Enhancing Personalised Dementia Risk Prediction through Transfer Learning and SHAP Interpretability

Author

Stephen Malcolm

Supervisor

Vahid Rafe

This Final Report/ CW2 is submitted for the degree of

Master of Science in Data Science and Artificial Intelligence

March 2024

Abstract

This study evaluated the performance of interpretable machine learning models, specifically the transfer learned XGBoost ML model and the Logistic Regression model, for personalised dementia risk prediction. Dementia poses immense societal burdens, which underlines the need for early risk assessment to enable timely interventions. However, technological advances like Machine learning provide powerful capabilities for multivariate risk modelling from heterogeneous health data that have significant potential in optimising dementia risk prediction while also ensuring interpretability.

The transfer learned gradient boost model consistently outperformed the Classical Logistic Regression model in terms of overall accuracy, weighted average precision, recall, and F1-scores on both the SHAREWave8 and ADNI datasets. Notably, incorporating hyperparameter tuning significantly improved the XGBoost model's performance, achieving an exceptional 99.02% overall accuracy and near-perfect 0.99 weighted averages on the other performance measures (recall, precision and f1-score), on the SHAREWave8 dataset.

A detailed comparative analysis shows that the transfer learned XGBoost model achieved higher performance metrics (approximately 1.38% higher overall accuracy, 7% higher weighted average precision, recall, and F1 scores) when trained on the larger SHAREwave8 dataset (20 input features, 79,144 instances) compared to the smaller ADNI dataset (10 input features, 16,421 instances). The same performance improvement pattern, on both the SHAREWave8 and the ADNI dataset, was also evident in the case of the Logistic Regression model, which recorded a 19%+ improvement across the performance metrics.

The interpretability of the models was assessed using the SHAP framework, which identified MRI-derived measurements, MMSE scores, and cognitive assessments as the most influential features for dementia risk prediction, aligning with known Alzheimer's disease pathophysiology.

The transfer learning enabled the XGBoost model to effectively adapt and generalize to heterogeneous real-world datasets, with improved performance on the larger SHAREWave8 dataset compared to the smaller ADNI dataset. This suggests that access to larger, more diverse training data can enhance the transfer learned ML model's ability to capture underlying patterns and relationships, leading to more accurate predictions. Nonetheless, there are still more validation studies to be conducted to ascertain the performance assessment in the context of varied clinical datasets.

Keywords: *SHAP, explainable AI, transfer learning, interpretable machine learning, XGBoost, Logistic Regression, Dementia risk prediction, Alzheimer's disease, Personalised medicine*

Table of Contents

List of Figures 4
List of Tables 4
List of Abbreviations 5
1.0 Introduction 6
1.1 Research Objectives 7
1.2 Research Questions 8
1.3 Minor Changes from the Initial Proposal 8
2.0 Stakeholders 8
3.0 Related Works 9
3.1 Multimodal Machine Learning for Dementia 9
3.2 Systematic Comparison of ML Methods 9
3.3 Interpretable Machine Learning 10
3.4 Reviewed Literature Summary 11
4.0 Methods 11
4.1 Datasets 11
4.2 Data Preprocessing 12
4.3 Machine Learning Model 13
4.4 Justification for the Sampled Models 15
5.0 Risk Assessment 15
5.1 Risk Mitigation Strategies 16
6.0 Ethical Consideration 16
7.0 Datasets and Pre-processing 16
7.1 Feature Selection 18
8.0 Results 18
8.1 ADNI DX and DX-bl Distribution 18
8.2 Feature Engineering-Correlation Analysis 19
8.3 XGBoost Performance Results 19
8.4 Logistic Regression Model Benchmark 23
8.5 Brier scores and Calibration plots 30
9.0 Discussion 30
9.1 Baseline Logistic Regression Vs Baseline Transfer Learned XGBoost Model 31
9.2 Results after Model Hyper-Parameter Tuning and Retraining 33
9.2.0 Bias and Fairness Assessment 33
9.2.1 Handling Skewed Class Distribution 33

9.2.2 Retrained XGBoost Model Performance 34
9.3 SHAP Explanations and Feature Importance 35
9.4 Survival analysis using Kaplan-Meier and Cox Proportional Hazards Models 36
9.5 Implications for Clinical Practice 38
10.0 Directions for Future Research 39
11.0 Conclusion 40
References 42
Appendices 46
Appendix 1: ADNI Input Feature Correlation Heat Map 46
Appendix 2: SHAREWave8 Input Feature Correlation Heatmap 47
Appendix 3: Cox Proportional Hazards Model Results for Alzheimer's Disease Progression-ADNI Dataset 48
Appendix 4: Concordance Score-ADNI Dataset 48
Appendix 5: First Kaplan-Meier plot-ADNI Dataset 48
Appendix 6: Second Kaplan-Meier plot- ADNI Dataset 49
Appendix 7: Fairlearn accuracy, Precision, and Recall scores by the ADNI Baseline model 49
Appendix 8: Work Plan 50

List of Figures

Figure 1: Visualisation for a gradient-boosting algorithm strategy (Chhetri et al., 2022) 14
Figure 2: SHAREwave8 Gender and Age Distribution 17
Figure 3: ADNI Gender and Age Distribution 17
Figure 4: DX and DX-bl Categorical Distributions 19
Figure 5: XGBoost Baseline AUC-ROC Curve-ADNI dataset 25
Figure 6: XGBoost Baseline ROC Curve-SHAREWave8 dataset 25
Figure 7: Logistic Regression Model ROC Curve for the ADNI dataset 26
Figure 8: Logistic Regression Model Roc Curve-SHAREWave8 26
Figure 9: XGBoost's Brier Score Calibration plot- the ADNI Dataset 27
Figure 10: XGBoost's Brier Score Calibration plot-the SHAREWave8 Dataset 28
Figure 11: Logistic Regression Model Brier Score Calibration Plot- the ADNI dataset 28
Figure 12: Logistic Regression Model Brier Score calibration plot- SHAREWave8 dataset 29
Figure 13: SMOTE gender balancing across diagnostic categories 34
Figure 14: XGBoost SHAP Summary plot 36
Figure 15: Predicted Survival Function for Three Selected Patients Based on Cox Proportional Hazards Model 37

List of Tables

Table 1: XGBoost Baseline Overall Results-ADNI Dataset 21
Table 2: XGBoost Overall Performance Result- ADNI dataset after Handling Imbalances and Fine-tuning 21
Table 3: XGBoost Baseline Overall Performance Results- SHAREWave8 Dataset 22

Table 4: XGBoost Overall Performance Results after Handling Imbalances and fine-tuning- SHAREWave8 Dataset ... 22
Table 5: Logistic Regression Model Overall Performance Results- SHARE dataset ... 24
Table 6: Logistic Regression Model Overall Performance Results- ADNI dataset ... 24
Table 7: Performance Comparative Analysis across Models and Datasets ... 30
Table 8: AUC-ROC performance across models and dataset classes ... 32
Table 9: Brier Score Comparative Analysis across Models and Dataset ... 32

List of Abbreviations

ADI - Alzheimer's Disease International

AI - Artificial Intelligence

AUC-ROC - Area Under the Receiver Operating Characteristic Curve

CAPI - Computer-Assisted Personal Interview

CPEC - Centre for Policy on Ageing

LSE - London School of Economics

ML - Machine Learning

MRI - Magnetic Resonance Imaging

PET - Positron Emission Tomography

SHAP - SHapley Additive exPlanations

SHARE - Survey of Health, Ageing and Retirement in Europe

WHO - World Health Organisation

XGBoost - Extreme Gradient Boosting

SHAP - SHapley Additive exPlanations

MMSE - Mini-Mental State Examination

ADNI - Alzheimer's Disease Neuroimaging Initiative

RAVLT - Rey Auditory Verbal Learning Test

1.0 Introduction

Dementia poses noticeable societal challenges, with over 50 million cases globally in 2022 (WHO, 2022). In the UK, dementia prevalence is projected to rise from 900,000 in 2019 to 1.6 million by 2040 as per an Alzheimer's Society report (CPEC & LSE, 2019). This underscores the urgent need for research and tools to address the impending impact and risks. Early and accurate risk prediction in the case of dementia, as is the case of many other disorders, offers immense clinical value by identifying high-risk individuals for timely interventions, monitoring, and planning. There are many diagnostic and early detection technologies, all of which are based on machine learning and artificial intelligence. Machine learning provides a powerful approach for developing personalised risk models, by discovering multivariate patterns from diverse health data. However, complex high-performing models like deep neural networks pose interpretability challenges that hinder trustworthiness and adoption in clinical practice. This research aims to balance accuracy and interpretability for dementia risk modelling, by proposing an integrated framework using transfer learning and SHAP (SHapley Additive exPlanations).

Machine learning has been increasingly applied for dementia risk modelling for years, capitalising on capabilities to handle heterogeneity and model complex nonlinear data relationships missed by traditional statistical techniques (Spooner et al., 2020). That is the reason many studies have converged efforts in developing predictive models using diverse training datasets including neuroimaging, genetics, demographics, cognition scores, and electronic health records (Wang et al., 2022). Deep machine learning methods like convolutional neural networks present flexibility to learn from multimodal datasets while accurately capturing multivariate interactions (Kumar et al., 2021). However, the complex inner workings of deep neural networks remain opaque, raising transparency issues around the logic behind the methods' prediction mechanisms. The black-box nature of predictive machine learning methods deters trust and adoption among stakeholders in the healthcare industry, particularly clinicians. As a result, the interpretability of the methods becomes a significant ethical concern for responsible clinical deployment of predictive machine learning methods.

Integrating transfer learning with interpretable models offers potential advantages for accurate and transparent dementia risk assessment. Transfer learning, in that context, presents an ideal strategy for integration. Transfer learning is a technique that transfers knowledge representations from an already trained model to boost performance on new related tasks with limited data (Tan et al., 2018). That form and type of machine learning method help mitigate data scarcity challenges, especially in the healthcare setting. Transfer learning has demonstrated promise for enhancing dementia prediction by transferring neural networks pre-trained on natural images compared to neural networks pre-trained on non-image medical data (Ding et al., 2019). The transferred representations reduce the need for extensive labelled medical data.

