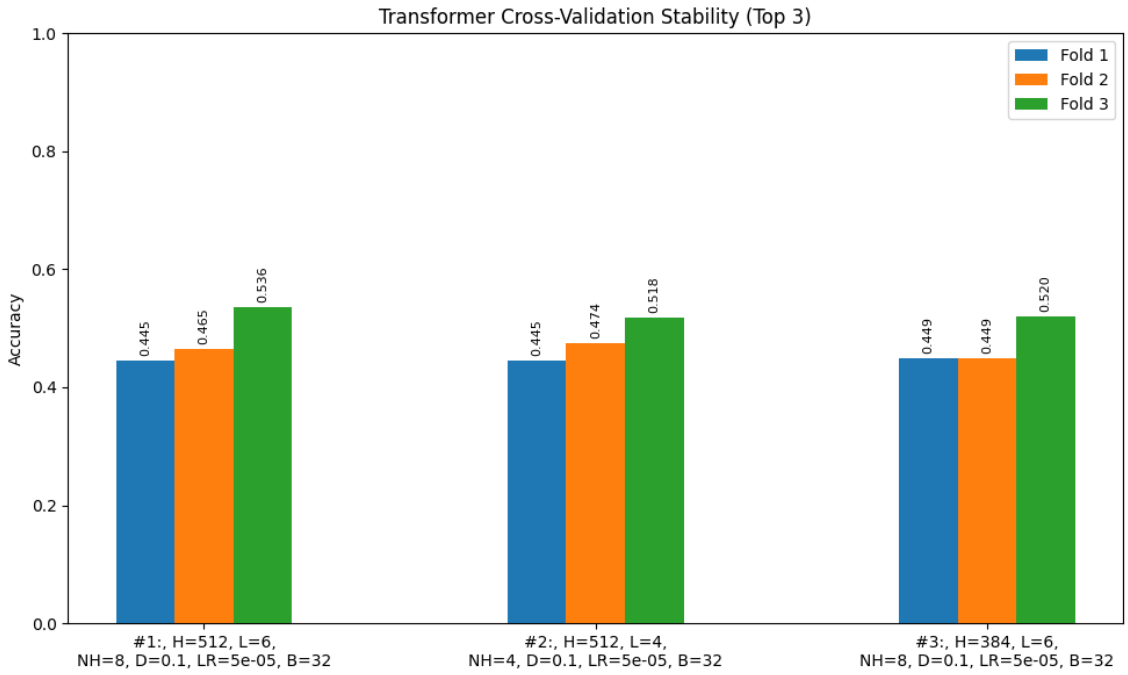

Figure 2: Transformer Cross-Validation Accuracy of Top Three Hyperparameter Configurations

3-fold cross-validation accuracy for the top Transformer hyperparameter configurations. Even though #1 had the highest average accuracy over the three folds, configuration #2 was chosen because of its better consistency over different folds.

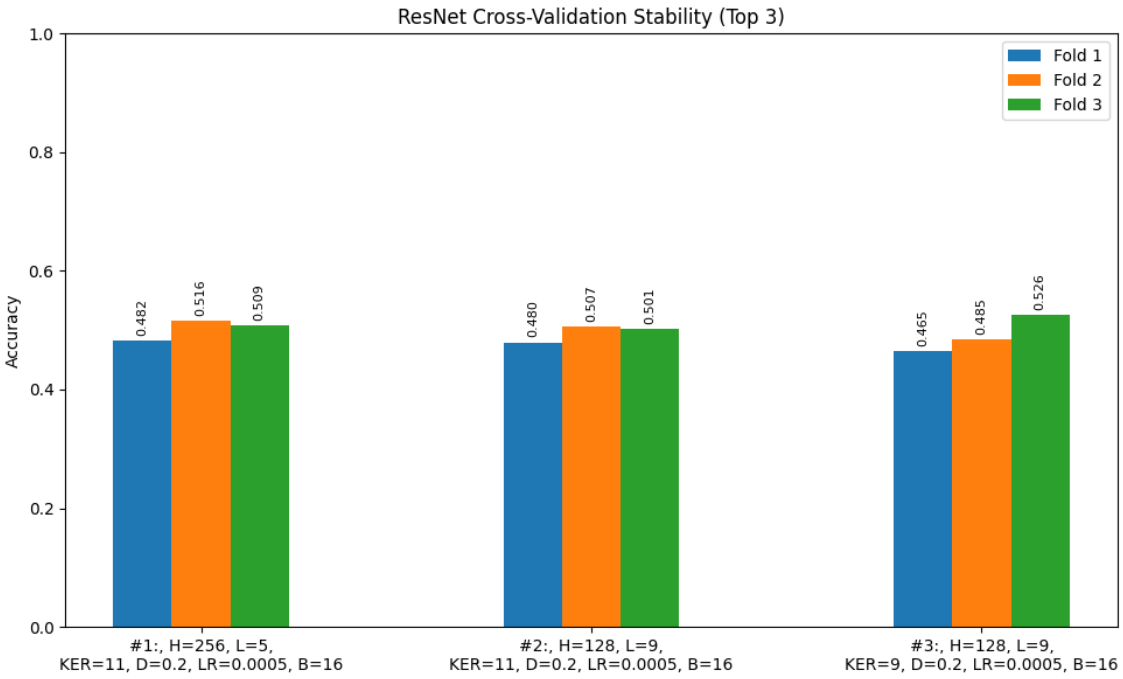

Figure 3: ResNet Cross-Validation Accuracy of Top Three Hyperparameter Configurations

3-fold cross-validation accuracy for the leading ResNet configurations. Even though #2 had smaller variance, 1 consistently achieved better results over all the folds.

2. Learning Curves

Figures 4–6 present the final training curves of the models with the selected hyperparameters.

Figure 4: Training and Validation curves of BiLSTM

This figure shows the training and validation loss and accuracy of the BiLSTM model across epochs. Validation performance improves until around epoch 7 and then the loss starts to increase with slight oscillations on the validation accuracy. Then after 5 epochs later the training ends with early stopping due to the lack of increase in performance.

Figure 5: Training and Validation curves of Transformer

This figure presents the training and validation loss and accuracy curves for the Transformer model. Validation metrics improve during early epochs and stabilize around epoch 4 and after a noticeable decrease in accuracy and increase in loss the training stops early at 11t epoch.

