Exploring Random Forests, K-Nearest Neighbors, and Multi-layer Perceptron Models for Health and Neuroimaging Datasets

Abstract

This report examines the application of machine learning (ML) to three health and neuroimaging datasets – Parkinson's, EEG Eye State, and EEG Database (addressed as EEG Alcohol) – to evaluate their potential for disease classification and prediction. Exploratory Data Analysis (EDA), preprocessing, and three classification models – Random Forests, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (addressed as MLP/ANN) – were applied. Central challenges included imbalanced and high dimensionality data, and dataset-specific preprocessing complexities. MLP performed well for Parkinson's disease classification, while EEG datasets highlighted the importance of feature scaling and structured workflows. The EEG Alcohol dataset posed significant challenges, underscoring the need for efficient preprocessing. This report reflects on lessons learned and proposes directions and improvements for future work.

Introduction

Machine learning (ML) is increasingly applied to health and neuroscience research, leveraging data-driven approaches to uncover patterns and predict outcomes. This project aimed to evaluate three distinct ML methods on three datasets, each presenting their unique challenges and opportunities.

The Parkinson’s dataset focused on classifying disease presence based on voice frequency features, such as jitter and shimmer. The small dataset allowed for quick experimentation but suffered from class imbalance. The EEG Eye State dataset aimed to distinguish between open and closed eye states using time-series EEG data, introducing challenges like variability and the need for normalization. Lastly, the EEG Alcohol dataset explored genetic predispositions to alcoholism using signals from 64 scalp electrodes, but its high dimensionality and organizational structure severely complicated analysis.

Methods

Dataset Overviews

1. Parkinson’s Dataset: A tabular dataset with features derived from voice recordings. Challenges included a small sample size and class imbalance, requiring attention to data splitting and model evaluation.
2. EEG Eye State Dataset: Time-series data designed to classify open vs. closed eye states. Challenges included feature scaling and participant variability.
3. EEG Alcohol Dataset: High-dimensional EEG data from 64 electrodes under various paradigms. Analysis was hindered by missing trials, inconsistent and complicated formatting, and high computational demands.

Exploratory Data Analysis (EDA)

1. Parkinson’s Dataset: Significant predictors were identified. Pair plots indicated mild clustering tendencies between Parkinsons and non-Parkinsons groups.

2. EEG Eye State Dataset: Histograms and time-series plots highlighted participant variability. Z-scores were used to normalize the features.
3. EEG Alcohol Dataset: Initial attempts at EDA focused on visualizing the data. However, the dataset's complexity stalled further analysis, despite concentrated efforts to debug code and make use of the data.

Models

1. Random Forests: Used for robustness to overfitting and its ability to handle small datasets effectively. Feature importance metrics provided insights into key predictors.
2. K-Nearest Neighbors (KNN): Simple and non-parametric, but sensitive to scaling and computationally expensive for larger datasets
3. Multi-Layer Perceptron (MLP): Artificial neural networks offered flexibility for learning complex patterns, but required careful tuning and substantial computational resources.

Results

Parkinson's Dataset

MLP achieved the highest accuracy (96%) and macro F1-score (0.94), followed closely by KNN (92% accuracy) and Random Forest (90% accuracy). As shown in Figure 2, the key predictors identified by Random Forest were MDVP:Flo (minimum vocal fundamental frequency), spread1, and HNR (harmonics-to-noise ratio), with shimmer-related features like Shimmer:DDA also contributing. While KNN excelled in accuracy, its macro F1-score lagged, reflecting challenges with class imbalance. Random Forest's interpretability and consistency make it valuable, though MLP emerged as the best-performing model overall.

