

School of Sciences | Department of Computer Science and Engineering

Replace this with the title of the thesis

Replace this with the subtitle (if applicable, if not delete the line)

Master of Science

in enter your discipline

Enter you name here

© Enter the month and year you are submitting

# Acknowledgements

# Abstract

*Write this last and remember what it is about = To provide a brief statement (no more than 1 page long) regarding the work performed. The statement should not go into too many specifics but should provide the reader with enough information to have a good idea of what the project is about.)*

# Table of Contents

[Acknowledgements 2](#_Toc63096425)

[Abstract 3](#_Toc63096426)

[Table of Contents 4](#_Toc63096427)

[1 Introduction 7](#_Toc63096428)

[1.1 Overview 7](#_Toc63096429)

[1.2 Aims and objectives 7](#_Toc63096430)

[1.3 Structure of thesis 7](#_Toc63096431)

[1.4 Summary 7](#_Toc63096432)

[2 Literature Review 8](#_Toc63096433)

[2.1 Introduction 8](#_Toc63096434)

[2.2 Title 1 8](#_Toc63096435)

[2.3 Title 2 8](#_Toc63096436)

[2.4 Title 3 8](#_Toc63096437)

[2.5 Summary 8](#_Toc63096438)

[3 Methodology 9](#_Toc63096439)

[3.1 Introduction 9](#_Toc63096440)

[3.2 Title 1 9](#_Toc63096441)

[3.3 Title 2 9](#_Toc63096442)

[3.4 Title 3 9](#_Toc63096443)

[3.5 Summary 9](#_Toc63096444)

[4 Results 10](#_Toc63096445)

[4.1 Introduction 10](#_Toc63096446)

[4.2 Title 1 10](#_Toc63096447)

[4.3 Title 2 10](#_Toc63096448)

[4.4 Title 3 10](#_Toc63096449)

[4.5 Summary 10](#_Toc63096450)

[5 Discussion 11](#_Toc63096451)

[5.1 Introduction 11](#_Toc63096452)

[5.2 Title 1 11](#_Toc63096453)

[5.3 Title 2 11](#_Toc63096454)

[5.4 Title 3 11](#_Toc63096455)

[5.5 Summary 11](#_Toc63096456)

[6 Conclusions and Future Work 12](#_Toc63096457)

[6.1 Introduction 12](#_Toc63096458)

[6.2 Title 1 12](#_Toc63096459)

[6.3 Title 2 12](#_Toc63096460)

[6.4 Title 3 12](#_Toc63096461)

[6.5 Summary 12](#_Toc63096462)

[Bibliography 13](#_Toc63096463)

[Appendices 14](#_Toc63096464)

[Appendix A 14](#_Toc63096465)

[Appendix B 15](#_Toc63096466)

[Appendix C 16](#_Toc63096467)

# Introduction

## Overview

Dementia is a progressive and universal condition that primarily affects memory, language, and cognitive function, often leading to significant impairment in daily activities. Among the various causes of dementia, Alzheimer’s disease (AD) is the most common, accounting for over 60% of documented cases globally. Several major risk factors for AD have been identified, including age, family history, genetics, and certain chronic diseases (Alzheimer’s Association, n.d.). At present, over 55 million people worldwide are living with some form of dementia. This number is expected to rise to 78 million by 2030 and to 139 million by 2050. Much of this increase will occur in developing countries, with the fastest growth in the elderly population projected in Asia, particularly in China, India, and their South Asian and Western Pacific neighbours (Alzheimer’s Disease International, n.d.).

Early identification of individuals at risk for Alzheimer’s disease is crucial for timely intervention, improved management, and the development of potential treatments. Currently, the diagnosis of AD typically involves a combination of clinical assessments, neuropsychological evaluations, and advanced neuroimaging techniques, including structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), computed tomography (CT), and positron emission tomography (PET) scans (Akhavan Aghdam et al., 2025). While the raw neuroimaging data generated by these modalities are inherently high-dimensional and complex, recent studies have increasingly focused on extracting quantitative imaging features, such as volumetric and texture-based measurements, to enable machine learning applications in Alzheimer’s disease research. These derived features preserve critical information regarding structural brain changes and have demonstrated promising predictive performance in multiple studies. However, due to the high dimensionality and complex relationships within these datasets, advanced computational approaches are required to identify meaningful patterns and improve early detection accuracy (Leandrou et al., 2023).

To address these limitations, various machine learning algorithms have been increasingly applied in AD research. These algorithms are capable of handling large, high-dimensional datasets and can identify complex, non-linear relationships that may not be evident through conventional statistical approaches. In particular, **Natural Gradient Boosting (NGBoost)** has emerged as a promising probabilistic machine learning algorithm based on the gradient boosting framework. Unlike conventional classifiers that output fixed class labels, NGBoost predicts full probability distributions over possible outcomes, capturing both the predicted value and the associated uncertainty (Duan et al., 2019). This allows NGBoost to model uncertainty in classification tasks, providing more informative and reliable predictions, which is especially valuable in medical decision-making contexts such as AD disease diagnosis. Therefore, the capacity to quantify predictive uncertainty is imperative where early detection of AD is critical yet frequently challenged by overlapping clinical symptoms and heterogeneous patient data.

