IAT481 Assignment 4: Speech Emotion Recognition DNN Analysis

Wei Xing Deng #301442155

Word count: 1334

School of Interactive Arts and Technology, Simon Fraser University

IAT 481W D100: Exploring Artificial Intelligence: Its Use, Concepts, and Impact

Professor: Dr. O. Nilay Yalcin

TA: Maryiam Zahoor

Feb 15, 2024,

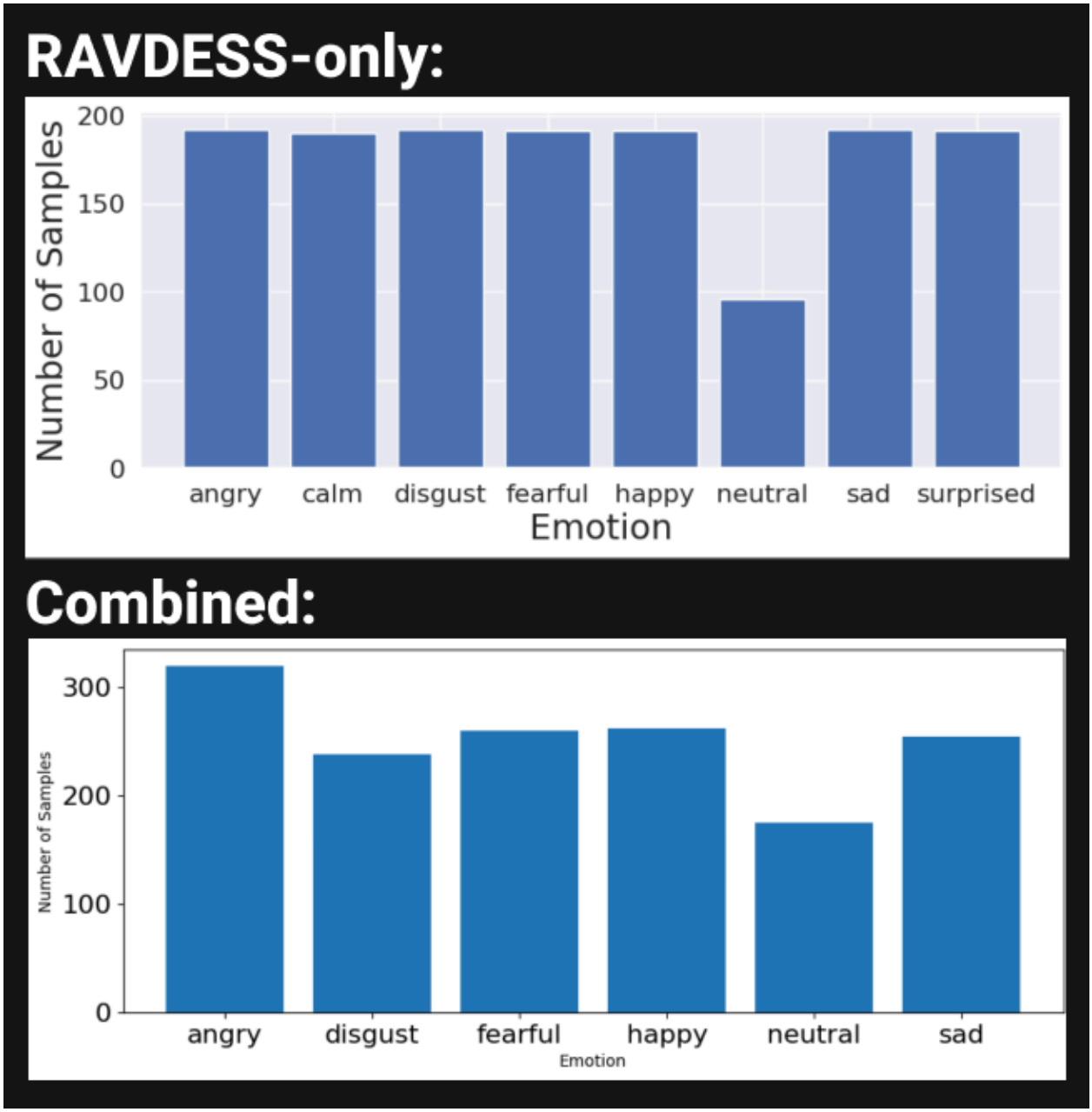
Introduction

In this report, I explored how different datasets influence the accuracy of Deep Neural Networks (DNNs) in recognizing emotions from speech. Specifically, I'm comparing the performance of DNNs trained on a combined dataset of EmoDB and RAVDESS against those trained solely on RAVDESS data. This comparison aims to uncover the benefits of dataset diversity on the model's ability to generalize across various emotional expressions. By analyzing these differences, I hope to contribute valuable insights into the optimal dataset composition for emotion recognition tasks in machine learning.

Comparing Models:

