Medical Diagnosis for Alzheimer's disease using Diverse Classification Methods

Goilean Giulia Carla Alexa
Advanced Methods in Data Analysis – Report 2
Group: 246/1
giulia.goilean@stud.ubbcluj.ro

Abstract

This study examines the application of diverse machine leaning paradigms for early detection of Alzheimer's disease, comparing the performance of linear models (logistic regression), ensemble learning (random forest) and deep learning (multi-layer perceptron). A clinical dataset of 2149 patients records characterized by 35 features was used for all three experiments. The models were evaluated using the same data cleaning and preprocessing pipeline, including Synthetic Minority Over-sampling Technique to address class imbalance and stratified cross validation. The results demonstrate that the Random Forest classifier has a significantly better performance, a ROC-AUC score of 0.9422 with a great balance between recall and precision. The MLP method struggled with misdiagnosing healthy patients, while logistic regression shows a great interpretable baseline.

1. Introduction

In the past few years, the rapid evolution of artificial intelligence has introduce its use in all different sectors and domains in our society. The integration of artificial intelligence into healthcare has shifted from experimental research to practical applications, promising a change towards high performance medicine [Top19]. Research and reviews highlight that AI algorithms can achieve or even surpass physician level accuracy in diagnostic tasks, from dermatology to radiology, offering the potential to reduce human error and improve the outcome for patients [Top19]. Alzheimer's Disease is a neurodegenerative disorder and given its progressive nature, early detection is essential for effective intervention and patient management. However, medical datasets can present a lot a challenges for machine learning algorithms: noise or severe class imbalance.

For this study, a dataset containing patient details for Alzheimer diagnosis was used. For medical data analysis, choosing the right classification algorithm is a complex decision, because it needs to have a balance between performance and interpretability. The goal of this study is to perform a comparative analysis of three advanced classification methods applied to the same dataset [Kha24]. The study covers three major families of algorithms: logistic regression for the baseline, random forest and multi-layer perceptron. While linear models offer transparency, they could fail to capture complex relationships between data. On the other hand, deep neural networks offer high capacity, but require careful tuning and large amounts of data to generalize in an efficient way.

The experiments follow a systematic procedure: data cleaning, resolving the issue of class imbalance and model specific preprocessing. The report follows and compares performances, understanding the advantages and disadvantages of each model for the medical diagnosis problem.

2. Theoretical Background

For this report three classification algorithms were chosen to test on the same dataset. Also, a data balancing technique was employed in the study. The selection of methods covers three distinct paradigms: linear probabilistic modeling (Logistic Regression), ensemble learning (Random Forest) and deep learning (a Multi-Layer Perceptron).

2.1. Logistic Regression

Logistic Regression can be considered a robust baseline for binary classification tasks, especially in the case of medical statistics where interpretability is crucial. In contrast with linear regression (which predicts a continuous outcome), logistic regression models the probability that a given instance belongs to a specific class. The model uses the logistic (sigmoid) function (as seen in figure 1) to map the output of a linear equation to a probability value between 0 and 1 [Has09]. The central premise is to model the log-odds (logit) of the probability p of the positive class. The parameters β are typically estimated using Maximum Likelihood Estimation (MLE). Logistic Regression is computationally efficient and provides interpretable coefficients indicating the direction of influence for each feature, it relies on the assumption of a linear decision boundary, which could limit its performance on complex and non linear data [Has09].

$$\log \frac{\Pr(G = K - 1|X = x)}{\Pr(G = K|X = x)} = \beta_{(K-1)0} + \beta_{K-1}^{T} x.$$

Figure 1. *Example of logistic regression model's form [Has09]*

2.2. Random Forest

The Random Forest algorithm was firstly introduced to address the limitations of single decision trees, as they struggled with high variance and the tendency to overfit [Bre01]. Random Forest is an ensemble learning method that operates on the principle of bagging (bootstrap aggregation). The algorithm constructs a number of decision trees during training. Each tree is built on a random subset of the training data (this is called bootstrap sample) and at each node split, only a random subset of features is taken into consideration. The final classification is determined by the majority vote of the individual trees. This randomization separates the trees, significantly reducing the variance of the model without increasing the bias. For medical datasets, random forests are a great suit because of their robustness to noise, they do not require feature scaling and they can capture complex, non-linear dependences between clinical variables [Bre01].

2.3. Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron represents the foundational architecture of Deep Learning. An MLP is defined as a feedforward artificial neural network consisting of at least three layers of nodes: an input layer, one or more hidden layers and an output layer [Goo16]. MLPs can break the limitations of linear models by introducing non-linearity through activation functions (e.g. ReLU) applied to the neurons in the hidden layers. The network learns a mapping $y = f\ (x;\ \theta)$ by adjusting the weights θ to minimize a loss function using the backpropagation algorithm. In theory, MLPs have a high capacity to approximate any continuous function, however they are sensitive to the scale of input data and require careful hyperparameter tuning to prevent overfitting [Goo16].