This thesis aims to explore the capabilities of NGBoost, an algorithm not previously evaluated on the ADNI dataset, and compare its performance with traditional machine learning models in predicting different diagnostic classes. Additionally, the study emphasizes the explainability of these models to better understand and quantify their behaviour, with NGBoost serving as the primary benchmark against which other algorithms are evaluated.

## Aims and objectives

The primary aim of this research is to develop a predictive framework for the early detection of Alzheimer’s disease using advanced machine learning techniques. This study is designed to evaluate the effectiveness of probabilistic and ensemble-based machine learning models in classifying Alzheimer’s disease cases, while also emphasizing model interpretability through explainability tools.

To achieve this aim, the following specific objectives have been established:

1. **To conduct an extensive exploratory data analysis** (EDA) prior to implementing any machine learning models. This is an essential step in understanding the underlying structure and distribution of the dataset, identifying potential anomalies or quality issues, and providing meaningful insights that will inform model selection and evaluation strategies.
2. **To develop and compare** the predictive performance of multiple machine learning algorithms for classifying Alzheimer’s disease into Normal control, Mild cognitive impairment, and Alzheimer’s disease stages. These include:
   1. Natural Gradient Boosting (NGBoost), a probabilistic machine learning algorithm capable of producing predictive distributions and quantifying uncertainty.
   2. XGBoost and Random Forest, two widely used ensemble learning methods known for their high predictive performance in various classification tasks.
   3. Logistic Regression, a simple, interpretable baseline model used to assess whether the adoption of more complex algorithms is justified based on performance improvements.
3. **To evaluate the predictive performance** of the developed models and apply machine learning explainability techniques to interpret and explain the models’ decision-making processes. This objective aims to enhance the transparency and clinical applicability of the predictive models by identifying the key features influencing classification outcomes.

## Structure of thesis

The thesis is structured as follows:

1. **Introduction**: outlines the research background, focusing on the challenges of early Alzheimer’s Disease (AD) detection and the need for explainable machine learning solutions in clinical practice. It highlights the significance of the ADNI-1 dataset and states the study’s objectives.
2. **Literature Review**: examines existing research on machine learning applications for Alzheimer’s diagnosis using clinical and neuroimaging data. It identifies gaps in model interpretability within healthcare and reviews recent developments in explainability methods, with emphasis on SHapley Additive exPlanations (SHAP).
3. **Theoretical Background**: Introduces key theoretical concepts underpinning the study. It covers the principles of supervised machine learning, classification algorithms relevant to clinical prediction tasks, and the mathematical foundation of explainability techniques such as Shapley values and SHAP. This chapter establishes the theoretical framework guiding the methodological choices made in the study.
4. **Methodology**: presents the ADNI-1 dataset, describing its clinical and cognitive assessment variables, along with the preprocessing steps undertaken. It details the exploratory data analysis, model selection, evaluation strategy, and the integration of SHAP for interpreting model predictions.
5. **Results**: reports the outcomes from exploratory analyses, model performance assessments, and SHAP-based explanations. Key predictive features are identified, offering clinical insights into factors linked to AD progression.
6. **Discussion**: contextualizes the findings in relation to existing literature and clinical understanding. It reflects on the implications of combining predictive models with explainability tools, addresses limitations, and suggests directions for future work.
7. **Conclusion and Future Work**: summarizes the main contributions and emphasizes the importance of interpretable machine learning in Alzheimer’s research. It proposes potential extensions, including testing alternative models, applying additional interpretability techniques, and validating results on external datasets.
8. **References**: : Lists all academic literature, datasets, and software tools cited throughout the thesis.
9. **Appendices**: Provides additional figures, tables, technical details, and supplementary materials that support the main content of the thesis.

## Summary

This introductory chapter established the context and motivation for the study, highlighting the growing global burden of dementia, particularly Alzheimer’s Disease (AD). It emphasized the importance of early detection for improving patient outcomes and discussed the current diagnostic approaches, including clinical assessments and advanced neuroimaging techniques, while noting the analytical challenges posed by high-dimensional neuroimaging data.

To address these challenges, the chapter introduced the role of machine learning, particularly ensemble-based and probabilistic algorithms, in uncovering complex patterns within such data. The Natural Gradient Boosting (NGBoost) algorithm was identified as a promising approach due to its capacity to model predictive uncertainty, an essential feature in clinical decision-making contexts.

This thesis will also present an extensive comparison of NGBoost with traditional classifiers such as XGBoost, Random Forest, and Logistic Regression, using a subset of the original ADNI dataset to differentiate between diagnostic groups. Finally, it aims to enhance the interpretability of these models using explainability techniques, specifically SHapley Additive exPlanations (SHAP), to better understand the factors influencing classification outcomes.

# Literature Review

## Introduction

AD is a progressive neurodegenerative disorder affecting millions of people globally. Its early stage, mild cognitive impairment (MCI), indicates a substantially elevated risk of AD conversion (Bloch & Friedrich, 2021). Accurate early detection where distinguishing cognitively normal controls (CN) from MCI or AD is therefore a critical research priority. To this end, the Alzheimer Disease Neuroimaging Initiative (ADNI) offers a rich, longitudinal, multimodal data set, including clinical tests, genetic markers, and MRI/PET imaging from CN, MCI, and AD individuals. In machine learning (ML), the ADNI dataset has instantiated numerous diagnostic classification models using features ranging from cognitive scores and biomarkers to imaging-derived measures (Weiner & Veitch, 2015).

