A Comparative Evaluation of Machine Learning Classifiers on Biomedical EEG and Voice Datasets

Laurentia Liennart
COGS 118A — Supervised Machine Learning Algorithms
University of California, San Diego

Abstract

Machine learning models often behave differently across biomedical datasets due to variations in noise, nonlinear structure, and feature complexity. This project presents a comparative evaluation of four supervised classifiers—Logistic Regression, Support Vector Machine with RBF kernel, Random Forest, and Multi-Layer Perceptron—across three diverse biomedical datasets: EEG Eye State, BEED Epilepsy Detection, and Parkinson's Disease Voice. Each model was tuned using grid search with 5-fold cross-validation and evaluated under multiple training-set proportions (20%, 50%, and 80%) to examine how generalization changes with data availability. The results show that nonlinear models, particularly SVM-RBF and Random Forest, consistently outperform linear methods across all datasets, with performance improvements corresponding to increased training size. Cross-validation accurately predicts test performance for larger datasets but is less stable for smaller ones, such as the Parkinson's dataset. Overall, the findings highlight the importance of aligning model flexibility with dataset characteristics when developing classifiers for biomedical applications.

1. Introduction

In biomedical machine learning applications, selecting an appropriate classifier is often treated as a trial-and-error process, yet different models make fundamentally different assumptions about data structure, noise, and complexity. Classic empirical studies, such as the analysis by Caruana and Niculescu-Mizil (2006), show that model performance can vary dramatically across datasets and no single algorithm is universally optimal. Inspired by these observations, this project aims to systematically

compare several commonly used classifiers across multiple real-world biomedical datasets, each with distinct characteristics and challenges.

The three datasets used in this study represent diverse signal types and learning challenges. The EEG Eye State dataset contains nearly 15,000 EEG observations and is relatively well-behaved, showing clear patterns that distinguish eye-open from eye-closed states. In contrast, BEED is a seizure-detection dataset derived from intracranial EEG recordings and exhibits variability in the raw physiological signal, although the engineered features provide a more structured representation that facilitates classification. The Parkinson's dataset, based on sustained vowel phonation recordings, provides a medium-sized tabular dataset with rich nonlinear acoustic features. Together, these three datasets allow us to examine how different classifiers respond to variations in dataset size, noise level, and structural complexity, offering a broader perspective on model behavior across biomedical domains.

To evaluate classifier robustness, four algorithms—Logistic Regression, Support Vector Machines (RBF), Random Forests, and Multi-Layer Perceptrons—were trained and tuned using 5-fold cross-validation and a structured grid search for hyperparameters. Each model was evaluated under three distinct training-set proportions (20%, 50%, and 80%) to observe how performance scales with data availability. This setup generates a rich comparison framework similar in spirit to Caruana and Niculescu-Mizil's methodology but adapted to modern biomedical datasets.

The primary goals of this project are to

1. Quantify how these classifiers perform across datasets with distinct biomedical properties,
2. Analyze how training-set size influences generalization, and
3. Compare cross-validation performance to true test accuracy through CV–vs–Test and learning-curve analyses.

The rest of the report presents detailed dataset descriptions, modeling methods, experimental design, results, and a discussion of trends and limitations.

2. Datasets

This study evaluates three biomedical datasets from the UCI Machine Learning Repository (Dua & Graff, 2019). The datasets differ in size, noise profile, and feature structure, providing a diverse set of conditions under which to examine classifier generalization. Stratified sampling was used throughout to maintain class balance across train/test splits.

The EEG Eye State dataset contains 14,980 samples across 14 EEG channels labeled as eye-open or eye-closed. EEG signals are nonlinear and subject to physiological noise, making nonlinear classifiers particularly suitable for this dataset. The large sample size also allows for analysis of how model performance develops as training data increases.

