Designing a Strategy for Model Comparison in the Non-Invasive Diagnosis of Alzheimer's Disease

Emanuele Tanzi - Student ID: 807406

A.Y. 2023-2024

1 Introduction

The integration of genetic data into diagnostics has improved early detection of complex diseases, with biomarkers serving as key indicators for monitoring and treatment. In Alzheimer's disease (AD), ideal biomarkers are non-invasive and sensitive to early pathological changes. MicroRNAs (miRNAs) are promising candidates, as they regulate gene expression and are stably detectable in fluids like blood or cerebrospinal fluid. Their involvement in AD-related processes, including neuroinflammation and amyloid-beta metabolism, supports their diagnostic potential, distinguishing mild cognitive impairment from normal aging and advanced AD (Petracci et al., 2025 [2]; Zhou & Pang, 2025 [3]). Advances in AI and multi-omics further enhance this field, with predictive miRNA panels and integrative approaches (e.g., combined with lipidomics or neuroimaging) showing promise for improving diagnostic precision (Latifi Navid et al., 2025 [1]).

The objective of this study is to identify the most suitable ternary classification strategy for the diagnosis of Alzheimer's disease, improving upon the baseline Random Forest trained on the original dataset. Beyond selecting a single best-performing model, the emphasis lies in developing a rigorous comparison framework that enables coherent and fair evaluation of different approaches. This framework explicitly accounts for the limitations of the dataset, such as its relatively small size and class imbalance, ensuring that the reported results reflect robust trends rather than artifacts of the evaluation process. The goal is therefore to provide a reliable indication of model validity for the task, rather than prescribing a specific model configuration for immediate practical deployment.

2 Dataset Description

The dataset is obtained from different files, containing information regarding 1256 patients. In particular, we will extract information from the following files:

- *df_concat_final.csv*: contains miRNA values collected for each patient, with some additional personal data
- *subj_embeddings_train.csv*: contains miRNA feature embeddings (128 features per patient)

The starting point will be a dataframe whose dimension are (1256 x 134), where the rows represent the patients while the columns represent both the miRNA embeddings and some personal info (country, age, sex, APOE4, disease).

In analyzing the target variable, it is important to note that the dataset is unbalanced. Specifically, 67% of the instances correspond to Alzheimer's Disease (AD), 24% to Normal Controls (NC), and only 9% to Mild Cognitive Impairment (MCI), a distribution that may influence model performance and must be considered when designing the classification strategy.

3 Preprocessing

The preprocessing pipeline began by merging patients' personal information extracted from the `df_concat_final.csv` file with the 128 miRNA-based biometric features, resulting in an initial dataframe of size 1256×134.

Data cleaning included the removal of 29 rows with invalid values in the `age`, `sex`, or `apoe4` features, as well as the elimination of the patient ID and the `country` attribute due to its very low variance.

To ensure comparability among features and avoid scale-related bias, all miRNA features were standardized using z-score normalization, which also facilitates downstream PCA, a variance-based technique sensitive to feature scale. Dimensionality reduction was then performed through PCA, projecting the 128 correlated features into 20 principal components that collectively explain approximately 99% of the total variance. This approach reduces the risk of overfitting, simplifies the feature space, and improves both computational efficiency and model stability, while retaining almost all the original information.

4 Methodology: Defining Models

4.1 Standardization of training procedures

The objective of this study is twofold:

- Highlight the importance of the pre-processing carried out, improving performance compared to the model in the previous study, trained on data processed in a naive manner.
- Identify model types that can improve baseline results solving the task.

To ensure a fair and rigorous comparison among the different models, we adopted the training strategy defined by the `evaluate_model` function. This function implements a nested cross-validation scheme based on a `RepeatedStratifiedKFold` with 5 folds and 5 repetitions, guaranteeing both class balance across folds and robustness of the estimates. Within each outer split, model selection was carried out through a `GridSearchCV` using a 3-fold inner cross-validation and optimizing for the `f1_macro` score. In this setup, the inner loop ensures that hyperparameters are tuned on data disjoint from the evaluation set, while the outer loop provides an unbiased assessment of generalization performance.