In my analysis of the combined dataset, which merges EmoDB with RAVDESS, I initially assessed the dataset balance, feature scaling, and implemented an 80/20 split for validation. I observed a disparity in data balance, notably with 'angry' emotions being more prevalent and 'disgust' less so, due to the integration with the EmoDB dataset (Figure 1).

Figure.1



Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Upon running the MLP initialization with sklearn's default settings for weight initialization—a method not highly recommended according to the tutorials—I found that the combined dataset's accuracy differed from that of the RAVDESS-only dataset.

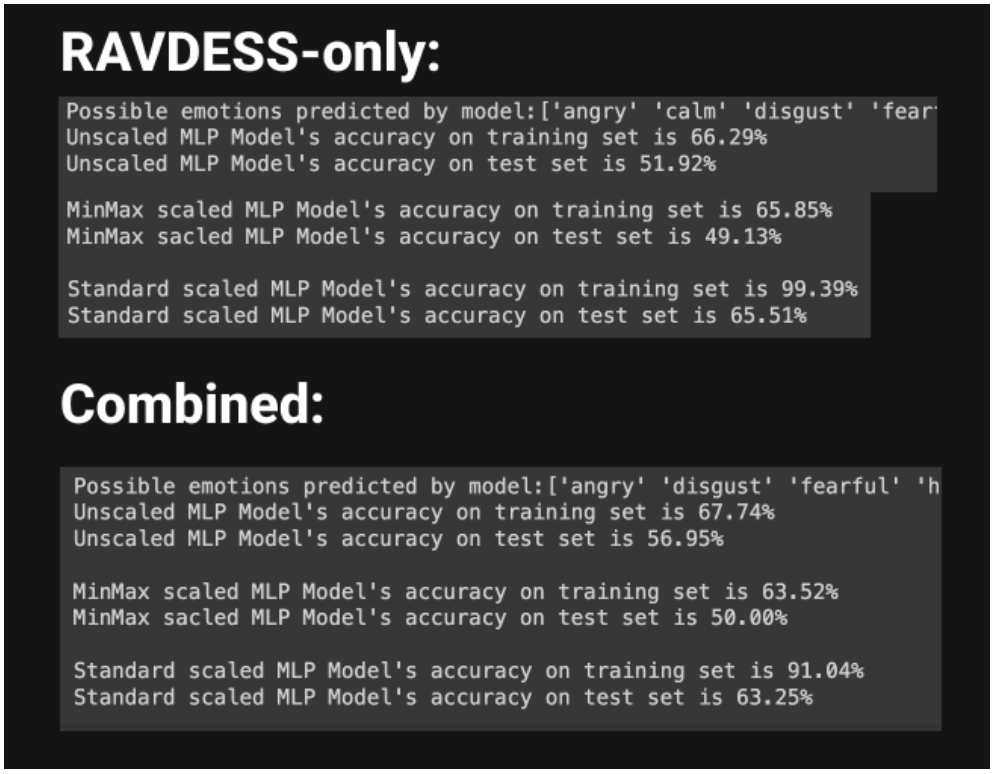
Specifically, the unscaled MLP model showed an increase in accuracy from the RAVDESS-only dataset, with training set accuracy rising from 66.29% to 67.74%, and test set accuracy from 51.92% to 56.95% in the combined dataset. However, for both MinMax and Standard scaled MLP models, the RAVDESS dataset outperformed the combined dataset, with the Standard scaled MLP model displaying particularly higher accuracy.