2.4. SMOTE

Class imbalance is a frequent issue for medical datasets, where the number of healthy controls can outnumber the diagnosed patients. Standard learning algorithms tend to be biased towards the majority class to maximize overall accuracy. To overcome this, the experiments in this study use Synthetic Minority Over-sampling Technique (SMOTE) [Cha02]. SMOTE generates new, synthetic instances. It does this by selecting a minority sample, finding its k-nearest neighbors in the feature space and creates a new point along the line segment joining the sample and one of its neighbors. This approach expands the decision region of the minority class without causing the overfitting typical for simple replication.

3. Proposed procedure

3.1. Dataset Analysis

For this study, the chosen dataset is Alzheimer's Disease Dataset [Kha24]. This contains a collection of health records designed to help in the predictive modelling of neurodegenerative disorders. The medical domain that explores factors associated with Alzheimer's is still very researched and of interest for scientist, so predictive models can represent a great tool to conduct statistical analyses. The dataset contains 2149 patient records, each characterizes by 35 features covering demographics, lifestyle, medical history and clinical assessments. The features are categorized into five domains, showing a great view of the patient's health status. The target variable (Diagnosis) is a binary label where 0 indicates a healthy control and 1 indicates the presence of Alzheimer's disease.

3.1.1. Class Distribution and Imbalance

A quantitative analysis of the target variable highlights a moderate class imbalanced. As shown in figure 2, the dataset includes 1389 healthy patients (around 64%) and 760 patients diagnosed with Alzheimer (around 35%). While the imbalanced is not extreme, this could be significant enough to negatively influence the classifiers and to create bias toward the majority class. To avoid this, the experiments used Synthetic Minority Over-sampling Technique (SMOTE) to rebalance the training data, making sure that the models learn to identify the minority class effectively.

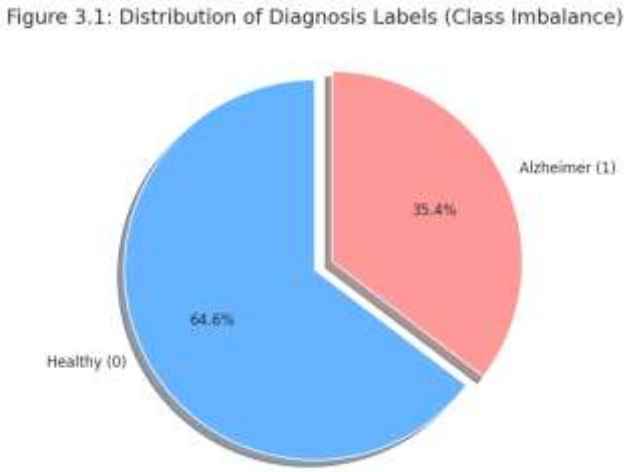

Figure 2. *Distribution of diagnosis labels*

3.1.2. Correlation Analysis

The Pearson correlation analysis was used to identify the most predictive features. The correlation matrix seen in figure 3 shows that the domains of Functional Assessments and Symptoms are the strongest linear predictors of the disease. Interestingly, demographic factors like Age and standard biomarkers like BMI exhibit very weak linear correlations with the diagnosis in this specific dataset, suggesting that the disease's presence is more strongly tied to cognitive decline and symptomatic presentation than to general physical metrics.

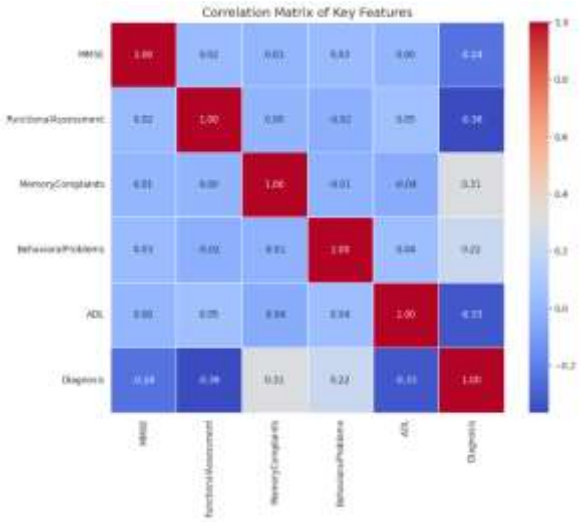

Figure 3. *Correlation Matrix of Key Features with Diagnosis*

3.2. Data Preprocessing

A preprocessing pipeline for the dataset was followed in order to ensure that the classifiers operate on an optimal input representation. The process included handling missing values, categorical encoding and feature scaling.

Firstly, non-predictive features such as PatientID or DoctorInCharge were removed to reduce noise. Missing values were handled using the SimpleImputer from sklearn library. For the numerical features missing values were filled using the median value of the column, as it should be robust against outliers. For categorical features, the missing entries were imputed using the most frequent value.

Categorical features were initially provided in a label encoded format, for example the EducationalLevel column had labels from 0 to 3. Algorithms like Logistic Regression and MLP might interpret these labels as having an ordinal relationship, assuming that for Ethnicity 3 is mathematically greater than 1, which is biologically invalid. So, treating them as numerical inputs creates a risk for linear models. To resolve this, one-hot encoding was applied to all categorical variables. This transformation creates binary "dummy" variables, like Ethnicity_0 and Ethnicity_1, making sure that the models treat each category as independent.