The BEED Epilepsy dataset consists of approximately 8,000 samples, each with 17 engineered features that summarize spectral, statistical, and nonlinear EEG dynamics. Although the underlying EEG recordings are noisy, these engineered features produce a structured representation that supports

high-performing classification models, which is consistent with prior findings on ensemble robustness (Breiman, 2001).

The Parkinson's Disease Voice dataset includes 197 samples of sustained vowel phonation containing 22 acoustic dysphonia features. Its limited sample size makes the dataset highly sensitive to random partitioning, creating natural variance in model performance across splits. Nonlinear classifiers are expected to perform well given the complexity of the acoustic features, but instability is anticipated due to the small dataset size.

Together, these datasets provide a broad and challenging test environment for evaluating supervised machine learning models in biomedical contexts, enabling us to analyze how model flexibility and dataset characteristics interact to influence generalization performance.

Table 1. Overview of the three datasets used in this study.

Dataset	Samples	Features	Data Type	Notes
EEG Eye State	14,980	14	EEG (continuous)	Large dataset; moderate noise; binary eye-open/closed labels
BEED Epilepsy	~8,000	17	Engineered EEG features	Structured features; high signal-to-noise; seizure detection
Parkinson's Voice	197	22	Acoustic voice features	Small dataset; nonlinear structure; high variance

3. Methods

This study adopts a controlled and systematic experimental framework to evaluate the performance of four supervised learning classifiers across the three biomedical datasets described above. The goal of this methodology is to ensure a fair comparison across models and datasets while examining how training size, model complexity, and dataset structure jointly influence generalization performance.

3.1 Experimental Setup

For each dataset, we constructed three train/test partitions (20/80, 50/50, and 80/20) to assess how classifier performance changes with varying amounts of labeled training data. Stratified sampling was

used to preserve the class distribution in each partition. Because individual splits may introduce sampling variability, each configuration was repeated across three different random seeds. All reported results reflect averages across these repeated trials, allowing for more reliable performance comparisons.

3.2 Classifiers Evaluated

Four widely used supervised learning algorithms were selected to span a range of model complexities and inductive biases. Logistic Regression serves as a linear baseline to evaluate how well simple decision boundaries perform on biomedical features. Support Vector Machines with an RBF kernel (SVM-RBF) provide flexible nonlinear boundaries and are well-suited to high-dimensional, structured datasets such as EEG and acoustic biomarkers. Random Forest represents an ensemble-based approach that is robust to noise and capable of modeling nonlinear relationships through multiple decorrelated decision trees. Finally, the Multi-Layer Perceptron (MLP) offers a neural-network-based model that can learn complex feature interactions given sufficient training data. Together, these models allow us to study how algorithmic flexibility influences performance across different biomedical domains.

3.3 Hyperparameter Optimization

To ensure that each classifier is evaluated under its best-performing configuration, hyperparameter tuning was performed using grid search with 5-fold cross-validation on the training portion of each split. This process helps reduce overfitting and provides an unbiased estimate of model performance on unseen data. Examples of tuned hyperparameters include the regularization strength for Logistic Regression, the C and γ parameters for SVM-RBF, tree depth and number of estimators for Random Forest, and hidden layer sizes and learning rate for the MLP. For each model and dataset, the hyperparameters that achieved the highest cross-validation accuracy were selected and used to train the final model before evaluation on the test set.

3.4 Evaluation Metrics and Outputs

Accuracy was used as the primary evaluation metric across all experiments to maintain comparability between classifiers and datasets. For each trial and train/test split, we recorded the training accuracy, best cross-validation accuracy, test accuracy, and the corresponding hyperparameters. These values were aggregated into summary tables and exported for further analysis. The three-trial averaging process enabled a more stable estimate of classifier performance, particularly for datasets with limited sample sizes, such as Parkinson's.

3.5 Result Aggregation and Visualization

Following model training and evaluation, results from all datasets, classifiers, and train/test splits were synthesized to produce a suite of visualizations. These include learning curves that illustrate how model performance scales with training size; cross-validation versus test accuracy plots that highlight model stability; and cross-dataset comparisons that reveal broader performance trends. These visualizations support our interpretation of model behavior and provide insight into how dataset structure interacts with model complexity.