1.1 Research Objectives

Dementia risk modelling using machine learning provides opportunities to uncover multidimensional patterns from heterogeneous data like neuroimaging, genetics, lab tests, demographics, and electronic records (Spooner et al., 2020). However, challenges persist around model interpretability and evaluation on diverse clinical cohorts and datasets. In practice, complex machine learning models tend to act as black boxes, limiting trust and adoption in high-stakes, life-and-death settings like dementia screening (Lundberg et al., 2020). However, overly simple models fail to capture complex, real-world data patterns, sacrificing accuracy over interpretability or simplicity. That presents a trade-off between performance and interpretability. To advance clinical deployment, techniques are needed to open these black boxes for patient-specific explanations without compromising flexibility and accuracy.

Emerging strategies aim to improve model transparency, including explainable AI methods like SHAP that attribute a prediction to contributing features based on cooperative game theory (Lundberg et al., 2020). By quantifying feature importance, SHAP enables personalised model explanations. For the prediction and accuracy method, transfer learning offers prediction accuracy and flexibility in healthcare diagnostics, using limited datasets (Lee et al., 2019). This research focuses on rigorously evaluating SHAP and transfer learning to balance accuracy and interpretability for dementia risk models. In that context, the research will be guided by the following objectives and questions:

- Investigate integration of transfer learning and SHAP for interpretable dementia modelling.
- Quantitatively demonstrate a minimum per cent improvement in discrimination, calibration, and net benefit at clinically relevant thresholds for the gradient boosting with transfer learning model, compared to the Logistic Regression model using AUC, Brier scores, and decision curve analysis- and across two heterogeneous real-world clinical datasets for dementia risk modelling.
- Provide insights on balancing accuracy and interpretability in patient-centred AI.

1.2 Research Questions

RQ1: How does the SHAP algorithm enhance the explainability of personalised dementia risk prediction models?

RQ2: By what per cent does the interpretable gradient boosting framework with transfer learning improve discrimination, calibration, precision, recall, and overall accuracy over the classical Logistic Regression model for heterogeneous real-world dementia risk data?

RQ3: How do transfer learning and ensemble learning techniques impact the accuracy and generalisability of personalised dementia risk predictions?

1.3 Minor Changes from the Initial Proposal

Minor changes related to the study's research question focus, case assessment model, and pre-processing steps were adopted in the final study for logical inference of the study's results. Research question two was particularly modified to find the quantitative performance difference between a transfer learned gradient boosting model and the classical Logistic Regression model instead of the Cox Proportional Hazard model, as the Cox Proportional Hazard model is not directly comparable to the gradient boosting model in terms of their underlying interpretability and assumptions. The Cox Proportional Hazard model was adopted to cross-validate feature selection and interpretability for assessed models. The study also dropped the LightGBM model for the XGBoost model primarily due to the researcher's familiarity with the XGBoost model, which has a wider community support base and documentation compared to the LightGBM model. Finally, hyper-parameter tuning, cross-validation, and transfer learning activities, which were originally planned for phase 3, were integrated into the pre-processing phase due to the overlapping nature of months 2 through 5 resulting in a consolidation of tasks across these phases.

2.0 Stakeholders

Dementia poses immense health, emotional, and economic burdens on patients, caregivers, families, healthcare systems, and societies worldwide (Alzheimer's Disease International [ADI], 2022). As populations age, dementia prevalence is projected to rise substantially, exacerbating these burdens. That signifies early risk assessment is the only logical strategy that has a significant chance of enabling timely interventions to delay progression and improve outcomes. In that case, statistical risk assessment techniques present practical and viable risk assessment modelling that would enhance operations while adding value in the healthcare space.

3.0 Related Works

3.1 Multimodal Machine Learning for Dementia

Several studies have developed multimodal frameworks that integrate diverse data types to gain a more comprehensive perspective on disease progression. Iddi et al. (2019) proposed a flexible approach for predicting Alzheimer's disease progression using longitudinal data like MRI and PET scans. Their model handled heterogeneous and missing data across modalities and time points. Definitively, multimodal learning is based on the assumption that observations from different modalities originate from a shared latent representation (Battineni et al., 2021). That implies each modality's generative process can be modelled to infer latent variables and reconstruct observations even with incomplete data. In experiments, Nguyen et al. (2020) demonstrated how their method outperformed single-modality baselines by capitalising on cross-modality relationships. A key advantage in the demonstrated experiments is that cross-modality relationship mapping facilitated predictions beyond a fixed horizon by inferring future latent state trajectories. However, limitations in the study's predictive methods included a lack of rigorous comparative analysis, diverse validation, and clinical interpretability.

Jo et al. (2020) also demonstrated the utility of multimodal data integration for Alzheimer's disease classification. They developed deep neural networks to combine PET imaging and cerebrospinal fluid biomarkers. Their approach integrated training datasets at multiple levels, enabling flexible representation learning. Even though classification performance improved over single modality models, confirming the benefits of complementarity, there was still underutilisation of potential interactions from multiple modalities. Most importantly, the lack of model interpretability also limited deep neural networks' optimal clinical utility. Particularly, multimodal approaches demonstrate promise but require more rigorous evaluation and transparency.

3.2 Systematic Comparison of ML Methods

While machine learning is increasingly applied to dementia risk modelling, few studies provide systematic comparisons validating performance against traditional statistical techniques. Spooner et al. (2020) presented an extensive head-to-head evaluation of twelve machine-learning algorithms for dementia risk prediction using two real-world clinical datasets. The study's heterogeneous, high-dimensional dataset included demographics, diagnoses, medications, healthcare utilisation, cognition, genetics, and MRI. After appropriate preprocessing, cross-validation, and hyperparameter tuning, the results demonstrated that certain machine learning methods like random survival forests significantly outperformed the conventional Cox proportional hazards model. In the same context, Wang et al. (2022) explain that the Cox model relies on restrictive assumptions like proportional hazards, which fail to capture complex multivariate relationships. Spooner et al.'s (2020) and Wang et al.'s (2022) findings provide valuable insights into the suitability of machine learning techniques for more accurate risk modelling compared to traditional statistical methods. Notable limitations in the reviewed studies included a lack of external validation on additional datasets, no calibration assessment, and minimal clinician involvement. More rigorous evaluation and user-centred design would strengthen the clinical applicability of machine learning in dementia risk modelling.

3.3 Interpretable Machine Learning

A major challenge for accepting machine learning risk modelling in clinical settings is the lack of transparency, especially for sophisticated deep neural networks. Lundberg et al. (2020) proposed TreeSHAP, an efficient algorithm to compute exact Shapley values for interpreting tree ensemble model predictions. TreeSHAP builds on the SHAP (SHapley Additive exPlanations) framework that attributes a model's output to the contribution of each input feature based on concepts from cooperative game theory. The SHAP local feature importance measures enable patient-specific explanations by revealing how different risk factors interrelate and drive the prediction. Previous SHAP implementations for tree models relied on approximations for computational efficiency. The TreeSHAP provides exact attributions in polynomial time by capitalising tree structure properties (Marcilio & Eler, 2020). Through quantitative metrics and benchmark datasets, Lundberg et al. (2020) and also Bussmann et al. (2021) demonstrate TreeSHAP's accuracy and consistency as superior compared to previous approximation methods and traditional feature importance techniques. The integration of solid theoretical foundations with algorithmic innovations represents an important advancement in interpretable machine learning.

Lombardi et al. (2022) provide a practical demonstration of SHAP's utility for dementia risk models. They apply SHAP to explain predictions from a random forest classifier for Alzheimer's disease progression in mild cognitive impairment patients. The feature attributions highlighted verbal memory scores and baseline diagnosis as most influential for predictions. Interactive visualisations demonstrated how each variable impacts the prediction locally. As suggested by Battista et al.(2020) such transparent and personalised explanations can increase clinician trust in the black-box model predictions, and therefore facilitating adoption. While SHAP is model-agnostic, TreeSHAP provides efficiency benefits for tree-based methods commonly used in healthcare applications. However, clinical SHAP application is still limited, especially regarding real-world evaluation of explanation fidelity and physician comprehension. Explainable strategies such as SHAP offer much promise but require further rigorous validation, which this research proposes to explore.

3.4 Reviewed Literature Summary

The reviewed literature suggests that while multimodal machine learning frameworks easily integrate diverse data types to gain more comprehensive dementia progression perspectives, handling missing data and capitalising on cross-modality relationships to significantly improve prediction accuracies, the lack of rigorous validation, comparative analysis, and transparent clinical interpretability remain as notable limitations (Iddi et al., 2019; Nguyen et al., 2020). Particularly, whereas certain case algorithms like random survival forests significantly outperform conventional Cox models by capturing complex multivariate risk relationships (Spooner et al., 2020; Wang et al., 2022), external validation across datasets, calibration assessment, and clinician involvement are still under-explored. Also, despite tackling prediction accuracy, sophisticated ML models suffer from opacity or lack of adequate transparency, which limits clinical acceptance. As a result of such limitations, prototype explanatory strategies such as the one based on cooperative game theory attempt to resolve (Lundberg et al., 2020), even though requiring further inquiry into physician comprehension and decision-making impact. According to the reviewed literature, while machine learning innovations promise more accurate dementia risk modelling capabilities, addressing recurrent limitations in transparency, rigorous real-world validation, and evaluation of clinical utility with interdisciplinary research remains more important than ever for improved patient outcomes.

4.0 Methods

4.1 Datasets

Real-world clinical training datasets were obtained from two main sources: the SHARE public dataset and the ADNI dataset with a high degree of overlap with the SHARE dataset. SHARE contains demographics, medications, diagnosis history, cognition scores, and genetics of over 140,000 elderly participants across Europe (Börsch-Supan et al., 2013). Using two distinct heterogeneous datasets with complementary variables enabled rigorous evaluation of model generalisation. That denotes any potentially sensitive identifiers in the datasets were removed to protect participant confidentiality. The model was trained on a split dataset with a test size of 20% and a specific random state for reproducibility. Appropriate preprocessing addressing issues like missing data and categorical encoding was implemented. The diverse clinical data particularly facilitated a comparative assessment of the predictive algorithms on real-world dementia risk modelling.