Figure 6: Training and Validation curves of ResNet

This figure illustrates the training and validation loss and accuracy of the ResNet model across epochs. Validation performance improves until epoch 4 and then starts to decline with stable accuracy but increasing loss.

After the Training is done the version with the least validation loss is selected for testing the models.

3. Test Guess Distributions

Figures 7–9 present the test-set guess distributions for the models. In each figure, horizontal axis shows the emotion guessed by the model, and the vertical axis indicates the total number of predictions for that emotion label. Each bar corresponds to the emotion label given by the model while each sub bar shows the actual emotion of the audio files with its corresponding color.

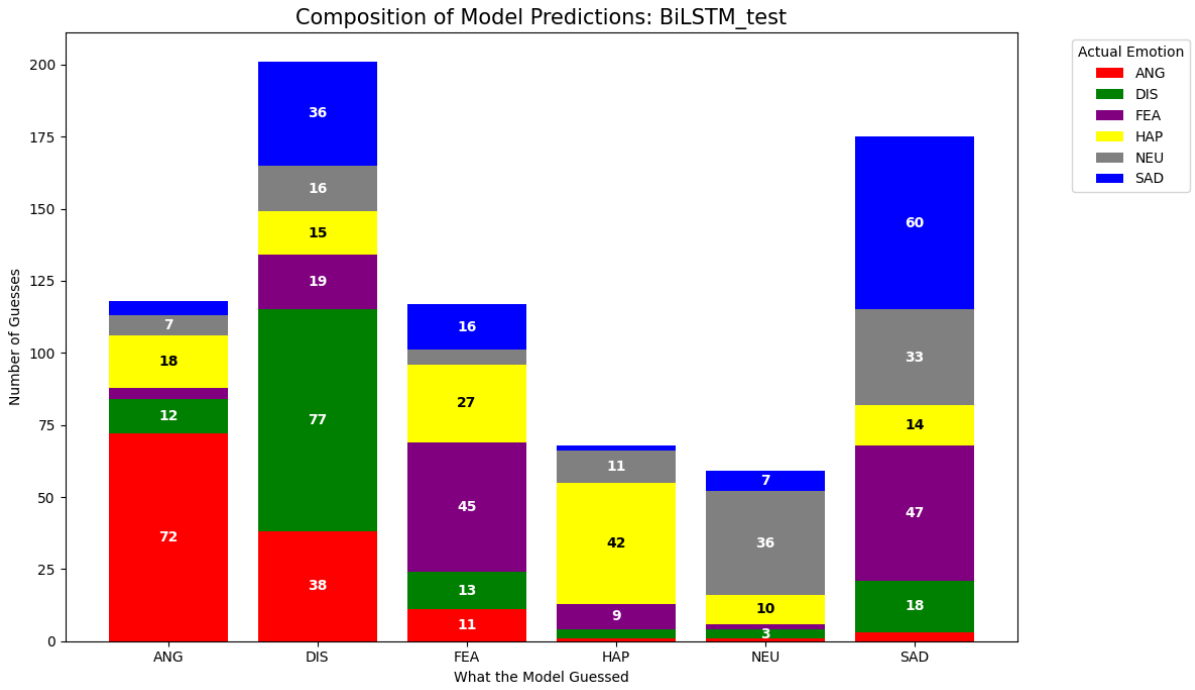

Figure7: Test guess breakdown for BiLSTM

This figure shows the distribution of BiLSTM predictions on the test set. Anger and Disgust are the most distinctive emotions, as they rarely get confused with other emotions. Fear and neutral on the other hand are less distinct and are continuously misclassified as sadness. When the model predicts happiness, the predictions are mostly correct, indicating high precision for this class. However, happiness is often under-predicted, as happiness audios often appear in other guesses.

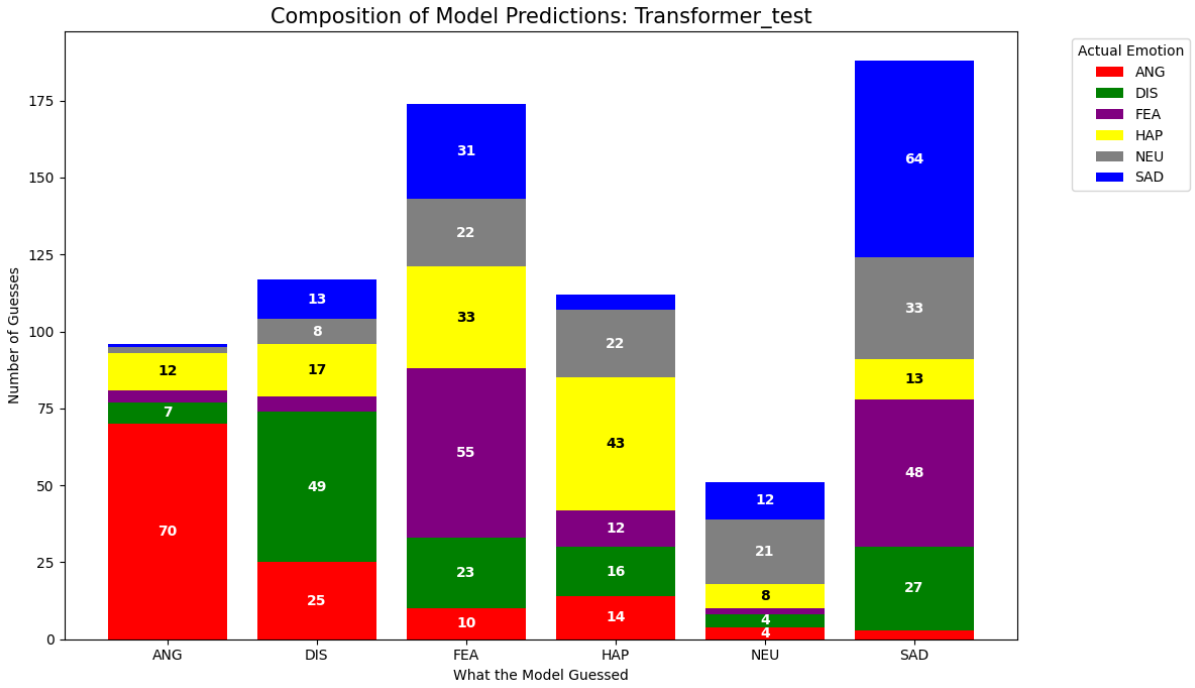

Figure 8: Test guess breakdown for Transformer

This figure presents the test-set prediction distribution of the Transformer model. Anger is the most distinctive emotion, showing a noticeable accuracy compared to other emotions, however neutral is commonly confused with other emotions. Like the BiLSTM model, sadness is confused with fear. For the Transformer, fear predictions are more frequent than disgust predictions unlike BiLSTM. Confusions are present across all emotion classes, except for neutral, most predictions remain correct.