To illustrate, Figure 2 details the comparison:

For the RAVDESS-only dataset, the unscaled MLP model achieved 66.29% accuracy on the training set and 51.92% on the test set. The MinMax scaled model reached 65.85% and 49.13%, respectively, while the Standard scaled model significantly outperformed with 99.39% on the training set and 65.51% on the test set.

Contrastingly, in the combined dataset, the unscaled MLP model's accuracy improved to 67.74% on the training set and 56.95% on the test set. However, the MinMax scaled model dropped to 63.52% and 50.00%, respectively, and the Standard scaled model saw a decrease to 91.04% on the training set, though still maintaining a high accuracy of 63.25% on the test set.

Figure.2



Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

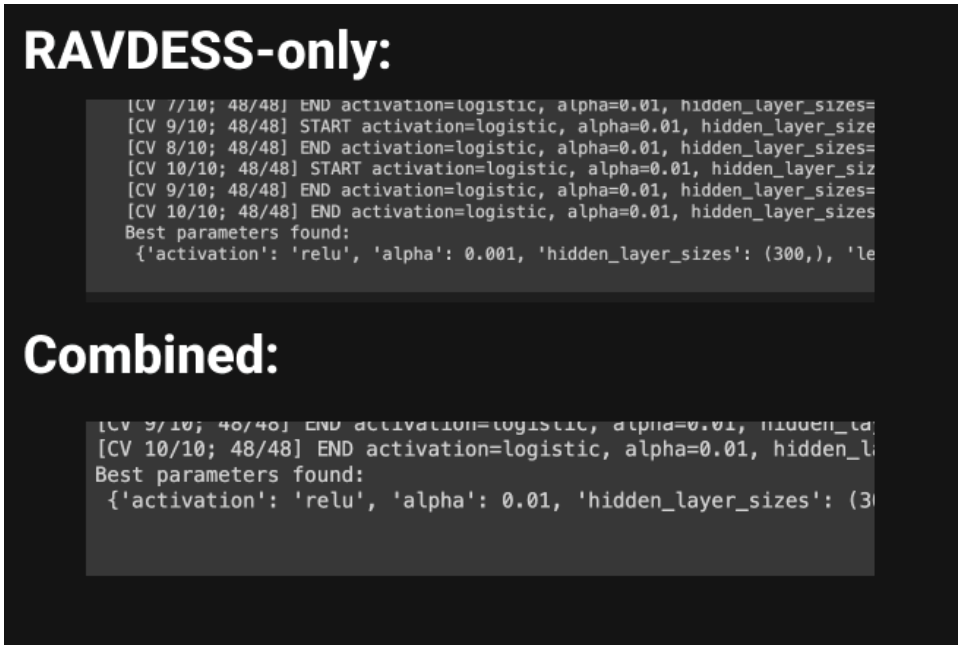
It seems curious that the RAVDESS-only dataset outperforms the combined dataset, especially after scaling. One possible explanation could be the nature of the RAVDESS dataset itself—it's more homogeneous and perhaps easier for the model to learn from due to less variability. In contrast, the combined dataset introduces a wider range of emotional expressions and acoustic features from EmoDB, which could increase the complexity of the learning task and potentially introduce noise, making it harder for the model to generalize. This suggests that while diversity in training data is often beneficial, too much variance without proper tuning or sufficient model complexity

might actually hinder performance, especially in nuanced tasks like emotion recognition.

Hyperparameter and Grid Search:

After opting for standard scaling on the combined dataset, I embarked on optimizing hyperparameters through grid search cross-validation, using the same candidate values suggested in the tutorial. Interestingly, the results mirrored those from the tutorial, with one notable exception: the optimal alpha value for the combined dataset was 0.01, diverging from the 0.001 preferred for the RAVDESS-only dataset. This subtle yet significant difference highlights the nuanced impact dataset composition has on model tuning.