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of microdata on health, socio-economic status, and social and family networks. The SHARE umbrella name covers over 30 European countries and Israel with its dataset collected via highly structured computer-assisted personal interview (CAPI) programs and additional self-completed paper-and-pencil questionnaires. SHARE's main questionnaire consists of 20 modules on different domains including health, socio-economics, and social networks. The data undergoes extensive quality checks and cleaning procedures as outlined in the SHARE Release Guide 8.0.0 (SHARE-Eric.eu, 2022), which ensures standardisation across participating countries and also through ex-ante harmonisation by translating generic questionnaires into different languages. Generic questionnaires are then processed automatically into the CAPI instrument used across study sites. The SHARE database spans multiple waves of data collection, with baseline data collected on over 140,000 individuals aged 50 or over in 11 European countries in the year 2004 (SHARE-Eric.eu, 2022). Since 2004, additional countries have joined the study and new waves of data have further expanded the longitudinal database. The multi-domain data available facilitates comparative research on the ageing population across Europe over time.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multisite observational study that archives comprehensive rows and columns of clinical, neuroimaging, biomarker, and genetic data on over 2,000 participants afflicted with Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (ADNI, 2021). The ADNI data is collected from multiple study sites across the United States and Canada, thus, providing a diverse and representative sample of the target population. The ADNI dataset used in the study consisted of approximately 16,421 rows and initially 15 columns, which was trimmed down to 11 columns after cleaning for missing data. The ADNI dataset was particularly selected due to its overlapping features to the SHARE dataset making it ideal for model assessment and cross-validation. Although another dataset: the PREVENT dataset had more overlapping features to the SHARE dataset compared to the ADNI dataset, it was not readily accessible. As a result, the study used the ADNI dataset for model cross-validation and assessment, relative to the SHARE dataset.

4.2 Data Preprocessing

Collected data was preprocessed to handle missing values and transform features as needed. Steps included:

- Deleting features where the percentages of the missing value exceed the threshold and removing duplicate rows.

- Encoding categorical variables.
- Preserving unique entries for analysis through the removal of redundant data during dataset management.
- Training the model on a split dataset with a test size of 20% and a specific random state for reproducibility.

4.3 Machine Learning Model

In the pursuit of developing an advanced predictive model for assessing dementia risk, this research endeavours to construct a gradient-boosting model utilising the XGBoost framework. Gradient boosting, an ensemble technique, amalgamates multiple weak base learner models iteratively to form a potent predictor, progressively refining performance by mitigating bias. XGBoost, an open-source implementation of gradient boosting, is particularly adept at handling large datasets characterised by high dimensionality.

To augment the capabilities of the gradient-boosting model in tackling the intricacies of medical prediction, transfer learning will be employed. As depicted in Figure 1, the neural network model will be initialised with parameters pre-trained on natural images, a technique demonstrated to enhance model performance in diverse domains. As explained and demonstrated by Chhetri et al. (2022), gradient boosting transfer learning involves the typical base decision tree learners from a model pre-trained on a source task being transferred to a new target task instead of initialising the gradient boosting model randomly. As illustrated in Figure 1 below (Chhetri et al., 2022), the gradient-boosting strategy begins from the transferred base models and continues the gradient-boosting iterative process by incrementally adding decision trees to rectify errors on the target task.

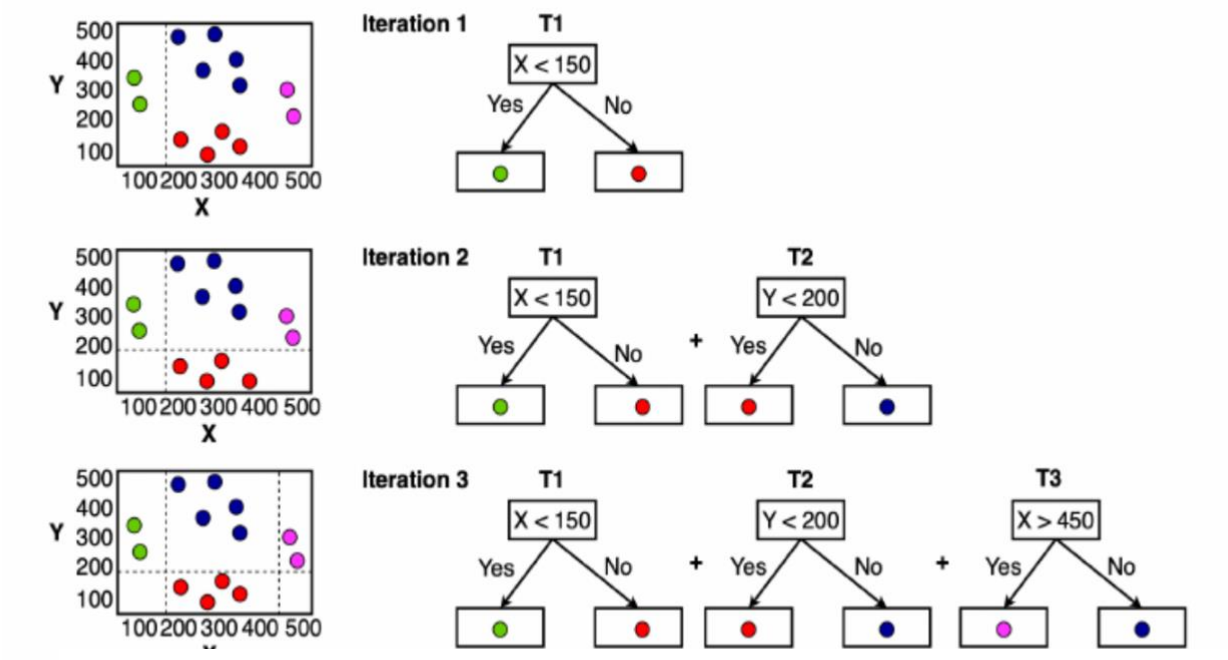


Figure 1: Visualisation for a gradient-boosting algorithm strategy (Chhetri et al., 2022)

However, given that the SHARE dataset primarily consists of non-image tabular data, the application of pre-trained image model parameters is inconsistent. Instead, the neural network will be initialised with random parameters and directly trained on the tabular features and labels from the SHARE dataset. The pre-trained parameters serve as an advantageous starting point, facilitating the transfer of valuable feature representations to the target task. Subsequent fine-tuning of the pre-trained model on the dementia risk datasets aims to achieve improved generalisation and faster convergence compared to training a model from scratch. The performance of the transfer-learned XGBoost model will be systematically compared to the classical Logistic Regression model without transfer learning, enabling a quantitative performance assessment of potential performance across the models and in the context of two datasets: the SHARE and the ADNI datasets.

Hyperparameter optimisation was conducted using random search cross-validation. Model performance was assessed using a range of metrics, including AUC-ROC, accuracy, sensitivity, specificity, and precision-recall curves. Strategies for handling missing data were analysed, with categorical columns imputed using mode and numerical columns imputed using mean to preserve dataset integrity. Through meticulous comparative analysis, this study aims to establish the suitability of the interpretable gradient-boosting approach for achieving accurate and robust clinical risk assessment. To enhance model interpretability, SHAP (SHapley Additive exPlanations) feature importance values were leveraged, offering local explanations by attributing predictions to contributing input features, thereby contributing to a more profound understanding of the model's decision-making process (Lundberg & Lee, 2017; Lundberg et al., 2020).

4.4 Justification for the Sampled Models

On one hand, gradient boosting models are characterised by multiple key performance strengths. Most notably, the models, as elaborated by Natekin and Knoll (2013), prove proficient at handling intricate prediction tasks like medical diagnosis- via ensemble integration of decision trees to enhance performance. Also, as demonstrated by Elith et al. (2008), ensemble-boosted decision trees include iterative error correction processes that tend to rectify predecessor tree shortcomings therefore contributing to high predictive accuracy. Transfer learning facilitates improved convergence and generalisation by initialising learners in a pre-trained model and transferring parameters to the target task (Tan et al., 2018). Recent applications demonstrate significant boosting model performance gains on medical tasks using transfer learning (Abbas, Abdelsamea, & Gaber, 2020). In that case, systematically comparing the transfer-learned XGBoost model against baselines will quantify potential advantages over training from scratch. Using real-world clinical data including demographics, medications, diagnoses, genetics, and cognition aims to rigorously evaluate predictive accuracy on demanding healthcare risk assessment tasks. Collectively, these properties seem to underpin the selection of XGBoost, which Ke et al. (2017) document as efficacious at large datasets with high dimensionality, for dementia risk modelling.

On the other hand, the TreeSHAP model enables seamless interpretation of boosted tree model predictions by attributing outputs to input features based on Shapley values from the game theory (Lundberg et al., 2020). As explained by Bogdanovic, Eftimov and Simjanoska (2022), SHAP explanation values can increase model transparency, clinical trust, and physician comprehension compared to opaque methods, which justifies the model's incorporation in the research method to help validate quality. Together, the XGBoost model balanced with careful validation and explainability from the TreeSHAP model presents a promising integrated strategy for patient-centric and accurate clinical forecasting.

5.0 Risk Assessment

This study involved minimal risk given the use of retrospective datasets and the absence of clinical implementation. However, several risks require consideration:

- Privacy risk: The datasets contain potentially sensitive personal health information. Any identifiers will be removed to protect confidentiality, but data leakage could still occur, exposing private details.

- Evaluation risk: If model limitations are not rigorously characterised, inappropriate reliance on predictions could ensue. The study will quantify uncertainties and refrain from overstating performance.
- Explanation risk: Interpretability does not guarantee trustworthiness if explanations do not match model reasoning. The study will evaluate local fidelity to avoid false assurances.

5.1 Risk Mitigation Strategies

The risks around data privacy, premature implementation, evaluation limitations, explanation fidelity, and algorithmic bias require thoughtful consideration. Specifically, while direct identifiers will be removed, unintended data leakage could still occur, exposing sensitive details. To address the risk, this study will implement strict data security protocols, and communicate the need for further evaluation. Adhering to ethical data management will help uphold data integrity and therefore conform to data privacy guidelines.