Another important step in the preprocessing pipeline was feature scaling. This step was customizes depending on each models' needs. For the Logistic Regression and MLP experiments standardization was applied (the StandardScalar method from sklearn library). This process transforms all numerical features to have a mean of 0 and a standard deviation of 1. This step is mandatory for gradient-based algorithms [Has09], because features with large values (like Cholesterol: 200) would otherwise dominate the objective function compared to features with smaller ranges (Clinical Scores: 0-10). For the Random Forest experiment feature scaling was omitted. In contrast with the other two experiments, this model process data by learning cut-off points based on the actual values, meaning it is not sensitive to the scale of numbers [Bre01]. Moreover, using the raw data allows clear decision rules and preserves the interpretability of the model.

3.3. Class Imbalance

In the dataset analysis step, a slight imbalance in the distribution of the target labels was observed. This could make the classifiers bias their predictions towards the majority class. To avoid this issue, SMOTE was applied in all three experiments [Cha02]. The random over sampling technique duplicates existing minority samples and can lead to overfitting. In comparison, SMOTE creates new instances by interpolating between existing minority samples in the feature space. SMOTE was used

with the help of imbalanced-learn library, ImbPipeline. This procedure ensures that SMOTE was applied exclusively to the training folds during cross-validation, so the results are realistic for the models' generalization capability.

3.4. Experimental Design and Configurations

For this study, to achieve a great overview and comprehensive analysis on the performance and capabilities of classifiers in the medical domain, three experiments covering different levels of complexity in machine learning were designed: a liner model, ensemble methods and neural networks. All models were tuned using GridSearchCv with stratified 5-fold cross-validation. The goal was to understand how certain configurations could improve the model's ability to discriminate between healthy and diagnosed patients.

3.4.1. Logistic Regression Experiment

Logistic Regression could be considered a baseline classifier due to its transparency and mathematical simplicity. As a linear model, it allows a direct interpretation of feature influence through regression coefficients. For the medical dataset used in this study this model could highlight the clinical relevance of certain features for the diagnosis.

The model used in the experiments is from the sklearn library. To prevent overfitting, regularization was an important step. Two types were tested: L1 (Lasso) and L2 (Ridge). They represent different types of penalties. L2 shrinks coefficients to prevent overfitting, while L1 regularization forces coefficients of irrelevant features to exactly zero. This allows the model to perform embedded feature selection, automatically removing non-predictive symptoms from the equation. For the regularization strength different four values were tested: 0.01, 0.1, 1.0, 10.0. The smaller the parameter the simpler the model should be. A larger value would allow the model to fit the data more closely. Typically, logistic regression classifies an instance as positive it the predicted probability is more than 0.5. As an experiment, an optimal decision threshold was calculated, because in the case o medical screening a false negative (missing a patient with Alzheimer's) could be more costly than a false positive. So the goal was to maximize the F1-Score or Recall, trading a small amount of precision to achieve a significantly higher sensitivity in detecting the disease.

3.4.2. Random Forest Experiment

Medical data can often involve non-linear and complex interactions, so the random forest method could represent a great solution to those obstacles, by capturing those patters. For the data preprocessing pipeline, scaling was skipped, as it is not an relevant step for this method.

For the algorithm, different hyperparameter optimization strategies were tested. For the ensemble size, forests with 100 and 200 trees were testes. In general, increasing the number of trees should stabilize the prediction error and improve accuracy, but it comes with a computational cost. For the tree depth 10, 20 and None were evaluated. The goal is to identify the best spot between generalization and memorization. The leaf regularization parameter was tuned over 1, 2 and 4. A minimum of 2 or 4 samples per leaf node can prevent the model from creating specific rules for outlier patients. Also, for imbalance to compare and contrast with SMOTE, the configuration class_weight="balanced" was tested. This automatically adjusts weights inversely proportional to class frequencies, effectively penalizing misclassifications of the minority class (Alzheimer's) more severely during the tree building process.

3.4.3. Multi-Layer Perceptron Experiment

For the third experiment, the efficacy of artificial neural networks was tested and observed in the context of medical diagnosis. Neural networks amp inputs to output through continuous, differentiable functions composed of weighted connections. For this experiment, a multi-layer perceptron (MLP) was used, to test whether a deep learning approach could extract non-linear patterns from clinical data in a more productive way in comparison to the ensemble method.

This experiment followed the same data preprocessing steps as the other two, including feature scaling. The backpropagation algorithm relies on gradient descent to minimize the loss function. For this dataset features have vastly different scales, so they could cause serious struggles for the model.