Figure.3

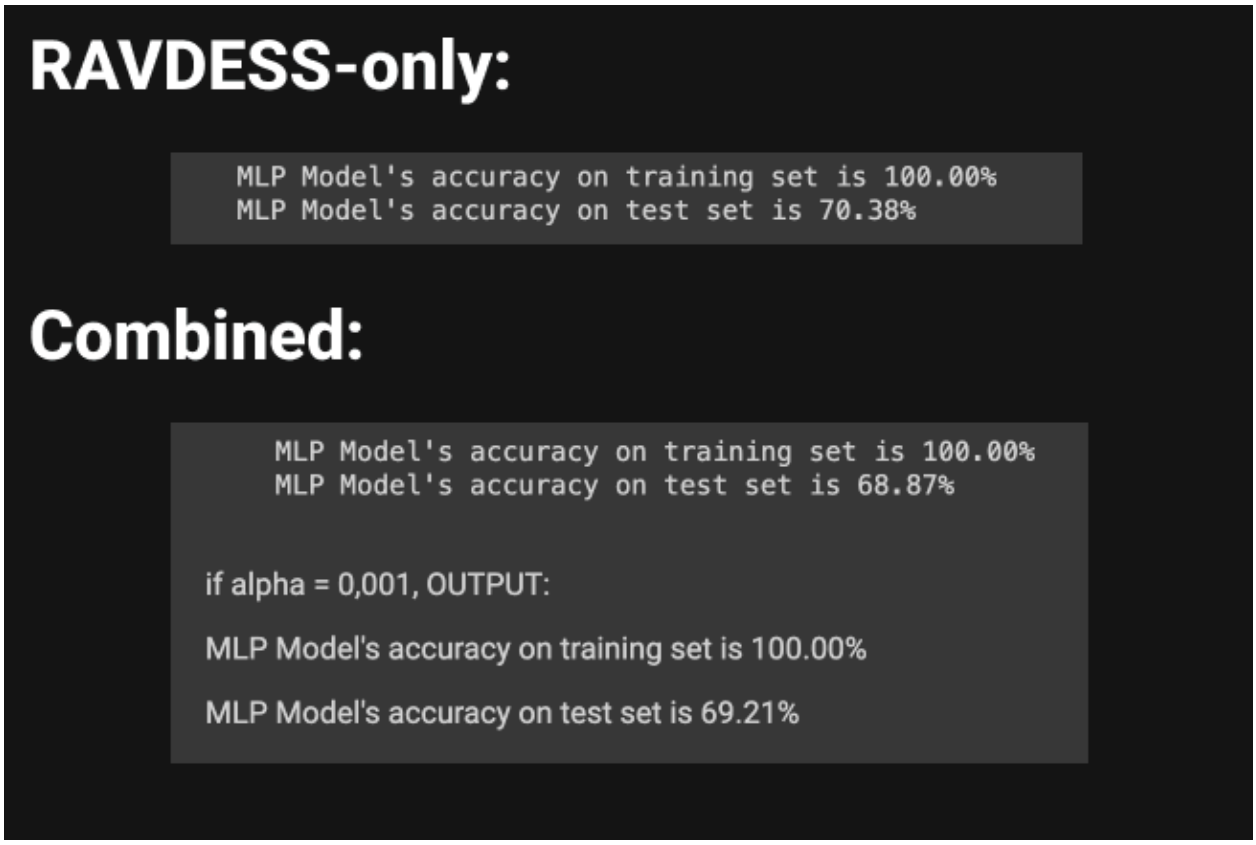


Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Training and Evaluating the MLP Model

Despite the grid search indicating alpha 0.01 as optimal, experimenting with alpha 0.001 actually led to improved performance, reaching a test set accuracy of 69.21% compared to 68.87% with alpha 0.01 (Figure 4). Contrary to the RAVDESS-only dataset, both alpha settings performed slightly worse in the combined dataset, with about a 1% difference, which aligns with the tutorial's expectation due to most hyperparameters selected by the grid search being sklearn's MLP defaults. This indicates our MLP model is overfitting the training data, demonstrating a challenge in generalizing effectively to the test set.