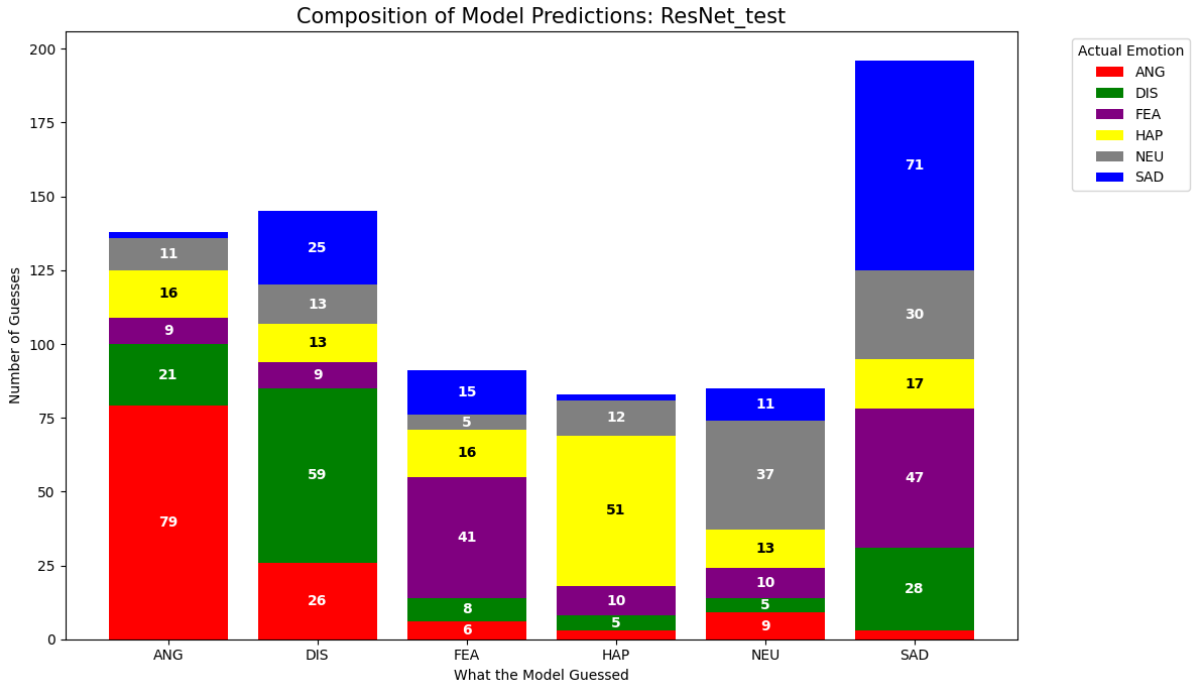

Figure 9: Test guess breakdown for ResNet

This figure illustrates the ResNet model's prediction distribution on the test set. Happiness accuracy is noticeably higher than the other two models though Anger remains the most distinctive emotion. And accuracy of neutral guesses is considerably higher.

Across all three models, anger is the most accurately predicted emotion, followed by disgust. In contrast, fear and sadness display confusion in all models frequently being misclassified as each other. Neutral speech is also commonly misclassified as sadness, indicating similar patterns between emotions.

Differences between models are also observed. BiLSTM predicts disgust more, while Transformer and ResNet more predict sadness, with Transformer also giving a high prediction count on fear. The Transformer shows more uniform confusion among the emotions while others confuse one or two specific emotions with each other.

4. Success Rates

Figures 10–12 present the emotion-wise success rates of the models on the test set. The bar charts show accurate prediction rates for each emotion class, alongside the mean recall and overall success rate. These two metrics differ because of the smaller number of neutral samples.

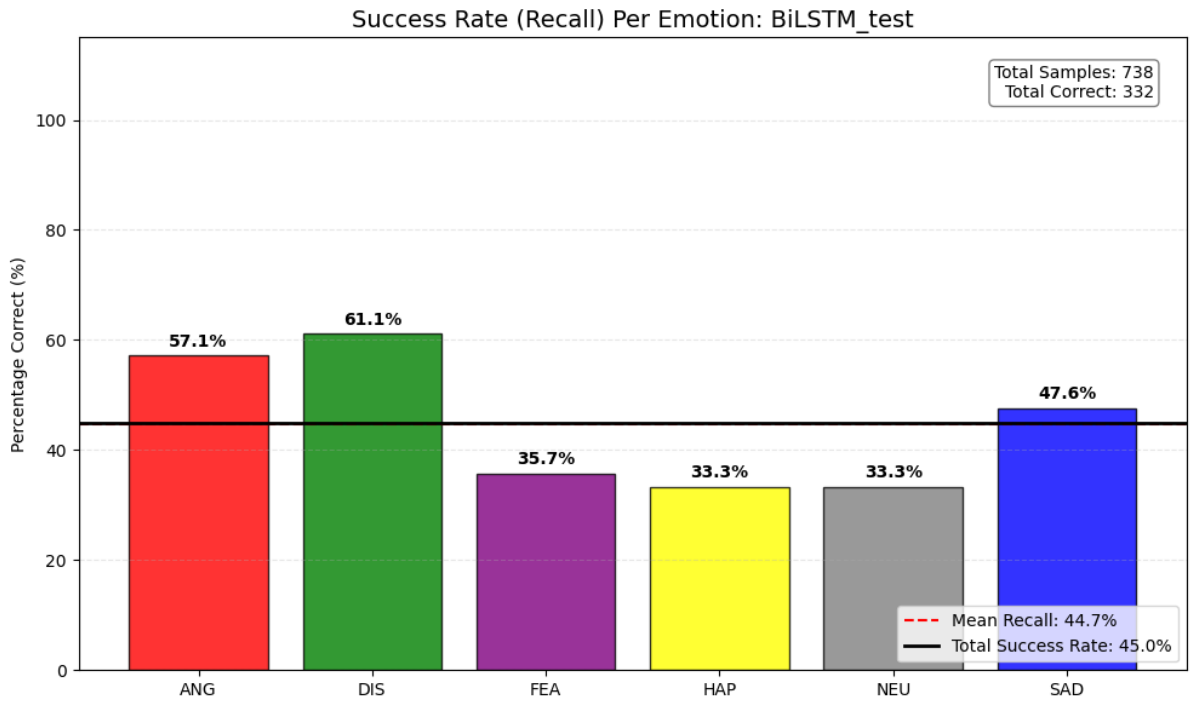

Figure 10: Test Set Success Rates for BiLSTM

The BiLSTM model achieves its highest recall on disgust, followed by anger and sadness. Performance on fear, happiness, and neutrality is lower. Overall a total success rate of 45%.