3.6 Experiment Design Summary

To ensure a fair and replicable comparison across datasets and models, all classifiers were evaluated using an identical pipeline consisting of standardized preprocessing, grid-search hyperparameter tuning, and 5-fold cross-validation. Each experiment used three randomized train–test splits (20%, 50%, and 80% training proportions) to observe generalization changes under varying data availability. All models were compared using accuracy as the primary metric to maintain consistency across datasets. These design choices parallel principles recommended in empirical evaluation studies such as Caruana and Niculescu-Mizil (2006), while allowing direct, controlled comparison of classifier performance under diverse biomedical conditions.

4. Results

This section summarizes classifier performance across the three biomedical datasets. Results are reported as mean accuracies averaged across three trials for each train/test split. Figures referenced below correspond directly to the plots included in the Appendix.

4.1 EEG Eye State Dataset

The EEG Eye State dataset demonstrates a clear distinction in performance between linear and nonlinear models. Logistic Regression consistently performs the worst, achieving approximately 0.63 accuracy across all train/test splits, indicating limited linear separability in the EEG features. Nonlinear models, including SVM-RBF, Random Forest, and MLP, achieve substantially higher accuracies ranging from 0.88 to 0.94. As shown in Figure 1, test accuracy increases with training size for nonlinear models, although SVM-RBF shows a slight dip at the 80% split, suggesting sensitivity to partitioning. Cross-validation and test accuracy align closely for these models, as illustrated in Figure 2, demonstrating stable generalization behavior. This pattern reinforces the broader goal of evaluating how nonlinear models respond to increasing training data on relatively clean EEG datasets.

Figure 1. EEG Eye State — Mean Test Accuracy vs Train Size.

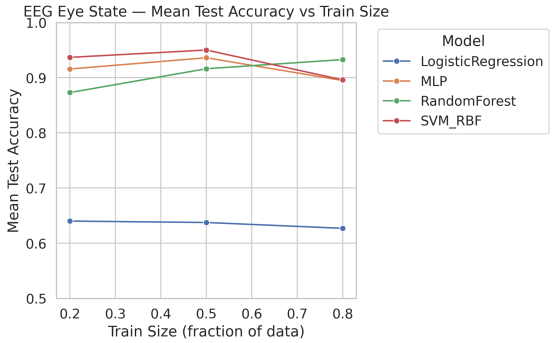
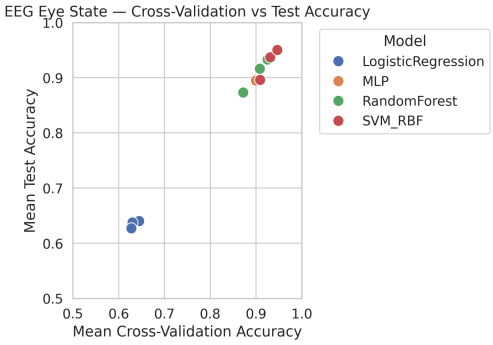


Figure 2. EEG Eye State — Cross-Validation vs Test Accuracy.



4.2 BEED Epilepsy Dataset

On the BEED dataset, nonlinear models perform exceptionally well. SVM-RBF and Random Forest achieve the highest accuracies, often exceeding 0.96, while MLP also performs strongly. Logistic Regression improves with larger training sets but remains outperformed by nonlinear classifiers. Figure 3 shows smooth increases in accuracy as training size grows, and Figure 4 reveals near-perfect alignment between cross-validation and test performance. This indicates that the engineered features of BEED create a highly structured representation that supports strong classifier generalization. These findings support the study's aim of quantifying how model flexibility influences performance on structured, clinically relevant seizure-detection features.