Alternative validation strategies, such as a simple hold-out set or a single k-fold CV with embedded grid search, could have been considered. However, these methods either waste too much data in the case of small datasets (hold-out) or introduce optimistic bias in performance estimation (single k-fold). Given the relatively limited size of our dataset ($\sim$1,200 samples) and the complexity of the models under evaluation, nested cross-validation with repeated stratification represented the most reliable and necessary choice, enabling robust hyperparameter tuning and providing a fair and consistent basis for comparing the classifiers.

4.2 Baseline

As a baseline, we adopted a Random Forest classifier trained on the full feature set without dimensionality reduction. The model was configured with 250 trees, unrestricted depth, and the balanced class weight option to address class imbalance. Hyperparameter tuning was intentionally kept minimal, with a fixed parameter grid, to ensure that this baseline primarily reflects the inherent performance of the Random Forest rather than the effect of extensive optimization.

The same structure was adopted to adapt to the two training cases on the original dataset and on the optimized dataset. They were then compared in a preliminary evaluation phase.

4.3 Enhanced Random Forest

Random Forest was considered a valid choice for this task because of its robustness in handling heterogeneous data, its ability to manage non-linear relationships, and its natural support for multi-class classification.

In this implementation, hyperparameters explored through grid search included maximum depth, minimum samples per split, minimum samples per leaf, and number of features considered at each split.

A specific adjustment compared to the baseline was the inclusion of a resampling strategy combining random oversampling and undersampling to mitigate class imbalance. This adjustment was applied only to the training folds within cross-validation, ensuring that model comparisons remained fair while improving the classifier's ability to learn minority classes.

4.4 XGBoost

XGBoost was selected as a candidate model due to its effectiveness in handling complex classification tasks, especially in imbalanced and heterogeneous datasets. Its ensemble boosting strategy often provides higher predictive power compared to bagging methods like Random Forest.

The grid search explored hyperparameters such as number of estimators, maximum tree depth, learning rate, subsampling ratio, and column sampling ratio.

A distinctive adjustment for this model was the use of class weighting, which compensates for the imbalance between classes by assigning higher penalties to misclassified minority samples. This mechanism is particularly valuable in this ternary classification task, as it allows the model to avoid bias toward the majority class and improves overall macro-level performance.

4.5 SVM + RBF

SVM was included as a candidate model because of its robustness in high-dimensional settings and its ability to handle heterogeneous data distributions. Given the ternary classification problem, only the RBF kernel was considered, as it is particularly suitable for capturing non-linear relationships between features—unlike linear or polynomial kernels, which would likely underperform in this context.

The grid search focused on the regularization parameter (C) and the kernel coefficient (γ), both crucial in controlling the decision boundary's flexibility and the model's generalization.

A specific adjustment was the inclusion of a feature standardization step (z-score normalization), applied through a preprocessing pipeline. This step is essential for SVM since its optimization is sensitive to feature scales. Moreover, class weighting was applied to mitigate imbalance across classes, ensuring fairer treatment of minority classes in the classification process.

4.6 Hierarchical strategy

The hierarchical model was designed for the ternary classification task in a structured two-phase approach. Rather than addressing all three classes simultaneously, the first phase separates NC from diseased subjects (AD/MCI), and the second phase distinguishes between AD and MCI, reflecting clinical reasoning and simplifying each sub-task.

XGBoost was chosen for the first phase due to its robustness with heterogeneous and imbalanced data, while Random Forest was used in the second phase to capture complex feature interactions in AD vs. MCI. Each sub-model was tuned independently via nested cross-validation, optimizing key hyperparameters for each algorithm.

Predicted probabilities from both phases were combined such that the AD/MCI probability was conditioned on first being classified as diseased, ensuring probabilistic consistency. This hierarchical strategy, although more complex than single-model approaches, leverages the problem's inherent structure to potentially improve classification performance.

5 Evaluation

The evaluation of the classification models was performed using a two-step strategy. First, we collected all metrics obtained from the nested cross-validation for each model, reporting *mean* ± *standard deviation* to provide a general overview of performance. In a second step, statistical tests were applied to assess whether differences between models were significant. Specifically, an *independent t-test* was used to compare the baseline model on the original dataset versus the baseline on the PCA-reduced dataset. For all other pairwise model comparisons, we employed a combination of *paired t-tests* and *Wilcoxon signed-rank tests*, ensuring a robust assessment that accounts for possible dependency among folds in nested cross-validation.