Different hyperparameter configurations were tested. For the network depth a single hidden layer with 100 neurons was tested alongside two hidden layers architectures, one with 50 and 25 neurons and one with 100 and 50. This allowed to explore if the problem could benefit from hierarchical features extractions, where deeper layers aggregate simple features into complex abstract representations. For the activation functions ReLU (Rectified Liniar Unit) and Tanh were used and compared. The regularization step was done using the L2 penalty parameter to prevent the weights from growing too large. As the dataset has a smaller size, the risk of overfitting was significant. To avoid this, the Adam adaptive learning rate optimizer was used, along with and early stopping mechanism. This technique sets aside 10% of the training data as an internal validation set. The training process is automatically stopped if the validation score does not improve for a specified number of consecutive epochs. This ensures the model stops learning exactly at the point of optimal generalization, before it begins to memorize the training noise.

4. Results and Analysis

4.1. Used Evaluation Metrics

In order to properly understand the performance and results of each model, different types of evaluation metrics were used. For the Alzheimer's Disease dataset the first challenge was its class imbalance, so relying only on the accuracy alone would have been insufficient and potentially misleading. The analyses focuses not only on raw performance metrics, but also on the interpretability of the models and their clinical applicability.

The baseline for each evaluation in all three experiments was the confusion matrix. It categorizes predictions into four distinct outcomes: True Positive (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). It is a great tool that shows exactly how and in which was the model makes mistakes.

Accuracy represents the ratio of correctly predicted observations to the total observations, as seen in figure 4. It can be a useful general indicator for the performance, however it can be biased towards the majority class in imbalanced datasets, so the accuracy could be high even if the model struggles to correctly classify the diagnosed patients [Pow11].

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP}$$

Figure 4. *Accuracy formula*

To have a better view on the models' performance recall (sensitivity) and precision were also used. Their formulas can be seen in figure 5. Recall is the most critical metric for medical screenings [Pow11], because a low value shows a high rate of false natives, meaning that patients with Alzheimer's' are missed, so they lose on important time for an intervention. A high precision indicates that patients are rarely misdiagnosed. This is also important because being wrongfully diagnosed with a disease could be a very costly and upsetting for healthy patients. Moreover, F1-Score, the harmonic mean of precision and recall, was calculated, representing a balanced view on the model's performance and results.

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP}$$

Figure 5. *Recall and Precision formulas*

The ROC curve plots the true positive rate (Recall) against the false positive rate at various settings. The Area Under the Curve (AUC) represents the probability that the classifier will rand a randomly chosen positive instance higher than a randomly chosen negative one. In comparison with accuracy, AUC is independent of the classification threshold, making it an ideal metric for model selection in probabilistic classifiers [Pow11].

4.2. Logistic Regression Experiment Results

For this study, the logistic regression model was the baseline. Its performance was evaluated with the previously mentioned metrics, for different settings and hyperparameters.

For the first experiment, L2 regularization was used with four different strength parameters. In this process the strength parameter 1.0 was identified as optimal, as seen in figure 6. However, the mean AUC score between 0.1 and 1.0 is negligible. This suggest that the model is robust and not too sensitive to the exact regularization strength within this range. In contrast, for a strong regularization the performance drops noticeably, revealing that to much constraint makes the model underfit the data.

Param-clf-C	Mean ROC-AUC	Std Dev	Rank
1.00	0.901012	0.013835	1
0.10	0.900808	0.013960	2
10.00	0.900798	0.014007	3
0.01	0.891583	0.015157	4

Figure 6. *Comparison between regularization settings for Logistic Regression*

4.2.1. Quantitative Performance

Moving forward, for the optimal model (C = 1.0) the performance was also evaluated for the held-out test set. The model achieved a ROC-AUC of 0,8858 as seen in figure 7, proving a strong discriminative ability. This suggests the risk factors together create a strong linear component in predicting the diagnosis. The model has a recall of 0.8355, a direst result of SMOTE, because now the priority is to find the minority class. In comparison, precision is much lower: 0,6940, indicating that the model is sometimes over cautious, so it flags healthy patients that present borderline symptoms. This behavior is also shown by the confusion matrix in figure 8, where the number of healthy patients classified as diagnosed is larger than the Alzheimer's patients labeled as healthy.