Figure.4

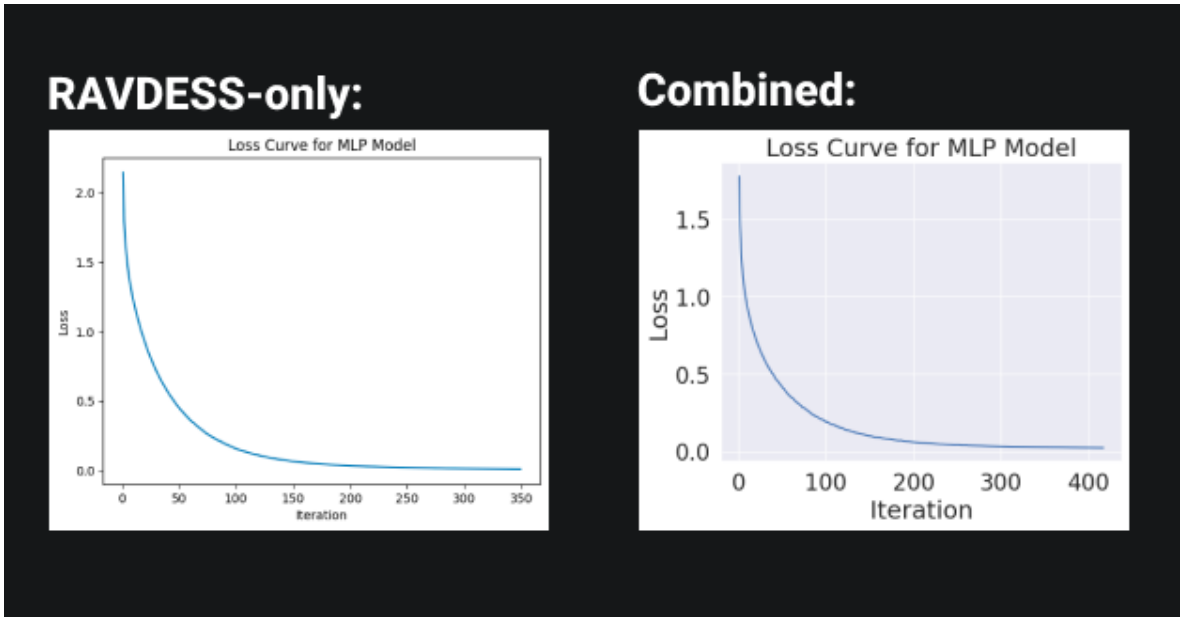


Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Loss Curve:

After digesting the tutorial notes, I can see that the loss curves indicate a solid learning rate; it's not too fast to cause the model to overfit to the last batches of data, nor too slow to delay convergence. The RAVDESS-only dataset's gradual loss descent and the combined dataset's steeper drop both seem to be avoiding these extremes. However, considering the perfect training performance and weaker test results, it seems my model might be too variance-heavy, fitting too closely to the training data and not generalizing well. To combat this, reducing the number of features or augmenting the dataset size could help, as well as implementing data regularization techniques like adding noise to the audio samples to reduce overfitting and improve the model's ability to generalize (Figure 5).