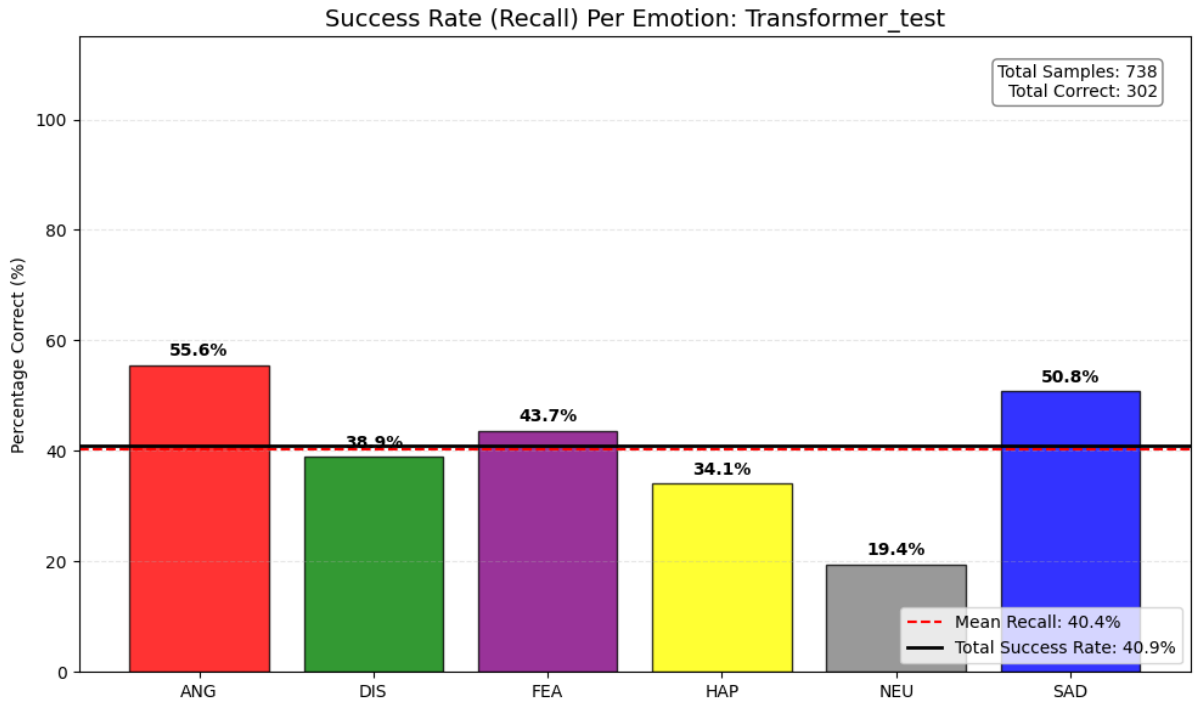

Figure 11: Test Set Success Rates for Transformer

The Transformer model performs best on anger, followed by sadness, where it exceeds the BiLSTM. It also shows better recall on fear, disgust, and happiness compared to the BiLSTM. However, neutral emotion recognition is noticeably weak, with recall dropping close to random performance. Overall Transformer shows the most uniform success rates apart from the noticeably weak neutral and achieves a mean recall of 40.4% and an overall success rate of 40.9%.

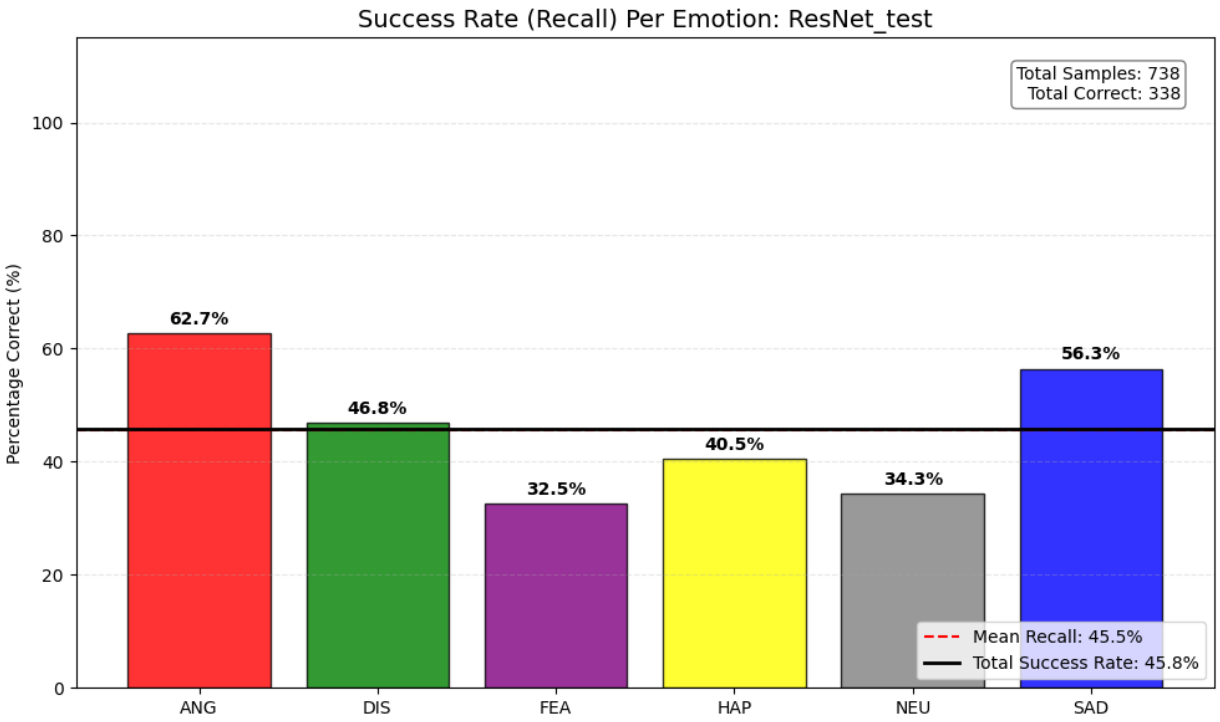

Figure 12: Test Set Success Rates for ResNet

The ResNet model yields the strongest overall performance. Achieving the best recall rates for anger, sadness, neutral and happiness. Neutral emotion recognition is stronger than fear, with fear being the weakest class for this model. ResNet achieves a mean recall of 45.5% and an overall success rate of 45.8%.