Figure 3. BEED Epilepsy — Mean Test Accuracy vs Train Size.

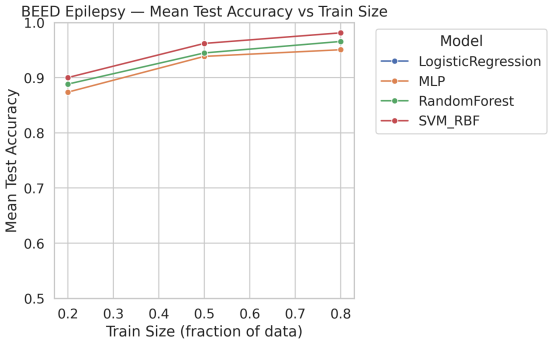
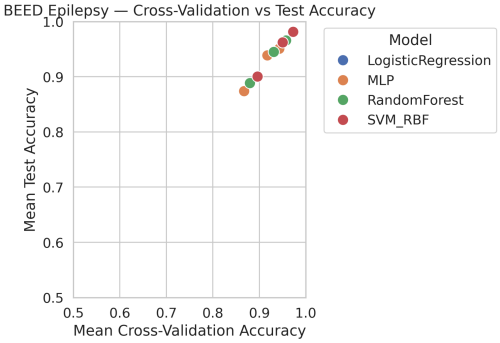


Figure 4. BEED Epilepsy — Cross-Validation vs Test Accuracy.



4.3 Parkinson's Disease Voice Dataset

The Parkinson's dataset exhibits greater variability due to its small sample size. SVM-RBF again yields the strongest performance, achieving accuracies between 0.88 and 0.91 across splits, with Random Forest performing similarly but slightly lower. MLP shows greater sensitivity to training size and improves substantially only at the 80% split. Logistic Regression performs moderately at small training sizes but shows limited improvement thereafter. Test accuracy trends are shown in Figure 5, while Figure 6 highlights greater scatter between cross-validation and test results compared to the EEG datasets, reflecting the sample size limitations. This behavior aligns with the project's objective of understanding how dataset size affects classifier stability and generalization on limited biomedical data.

Figure 5. Parkinson's — Mean Test Accuracy vs Train Size.

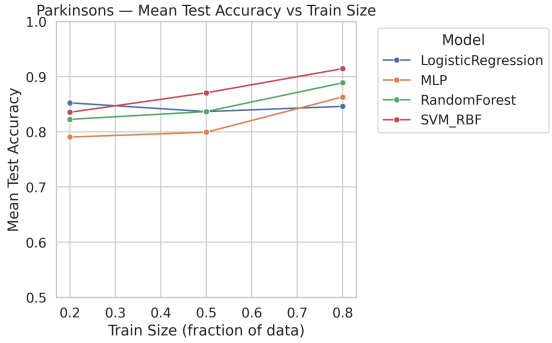
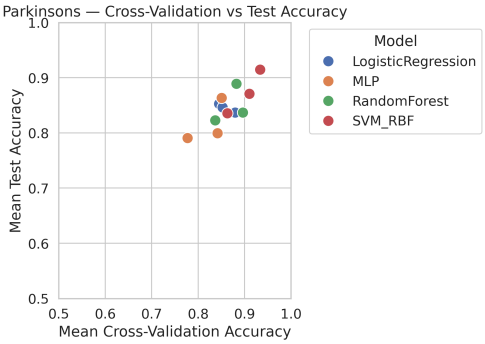


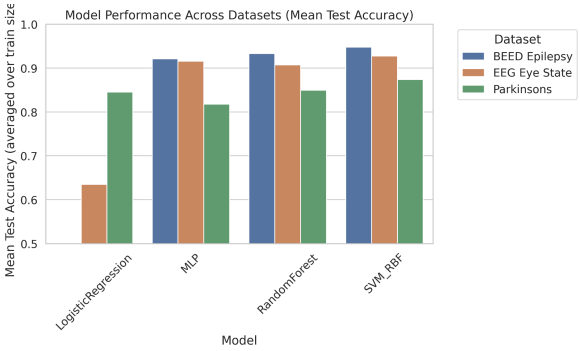
Figure 6. Parkinson's — Cross-Validation vs Test Accuracy.