Figure.5

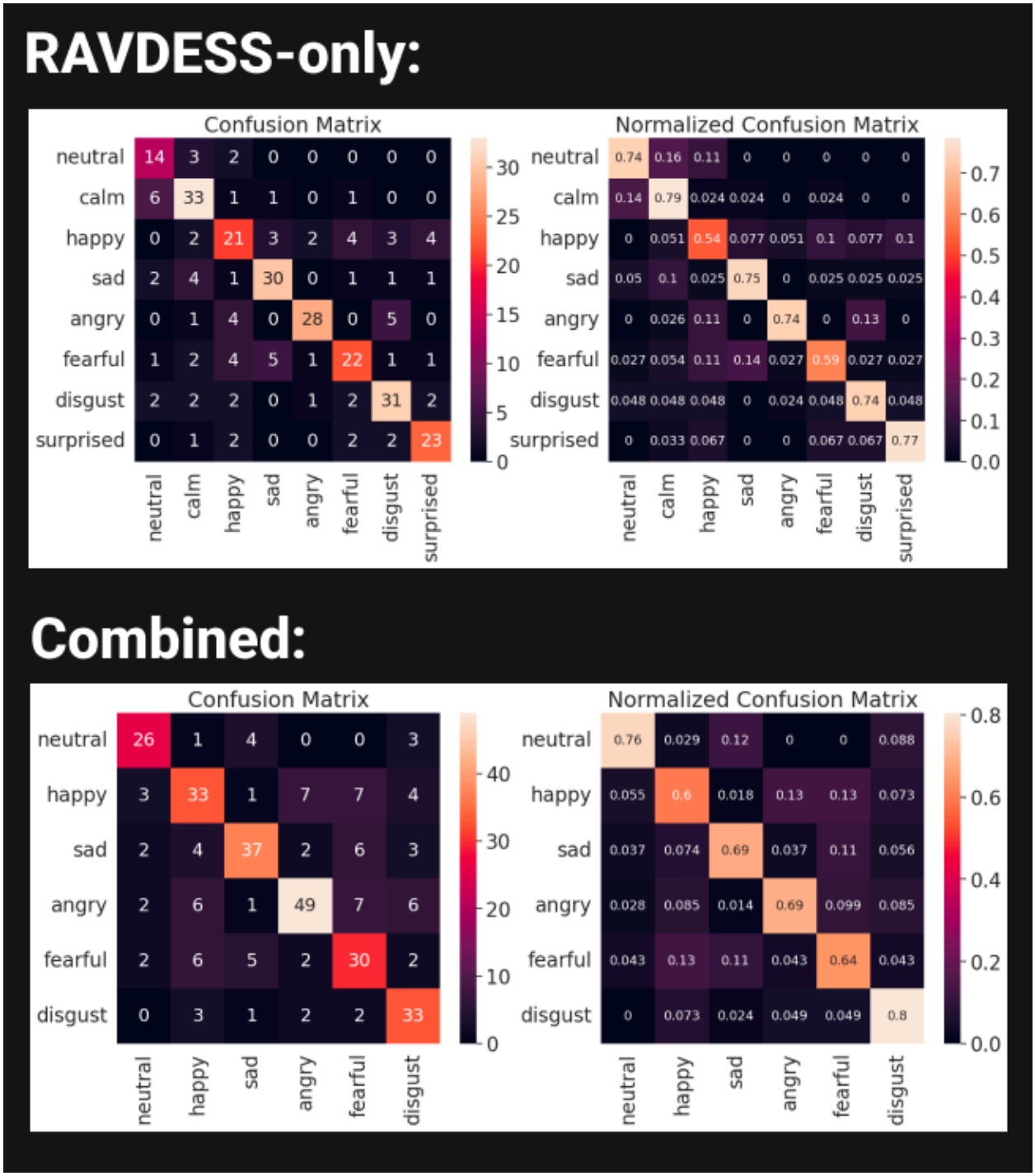


Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Confusion Matrix:

The confusion matrices (Figure 6) show that the RAVDESS-only dataset had consistent true positive rates for most emotions, excelling with 'Calm' at 79% accuracy, but falling short with 'Happy' at 54%. Conversely, the combined dataset improved 'Disgust' recognition to 80%, suggesting that more diverse data enhances specific emotion detection. Yet, the accuracy for 'Happy' remained lower, indicating persistent challenges in classifying this emotion. Across both datasets, emotions like 'Angry' and 'Sad' maintained similar true positive rates, hinting at a certain level of model consistency. These matrices are crucial for identifying where the model confuses emotions, guiding further improvements in its predictive accuracy.