Discussion

Overall, the results align with expectations from previous studies in literature. Convolutional architecture is known to perform strongly on audio-based classification tasks. As expected, the 1D-ResNet achieved the best performance among the evaluated models.

The BiLSTM model produced more balanced and stable performance compared to the Transformer. This behavior is consistent for the size of the dataset and previous findings of the matter. But the unexpected result was Transformer performing worse on the high energy emotions specifically anger. A possible reason for this behavior is the size of the dataset. Transformers usually need a larger dataset size to perform better.

The Transformer model struggled to recall the neutral speeches as expected since neutral speech lacks strong patterns or cues, as it represents the lack of emotion. This makes neutral emotion particularly challenging for Transformer architecture. Thus, the Transformers performance on neutral classification is on par with the other studies in literature.

Confusions between certain emotions such as fear and sadness, were observed across all models. These confusions are consistent with prior research and consistent with human perception. Such confusion patterns suggest that the performance is not a model selection problem but possibly a deficiency in representation of the data.

Some aspects of the work may be improving. Expanding and changing the hyperparameter search space and using a bigger dataset will most probably present better and more accurate results, particularly for the Transformer model. Also using different number of dimensions while extracting MFCCs or using a different representation may substantially affect the results.

Overall, the goals stated were largely achieved. The models were successfully trained and evaluated, comparisons gave a better understanding of SER and further supported the existing literature, the expected dominance of ResNet architecture was observed. While some outcomes such as the BiLSTM outperforming the Transformer in anger recognition were unexpected, these results provide valuable insight into the interaction between model architecture and dataset size.

References

[1] https://www.kaggle.com/datasets/ejlok1/cremad?resource=download

[2] D. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980. https://ieeexplore.ieee.org/document/1163420

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. https://www.bioinf.jku.at/publications/older/2604.pdf

[4] A. Vaswani et al., “Attention is all you need,” *NeurIPS*, pp. 5998–6008, 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CVPR*, pp. 770–778, 2016. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Appendices:

Appendix A: Hyperparameter Spaces

Table A 1: Hyperparameter Search Space for BiLSTM Training

Hidden Size	Layer Count	Dropout	Learning Rate	Batch Size
32	1	0.0	3×10^{-4}	32
64	2	0.1	5×10^{-4}	64
128	3	0.2	1×10^{-3}	
192	4	0.4		
256				

Table A 2: Hyperparameter Search Space for Transformer Training

Hidden Size	Layer Count	Head Number	Dropout	Learning Rate	Batch Size
128	2	4	0.1	5×10^{-5}	32
256	4	8	0.2	1×10^{-4}	64
384	6	16	0.3	3×10^{-4}	
512	8				

Table A 3: Hyperparameter Search Space for ResNet Training

Hidden Size	Layer Count	Head Number	Dropout	Learning Rate	Batch Size
64	3	3	0.2	5×10^{-4}	16
128	5	5	0.3	1×10^{-3}	32
256	7	7	0.4		
384	9	9			
512	11	11			

Appendix B: Tuning Tables

Table B 1: Hyperparameter tuning summary of BiLSTM

BiLSTM Full Hyperparameter Summary

Hidden	Layers	Drop	LR	Batch	Mean Acc
256	2	0.1	0.001	32	0.5169
192	3	0.1	0.001	32	0.5088
256	2	0.0	0.001	32	0.5057
192	2	0.1	0.001	32	0.5025
256	3	0.1	0.001	32	0.4986
128	3	0.0	0.001	32	0.4973
128	3	0.1	0.001	32	0.4964
256	3	0.0	0.001	32	0.4959
192	1	0.0	0.001	32	0.4926
128	2	0.1	0.001	32	0.4916
192	3	0.0	0.001	32	0.4883
256	1	0.1	0.001	32	0.4882
128	2	0.0	0.001	32	0.4808
192	1	0.1	0.001	32	0.4779
128	1	0.1	0.001	32	0.4713
128	1	0.0	0.001	32	0.4657
192	2	0.0	0.001	32	0.4556
192	2	0.4	0.0005	64	0.4360
256	1	0.0	0.001	32	0.4333
64	2	0.0	0.0005	32	0.4238
32	4	0.1	0.001	32	0.3919
32	3	0.2	0.001	64	0.3896
32	2	0.2	0.0005	32	0.3692
64	1	0.2	0.001	64	0.3652

Table B 2: Hyperparameter tuning summary of Transformer

Transformer Full Hyperparameter Summary

H	L	NH	D	LR	B	Mean Acc
512	6	8	0.1	5e-05	32	0.4819
512	4	4	0.1	5e-05	32	0.4792
384	6	8	0.1	5e-05	32	0.4724
512	4	8	0.1	5e-05	32	0.4718
512	6	4	0.1	5e-05	32	0.4715
384	2	8	0.1	5e-05	32	0.4645
384	6	4	0.1	5e-05	32	0.4621
256	4	4	0.1	5e-05	32	0.4587
256	2	8	0.1	5e-05	32	0.4584
384	4	8	0.1	5e-05	32	0.4575
256	6	8	0.1	5e-05	32	0.4566
384	2	4	0.1	5e-05	32	0.4551
512	2	4	0.1	5e-05	32	0.4537
512	2	8	0.1	5e-05	32	0.4523
256	4	8	0.1	5e-05	32	0.4519
256	2	4	0.1	5e-05	32	0.4478
384	4	4	0.1	5e-05	32	0.4390
256	6	4	0.1	5e-05	32	0.4386
256	4	4	0.1	0.0001	32	0.4346
384	8	8	0.1	5e-05	64	0.4316
256	6	16	0.2	0.0001	64	0.4306
512	6	16	0.3	0.0001	32	0.4275
384	8	16	0.2	5e-05	64	0.4214
256	2	8	0.2	5e-05	64	0.4096
128	2	16	0.1	0.0001	64	0.4075