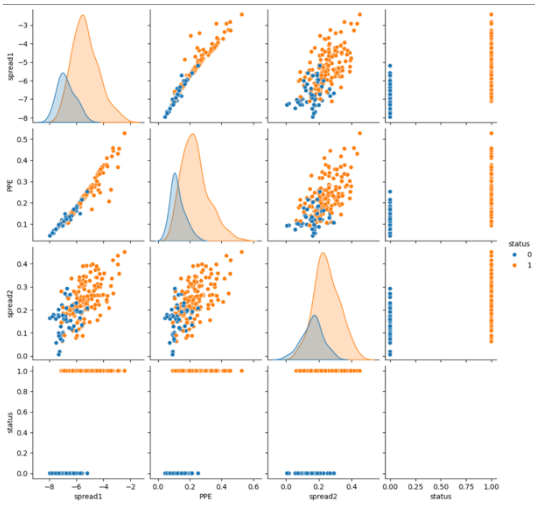

Figure 1. Pair plots of the Parkinson's dataset, highlighting clustering of features by disease status.

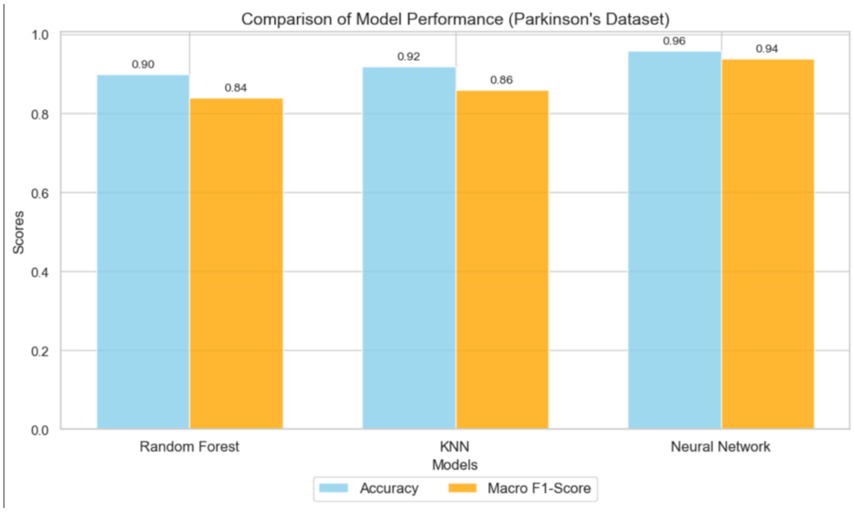

Figure 2. Comparison of Model Performance for Parkinson's Dataset: Random Forest, KNN, and ANN.

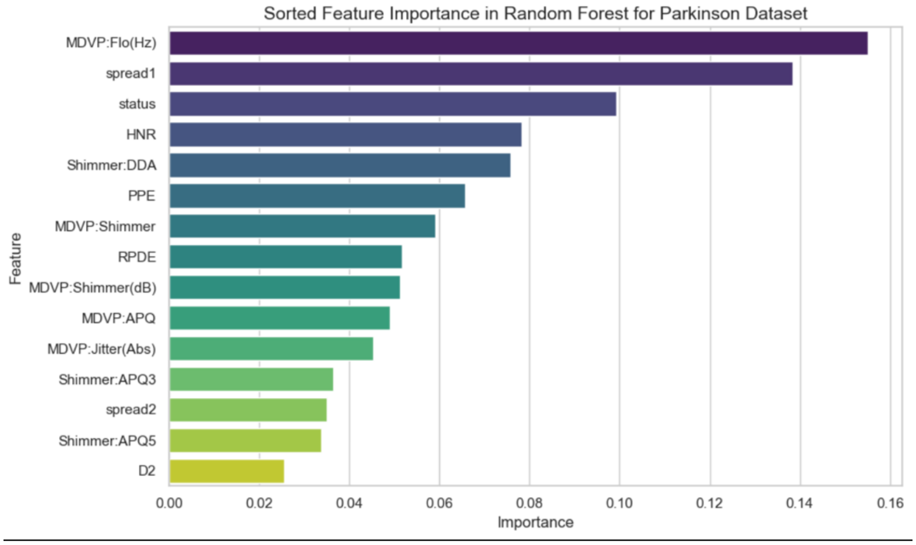

Figure 3. Sorted Feature Importance in Random Forest for Parkinson Dataset.