6.0 Ethical Consideration

After obtaining formal permission from the respective organisations-ADNI and SHARE, the study used historical clinical data on demographics, medications, diagnoses, genetics, and cognition for over 140,000 patients. While the data was expected to mask personally identifying markers, there was still potential privacy risk from unintended data leaks that could expose sensitive health details. In that case, strict data security protocols including access controls, encryption, and secure data storage were implemented to uphold participant confidentiality. The study also clearly communicated the need for further evaluation before clinical implementation, outlining uncertainties around real-world model performance. Model limitations and potential biases were assessed and transparently reported to avoid inappropriate reliance on predictions. This study aims to develop a patient-centric tool for early dementia interventions while respecting participant privacy by adhering to these ethical data practices.

7.0 Datasets and Pre-processing

The study sampled two datasets for model evaluation and analysis, the SHAREwave8 and ADNI datasets. The two datasets were readily accessible and had characteristic overlapping best features for Dementia risk prediction modelling. Particularly, the age and gender distribution in the sampled SHAREwave8 dataset ranged between 60-80 years, with a median age of 68 years and 54% female relative to 46% male respectively (Figure 2 below). Comparatively, the ADNI's age and gender distribution were 60-80 years, with a median age of 70 years, and 55% male relative to 45% female respectively (Figure 3 below).

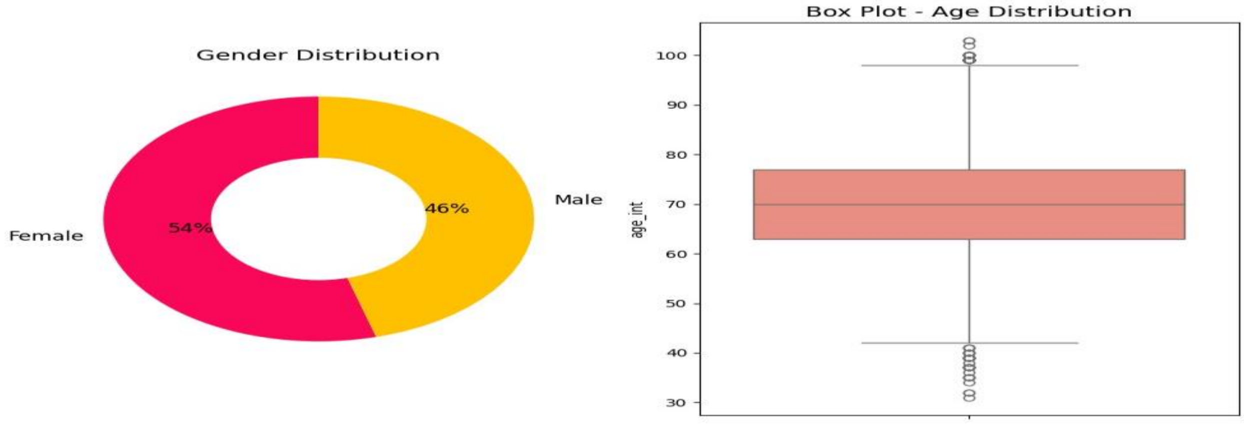


Figure 2: SHAREwave8 Gender and Age Distribution

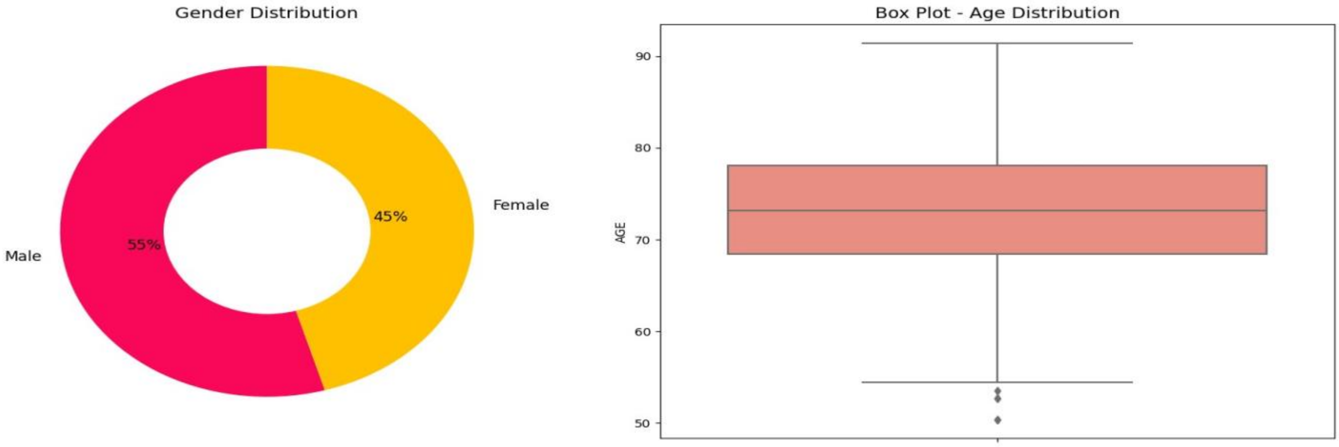


Figure 3: ADNI Gender and Age Distribution

Data pre-processing for both the SHAREwave8 and the ADNI datasets involved redundant rows, which were scanned and removed, while missing values were replaced with mode for categorical data and mean for numerical data variables. For the SHAREwave8 dataset, columns containing over 50% missing values were removed to avoid features lacking substantial data. On the same dataset, rows with greater than 75% missing values were dropped to ensure overall sample completeness. Comparatively, the ADNI dataset had 0% missing values and redundant rows. For both datasets, categorical features were label-encoded into numerical formats utilising the LabelEncoder class from Python’s sklearn pre-processing package. The label encoding to the categorical variables including categorical diagnostic features, gender, country, language, and education level aimed at converting categorical labels into numerical values as part of pre-processing. The encoded categories into numerical designations make them interpretable for modelling

7.1 Feature Selection

Feature selection for both datasets included input variables pertinent to Alzheimer's disease or dementia and related to health and medical history, biomarkers, cognitive function, socioeconomic factors, lifestyle, and demographics. Specifically, 15 and 8 input features were extracted from the ADNI and SHAREwave8 datasets respectively (Appendices 1 and 2). As explained in studies conducted by Jessen et al. (2014), Ritchie et al. (2020), and Jessen et al. (2011), features most related to cognitive function including memory, attention, execution of function among others, and those associated with mental health form a reliable predictive base for dementia. In that context, the target variables on both datasets were identified based on cognitive health indicators, reported mental health issues, and dementia diagnosis proxies.

8.0 Results

8.1 ADNI DX and DX-bl Distribution

The ADNI dataset modelling involved two target diagnosis variables: DX and DX-bl and patients with diverse background diagnoses. The DX-bl referred to a baseline diagnosis at the baseline visit, while the Dx referred to the diagnosis at the last visit, also called the current diagnosis. The two target diagnosis variables encompass five diagnostic categories including; CN (Cognitively Normal), EMCI (Early Mild Cognitive Impairment), LMCI (Late Mild Cognitive Impairment), AD (Alzheimer's Disease), and SMC (Significant Memory Concern). The Baseline Dx distribution, as presented in Figure 4 below, reveals that the majority of the patients were either CN (27.2%) or LMCI (33.0%) followed by EMCI (18.6%), AD (11.6%), and SMC (9.5%). That suggests that most of the patients had low mild or no cognitive impairment at the beginning of the study. Comparatively, the Dx distribution indicates that the majority of the patients were either MCI (51.5%), CN (30.1%) or Dementia (18.4%), which suggests that some of the patients had progressed to more severe cognitive impairment or Alzheimer's disease, while others remained stable or improved. For classification purposes, the study focused on two classes: dementia and no dementia classifications.

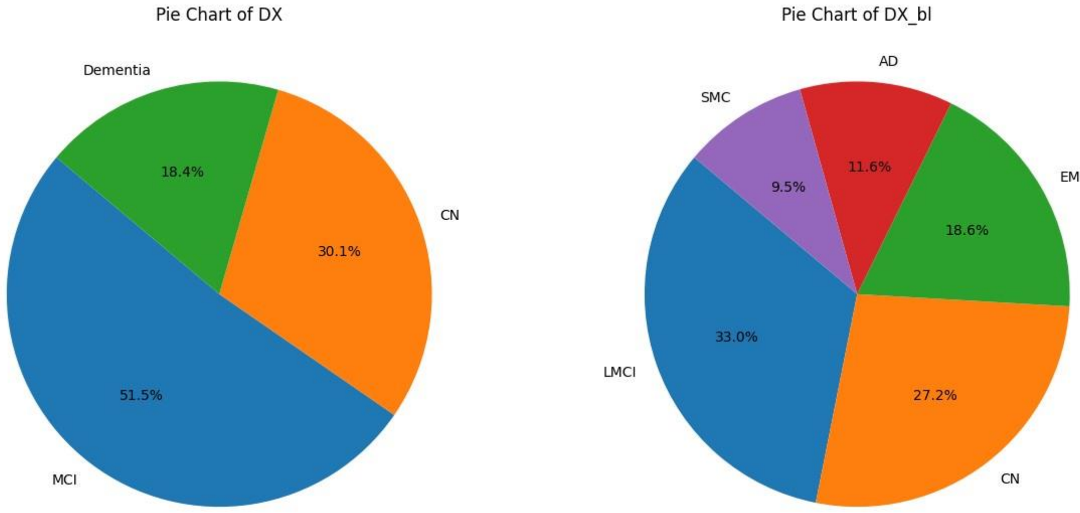


Figure 4: DX and DX-bl Categorical Distributions

8.2 Feature Engineering-Correlation Analysis

A correlation analysis was done to measure the strength and direction of the linear relationships between pairs of input features, as well as between input features and the target variable (DX). Feature engineering on both the ADNI and the SHAREWave8 datasets was done such that the input feature pairs with correlation above 0.75 including; RAVLT_immediate and RAVLT_learning, RAVLT_learning and RAVLT, and RAVLT_forgetting and RAVLT_perc_forgetting were carefully selected or transformed to reduce overfitting and multicollinearity issues. Specifically, the selected XGBoost model input features from the ADNI dataset included: AGE, PTGENDER, PTEDUCAT, APOE4, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting and RAVLT_perc_forgetting.