In this review, we focus on two binary classification problems using ADNI Group one (CN vs. MCI), and Group two (CN vs. AD). We survey recent literature on four popular ML models such as logistic regression (LR), random forest (RF), XGBoost (XGB), and the emerging Natural Gradient Boosting (NGBoost). We also review common data-preparation techniques needed for ADNI, including strategies for handling **missing-data** and **dimensionality reduction**. Classifier performance will also be covered, as it is typically evaluated by metrics such as area under the ROC curve (AUC), accuracy, sensitivity, and specificity. In general, tree-based ensembles (RF, XGB) tend to outperform simple LR for these tasks (Yi et al., 2023; Velazquez & Lee, 2021).

NGBoost is a newer probabilistic boosting method that provides uncertainty estimates and has not yet been applied to the ADNI dataset. Therefore, by including NGBoost in ADNI analyses represents a novel contribution. We contextualize the strengths and weaknesses of each model and conclude by identifying open gaps and proposing the objectives of our thesis.

## ADNI Dataset Overview

The ADNI consortium was launched in 2004 to establish standardized neuroimaging and biomarker data on aging, mild cognitive impairment (MCI), and Alzheimer's disease (AD) subjects. ADNI-1 enrolled cognitively normal (CN) individuals, individuals with MCI, and early AD cases. Subsequent phases such as ADNI-GO and ADNI-2 added larger MCI cohorts, including both early and late MCI, and expanded the range of imaging modalities. Data from over a thousand participants, ranging in age from approximately 55 to 90, have been collected across sites in the United States and Canada. The ADNI dataset is richly multimodal where each subject may have neuropsychological test scores, demographic and genetic information such as APOE status, cerebrospinal fluid biomarkers, and MRI or PET imaging features (Spooner et al., 2023; Weiner & Veitch, 2015).

The wide variety of ADNI features enables diverse modelling approaches but also introduces high dimensionality where typical study cohorts include only several hundred subjects. For example, one analysis used data from 509 ADNI-1 participants, comprising 137 AD, 210 MCI, and 162 CN cases, with hippocampal MRI features (Bloch & Friedrich, 2021). Given the class imbalance and the presence of missing data in ADNI, careful pre-processing is essential. ADNI-based classification studies often address multiple related tasks such as data imputation, feature selection and then progressively explore grouping classification such as CN vs. AD, CN vs. MCI, and MCI vs. AD to fully capture model performance across the clinical continuum (Yi et al., 2023). Below, we will reiterate the common challenges found in ADNI research.

## Handling Missingness in ADNI

Missing data is **common in ADNI**, because subjects may skip assessments or drop out of longitudinal follow-up. As Chandrasekaran and Xie (2024) note, ADNI often has **substantial missingness**, and simply discarding incomplete cases can cause efficiency loss or bias. Appropriate imputation methods is therefore critical. In practice, imputation approaches range from simple to highly sophisticated methods. For example, **simple imputation** such as mean or median substitution is often used as a baseline. One review stated that one can replace missing values with the overall feature mean, median, or even a winsorized (trimmed) mean to mitigate outliers. However, simple imputation ignores correlations among features and may under-estimate variability (Campos et al., 2024).

More advanced strategies iteratively predict missing entries. An example would be **k-Nearest Neighbors (k-NN)** imputation where it fills a missing value by averaging the corresponding values from the nearest subjects. This leverages data structure, since nearby subjects in feature space tend to have similar values. Another approach, **MissForest**, uses random forests (RF) to iteratively predict missing entries where the algorithm trains an RF on observed data to predict missing values, and then cycles through features until convergence. It is observed that MissForest as a flexible method that often outperforms basic imputation in complex biomedical data (Campos et al., 2024).

A widely-used class of methods is **multiple imputation**, of which Multivariate Imputation by Chained Equations (MICE) is a popular example. In MICE, each feature with missingness is imputed by a ML model using the other variables where these imputations are iterated multiple times to stabilize the estimates (Campos et al., 2024). The result is a set of plausible complete data sets that reflect imputation uncertainty. Chandrasekaran and Xie specifically advocate MICE for ADNI, showing that it yields valid, more efficient regression estimates compared to complete-case analysis. In their ADNI study, MICE-based imputation produced tighter confidence intervals and avoided bias that would arise from dropping subjects with missing tests.

Overall, imputation in ADNI typically involves either a basic substitution (mean/median) for simplicity, k-NN or MissForest for tree-based, similarity-driven estimates, or fully iterative methods like MICE to account for uncertainty. The chosen method can affect downstream classifier performance, so studies often experiment with several.