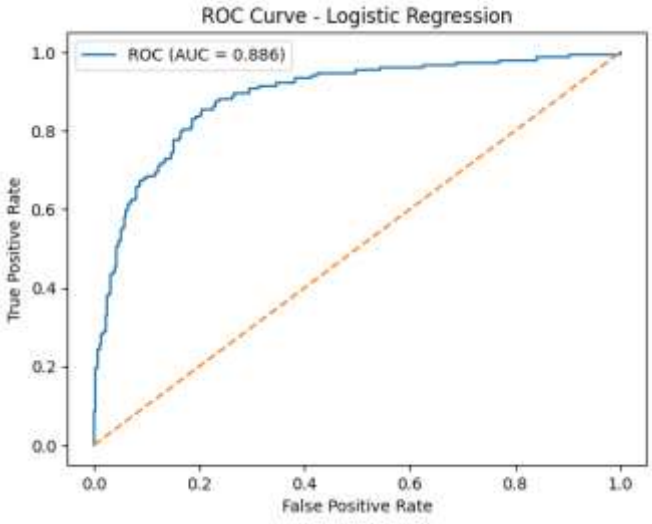

Figure 7. *ROC Curve for Logistic Regression model*

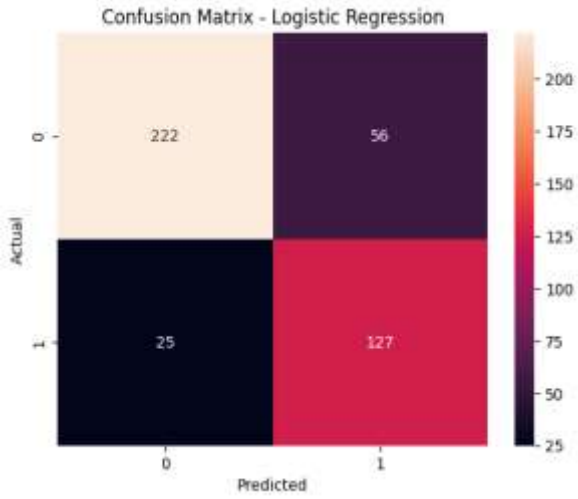

Figure 8. *Confusion Matrix Logistic Regression*

4.2.2. Feature Interpretability and Sparsity

To be able to validate the clinical relevance of the model, the regression coefficients were analyzed. As visualized in figure 9, the model found the predictors that show an increasing risk of Alzheimer's: memory complaints, behavior problems or family history. Furthermore, another experiment regarding L1 and L2 regulation was done. The L1 regularized model worked the best with 0.1 and automatically reduced 33 out of 53 feature coefficients to exactly zero without a major loss in accuracy. This suggests that a panel involving only the top 20 markers could be sufficient for an effective screening.

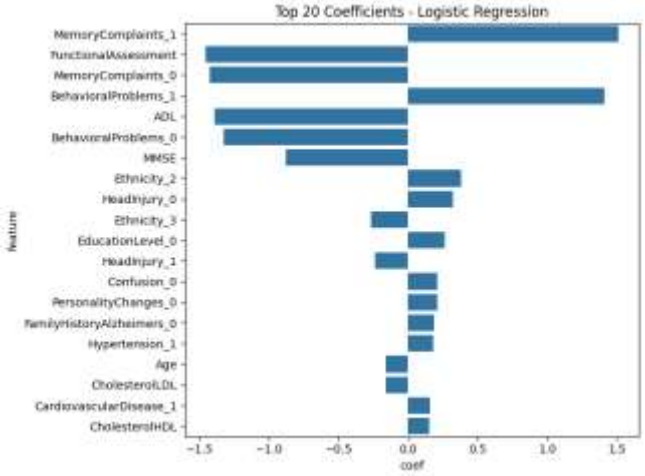

Figure 9. *Top 20 Coefficients Logistic Regression*

4.2.3. Clinical Threshold Optimization

For the last experiment involving this model, the medical priority of minimizing false negatives was addressed. The standard threshold of 0,5 created a recall of 83,55%. The precision-recall curve and the optimal F1-score threshold was calculated, as seen in figure 10, so the cut-off became 0,49. The adjustment increased the recall to 85,53%, and this could be a success for all at risk patients that could have been missed.

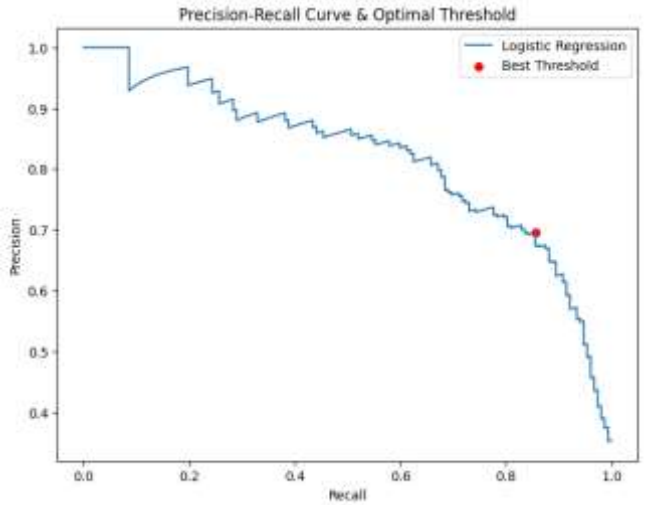

Figure 10. *Precision-Recall Curve*

4.3. Random Forest Experiment Results

The Random Forest experiment aimed to capture the non-linear interactions within the clinical data. The model demonstrated superior performance in comparison to the logistic regression model,

used as the baseline model. This proves the initial assumption about the dataset: the decision boundaries in Alzheimer's diagnosis are complex and hierarchical.

4.3.1. Hyperparameter Stability and Tuning

The GridSearchCV showed a great overall stability for the architecture. As illustrated in figure 11, the top 15 model configurations achieved ROC-AUC scores within a very narrow range (0,954 – 0,955). This proves that the ensemble method is highly robust to hyperparameter variations, reducing the risk of overfitting to a specific configuration. The optimal configuration was a forest with 200 trees, balanced class weights and a minimum leaf size of 4 samples. This setup prioritizes generalization and penalizes errors on the minority class.