8.3 XGBoost Performance Results

The ADNI dataset was split into 80% training data and 20% testing data. Hyperparameter fine-tuning was performed using a hyperparameter grid for grid search to find the optimal hyperparameter combination for the XGBoost model. As a result, the best hyperparameters fine-tuned into the model included: max_depth- 7, learning_rate- 0.1, n_estimators-50, subsample- 0.8 and colsample_bytree-1. Further hyperparametric fine-tuning was done using a grid search strategy with cross-validation to split the training data into multiple folds and enable the use of one-fold as validation data while the rest as training datasets. The best hyperparameter combinations were thereafter selected based on accuracy scores.

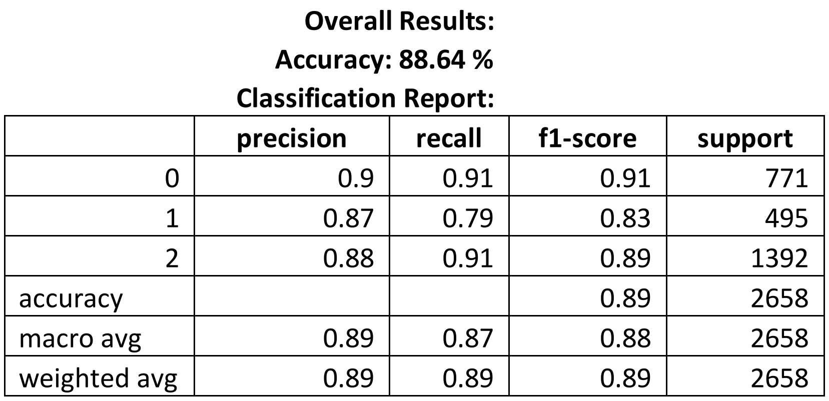
After hyperparametric fine-tuning and model training, the XGBoost classifier demonstrated strong performance on the test sets across multiple evaluation metrics- accuracy, precision, recall, and F1-score metrics. The XGBoost baseline classifier was evaluated on a test set comprising 2,658 patient case samples with input features including; AGE", "PTGENDER", "PTEDUCAT", "APOE4" and "MMSE". As highlighted in Table 1 below, the XGBoost model achieved an overall accuracy of 88.64% across the three diagnostic categories exhibiting a commendable ability to correctly predict the outcomes on the test data and successfully classifying nearly 9 out of 10 instances. As presented in Table 1 below, the model reported strong precision and recall of over 0.88 for classes 0 (dementia) and 2 (Mild Cognitive Impairment), which indicates that the model could accurately predict and flag dementia input features. However, class 1 (the no dementia class) recall was at 0.79, suggesting that the classifier overlooked some positive cases. Nonetheless, the model's macro-average and weighted averages aligned well with 88.64% overall accuracy.

Looking closely into individual class performance, Class 0 exhibited remarkable precision and recall values of 0.90 and 0.91, respectively. That implies 90% of the instances classified as Class 0 or dementia cases were indeed correct, while the model successfully identified 91% of the actual instances of Class 0. Notably, the harmonious balance between precision and recall is reflected in the 0.91 F1-score, suggesting that the model's robust performance in the dementia class. For assessment contextualisation purposes, it is worth noting that the model was evaluated on about 771 instances of Class 0, lending credibility to the reported metrics.

Same as in the above context, class 1 (no dementia class) exhibited a precision of 0.87 and a recall of 0.79, resulting in an F1-score of 0.83. While the precision score indicates that 87% of the instances classified as Class 1 were correct, the recall score indicates that the XGBoost model identified 79% of the actual instances of Class 1. In that case, the lower recall value compared to precision suggests an area for potential improvement in detecting instances of no dementia cases. Nevertheless, the 0.83 F1-score represents a significant balance between precision and recall, considering the model was tested on 495 instances of Class 1.

In the case of Class 2 or the Mild Cognitive Impairment class, the XGBoost model achieved a precision of 0.88 and a remarkable recall of 0.91 that aggregated to an F1-score of 0.89. The performance results mean that 88% of the instances classified as Class 2 were indeed correct, while the model successfully identified 91% of the actual instances of Class 2. The resulting high recall value together with a strong precision score contributes to the model's comparatively higher performance in this class compared to previous classes, considering that the model was evaluated on approximately 1392 instances.

When considering all assessed classes equally, the macro average provides an unbiased assessment. The macro average precision of 0.89 indicates that, on average, 89% of the instances were correctly classified across all classes. The 0.87 macro average recall suggests that, on average, 87% of the actual instances were correctly identified. In the same context, the 0.88 macro average F1-score represents a balanced measure of performance, taking into account both precision and recall across all classes. Lastly, the weighted average considering the number of instances in each class, presents a more representative view of the model's overall performance. The weighted average precision of 0.89 aligns closely with the model's overall accuracy, suggesting that the model's predictions were highly precise across the ADNI dataset. Similarly, the weighted average recall of 0.89 further supports the model's ability to correctly identify instances across the classes. The XGBoost model's performance on the ADNI dataset exhibits strong accuracy and precision across all classes, with particularly remarkable performance in Classes 0 (dementia) and 2 (Mild Cognitive Impairment class).



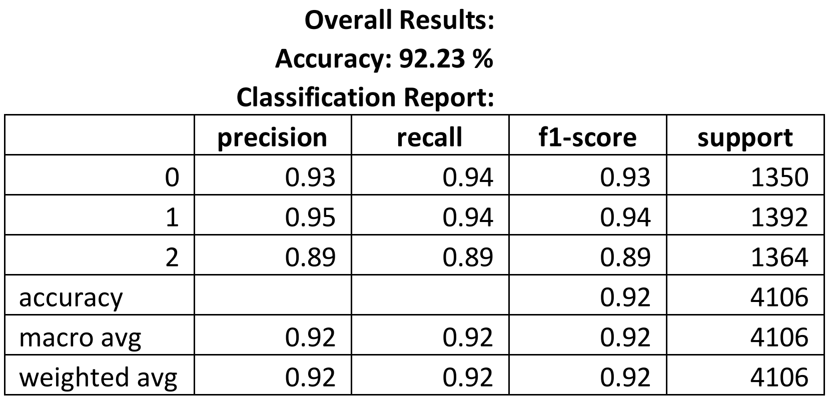
Overall Results:

Accuracy: 88.64 %

Classification Report:

	precision	recall	f1-score	support
0	0.9	0.91	0.91	771
1	0.87	0.79	0.83	495
2	0.88	0.91	0.89	1392
accuracy			0.89	2658
macro avg	0.89	0.87	0.88	2658
weighted avg	0.89	0.89	0.89	2658

Table 1: XGBoost Baseline Overall Results-ADNI Dataset



Overall Results:

Accuracy: 92.23 %

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	1350
1	0.95	0.94	0.94	1392
2	0.89	0.89	0.89	1364
accuracy			0.92	4106
macro avg	0.92	0.92	0.92	4106
weighted avg	0.92	0.92	0.92	4106

Table 2: XGBoost Overall Performance Result- ADNI dataset after Handling Imbalances and Fine-tuning

Overall Results:
Accuracy: 97.94 %
Classification Report:

	Precision	recall	f1-score	support
0	0.98	1.00	0.99	9096
1	0.71	0.30	0.42	233
accuracy			0.98	9329
macro avg	0.84	0.65	0.71	9329
weighted avg	0.98	0.98	0.98	9329

Table 3: XGBoost Baseline Overall Performance Results- SHAREWave8 Dataset

Overall Results:
Accuracy: 99.02 %
Classification Report:

	precision	recall	f1-score	support
0	0.99	1	0.99	9287
1	0.99	0.99	0.99	8906
accuracy			0.99	18193
macro avg	0.99	0.99	0.99	18193
weighted avg	0.99	0.99	0.99	18193

Table 4: XGBoost Overall Performance Results after Handling Imbalances and fine-tuning-SHAREWave8 Dataset

The initial XGBoost model was tested on the SHAREWave8 dataset comprising 46,733 instances, and taking into consideration input variables such as “ep….”, “hc…”, “age”, and “ph…” among other featured input variables. The variables were selected based on their relevance and availability in the SHAREw8 dataset, thus ensuring a comprehensive analysis of the factors influencing the outcomes. As presented in Table 3 above, the XGBoost model's performance was impressive with an overall accuracy of 97.94%- meaning the model correctly predicted the outcome in nearly 98 out of 100 of the fitted instances. Narrowing down to classification performance, the model, having been assessed on approximately 9096 instances of Class 0, exhibited exceptional precision on Class 0 (Cognitively Normal) with a value of 0.98, which indicated that 98% of the instances classified as Class 0 or cognitively normal were indeed correct. Furthermore, on the same cognitively normal class, it achieved a perfect recall score of 1.00, demonstrating the model’s ability to identify all instances of Class 0 without failing to detect any positively identified as cognitively normal. The presented balance between precision and recall is reflected in the outstanding F1-score of 0.99, a metric that confirms the model's remarkable performance on Class 0. Still, on the SHAREWave8 dataset, the XGBoost’s performance on Class 1 (Dementia) was comparatively less remarkable with a precision score of about 0.71 and a recall score of 0.30. The discrepancy between precision and recall is reflected in the relatively low F1-score of 0.42 possibly as the XGBoost model was tested on a smaller number of instances, 233, for Class 1, which may have contributed to the lower performance compared to Class 0.

The macro average provides an unbiased evaluation when considering both classes equally. The macro average precision of 0.84 indicated that, on average, 84% of the instances were correctly classified across both classes. On the same note, the macro average recall of 0.65 and the macro average F1-score of 0.71 represent a balanced measure of performance. That is true, as the weighted average, which considers the number of instances in each class, offers a more representative view of the model's overall performance. The weighted average precision of 0.98 aligns closely with the overall accuracy and the weighted average recall and F1 scores of 0.98 which shows that the model's predictions were highly precise across the SHAREWave8 dataset. Following the optimisation of hyper-parameters in the training datasets-ADNI and SHAREWave8 datasets, the XGBoost classifier demonstrated strong performance improvements in all the performance metrics (weighted averages-precision, recall, and f1-score) with performance on the SHAREWave8 dataset being the highest- 0.99 against 0.92 on the ADNI dataset (Table 4 and Table 2 respectively).