Table B 3: Hyperparameter tuning summary of ResNet

ResNet Full Hyperparameter Summary

H	L	KER	D	LR	B	Mean Acc
256	5	11	0.2	0.0005	16	0.5020
128	9	11	0.2	0.0005	16	0.4959
128	9	9	0.2	0.0005	16	0.4921
384	7	9	0.2	0.0005	16	0.4914
128	7	9	0.2	0.0005	16	0.4912
256	5	9	0.2	0.0005	16	0.4910
256	7	9	0.2	0.0005	16	0.4907
128	7	11	0.2	0.0005	16	0.4898
128	5	9	0.2	0.0005	16	0.4898
384	9	11	0.2	0.0005	16	0.4833
384	5	11	0.2	0.0005	16	0.4833
256	9	11	0.2	0.0005	16	0.4831
256	7	11	0.2	0.0005	16	0.4827
384	7	11	0.2	0.0005	16	0.4824
384	9	9	0.2	0.0005	16	0.4774
256	9	9	0.2	0.0005	16	0.4758
128	5	11	0.2	0.0005	16	0.4752
384	11	3	0.4	0.001	16	0.4685
64	5	7	0.2	0.0005	32	0.4685
512	11	9	0.4	0.0005	32	0.4651
384	5	9	0.2	0.0005	16	0.4596
64	3	9	0.4	0.001	16	0.4593
384	9	9	0.3	0.001	32	0.4593
64	3	7	0.3	0.001	16	0.4590
256	9	3	0.4	0.001	32	0.4404