## Reducing Dimensionality

With hundreds of potential features, feature selection is often applied to reduce dimensionality. Feature selection methods in AD applications include wrapper algorithms, which evaluate subsets via a predictive model and filter methods, which score features by statistical criteria. Two popular wrappers are **Boruta** and **Recursive Feature Elimination (RFE)**. Boruta works by comparing each real feature’s importance, usually from a random forest, to that of randomized shadow copies. When initiated, Boruta identifies all features that are more important than noise, reducing features to those with significant signal. By contrast, RFE iteratively removes the least-important feature, as measured by a model coefficient or tree split weight, re-training until an optimal subset remains. For example, one study used cross-validated RFE and Boruta on ADNI features to try to eliminate irrelevant features. However, they found practical drawbacks where RFE was computationally expensive and unstable across folds, while Boruta tended to select a large set of features that did not improve accuracy. In that work, neither RFE nor Boruta ultimately yielded a better model than simpler filters such as Pearson filtering (Jahan et al., 2023).

Another common **filter-based** approach is to use mutual information(MI) to select features that have high statistical dependency with the outcome. MI-based selection, often implemented via maximum-relevance/minimum-redundancy methods, picks features that individually carry the most information about the class label. In ADNI studies, MI filtering has proven effective. For instance, Al Mansoori et al. (2024) applied MI-based ranking to blood biomarker features and trained a support vector machine (SVM). They reported that an SVM on MI-selected features achieved an AUC of 0.94 and 95% accuracy for differentiating controls vs early-stage AD cases. This shows that MI can yield a powerful compact feature set (AlMansoori et al., 2024).

Other strategies sometimes used an array of extensive univariate tests, but recent ADNI literature emphasizes Boruta/RFE and MI. In practice, feature selection in this domain is balanced against model complexity where aggressive selection may remove useful information, whereas poor selection leaves noisy features.

## Common Metrics in AD Research

**Selecting appropriate evaluation metrics is another crucial step prior to implementing any ML models, as it ensures accurate assessment and comparison of classifier performance in AD research.** Among these, the area under the receiver operating characteristic curve (AUC) remains the most widely cited metric because it summarizes a model’s ability to discriminate between classes across all thresholds (Blanco et al., 2023). AUC is particularly valuable in imbalanced datasets because it evaluates a model's ability to distinguish between classes across all thresholds, independent of class distribution. This makes it a robust metric when dealing with datasets like ADNI, where the prevalence of MCI or AD is lower than that of healthy controls. For instance, Richardson et al. (2023) highlight that the ROC curve is robust to class imbalance, making AUC a reliable metric in such scenarios. In addition, Relying solely on accuracy can be misleading in imbalanced datasets. A model might achieve high accuracy by predominantly predicting the majority class, thereby failing to identify instances of the minority class. This concern is reported in various studies, emphasizing the need for alternative metrics like AUC, precision, recall, and F1-score to obtain a more comprehensive evaluation of model performance. Therefore, sensitivity and specificity are heavily implied in many ADNI studies (Li & Hsu, 2022). More specifically, sensitivity quantifies how well a model correctly identifies individuals with MCI or AD, whereas specificity measures how well healthy controls are correctly excluded. By considering AUC alongside sensitivity and specificity, researchers obtain a more accurate and clinically relevant evaluation of model reliability, generalizability, and potential real-world impact.

## Machine Learning Approaches for AD Classification

In the context of AD, a wide range of supervised ML models have been used to assess both performance and explainability. Models ranging from highly complex to very simple have been covered. To keep it concise, we will only mention the most prominent and widely used classifiers in AD research, as the array of models is extensive.

A classical baseline, Logistic Regression (LR), models the log-odds of individual classes as a linear function of the features. In ADNI tasks, LR is often used as a reference method due to its simplicity and interpretability. However, LR has limited capacity to capture complex nonlinear relationships present in high-dimensional biomarker data. In many studies, LR yields lower accuracy and AUC than nonlinear classifiers. For example, Velazquez et al. (2021) found that their random forest model outperformed a comparison logistic regression classifier on predicting MCI vs AD conversion. Similarly, (Shastry & Sattar, 2023) ensemble (LRFB) paper reports that LR alone underperforms compared to RF/XGBoost within their framework. Strengths of LR include direct probability outputs and ease of interpretation, where feature weights are meaningful, but its weakness is oversimplicity. It may underfit when features interact or effects are nonlinear. Therefore, while LR can serve as a sanity check, more flexible models are typically preferred for ADNI classification.

**Another classifier is Random Forest (RF),** an ensemble of decision trees trained on bootstrapped samples with random feature subsets. It is nonparametric and inherently handles mixed data types and missing values to some extent. RF is robust to outliers and captures nonlinear feature interactions without much tuning. It also provides built-in measures of feature importance. Many ADNI studies use RF as a strong benchmark. For instance, Velazquez et al. (2021) reported that their RF achieved 93.6% accuracy (and 0.96 AUC) in predicting Early Mild Cognitive Impairment EMCI-to-AD conversion, which outperformed other methods including SVM and logistic regression. In classification tasks, RF often yields high sensitivity and specificity. However, RF can overfit if trees are too deep and class imbalance is present. Its main weaknesses are that it is less interpretable than LR, though variable importance helps, and it can be computationally intensive for large feature sets. Overall, RF is favored for its strong out-of-the-box performance on ADNI classification.