EEG Eye State Dataset

After normalization, both Random Forests (accuracy 92%) and KNN (accuracy 97%) achieved strong performance on the validation data of the EEG Eye State dataset, with KNN slightly outperforming Random Forests (Figure 5). Normalization reduced overfitting, further improving generalization. MLP (accuracy 87%) showed promise but required more hyperparameter tuning to match the performance of the simpler models.

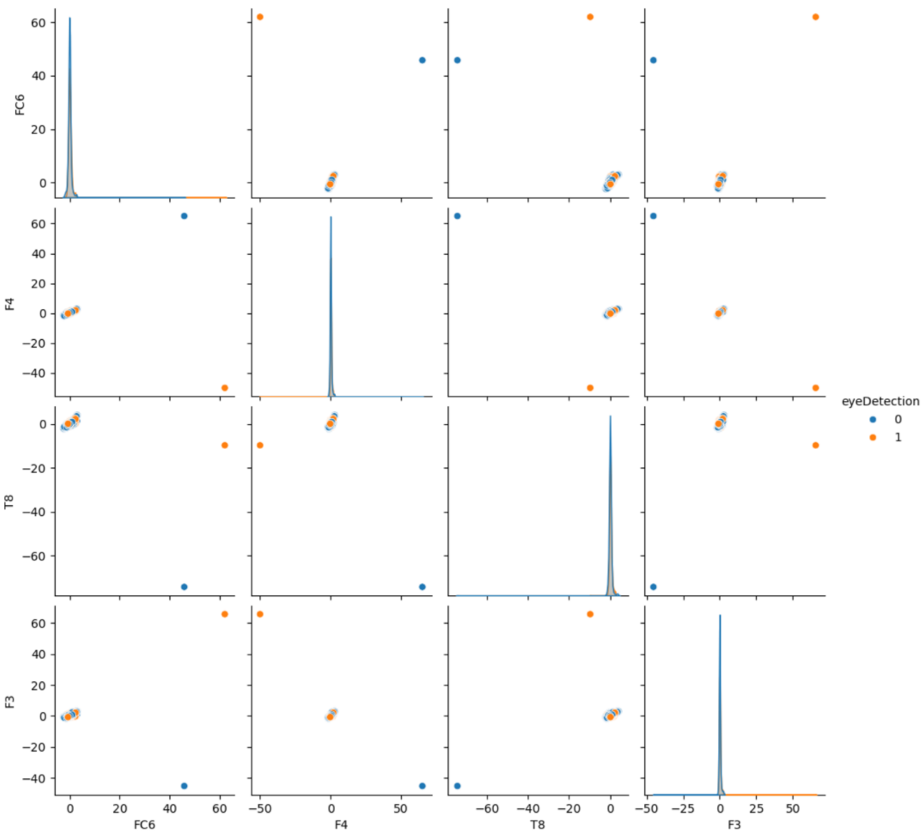

Figure 4. Pair plot for scaled features from the EEG Eye State Dataset.