8.4 Logistic Regression Model Benchmark

However, the same performance metrics for the Logistic Regression model applied on the SHAREWave8 dataset were the second lowest with an overall accuracy of 97.56%, 0.97 weighted average precision, 0.98 weighted average recall and 0.97 f1-scores for the SHAREWave8 dataset (Table 5 below). In the same context, the baseline logistic regression classifier trained on the ADNI dataset attained 77.46% overall accuracy on the test set, correctly classifying just over 77% of cases across the 3 Alzheimer’s disease diagnostic categories. The model’s precision for class 0 was 0.73, f1-score 0.71 and recall reached 0.70 (Table 6 below).

Overall Results:

Accuracy: 97.56 %

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	9079
1	0.59	0.28	0.38	250
accuracy			0.98	9329
macro avg	0.79	0.64	0.68	9329
weighted avg	0.97	0.98	0.97	9329

Table 5: Logistic Regression Model Overall Performance Results- SHAREWave8 dataset

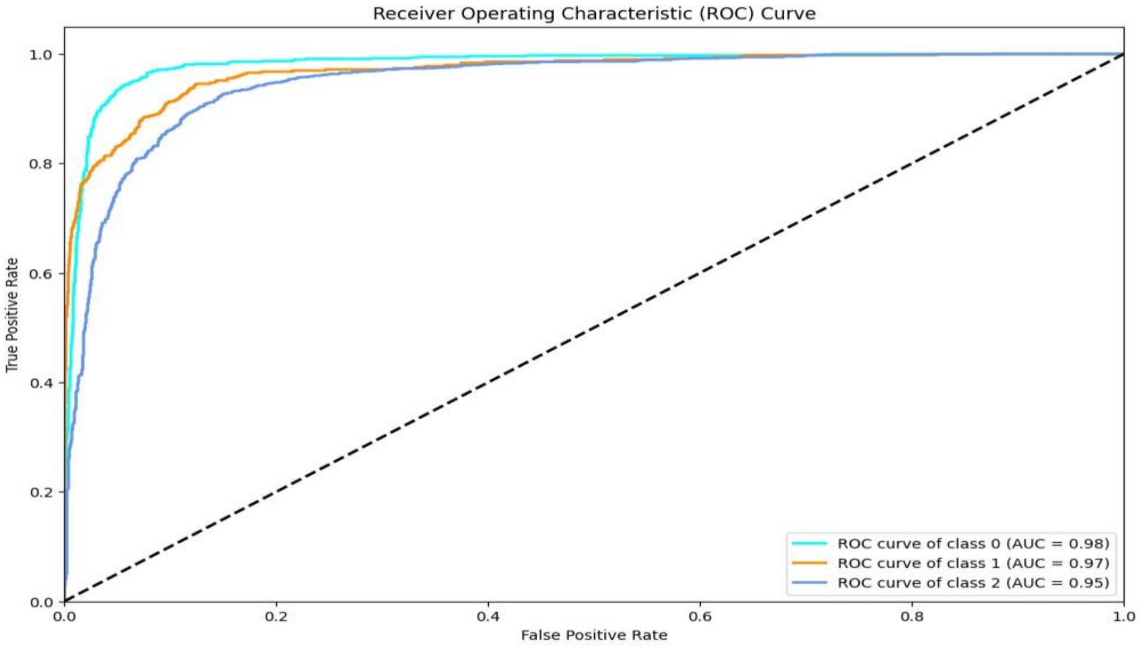
Overall Results:

Accuracy: 77.46 %

Classification Report:

	precision	recall	f1-score	support
0	0.73	0.7	0.71	771
1	0.86	0.76	0.81	495
2	0.77	0.82	0.79	1392
accuracy			0.77	2658
macro avg	0.79	0.76	0.77	2658
weighted avg	0.78	0.77	0.77	2658