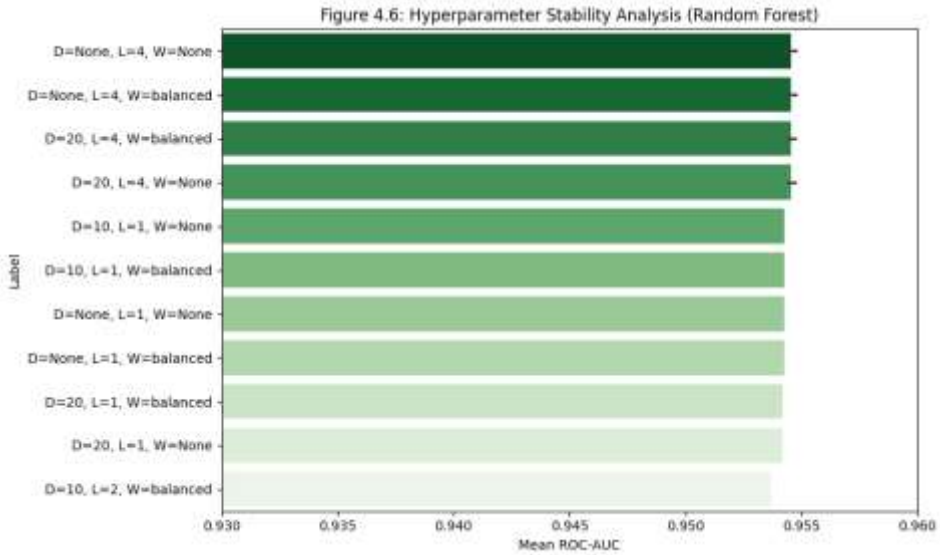

Figure 11. Hyperparameter Stability for Random Forest

4.3.2. Quantitative Performance Analysis

When classifying on the test set, the model achieved a ROC-AUC of 0,9422 (seen in figure 12), a significant improvement over the logistic regression experiment. The model had a recall of 91,45% and precision of 92,05%. This makes the model a highly reliable screening tool and it also reduces false alarms in an important percent. The confusion matrix from figure 13 indicates that the model rarely confuses the two classes.

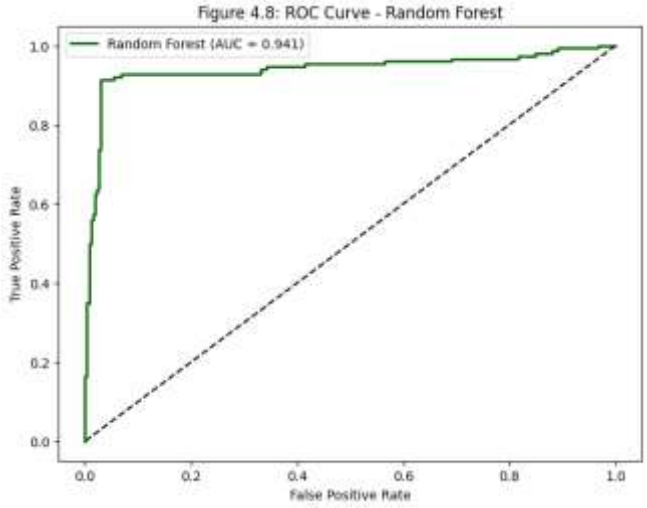

Figure 12. *ROC Curve- Random Forest*

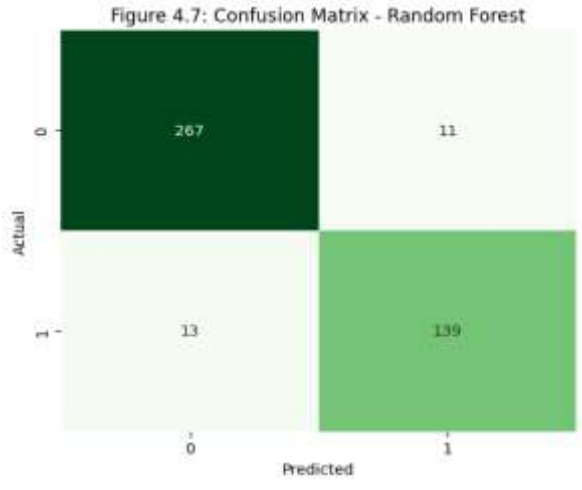

Figure 13. *Confusion Matrix- Random Forest*

4.3.3. Feature Importance Analysis

For this experiment, the regression coefficients that show direction are no longer present, so the future importance seen in figure 14 was based on mean decrease in impurity and it reveals non linear influences of the diagnosis. The analysis shows that activities of daily living and functional assessment are the most critical determinants. This model relies more on objective functional decline to establish decision rules.

