Performance Analysis

The supervised models for learning were analysed against parameters such as accuracy, precision, recall, and F1-score. The Random Forest and KNN models showed excellent accuracy among all models, with accuracies of 98.7% and 97.9%, respectively. The rest also presented high accuracy (~95%) and could be considered as strong baseline models. The best model did showcase a bit of overfitting.

The clustering quality of unsupervised learning models was assayed in terms of silhouette scores. K-Means and GMM provided the most distinct clustering (~0.59), which was considered good clustering quality. Agglomerative Clustering was on a similar level. However, DBSCAN and Spectral Clustering proved much more sensitive to parameter selection and gave rise to moderate clustering quality.

The Neural Network models, were better than the traditional models, with the most stable accuracy of 91.67%. CNN proved its strength in capturing the complexity of the data; models of similar complexity can take advantage of the small dataset. The MLP also showed strong performance, with the same accuracy of 91.67%, due to its deep architecture and learning rate decay effects. Concluding that both models gave the same, but highest, overall performance on the dataset.

Impact of Data Size and Complexity

2.1 Challenges of Small Datasets

Training the models with the small, and unbalanced, dataset presented serious issues. First and foremost, the dataset was very small, which did not allow complex models, including the neural network, to generalize well. Learning techniques, namely regularization, learning rate schedules, and early stopping, were crucial to controlling overfitting and having more generalization. The learning rate schedule is a technique used to gradually decrease the learning rate that the model follows to learn to converge more smoothly. Early stopping helped to monitor model performance on validation data and stopped training when the validation performance dropped, preventing overfitting.

2.2 Model Complexity and Overfitting

Model complexity plays a significant role in model performance and overfitting. This made simple models, such as Logistic Regression and Decision Trees, more forgiving to model and interpret; however, this usually caused overfeeding of the small dataset with the addition of noise rather than really capturing the pattern of the data. On the other hand, high-complexity models like the neural networks, namely MLP and CNN, coped well with the small dataset yet demanded both meticulous tuning and regularization. In the meantime, the ensemble approach of the Random Forest overfit the dataset. Neural networks can capture more complex relationships with deep architectures, but the overfitting problem should be avoided by proper techniques.

Conclusion

A comparative analysis emphasizes the fact that though traditional machine learning models score robustly when using minor datasets, neural networks, particularly MLPs and CNNs, can attain better performance by capturing complex patterns. On the other hand, their proficiency depends on how adroitly they are handled to prevent overfitting. The present study is a prominent indication of the importance of model selection, hyperparameter optimization, and regularization techniques to achieve optimal performance when dealing with limited and unbalanced data. Great care must be taken to strike the right balance between increasing model complexity and the danger of overfitting it: make it as complex as possible to capture the underlying patterns but not too complex to memorize noises.