Table 6: Logistic Regression Model Overall Performance Results- ADNI dataset



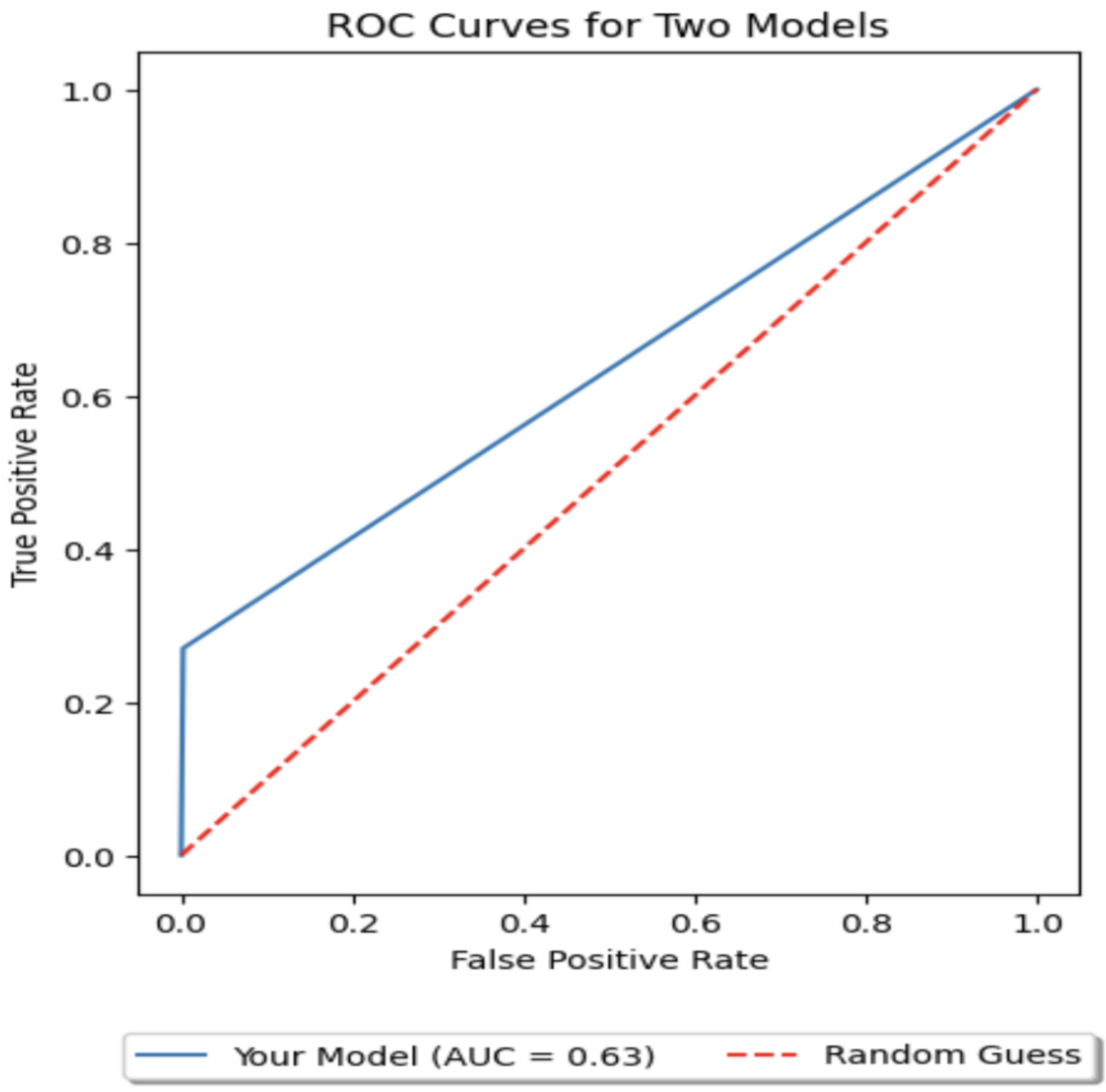
AUC-ROC (0): 0.98

AUC-ROC (1): 0.97

AUC-ROC (2): 0.95

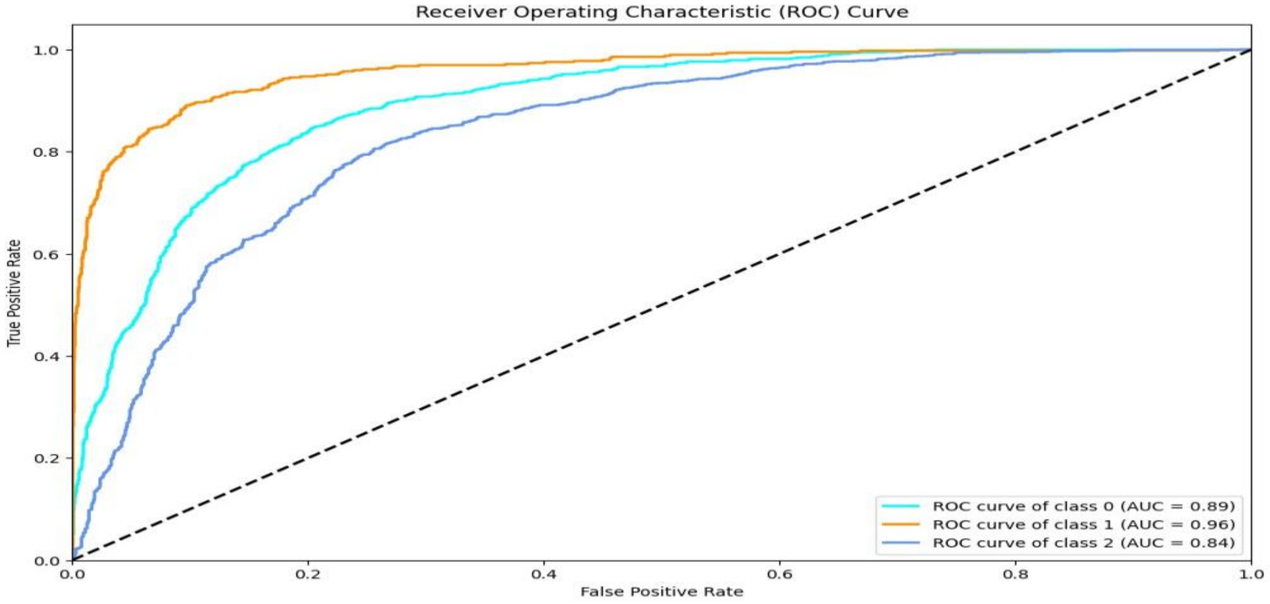
Average AUC-ROC: 0.97

Figure 5: XGBoost Baseline AUC-ROC Curve-ADNI dataset



Average AUC-ROC: 0.63

Figure 6: XGBoost Baseline ROC Curve-SHAREWave8 dataset



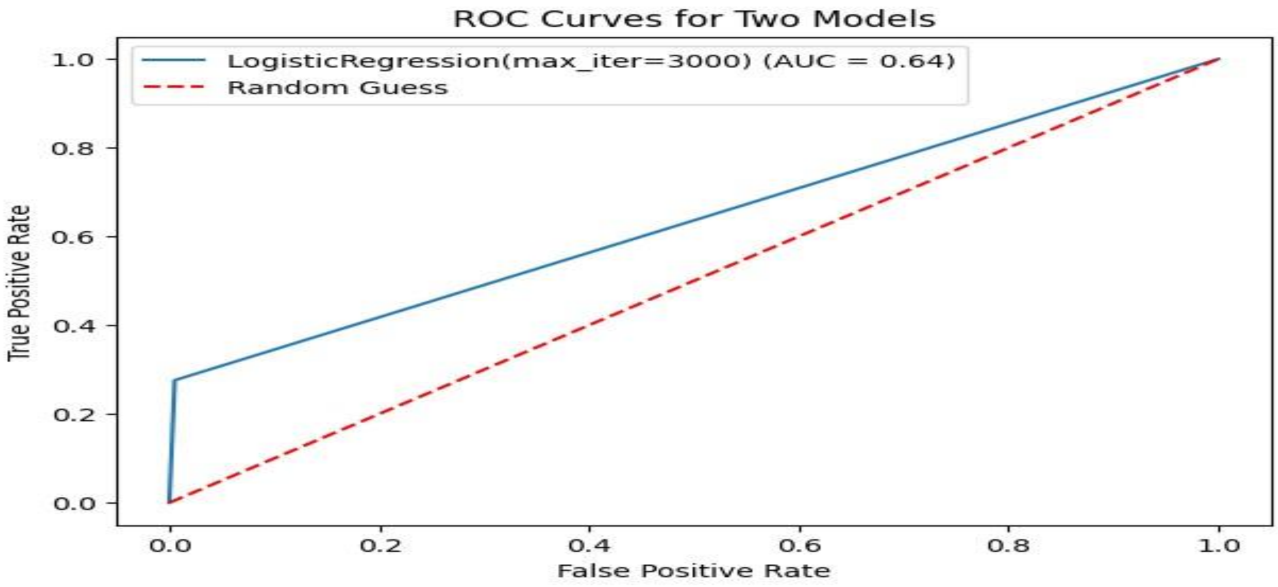
AUC-ROC (0): 0.89

AUC-ROC (1): 0.96

AUC-ROC (2): 0.84

Average AUC-ROC: 0.90

Figure 7: Logistic Regression Model ROC Curve for the ADNI dataset

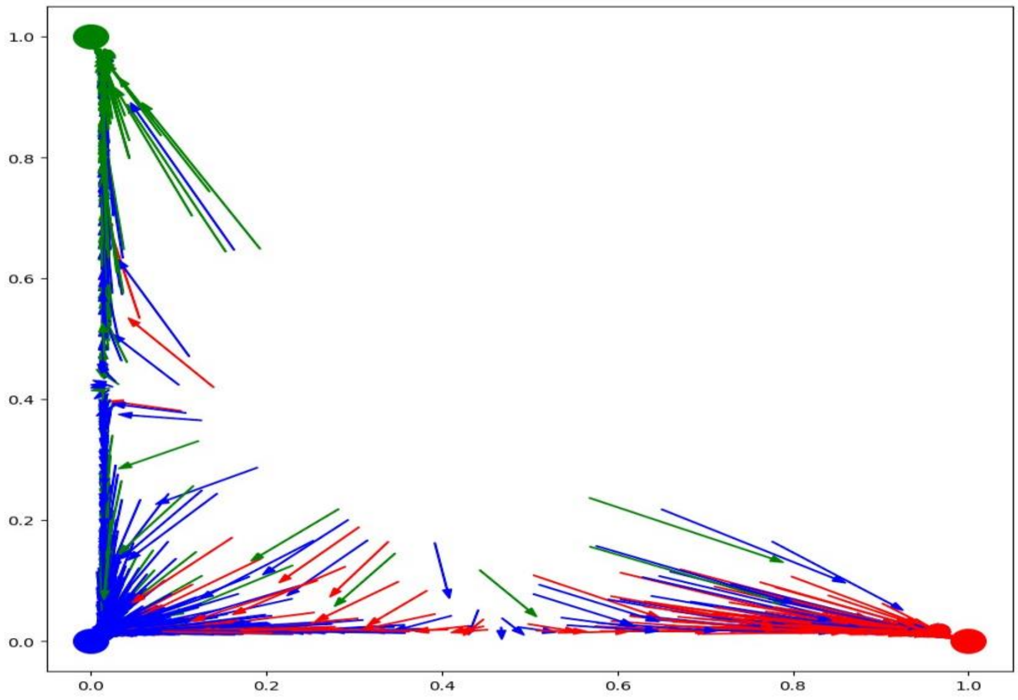


AUC-ROC (0): 0.64

Average AUC-ROC: 0.64

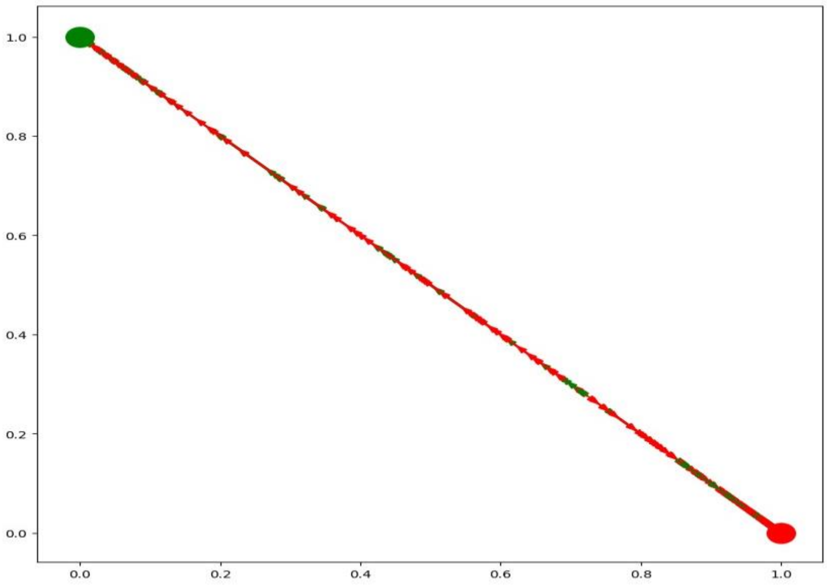
Figure 8: Logistic Regression Model Roc Curve-SHAREWave8

The Receiver operating characteristic (ROC) curve, in Figure 5, Figure 6, Figure 7, and Figure 8 above, showcases the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across thresholds for each diagnostic class. The area under the curve, the AUC-ROC curve, measures separability where higher scores indicate better discrimination capacity among categories. In that context, the XGBoost classifier recorded the highest performance of 0.97 on the ADNI dataset with the Logistic Regression model following closely with 0.90. As presented in Figure 5 and Figure 7, the XGBoost model registered the best discrimination score of -0.98 on the dementia class (class 0) while the Logistic Regression model had the best discrimination performance of 0.96 on the no dementia class (class 1). That signifies that both the XGBoost and the Logistic Regression models could reliably distinguish each dementia diagnostic category from the others with both high sensitivity and specificity-but more in the context of the ADNI dataset (0.90-0.97) than in the SHAREWave8 dataset (0.63-0.64).