Selected metrics. Three main metrics were selected as central to evaluation:

- **Balanced Accuracy:** particularly relevant for multiclass classification with imbalanced classes, as it gives equal weight to each class.

- **F1-Macro:** captures the trade-off between precision and recall across classes, averaging the scores to reflect overall performance.
- **PR-AUC Macro:** evaluates the precision-recall curve for each class, providing a more nuanced measure than accuracy when dealing with class imbalance.

These metrics were chosen because they provide complementary perspectives on classification performance, especially in a ternary, imbalanced setting.

Table 1: Model performance with mean ± std for different metrics.

Model	Balanced Acc.	F1-Macro	PR-AUC	Precision	Recall
Original	0.581 ± 0.039	0.629 ± 0.040	0.700 ± 0.046	0.824 ± 0.038	0.581 ± 0.039
Baseline	0.626 ± 0.040	0.684 ± 0.040	0.731 ± 0.046	0.836 ± 0.031	0.626 ± 0.040
Hierarchical XGB+RF	0.653 ± 0.042	0.710 ± 0.041	0.739 ± 0.043	0.837 ± 0.024	0.653 ± 0.042
RF resampling	0.671 ± 0.043	0.716 ± 0.040	0.730 ± 0.044	0.811 ± 0.025	0.671 ± 0.043
SVM RBF	0.633 ± 0.039	0.692 ± 0.039	0.710 ± 0.037	0.840 ± 0.021	0.633 ± 0.039
XGB	0.685 ± 0.038	0.722 ± 0.037	0.743 ± 0.040	0.792 ± 0.035	0.685 ± 0.038

Results overview. From the averaged metrics across folds, all tuned models improved over the baseline.

Independent t-tests confirmed that PCA-reduction significantly enhanced performance across key metrics—balanced accuracy, F1-macro, and PR-AUC macro (e.g., F1-macro: $t = 4.888$, $p = 0.000012$). PCA reduced 128 biometric features to 20 components, retaining 99% of variance, simplifying the feature space, and potentially improving model robustness.

Comparisons between baseline and other models reveal:

- **Random Forest with resampling**: consistently outperformed baseline in balanced accuracy and F1-macro (e.g., balanced accuracy: $t = -10.789$, $p \approx 0$; $W = 0$, $p \approx 0$), effectively handling class imbalance through combined over- and under-sampling. PR-AUC macro gains were not statistically significant.
- **SVM with RBF Kernel**: showed modest improvements, not statistically significant across all metrics (e.g., F1-macro: $t = -1.313$, $p = 0.202$; $W = 112$, $p = 0.182$). Its strength is capturing non-linear decision boundaries, though performance varied across folds.
- **XGBoost with class weighting**: top-tier performance, with significant improvements over baseline across most metrics (e.g., F1-macro: $t = -8.514$, $p \approx 0$; $W = 6$, $p \approx 0$). Class weighting improved detection of minority classes; PR-AUC macro gains were modest but significant.
- **Hierarchical classifier**: combines XGB for AD/MCI vs NC and RF for AD vs MCI. Significant improvements were observed for most metrics (e.g., balanced accuracy: $t = -5.710$, $p < 0.00001$; $W = 19$, $p \approx 0$). PR-AUC macro gains were slightly lower than XGB, reflecting some variability in the second-step predictions.

Direct comparison between XGB and Hierarchical models using paired t-tests and Wilcoxon tests indicates that XGB slightly outperforms Hierarchical in F1-macro and balanced accuracy (e.g., F1-macro: $t = -3.213$, $p = 0.0037$; $W = 64$, $p = 0.0067$). Comparing the two models, XGB proves to be the most reliable model, with a reasonable level of accuracy even when representing less frequent classes, while the hierarchical strategy, despite generally positive results, is somewhat more limited in performing the same task.

Overall, this evaluation provides a quantitative comparison of average performance and a statistically sound assessment of whether observed differences are meaningful.