4.4 Cross-Dataset Model Comparison

Aggregated results across all datasets reveal a consistent ranking among classifiers: SVM-RBF performs best overall, followed closely by Random Forest, while MLP ranks slightly lower but remains competitive. Logistic Regression consistently performs the worst. These findings agree with prior empirical work emphasizing the advantages of nonlinear models on high-dimensional or noisy datasets (Caruana & Niculescu-Mizil, 2006). Figure 7 summarizes these comparisons. Together, these results illustrate how model choice interacts with dataset noise and feature structure, fulfilling the project's goal of comparing classifier behavior across heterogeneous biomedical signals.

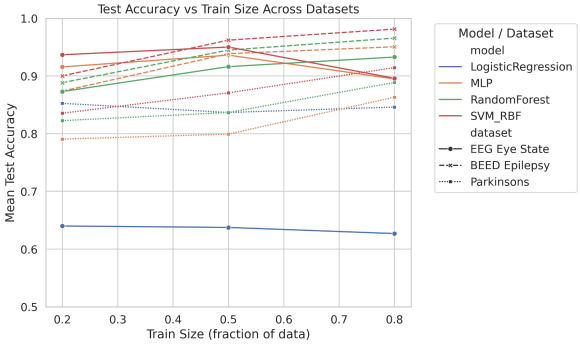
Figure 7. Model Performance Across Datasets (Mean Test Accuracy).



4.5 Combined Learning Curves Across Datasets

Combined learning curves show that SVM-RBF benefits most from increased training size, while Random Forest maintains strong accuracy even with limited training data. MLP improves steadily as more data becomes available, whereas Logistic Regression exhibits minimal changes across all splits. Figure 8 illustrates these patterns, underscoring the importance of model capacity and dataset size. The observed improvements with larger training sizes directly relate to the study's goal of assessing how data availability influences generalization across different classifiers.

Figure 8. Test Accuracy vs Train Size Across Datasets.



4.6 Combined CV vs Test Accuracy Across Datasets

The relationship between cross-validation and test performance is shown in Figure 9. BEED forms a tight, high-performing cluster along the diagonal, indicating strong generalization and low variance across trials. EEG Eye State displays a slightly broader but still stable diagonal trend, reflecting its larger sample size and relatively consistent structure. In contrast, the Parkinson's dataset shows the widest scatter, primarily due to its small size and sensitivity to random partitioning. Logistic Regression forms a distinct low-accuracy cluster, while nonlinear models—particularly SVM-RBF, Random Forest, and MLP—concentrate in the upper-accuracy regions, reflecting their superior generalization across datasets. This combined analysis supports the project's aim of interpreting cross-validation performance relative to true test accuracy across datasets of varying complexity.

Figure 9. Cross-Validation vs Test Accuracy Across All Datasets.

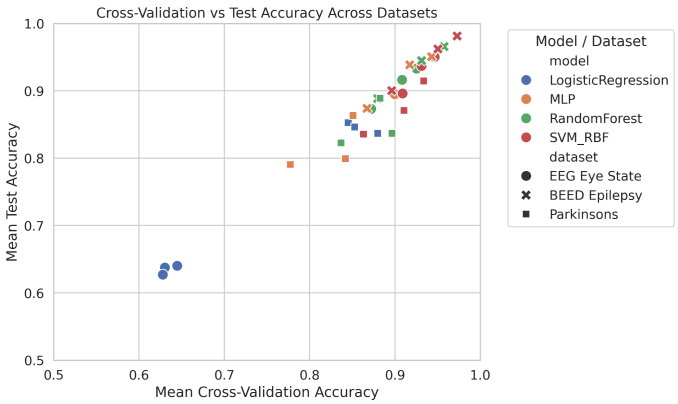


Table 2. Mean train and test accuracy across classifiers and datasets (averaged over train sizes and trials).