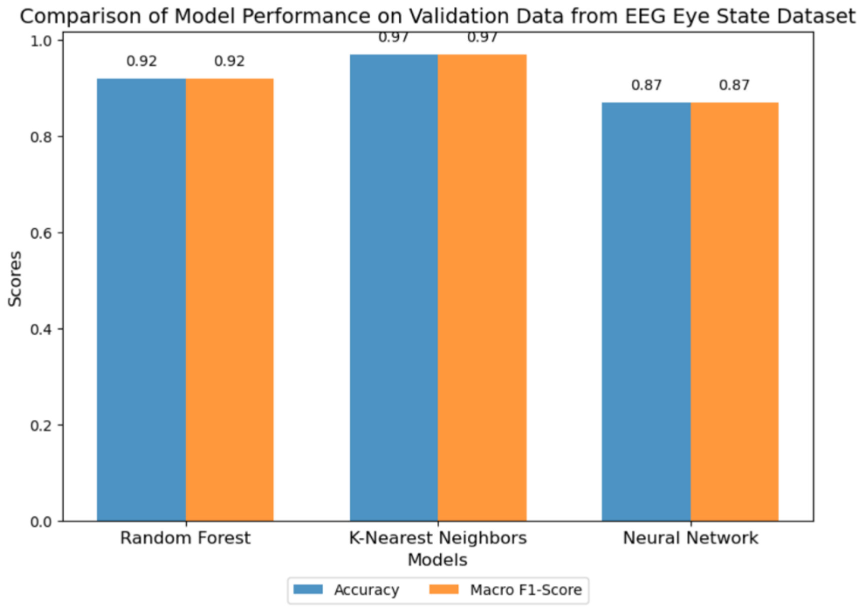

Figure 5. Comparison of Model Performance on Validation Data from EEG Eye State Dataset.

As seen in figure 6., feature importance analysis highlighted key contributions of channels such as O1 and P7. This underscores the role of feature scaling and iterative testing for hyperparameter tuning in order to achieve reliable classifications in EEG datasets.

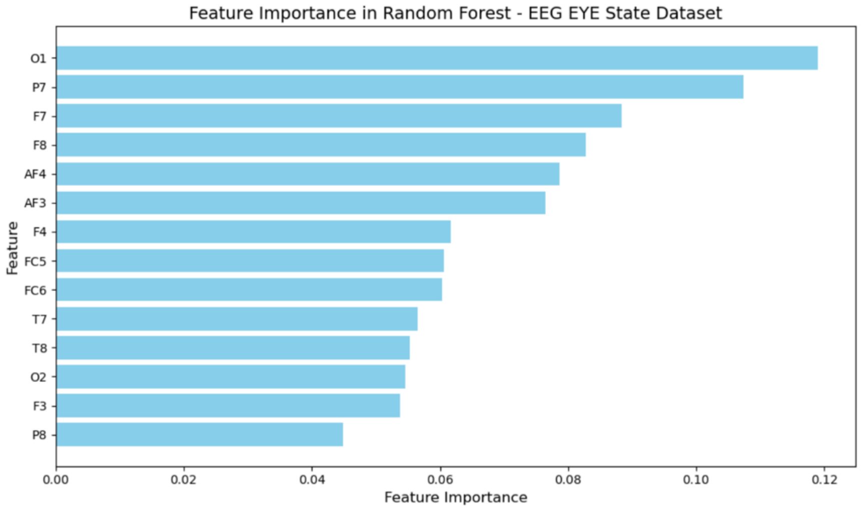

Figure 6. Feature Importance in Random Forest – EEG Eye State Dataset.

As seen in figure 7., KNN maintained its superior performance on the test data with 97% accuracy, followed by Random Forests at 92% and MLP at 88% (Figure 5). The consistency of KNN highlights the robustness after normalization and also the potential for Random Forests to provide interpretable results, while MLP required some optimization to improve generalization.