Similarly, XGBoost is a gradient boosting framework that builds an ensemble of trees sequentially, each new tree correcting errors of the prior ensemble. It is highly optimized for speed and performance, with regularization to prevent overfitting. XGBoost has become ubiquitous in ADNI studies due to its state-of-the-art predictive power. It often surpasses RF in accuracy and can handle heterogeneous features well. In the XGBoost-SHAP framework of Yi et al. (2023), XGBoost achieved impressively high performance on multiclass ADNI data around 0.88–0.91 AUC for classifying between normal, MCI, and AD.

Notable papers mention that ensembles like RF and XGBoost typically outperform LR. For example, Yi et al. highlight that complex models like XGBoost achieve higher accuracy at the cost of interpretability. In addition, Chen et al. observed that XGBoost gave better performance than RF and similar trends hold in ADNI analyses. Common performance metrics for classification between CN individuals and those with AD include sensitivity, specificity, accuracy, and ROC AUC. In summary, ensemble methods consistently yield high AUC and accuracy, whereas logistic regression often falls toward the lower end of reported performance ranges.

## 2.6 Probabilistic Boosting

Natural Gradient Boosting (NGBoost) is a recently developed algorithm for probabilistic prediction. Unlike standard boosting (which predicts a point estimate of the target), NGBoost models the entire conditional distribution of the target variable. It does so by choosing a parametric distribution family (e.g. Gaussian for regression, Bernoulli for binary classification) and then using gradient boosting to estimate the distribution parameters. Crucially, NGBoost employs natural gradients for more efficient, stable learning of these parameters.

The result is a model that can output not just a class label but also an uncertainty measure for each prediction. In practical terms, NGBoost can be seen as a modular framework where one specifies a base learner and a parametric family such as Gaussian or logistic, and NGBoost iteratively fits trees to update the distribution’s parameters. The authors of NGBoost demonstrate that it achieves competitive point-estimation accuracy on UCI benchmarks while also providing well-calibrated predictive distributions Duan et al. (2020). For classification tasks, NGBoost can output the probability of each class along with credible intervals, which is valuable in high-stakes domains like medicine.

**As such and** to our knowledge, NGBoost has notyet been applied to ADNI classification in the literature. Prior ADNI studies have largely used classifiers that predict only a point label or probability without modeling uncertainty compared to RF, XGB, SVM, LR. Incorporating NGBoost thus represents a novel contribution. It may offer advantages in evaluating AD risk with uncertainty bounds, which could improve decision-making.

However, NGBoost also brings increased complexity. It requires choosing a suitable distribution family and can be slower to train. It also inherits tree-based models’ interpretability issues. In the context of ADNI, we would likely configure NGBoost with a Bernoulli distribution for the binary tasks. Its promise lies in potentially matching or exceeding XGBoost in accuracy, as NGBoost can mimic gradient boosting, while giving probabilistic calibration. Nonetheless, its novelty means we must critically compare it with established methods.

ADD Tuning

ADD Explainability

Mention Pearson and remove mutual info.

Refine.

## 2.8 Summary

In summary, despite extensive work on ADNI classification, several gaps remain. First, Natural Gradient Boosting (NGBoost) has not been explored in this context. Most ADNI studies use point-estimate classifiers. NGBoost’s probabilistic output is novel for ADNI where our literature search found no application to ADNI data. Incorporating NGBoost could reveal whether modeling predictive uncertainty improves classification or clinical utility.

Second, many prior studies focus on CN vs. AD, with less emphasis on the CN vs. MCI task. We will ensure both tasks are addressed. Third, although many papers report high AUC or accuracy, few compare models systematically using all four metrics (AUC, accuracy, sensitivity, specificity) on the same splits. Our thesis will fill this gap by rigorously evaluating LR, RF, XGB, and NGBoost on standardized CN/MCI and CN/AD splits, using consistent cross-validation and reporting comprehensive metrics.

The objective of this literature review is therefore to lay the groundwork for such a comparative study. We will gather and synthesize findings on how data preparation such as imputation, feature selection and model choice affect performance on ADNI tasks. In particular, we will highlight NGBoost as a contribution where we will explain its theory, why natural gradients and probabilistic boosting matter, and contrast it with XGBoost. By identifying what has and has not been done in the literature, we motivate our approach of applying NGBoost to ADNI for the first time, alongside standard models, to see if it can match or exceed existing benchmarks.

# Methodology

## Introduction

This chapter describes the end-to-end workflow used to develop and interpret four ML models for Alzheimer’s classification on the ADNI-1 dataset. We begin with cohort-specific exploratory data analyses first comparing CN vs MCI, then CN vs AD to assess data quality, class balance, feature distributions, and inter-feature correlations. Next, we systematically compare four imputation methods to handle missing values within each cohort pair. We then evaluate multiple feature-selection techniques to reduce dimensionality and improve generalization. After that, we benchmark and tune four classifiers Logistic Regression, Random Forest, XGBoost, and NGBoost using a stratified cross-validation framework with a Tree-structured Parzen Estimator (TPE) search. Finally, we apply SHAP-based explainability analyses to interpret model predictions and identify key biomarkers for each classification task. Each step is detailed below to ensure reproducibility and clarity.