Appendix C: Training Guess Breakdowns

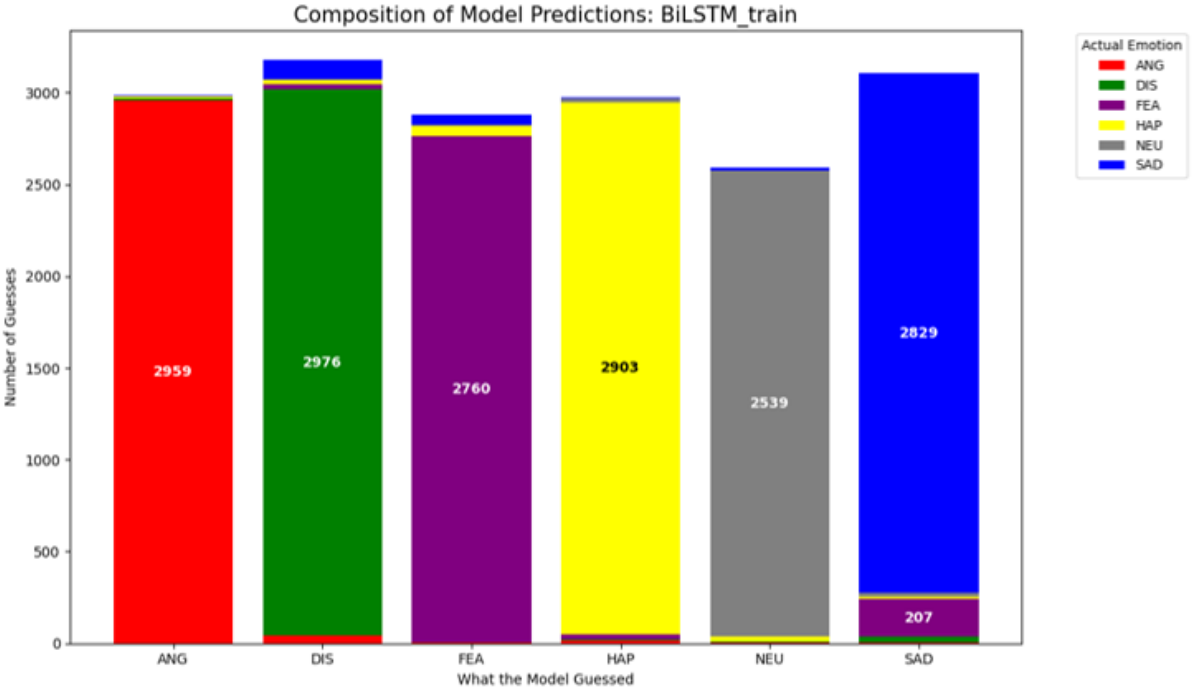

Figure C 1:Training guess breakdown for BiLSTM

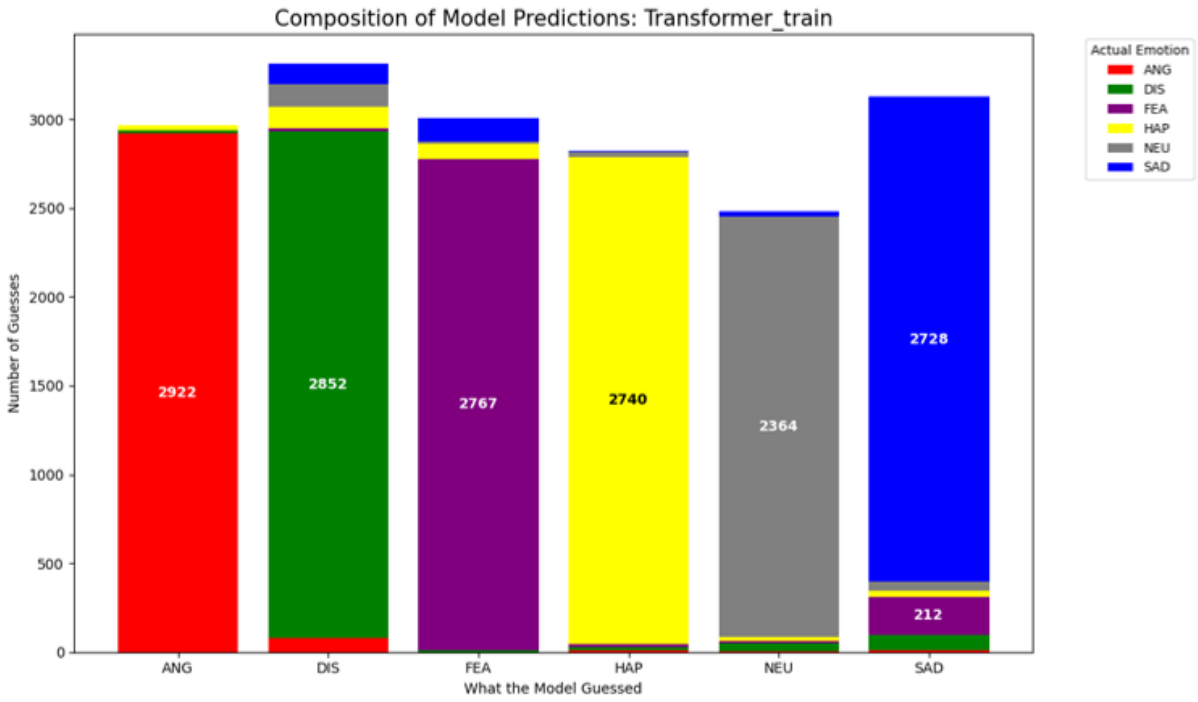

Figure C 2:Training guess breakdown for Transformer

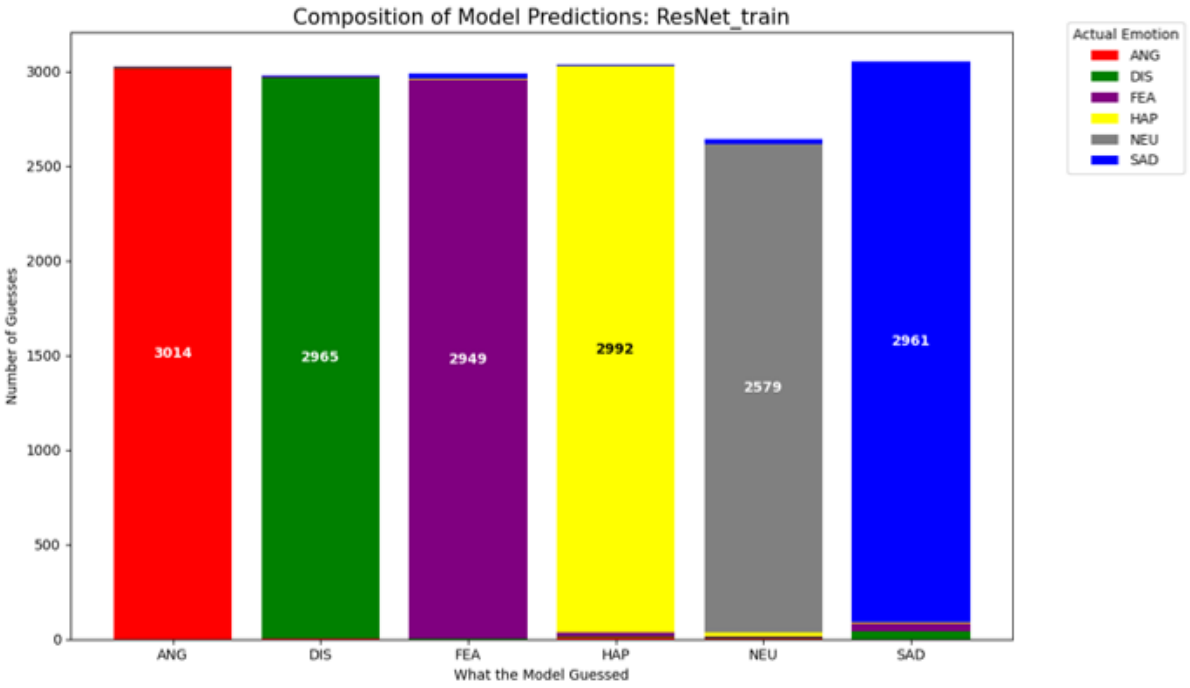

Figure C 3:Training guess breakdown for ResNet

Appendix D: Training Succes Rates

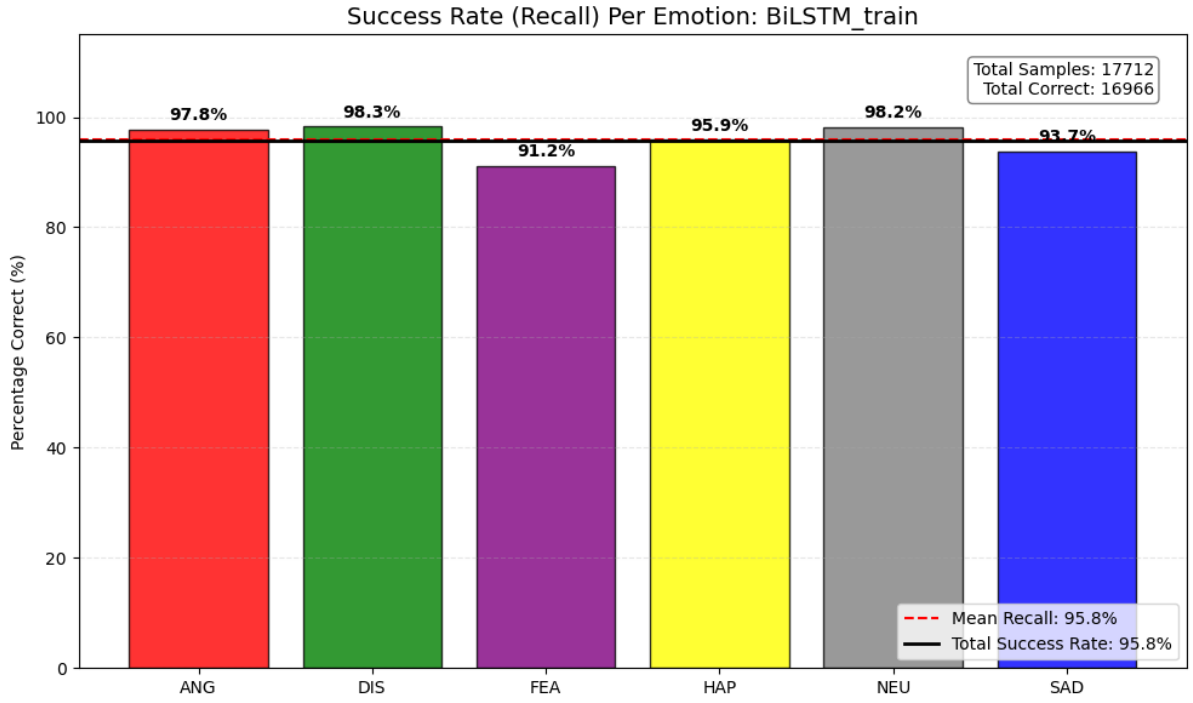