Figure.6

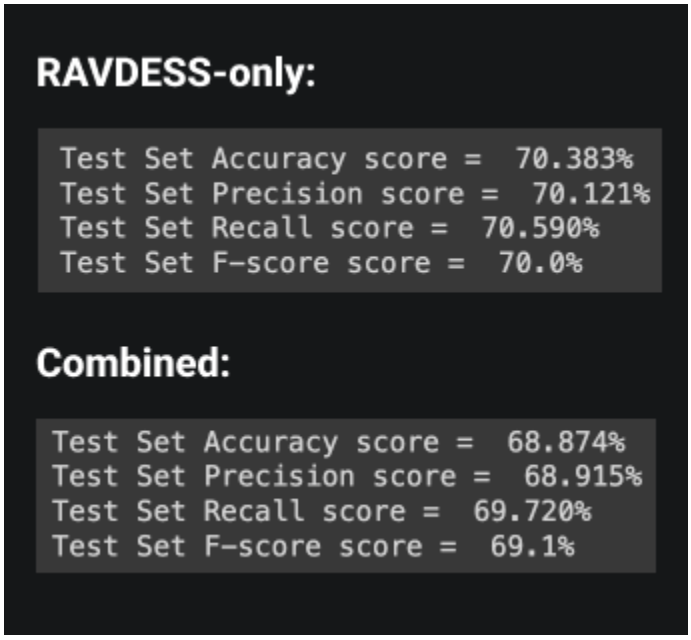


Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Precision, Recall, F-Score

The RAVDESS-only dataset seems to have a balanced performance with similar rates of precision, recall, and F-score, all hovering around 70%, indicating that the model has a comparable false positive and false negative rate across all emotion classes (Figure 7). In comparison, the combined dataset shows a slight dip in precision and accuracy, but an increase in recall, suggesting it's better at capturing all positive samples but at the cost of a few more false positives (Figure 7). This could mean that while the combined dataset provides a richer variety of data, it may also introduce complexity that slightly hinders the model's precision. From a student's perspective, these metrics are crucial as they give a clearer picture of the model's predictive strengths and weaknesses, guiding potential strategies for fine-tuning.