## Explanatory Data Analysis

In this section, we present a comprehensive exploratory analysis of the ADNI-1 dataset to assess its suitability for two binary classification tasks: CN vs. MCI and CN vs. AD. We begin by evaluating data completeness, identifying missing values, duplicate records, and summarizing the distribution of diagnostic labels to assess class imbalance. To examine the distributional properties of the features, we employ the Shapiro-Wilk test, as well as measures of skewness and kurtosis. For each cohort pair, we compute univariate statistics and generate graphical visualizations such as histograms, Q-Q plots and boxplots to characterize feature distributions and identify potential outliers.

Pairwise correlation analyses using heat-map was implemented to check multicollinearity among variables and variables exhibiting correlations above ±90% were specifically mapped to identify strongly collinear feature pairs. Further correlation check was employed to assess the relationship between each individual feature and the target variable. RID feature was dropped since correlation between variables was very minimal, thereby contributing less to improving model performance. To validate our sampling strategy and reduce overfitting risk, we then perform 70/30 train-test splits separately for CN/MCI and CN/AD, verifying that each partition preserves class proportions and maintains sufficient sample sizes. Insights from this analysis inform all subsequent data-cleaning steps, feature transformations, and the selection of candidate predictors.

## Missing-Data Imputation Comparison

To assess the impact of different imputation strategies on model performance, several widely used methods in Alzheimer's Disease (AD) research were evaluated. The objective was to determine which approach yields the most accurate and reliable results. Using a 70/30 train-test split, the following imputation techniques were applied: K-Nearest Neighbors (KNN) imputation, Multiple Imputation by Chained Equations (MICE) with various internal models, Predictive Mean Matching (PMM), and a baseline comparison using no imputation. For MICE, multiple regression-based internal models were tested, including Bayesian regression, decision tree regression, KNN regression, and MissForest, chosen for their ability to handle multicollinearity and non-linearity among features. For PMM, different donor draw values (5, 10, 15, and 50) were explored to investigate whether increasing the number of donors improves imputation quality or leads to diminishing returns. To evaluate the effectiveness of each imputation method, a Random Forest classifier was used as the preliminary modeling algorithm. Performance was assessed using 10-fold stratified cross-validation, ensuring class balance within folds, improving generalizability, and minimizing overfitting while keeping the evaluation process consistent and interpretable.

## Feature-Selection Methods Comparison

In this section, we compare multiple feature selection techniques to identify the most informative subsets of features for the classification tasks. Specifically, we evaluated Pearson correlation, Recursive Feature Elimination (RFE), and Boruta. Prior to applying these methods, a preliminary filtering step was conducted to eliminate features with very low variance, ensuring that uninformative variables were excluded early in the pipeline. Each feature selection method produced a subset of features deemed most relevant, and these subsets were then evaluated using four classifiers: Logistic Regression (LR), Random Forest (RF), NGBoost, and XGBoost. For benchmarking purposes, we also included a model trained on the full feature set (i.e., no selection) to assess the impact of feature selection on performance and computation time. Model performance was evaluated using a comprehensive set of metrics: accuracy, balanced accuracy, precision, recall, F1-score, and ROC-AUC. This multi-metric evaluation ensures a robust assessment of each method, capturing both general and class-specific performance to avoid overlooking any aspect of model behavior.

## Hyperparameter Tuning

To enhance model performance, Bayesian Optimization was implemented to intelligently search the hyperparameter space, improve explainability, and reduce computational complexity when exploring high-dimensional search spaces. Specifically, using Optuna, the Tree-structured Parzen Estimator (TPE) algorithm was employed with 20 optimization iterations for each model. This method allows for a more guided and efficient exploration compared to exhaustive grid or random search. The hyperparameter tuning was performed for four classification models: Logistic Regression (LR), Random Forest (RF), XGBoost, and NGBoost. Below is a summary of the hyperparameters optimized for each model. Below is a summary of the hyperparameters optimized for each model.

Table 1. Logistic Regression parameters and intervals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Minimum** | **Maximum** | **Grid** |
| **C** | 0.001 | 100 | {0.001, 0.032, 1.000, 31.622, 100.000} |
| **tol** | 0.000001 | 0.01 | {0.000001, 0.000316, 0.001, 0.00316, 0.01} |
| **solver** | *Categorical only* | *Categorical only* | {lbfgs, newton-cg, newton-cholesky, liblinear} |
| **max\_iter** | 300 | 500 | {300, 350, 400, 450, 500} |

Table 2. Random Forest parameters and intervals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Minimum** | **Maximum** | **Grid** |
| **n\_estimators** | 0.001 | 100 | {100, 200, 300, 400, 500} |
| **Criterion** | *Categorical only* | *Categorical only* | {gini, entropy} |
| **max\_depth** | *3* | *4* | {3, 4} |
| **min\_samples\_split** | 0.001 | 0.2 | {300, 350, 400, 450, 500} |
| **min\_samples\_leaf** | 0.001 | 0.1 | {0.001, 0.025, 0.05, 0.075, 0.1} |
| **max\_features** | *Categorical only* | *Categorical only* | {sqrt, log2} |
| **class\_weight** | *Categorical only* | *Categorical only* | {balanced, balanced\_subsample} |