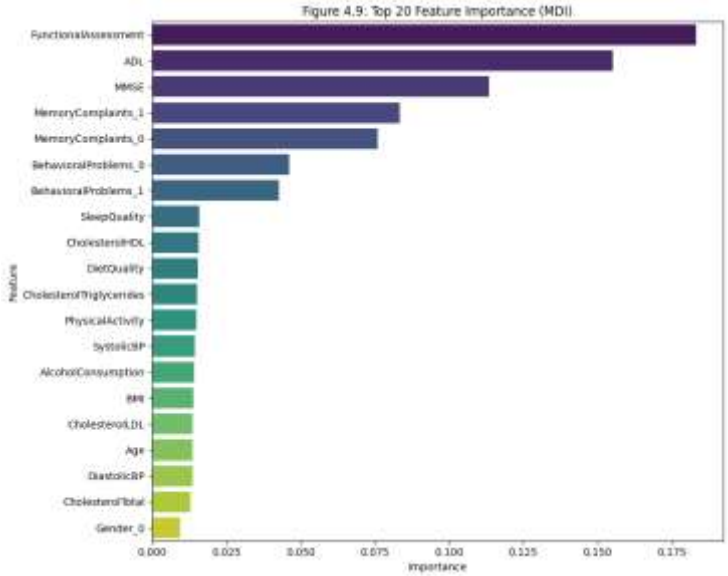

Figure 14. Top Features Importance for Random Forest

4.4. Multi-Layer Perceptron Experiment Results

The third experiment evaluated the performance of a Multi-Layer Perceptron (MLP) to find out if a deep learning approach could extract patterns missed by the ensemble method. The results indicate that MLP is not always the best model, as it can face certain challenges.

4.4.1. Architecture Search and Stability

The GridSearchCV results see in in figure 15 highlight a preference for simpler network architectures. The top performing configuration used a single hidden layer with 100 neurons and the ReLU activation function. When the depth of the network was increased, no significant improvement was seen. This shows that a dataset of moderate size does not have a feature hierarchy deep enough to justify adding complexity to the model.

activation	alpha	Hidden layer	Learning rate	Mean ROC-AUC	Std	rank
ReLU	0.0100	(100,)	0.001	0.901157	0.013697	1
ReLU	0.0001	(100,)	0.001	0.900497	0.013191	2
Tanh	0.0100	(100,)	0.010	0.899896	0.016328	3
Tanh	0.0100	(100,50)	0.001	0.899828	0.013579	4
tanh	0.0001	(100,50)	0.001	0.899761	0.013584	5

Figure 15. *Top 5 configurations MLP experiment*

4.4.2. Quantitative Performance

On the test set, the optimal MLP configuration achieved a ROC_AUC (figure 16) of 0.8940. Initially, the score shows an excellent discrimination, however the other metrics reveal certain issues. The recall is 86,18% and the precision is only 70,81%. Therefore, the MLP has a more aggressive strategy, struggling to differentiate borderline healthy cases leading to a high rate of false positives. The confusion matrix seen in figure 17 confirms the trade off, more healthy patients are misdiagnosed.

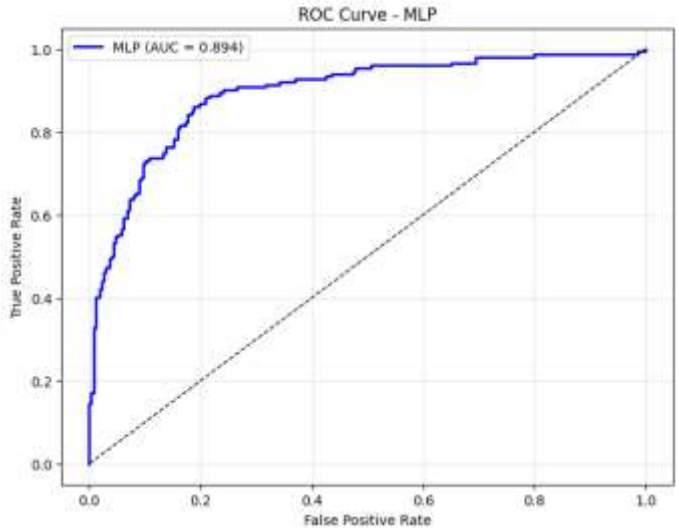

Figure 16. *ROC Curve- MLP*

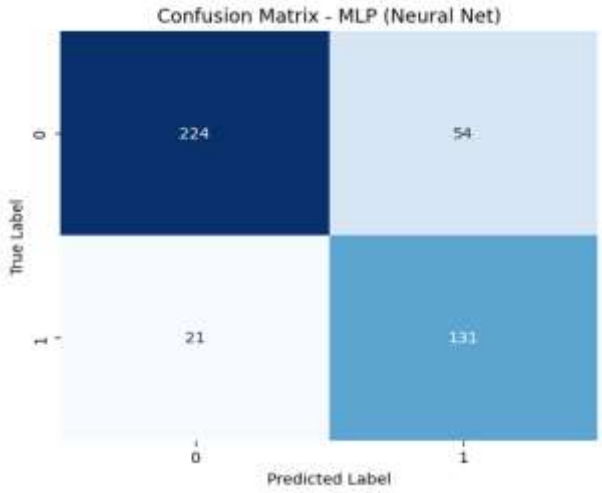

Figure 17. *Confusion Matrix- MLP*

4.4.3. Training Dynamics

In order to understand that the model learned correctly, the learning curve (loss over epochs) was analyzes, shown in figure 18. The curve demonstrates a good convergence profile: an initial drop in loss followed by stabilization. Also, it proves that the early stopping mechanism was successfully implemented and used.