Figure.7

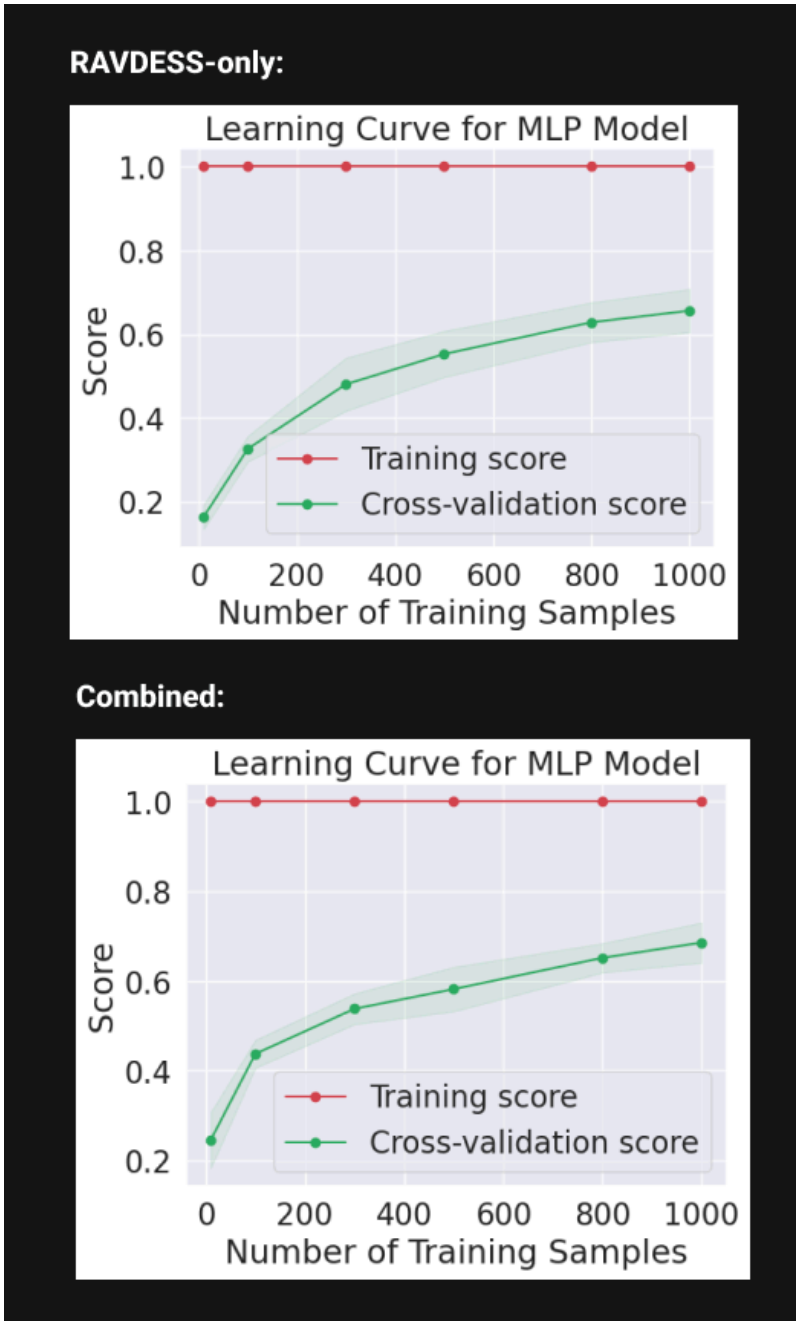


Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Learning-Curve

Comparing the learning curves for the MLP models trained on the RAVDESS-only and combined datasets, it's noticeable that the combined dataset exhibits a closer gap between the training and cross-validation scores. This suggests that the model trained on the combined dataset is less overfitted and generalizes better than the one trained on RAVDESS-only, despite both models starting with perfect training scores. The cross-validation score for the combined dataset also increases steadily with more data, which could imply that additional data might still be beneficial, in contrast to the RAVDESS-only model where the benefit of more data seems limited. These observations lead me to infer that the combined dataset, with its greater complexity and diversity, provides a better training environment for the MLP model than the more homogeneous RAVDESS-only dataset.

Figure.8



Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Conclusion and Comparison with the ML models in Assignment 3

The DNN's learning curves suggest that it has a stronger capacity for handling complex patterns in the combined dataset compared to traditional ML models like SVC, KNN, and RandomForest, which tend to plateau. The DNN models benefit from the increased data diversity, showing less overfitting than when trained on the more uniform RAVDESS-only dataset. In contrast, the ML models demonstrated improvements with the combined dataset but still showed signs of reaching their learning limits, as the complexity they can capture is inherently restrained by their simpler algorithms. The high variance observed with the DNN on RAVDESS-only data indicates a potential overfit to the training data, a problem less pronounced in the ML models. These ML models seem to offer a balance between performance and generalization that might be more suitable for datasets with less variety.

In conclusion, while DNNs show promise in extracting intricate relationships within complex datasets, their sophistication comes with the challenge of overfitting, especially when the data lacks diversity. ML models, on the other hand, offer a more stable learning curve but might not fully capitalize on the richness of complex datasets. Understanding the strengths and limitations of both approaches is key to choosing the right model for the task at hand, considering both the dataset's characteristics and the desired outcome of the analysis.

Reference

Python 2.7 Tutorial. (2024). Pitt.edu. https://sites.pitt.edu/~naraehan/python2/tutorial9.html

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Simon Fraser Universty. (2019). *Week4 Neural Networks Lab materials*. Google Colab. https://colab.research.google.com/drive/1gLprHODuNr-YOp9bI7JY6sksKxUP78H3#scrollTo=2RaC-LS5L9hs