Table 3. XGBoost parameters and intervals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Minimum** | **Maximum** | **Grid** |
| **n\_estimators** | 200 | 500 | {200, 275, 350, 425, 500} |
| **max\_depth** | 2 | 3 | {2, 3} |
| **learning\_rate (η)** | 0.01 | 0.2 | {0.01, 0.058, 0.105, 0.153, 0.2} |
| **subsample (s)** | 0.5 | 0.8 | {0.5, 0.575, 0.65, 0.725, 0.8} |
| **colsample\_bytree (c)** | 0.6 | 0.8 | {0.6, 0.65, 0.7, 0.75, 0.8} |
| **min\_child\_weight (w)** | 1 | 6 | {1, 2.25, 3.5, 4.75, 6} |
| **gamma (γ)** | 0.0 | 2.0 | {0.0, 0.5, 1.0, 1.5, 2.0} |

Table 4. NGBoost parameters and intervals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Minimum** | **Maximum** | **Grid** |
| **n\_estimators** | 100 | 200 | {100, 200, 300, 400, 500} |
| **learning\_rate (η)** | 0.01 | 0.1 | {0.01, 0.032, 0.055, 0.078, 0.1} |

Bayesian Optimization with the Tree-structured Parzen Estimator (TPE) enabled a principled and computationally efficient search of high-dimensional hyperparameter spaces across four models. The tuning process was performed individually for each model, rather than using nested hyperparameter search, to reduce bias and enhance explainability. Additionally, various plots were generated to assess hyperparameter importance and to identify the most influential regions within the search space.

## Explainability Analysis

To better interpret the predictions of each model and determine the contribution of individual features, SHAP (SHapley Additive exPlanations) was applied at both local and global levels. At the local level, class-specific SHAP beeswarm plots were generated to visualize how feature values influenced predictions for each individual class. This allowed for a more detailed, class-wise understanding of model behavior. At the global level, aggregated SHAP beeswarm plots were used to provide an overall view of feature importance across all classes, offering a comprehensive summary of how each feature contributed to the model’s predictions throughout the dataset.

## Summary

In this chapter, we presented a fully reproducible workflow for developing and interpreting four machine learning models on the ADNI‑1 dataset. We began with exploratory data analysis to characterize data quality, feature distributions, and inter‑feature relationships for CN vs. MCI and CN vs. AD groups. We then compared four missing‑data imputation techniques using KNN, MICE with various internal models, PMM, and a no‑imputation baseline using stratified cross‑validation to identify the most reliable approach. Next, we evaluated three feature‑selection methods such as Pearson correlation, RFE, and Boruta against a full‑feature benchmark, using multiple classifiers and a suite of performance metrics to determine the optimal feature subsets. We proceeded with Bayesian hyperparameter tuning via Optuna’s Tree‑structured Parzen Estimator (TPE), conducting 20 iterations per model and documenting the search spaces for Logistic Regression, Random Forest, XGBoost, and NGBoost. Finally, we applied SHAP for both local for class‑specific outcomes and global explainability, generating beeswarm plots to reveal feature contributions at individual and aggregate levels. Collectively, these steps ensure that our modeling pipeline is transparent, efficient, and optimized for robust AD classification.

# Results

## Introduction

This chapter presents the results obtained from applying the ML workflow outlined in the previous chapter to two datasets: (1) the ADNI 1 cohort alone, and (2) a merged cohort combining ADNI 1 with the OASIS 3 Freesurfer‑derived features. For each dataset, we follow an identical analysis pipeline end to end, allowing us to directly compare whether augmenting ADNI 1 with OASIS 3 data yields improved predictive performance.

We begin by summarizing findings from the exploratory data analysis on both datasets, highlighting data completeness, the three‑class distribution (NC, MCI, AD) and key feature properties for the AD classification task. Next, we report on the performance of various missing‑data imputation techniques, evaluating their impact on model accuracy and reliability in both the ADNI 1 and merged cohorts. This is followed by a comparative assessment of feature selection methods, detailing their influence on classifier performance across the two datasets.

Subsequently, we present the outcomes of hyperparameter optimization for each classifier using Bayesian optimization to identify optimal parameter configurations in both the standalone and merged datasets. Finally, we provide explainability analyses using SHAP values to interpret model predictions and identify the most influential biomarkers driving multiclass classification in each cohort. All results are reported using a consistent set of evaluation metrics (accuracy, balanced accuracy, precision, recall, F1‑score and ROC‑AUC) to ensure comparability across datasets, models and methodological choices.

## Exploratory Data Analysis Results

The analysis revealed in the ADNI-1 dataset that out of 608 observations, missing values were present in 52.63% of cases, with 20 out of 24 variables containing at least one missing entry. The variables with the highest proportion of missingness were ERCs\_thicknessbaseline (6.09%), ERCsClusterShadebaseline (5.59%), and ERCsCorelationbaseline (4.61%). The remaining features exhibited missing values ranging from 0.16% to 4.11%.