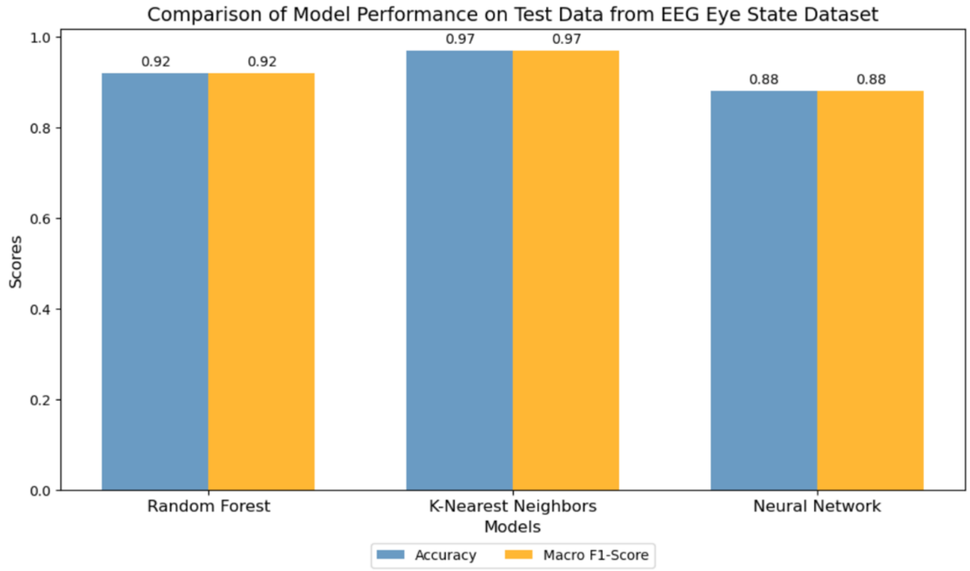

Figure 7. Comparison of Model Performance on Test Data from EEG Eye State Dataset.

EEG Alcohol Dataset

Despite multiple preprocessing attempts, including processing the data in multiple different attempts and addressing missing trials, I was unable to get past the preprocessing step in order to successfully apply ML models. The large volume of data and the high dimensionality made both debugging and meaningful feature extraction overwhelming.

Challenges and Learning

My initial work on the Parkinson's dataset revealed key mistakes, such as scaling data before splitting into training and test sets and failing to include a validation set. These errors emphasize the importance of a structured workflow and proper evaluation. Continuing on the EEG Eye State dataset, I applied these lessons and achieved better results. However, the EEG Alcohol dataset proved the most challenging, resulting in many hours trying to debug without yielding any usable results. Its complexity underscored the need for setting clear objectives, establishing an efficient preprocessing pipeline, and maintaining a well-organized notebook.

1. Pipeline Workflow

A clear and reproducible workflow is essential to avoid data leakage and misleading results. Introducing validation sets and adhering to best practices ensures reliable analyses.

2. Feature Scaling

The EEG Eye State dataset demonstrated the importance of feature scaling for models like KNN, particularly with noisy EEG data. Normalization significantly improved performance.

3. High Dimensionality Data

The EEG Alcohol dataset highlighted challenges with high-dimensional data. Techniques like PCA should have been used to simplify preprocessing and analysis.

4. Model Selection and Tuning

Random Forests performed well due to its robustness and interpretability. KNN gave good results with normalized data but struggled otherwise, while MLP required extensive tuning and resources, limiting its efficiency for quick iterations.

Discussion

The results of this study emphasized the importance of model selection, feature preprocessing, and structured workflows when applying machine learning techniques. The Parkinson's dataset demonstrated the use of Random Forests for feature importance analysis, while MLP achieved the highest accuracy, highlighting the strengths of applying neural networks for complex classification problems. The EEG Eye State dataset underscored the necessity of normalization for time-series data, showing that even simple models like KNN could perform well with proper preprocessing. And the EEG Alcohol dataset, while incomplete, demonstrated the importance of clear goals and efficient workflows for high-dimensional datasets. Overall, the process reinforced the importance of preprocessing, iterative testing, and model interpretability for ML applications in health and neuroscience.

Conclusion

This study explored the application of Random Forests, KNN, and MLP models to health and neuroimaging datasets, revealing both the promise and limitations of these approaches. While Random Forests excelled in feature importance analysis for the Parkinson's dataset, EEG datasets benefited significantly from normalization and iterative preprocessing. The challenges faced reinforced the importance of methodology, clear objectives, and detailed preprocessing pipelines.

In future projects, refining neural network architectures and introducing dimensionality reduction techniques for high-dimensional EEG datasets will be emphasized. This project served as a foundation for applying ML in health and neuroscience, providing and improving technical skills and also lessons for future exploration and analyses.

References

Begleiter, H. (1995). EEG Database [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TS3D

Little, M. (2007). Parkinsons [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59C74.

Roesler, O. (2013). EEG Eye State [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57G7J.