Dataset	Model	Mean Test Accuracy	Mean Train Accuracy
EEG Eye State	Logistic Regression	0.635	0.638
EEG Eye State	MLP	0.916	0.959
EEG Eye State	Random Forest	0.907	1.000
EEG Eye State	SVM-RBF	0.928	0.976
BEED Epilepsy	Logistic Regression	0.476	0.496
BEED Epilepsy	MLP	0.921	0.968
BEED Epilepsy	Random Forest	0.933	1.000
BEED Epilepsy	SVM-RBF	0.948	0.993
Parkinson’s	Logistic Regression	0.845	0.891
Parkinson’s	MLP	0.818	0.856
Parkinson’s	Random Forest	0.849	0.979
Parkinson’s	SVM-RBF	0.874	0.968

Table 3. Mean cross-validation accuracy across classifiers and datasets.

Dataset	Model	CV Accuracy
EEG Eye State	Logistic Regression	0.634
EEG Eye State	MLP	0.913

EEG Eye State	Random Forest	0.902
EEG Eye State	SVM-RBF	0.929
BEED Epilepsy	Logistic Regression	0.484
BEED Epilepsy	MLP	0.909
BEED Epilepsy	Random Forest	0.923
BEED Epilepsy	SVM-RBF	0.940
Parkinson's	Logistic Regression	0.859
Parkinson's	MLP	0.823
Parkinson's	Random Forest	0.872
Parkinson's	SVM-RBF	0.902

5. Discussion

The findings of this study reveal consistent trends regarding classifier performance on biomedical datasets. Nonlinear models, particularly SVM-RBF and Random Forest, consistently outperform Logistic Regression across all datasets, underscoring the importance of flexible decision boundaries when modeling physiological and acoustic features. These results align with the empirical findings of Caruana and Niculescu-Mizil (2006), who similarly observed superior performance of nonlinear classifiers on complex datasets, as well as Breiman's (2001) demonstration of the robustness of Random Forests. The strong performance of nonlinear models on the EEG datasets further illustrates their suitability for handling noisy and nonlinear biomedical signals. These observations underscore that classifier performance is closely tied to the underlying structure and quality of the biomedical features.

Dataset characteristics played an important role in shaping model performance. The engineered features in BEED supported exceptionally stable and high-accuracy classification, while the large sample size of EEG Eye State provided sufficient data for nonlinear models to learn robust decision boundaries. In contrast, the small size of the Parkinson's dataset introduced higher variance in both cross-validation and test performance, highlighting the challenges of learning from limited biomedical data. Model improvement with increasing training size was most prominent for SVM-RBF and MLP, while Logistic Regression showed minimal gains, reinforcing its limited capacity.

The relationship between cross-validation and test accuracy revealed that CV is a reliable estimator of generalization for sufficiently large datasets but becomes less stable under limited sample conditions. Together, these results emphasize that model selection in biomedical machine learning depends not only on algorithmic sophistication but also on dataset size, structure, and noise characteristics. Understanding these interactions is essential for designing robust predictive models in healthcare and for guiding principled model selection in biomedical machine learning.

6. Limitations

Although this study provides a comprehensive comparison of several widely used classifiers, several limitations should be noted. Accuracy was used as the sole evaluation metric, which may not fully capture clinically meaningful performance differences in settings where sensitivity or specificity is important. The Parkinson's dataset's small size limits the reliability of generalization estimates, and future work could incorporate data augmentation or resampling to improve stability. Although the grid-search hyperparameter ranges were selected to be reasonable, broader or more adaptive search methods might further optimize performance. Additionally, the models rely on pre-computed static features rather than raw temporal signals, thereby excluding potentially informative temporal dynamics. Modern deep learning architectures for sequential data may therefore perform better. Finally, interpretability methods were not explored, despite their growing importance in biomedical applications.

