Daniel Chung
BioE 145
May 4, 2023

Final Project Write-Up

Part 1:

After visualizing the data using PCA, t-SNE, and my autoencoder, I found that the autoencoder produced the most meaningful visualization. One reason for this may be that autoencoders are designed to learn a compact and meaningful representation of the input data by reconstructing it from a lower-dimensional latent space. This means that the autoencoder is better able to capture the non-linear relationships between the features and create a low-dimensional representation that preserves the most important information. In contrast, PCA and t-SNE are linear methods that may not be able to capture complex non-linear relationships as effectively. The PCA visualization was also useful, but it did not capture as much of the underlying structure as the autoencoder did. The t-SNE visualization seemed to be less informative than the other two methods, as it produced a dense cloud of points with less discernible patterns. Overall, the autoencoder was the most effective method for visualizing the data, followed by PCA and then t-SNE.

Part 2:

I was able to achieve an accuracy score of 0.80714 and 0.81429 with Bagging Random Forest Classifier and Boosting Random Forest Classifier, respectively.

Based on the classification results, it seems that ensemble methods, specifically bagging and boosting with random forest, were the most effective in improving the accuracy of the model. This is because random forest already performs well as a standalone classifier, but by adding bagging or boosting, it further improves the performance by reducing the variance and bias of the model. Bagging with KNN also showed some improvement, but not as much as with random forest, possibly due to the nature of the KNN algorithm being more prone to overfitting. The feedforward neural network had a relatively lower accuracy score compared to the other models, which may suggest that it was not the best fit for this particular classification problem. Overall, the results indicate that ensemble methods, in particular bagging and boosting with random forest, can be effective in improving the accuracy of classification models.