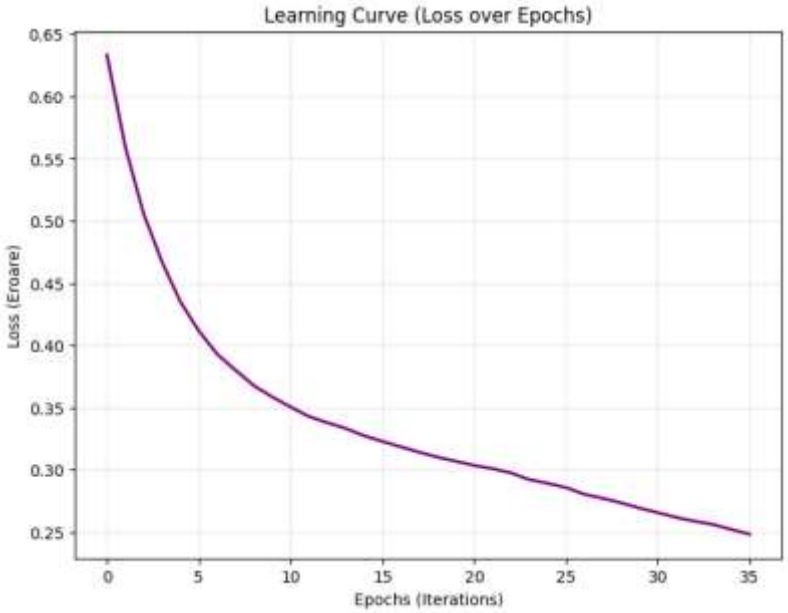

Figure 18. *Learning curve*

4.5. Comparative Analysis between the models

After understanding and analyzing the results of each model, a comparison between them can be made. It reveals that working on the same dataset, each paradigm encounter different types of obstacles. Among the evaluated models, Random Forest classifier was the superior solution, outperforming both the logistic regression model and the MLP across al metrics. It achieved the highest discriminative power with ROC-AUC of 0,9422, while MLP 0,8940 and the baseline model 0,8858. Also, the ensemble methods showed both recall and precision of over 90%. These results are especially important in the healthcare system as misdiagnosing people could have huge repercussions. In contrast, the multi-layer perceptron (MLP) demonstrated a mixed performance. It surpassed the logistic model in sensitivity, but had a much smaller precision. This indicates that the neural network

adopted an overly aggressive classification strategy, and this highlights the struggle of deep learning methods when working with smaller datasets. From a clinical utility and interpretability perspective, logistic regression offers transparency, allowing medical staff to directly validate risk factors through regression coefficients. On the other hand, random forest provides a balanced alternative and a great overall performance. Consequently, all three experiments show different sides of the same problem: finding the right balance between misdiagnosing healthy patients and missing patients with Alzheimer's. Also, each model can be a great choice depending on the scale of the dataset and the users perspective on the interpretability side.

5. Conclusions

This study analyzed the efficacy of three distinct classification algorithms: Logistic Regression, Random Forest and Multi-Layer Perceptron for the prediction of Alzheimer's disease diagnosis. The comparison highlights that algorithm selection in the medical field must balance predictive performance with clinical safety and interpretability.

The Random Forest experiment showed the best results in all teste metrics. Its ability to model non linear data has shined in the medical diagnosis logic. Having a high precision, the negative consequences of misdiagnosing a healthy patient were reduced.

The MLP model outperformed the logistic regression experiment in sensitivity, however it did not match the stability of the ensemble model. It has a high rate of false positive, proving that complex neural architectures struggle to generalize based on datasets with limited size.

The use of artificial intelligence in the medical field is an innovative approach. Medical experts may not be familiar with the technology behind machine learning, so they may prefer a model that provides transparency, like logistic regression. The black box nature of certain model could represent a barrier for medical specialist, as they need to trust and understand the clinical decision making. This emphasizes the need for explainable AI approaches [Hol19].

Acknowledgement:
This work is the result of my own activity, and I confirm I have neither given, nor received unauthorized assistance for this work.
I declare that I did not use generative AI or automated tools in the creation of content or drafting of this document.

REFERENCES

[Bre01] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[Cha02] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[Goo16] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[Has09] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.
[Kha24] R. El Kharoua, "Alzheimer's Disease Dataset", Kaggle Repository, 2024. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset.
[Top19] E. J. Topol, "High-performance medicine: the convergence of human and artificial I ntelligence", Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.
[Pow11] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation", Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.
[Hol19] A. Holzinger et al., "Causability and explainability of artificial intelligence in medicine", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, 2019.