## Missing-Data Imputation Results

## Feature Selection Results

## Hyperparameter Tuning Results

## Explainability Results

## Summary

# Discussion

## Introduction

## Title 1

## Title 2

## Title 3

## Summary

# Conclusions and Future Work

## Introduction

## Title 1

## Title 2

## Title 3

## Summary

# Bibliography

1. Alzheimer’s Association. (n.d.). What is dementia? Alzheimer’s Association. https://www.alz.org/alzheimers-dementia/what-is-dementia
2. Alzheimer’s Disease International. (n.d.). Dementia statistics. <https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/>
3. Akhavan Aghdam, M., Bozdag, S., Saeed, F., & Alzheimer’s Disease Neuroimaging Initiative. (2025). Machine-learning models for Alzheimer’s disease diagnosis using neuroimaging data: Survey, reproducibility, and generalizability evaluation. Brain Informatics, 12, Article 8. <https://doi.org/10.1186/s40708-025-00252-3>
4. Leandrou, S., Lamnisos, D., Bougias, H., Stogiannos, N., Georgiadou, E., Achilleos, K. G., Pattichis, C. S., & Alzheimer’s Disease Neuroimaging Initiative. (2023). A cross-sectional study of explainable machine learning in Alzheimer’s disease: Diagnostic classification using MR radiomic features. Frontiers in Aging Neuroscience, 15, 1149871. <https://doi.org/10.3389/fnagi.2023.1149871>
5. Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y., & Schuler, A. (2019). NGBoost: Natural gradient boosting for probabilistic prediction [Preprint]. arXiv. <https://arxiv.org/abs/1910.03225>
6. **Bloch, L., & Friedrich, C. M.** (2021). Data analysis with Shapley values for automatic subject selection in Alzheimer’s disease data sets using interpretable machine learning. Alzheimer’s Research & Therapy, 13, Article 155. <https://doi.org/10.1186/s13195-021-00879-4>
7. **Weiner, M. W., & Veitch, D. P.** (2015). Introduction to special issue overview of ADNI. Alzheimer's & Dementia, 11(7), 730–733. <https://doi.org/10.1016/j.jalz.2015.05.007>
8. Yi, F., Yang, H., Chen, D., Qin, Y., Han, H., Cui, J., Bai, W., Ma, Y., Zhang, R., & Yu, H. (2023). XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer’s disease. BMC Medical Informatics and Decision Making, 23, Article 137. <https://doi.org/10.1186/s12911-023-02238-9>
9. Velazquez, M., & Lee, Y., for the Alzheimer’s Disease Neuroimaging Initiative. (2021). Random forest model for feature-based Alzheimer’s disease conversion prediction from early mild cognitive impairment subjects. PLOS ONE, 16(4), e0244773. <https://doi.org/10.1371/journal.pone.0244773>
10. Spooner, A., Mohammadi, G., Sachdev, P. S., Brodaty, H., & Sowmya, A.; for the Sydney Memory and Ageing Study and the Alzheimer’s Disease Neuroimaging Initiative. (2023). Ensemble feature selection with data-driven thresholding for Alzheimer's disease biomarker discovery. Frontiers in Aging Neuroscience, 14, 9830744. <https://doi.org/10.3389/fnagi.2022.9830744>
11. Chandrasekaran, G., & Xie, S. X. (2024). Improving regression analysis with imputation in a longitudinal study of Alzheimer’s disease. Journal of Alzheimer’s Disease, 99(1), 263–277. <https://doi.org/10.3233/JAD-231047>
12. Campos, S., Zamora, J., & Allende, H. (2024). Block-wise imputation EM algorithm in multi-source scenario: ADNI case. Pattern Analysis and Applications, 27, Article 44. <https://doi.org/10.1007/s10044-024-01268-x>
13. Jahan, S., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S. M. S., & Ra, I.-H. (2023). Explainable AI-based Alzheimer’s prediction and management using multimodal data. PLOS ONE, 18(11), e0294253. <https://doi.org/10.1371/journal.pone.0294253>
14. AlMansoori, M. E., Jemimah, S., Abuhantash, F., & AlShehhi, A. (2024). Predicting early Alzheimer’s with blood biomarkers and clinical features. Scientific Reports, 14(1), 6039. <https://doi.org/10.1038/s41598-024-56489-1>
15. Blanco, K., Salcidua, S., Orellana, P., Sauma-Pérez, T., León, T., López Steinmetz, L. C., Ibañez, A., Duran-Aniotz, C., & de la Cruz, R. (2023). Systematic review: Fluid biomarkers and machine learning methods to improve the diagnosis from mild cognitive impairment to Alzheimer’s disease. Alzheimer's Research & Therapy, 15, Article 176. <https://doi.org/10.1186/s13195-023-01351-2>
16. Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns, 5(3), 100994. <https://doi.org/10.1016/j.patter.2024.100994>
17. Li, Y., & Hsu, W.-W. (2022). A classification for complex imbalanced data in disease screening and early diagnosis. BMC Medical Informatics and Decision Making, 22, Article 142. <https://doi.org/10.1186/s12911-022-01933-z>
18. Shastry, K. A., & Sattar, S. A. (2023). Logistic random forest boosting technique for Alzheimer’s diagnosis. International Journal of Information Technology, 15, 1719–1731. <https://doi.org/10.1007/s41870-023-01283-6>
19. Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020). NGBoost: Natural Gradient Boosting for Probabilistic Prediction. Proceedings of the 36th International Conference on Machine Learning (ICML). <http://arxiv.org/abs/1910.03225>

# Appendices

## Appendix A

## Appendix B

## Appendix C