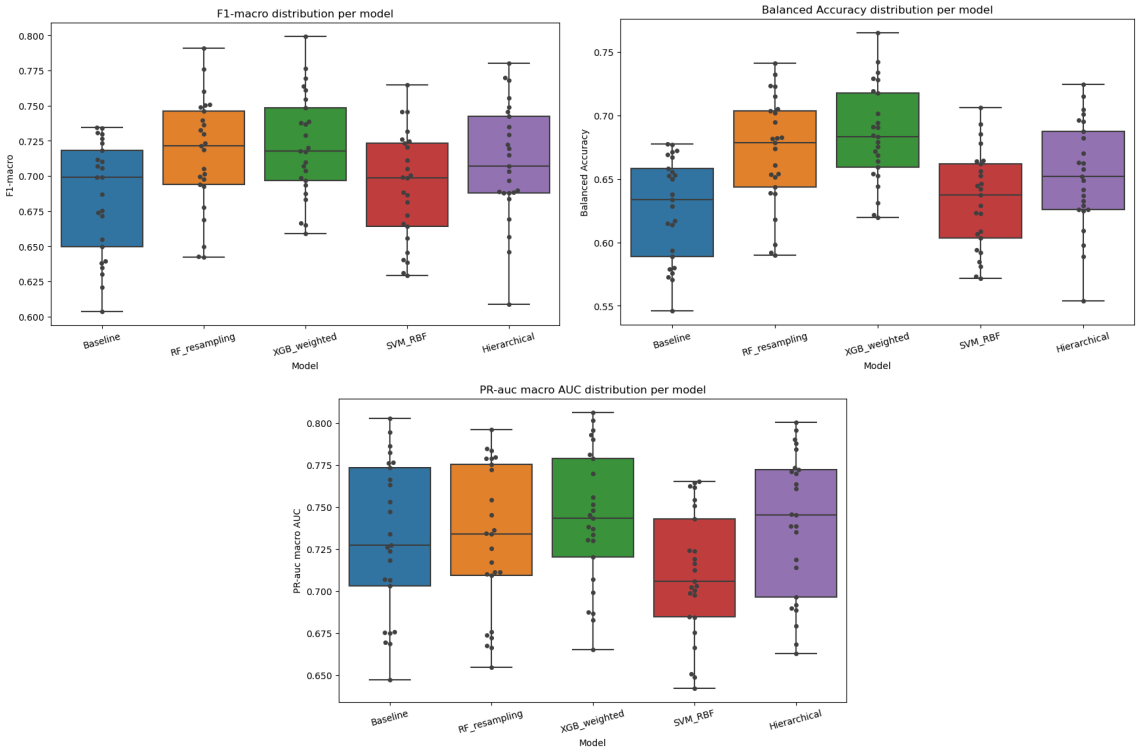

Figure 1: Boxplots showing model performance across main metrics.

6 Conclusion

The results indicate that XGBoost with class weighting consistently achieved the best overall performance across multiple metrics (balanced accuracy, F1-macro, and PR-AUC), while also displaying a relatively balanced ability to handle all three classes, as reflected in its average confusion matrix. In contrast, the Hierarchical XGB+RF model, despite reaching competitive aggregate scores, revealed less uniform class-wise performance, particularly showing weaker recall in the minority classes. RF with resampling also performed competitively, further highlighting the effectiveness of tree-based ensemble methods for this task. Statistical comparisons confirmed that XGB significantly outperformed both the baseline and SVM in most cases, while differences between the hierarchical strategy and other ensemble methods were less clear-cut. Taken together, these findings suggest that weighted XGB is the most robust candidate for further fine-tuning or deployment, whereas hierarchical strategies may require additional refinement to achieve consistent improvements.

Future work could explore advanced feature selection methods beyond PCA or hybrid hierarchical approaches integrating clinical or genetic information, potentially enhancing both interpretability and predictive accuracy.

References

[1] H. Latifi-Navid, S. Mokhtari, S. Taghizadeh, F. Moradi, D. Poostforoush-Fard, S. Alijanpour, and M.-R. Aghanoori. Ai-assisted multi-omics analysis reveals new markers for the prediction of ad. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, page 167925, 2025.

[2] I. Petracci, S. Bellini, K. Goljanek-Whysall, L. R. Quinlan, A. Fiszer, A. Cakmak, C. M. Njume, B. Borroni, and R. Ghidoni. Exploring the role of micrornas as blood biomarkers in alzheimer's disease and frontotemporal dementia. *International Journal of Molecular Sciences*, 26(7):3399, 2025.

[3] M. Zhou and X. Pang. Polyphenols and mirna interplay: a novel approach to combat apoptosis and inflammation in alzheimer's disease. *Frontiers in Aging Neuroscience*, 17:1571563, 2025.