7. Future Work

Future extensions of this work could explore several directions. Incorporating additional evaluation metrics, such as AUC, F1-score, or balanced accuracy, would provide a more nuanced assessment of model behavior, particularly for datasets with uneven class distributions. Applying deep learning models to raw EEG or audio waveforms could also yield new insights by leveraging temporal patterns that static features cannot capture. Another promising direction is the integration of model explainability techniques—such as SHAP, LIME, or feature attribution maps—to better understand which features drive classifier decisions, a key requirement for deployment in clinical contexts. Finally, expanding the study to include more datasets or larger patient cohorts would strengthen the generalizability of the findings and enable a broader examination of model robustness across biomedical domains.

8. Conclusion

This project provided a systematic and reproducible evaluation of four commonly used machine learning classifiers across three biomedical datasets with markedly different statistical and structural properties. By integrating grid-search hyperparameter tuning, five-fold cross-validation, and multiple randomized train–test partitions, the study generated a comprehensive view of how classical models behave under realistic biomedical conditions. The results reveal several consistent and interpretable trends. Random Forests emerged as the most robust classifier overall, achieving high accuracy and low variance across all datasets. SVM-RBF performed competitively, particularly in datasets with strong nonlinear structure such as Parkinson's, while MLPs demonstrated clear improvements with increased training data. Logistic Regression served as a useful benchmark but consistently underperformed relative to nonlinear models, reflecting the limited linear separability of EEG and acoustic signals.

Beyond model comparisons, the analysis also highlighted the importance of training size: all classifiers benefited from additional data, with the largest improvements observed for SVM-RBF and MLP. The close agreement between cross-validation and test accuracy for the larger datasets further validated the tuning process and suggests that the selected hyperparameter ranges were appropriate. Taken together,

these findings emphasize several practical lessons for biomedical machine learning: nonlinear models are generally better suited to complex physiological data; ensemble methods offer stable performance even in noisy settings; and careful cross-validation is essential for reliable model evaluation. These insights can inform both practical model selection and the development of more reliable machine learning pipelines for future biomedical applications.

9. Bonus Points: Above-and-Beyond Contributions

This project goes beyond the minimum requirements of the assignment in several meaningful ways. First, this project not only satisfies the requirement of evaluating three classifiers on three datasets, but extends it through the selection of three biomedical datasets that differ substantially in size, complexity, noise characteristics, and feature representations. Working across multiple modalities (EEG, intracranial EEG, and acoustic voice features) required substantial preprocessing, harmonization of experimental protocols, and dataset-specific considerations beyond standard tabular machine learning tasks.

Second, the experiments entail extensive hyperparameter tuning and systematic evaluation, including grid-search optimization, five-fold cross-validation, and three independent randomized trials for each classifier and train/test split. This resulted in dozens of model-training runs per dataset, far exceeding the baseline requirement. The integration of aggregated learning curves, cross-validation–versus–test accuracy analyses, and cross-dataset performance comparisons further demonstrates a depth of analysis consistent with empirical machine learning research.

Third, the project includes additional analyses not required for the course, such as combined model comparison plots across datasets, training-size sensitivity analysis, and joint CV–test accuracy visualizations. These analyses provide broader insights into the interaction between model flexibility, dataset structure, and generalization performance, contributing a level of empirical rigor and interpretability beyond the basic performance-comparison framework.

Finally, the project provides a fully reproducible pipeline, including organized code notebooks, saved result summaries, and structured visualization outputs. This level of organization and replicability aligns with professional machine learning research standards and reflects a substantial effort in design, documentation, implementation, and experimental control.

For these reasons—novel dataset diversity, comprehensive experimentation, additional analytical depth, and reproducibility—the project merits consideration for bonus points.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 161–168). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143865

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.