Figure D 1: Training Set Success Rates for BiLSTM

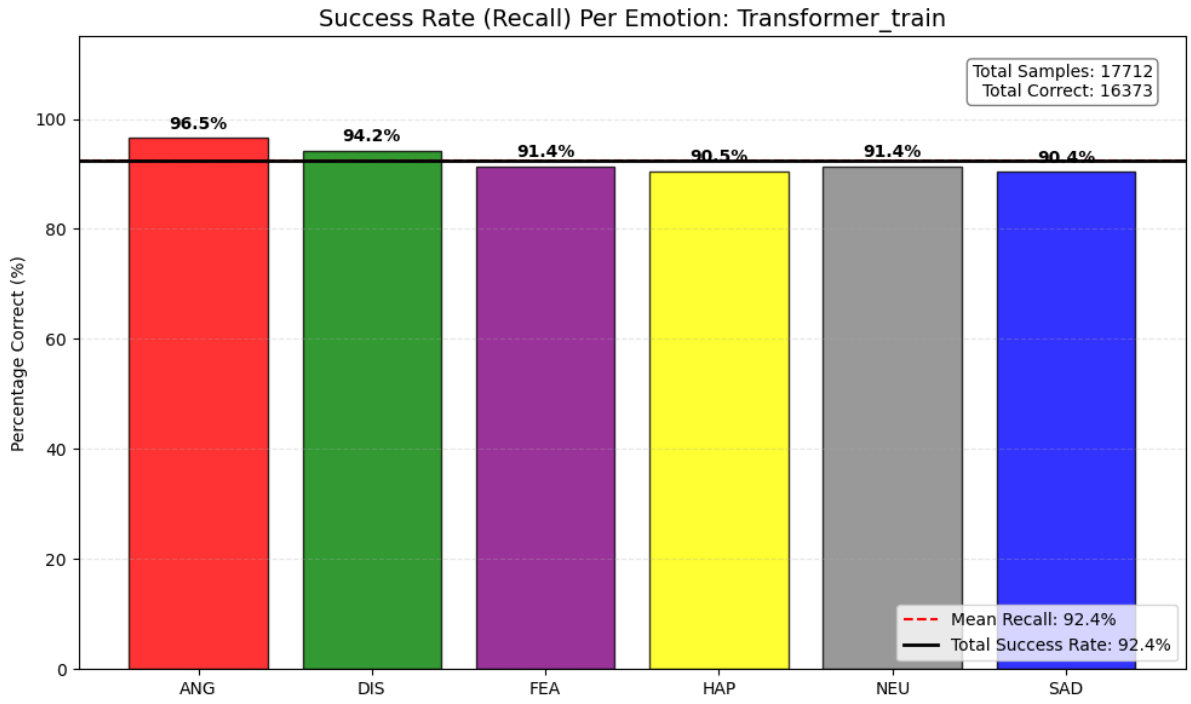

Figure D 2: Training Set Success Rates for Transformer

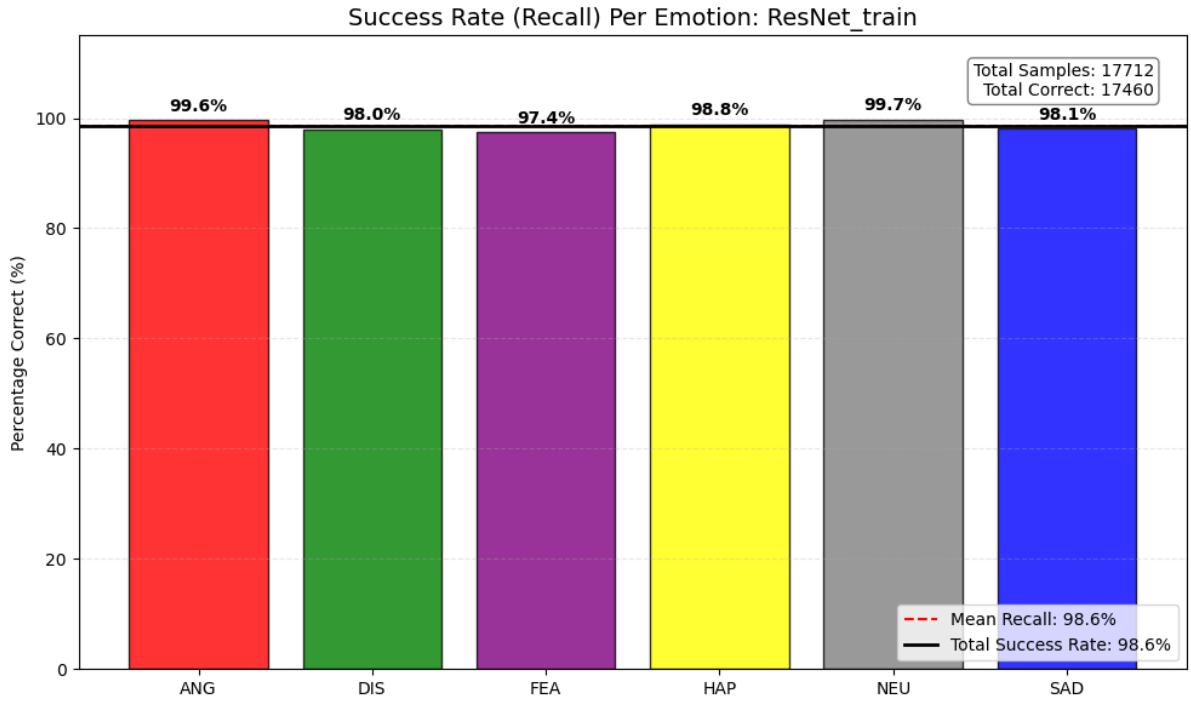

Figure D 3: Training Set Success Rates for ResNet

Appendix E: Dataset Distribution

Tabel E 1: Emotion Distribuion of Crema-D

Emotion	Count
Anger	1271
Disgust	1271
Fear	1271
Happiness	1271
Neutral	1087
Sad	1271