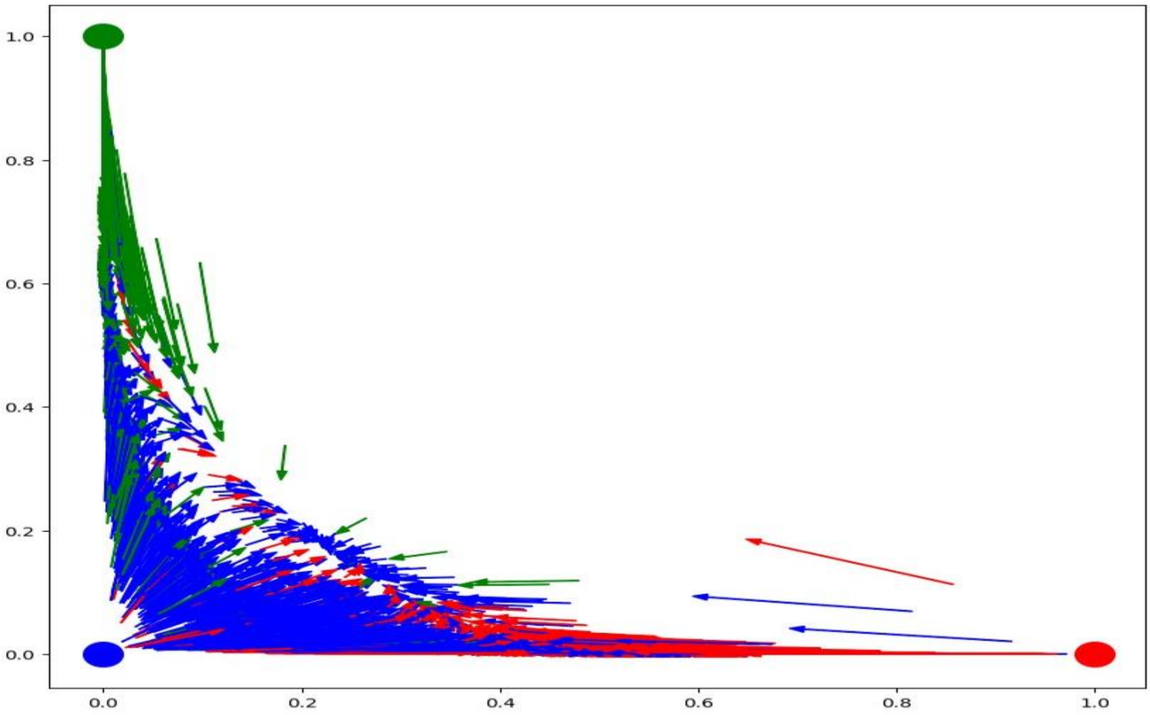
0.11755031570984512

Figure 9: XGBoost's Brier Score Calibration plot- the ADNI Dataset



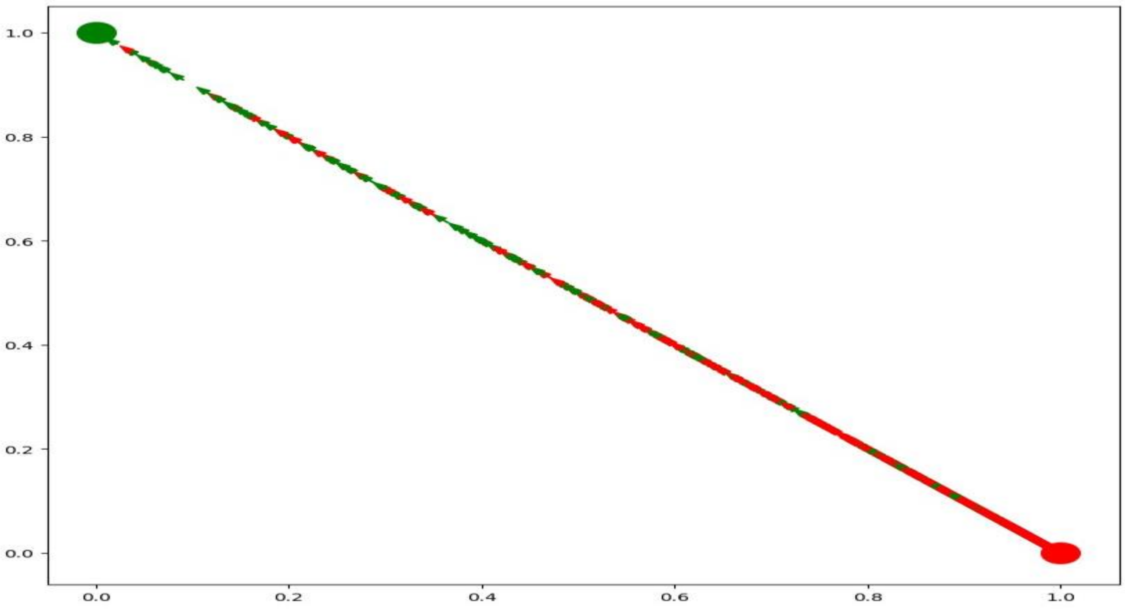
0.017331801771192992

Figure 10: XGBoost's Brier Score Calibration plot-the SHAREWave8 Dataset



0.33158358634903945

Figure 11: Logistic Regression Model Brier Score Calibration Plot- the ADNI dataset



0.019553400840217044

Figure 12: Logistic Regression Model Brier Score calibration plot- SHAREWave8 dataset

8.5 Brier scores and Calibration plots

The XGBoost model assessed on the ADNI dataset had a Brier score of about 0.117 (figure 9 above), which is relatively low and therefore better performing compared to logistic regression's 0.33 (figure 11 above). However, the XGBoost trained and assessed on the SHAREWave8 dataset had the best-performing Brier score of 0.017 (figure 10 above), which suggests the best performance, in terms of probability estimates and reliability, among the four case scenarios. The second model case performance was the Logistic Regression Model trained and fitted with the SHAREWave8 dataset (figure 12 above) with a Brier score of 0.019. The contrasting Brier scores imply XGBoost's better performance in terms of the model's predicted probabilities that match with the actual diagnostic as defined in the dataset, and in comparison, to the logistic regression model which recorded the second-best calibration score.

9.0 Discussion

Model/ Dataset	**Overall Accuracy**	**Precision (Weighted Avg)**	**Recall (Weighted Avg)**	**F1-Score (Weighted Avg)**
XGBoost - ADNI (Table 1)	88.64%	0.89	0.89	0.89
XGBoost - ADNI after adjustments (Table 2)	92.23%	0.92	0.92	0.92
XGBoost - SHAREWave8 (Table 3)	97.94%	0.98	0.98	0.98
XGBoost - SHAREWave8 after adjustments (Table 4)	99.02%	0.99	0.99	0.99
Logistic Regression – SHAREWave8 dataset (Table 5)	97.56%	0.97	0.98	0.97
Logistic Regression - ADNI dataset (Table 6)	77.46%	0.78	0.77	0.77

Table 7: Performance Comparative Analysis across Models and Datasets

The above comparative model performance (Table 7) suggests that the XGBoost model recorded the best performance across all metrics, proving to be a robust and effective solution for predicting dementia instances in both the ADNI and SHAREWave8 datasets. Notably, the adjustments or hyperparametric fine-tuning that went into handling class imbalances and fine-tuning led to substantial improvements in overall accuracy and weighted average metrics for the ADNI dataset. Even more remarkably, the adjustments propelled the XGBoost model to exceptional performance on the SHAREWave8 dataset, achieving an outstanding overall accuracy of 99.02% and a near-perfect weighted average precision, recall, and F1-score of 0.99. The notable improvement underlines the XGBoost's ability to adapt and excel when presented with varied data distributions and challenges. According to the comparative performance analysis table above, the transfer learned XGBoost model fine-tuned on multiple target input features (20 input features from about 79,144 instances for the SHAREWave8 dataset) achieved approximately 1.38% higher overall accuracy, 7% higher weighted average precision, and 7% higher weighted average recall and F1 scores compared to the same transfer model trained on a limited scale dataset (the ADNI dataset characterised by 10 input features from about 16,421 instances).

9.1 Baseline Logistic Regression Vs Baseline Transfer Learned XGBoost Model

While the baseline logistic regression model achieved moderately accurate predictions on the ADNI dataset with 77.46% overall accuracy, its weighted average precision of 0.78, recall of 0.77, and F1-score of 0.77 suggesting room for improvement (Table 7). In comparison, the baseline XGBoost transfer learned model demonstrated superior performance with 88.64% overall accuracy and higher weighted averages of 0.89 precision, 0.89 recall, and 0.89 F1-score on the same baseline ADNI dataset before handling imbalances.

On the SHAREWave8 dataset, the logistic regression model attained a far better performance compared to the ADNI dataset with a 97.56% overall accuracy, a weighted average precision of 0.97, recall of 0.98, and an F1-score of 0.97, which signifies an overall accuracy improvement of 20.1% and 19%+ improvement in the other performance metrics (Table 7). However, the transfer learned XGBoost model still outperformed with 0.98 weighted averages for precision, recall and F1-score, showing an edge over logistic regression. In that case, the XGBoost classifier consistently provided higher weighted average precision, recall and F1-scores than the baseline logistic regression approach across both the ADNI and SHAREWave8 datasets even though the Logistic Regression model recorded the highest per cent point (19%+) improvement across all the performance metrics. While the Logistic Regression model performance may not have matched the adjusted XGBoost model's performance on the ADNI dataset, it still demonstrated high accuracy (97.56%) and commendable weighted average scores, which suggest its viability as an alternative solution, especially in scenarios where interpretability and model simplicity are a priority over marginal performance gains.

Model/ Dataset	**AUC-ROC (Class 0)**	**AUC-ROC (Class 1)**	**AUC-ROC (Class 2)**	**Average AUC-ROC**
XGBoost - ADNI (Fig. 5)	0.98	0.97	0.95	0.97
XGBoost - SHAREWave8 (Fig. 6)	-	-	0.63	0.63
Logistic Regression - ADNI (Fig. 7)	0.89	0.96	0.84	0.90
Logistic Regression - SHAREWave8 (Fig. 8)	0.64	-	-	0.64

Table 8: AUC-ROC Performance across Models and Dataset Classes

The AUC-ROC comparative assessment above (Table 8) reveals that while both models: XGBoost and the Logistic Regression models perform well on the ADNI dataset, their ability to discriminate between classes in the SHAREWave8 dataset varied, with a noticeable performance drop. The comparative table highlights the need for model and dataset-specific fine-tuning to optimise performance, especially in complex application settings like medical diagnostics.

Model/Dataset	**Brier Score**
XGBoost - ADNI (Fig. 9)	0.117
XGBoost - SHAREWave8 (Fig. 10)	0.017
Logistic Regression - ADNI (Fig. 11)	0.332
Logistic Regression - SHAREWave8 (Fig. 12)	0.020

Table 9: Brier Score Comparative Analysis across Models and Dataset

As presented in Table 9 above, the XGBoost model recorded a 0.117 Brier Score, which signals reasonable calibration, although there is room for improvement- aligning its probabilistic predictions with actual outcomes. In comparison, the Logistic Regression model's 0.332 Brier Score on the same dataset suggests a greater discrepancy between predicted probabilities and actual outcomes compared to the XGBoost model and, therefore less reliable probabilistic predictions. The same Brier score on the SHAREWave8 dataset suggests more reliable probabilistic predictions for both the XGBoost model (0.017) and the Logistic Regression model (0.020). The low Brier scores imply that both models are highly accurate and reliable for the SHAREWave8 dataset with the XGBoost model having a slight edge in the calibration performance. These results are consistent with previous studies reporting superior performance of transfer learning-based gradient boosting models, such as XGBoost, in Alzheimer's disease classification tasks (Lee et al., 2019; Dara et al., 2023). The high accuracy levels, especially after fine-tuning the transfer learned models reflect the model’s effective learning of the underlying patterns and relationships between the input features and the diagnostic outcomes, therefore enabling accurate dementia risk classification.

Particularly, the XGBoost model demonstrated remarkable precision and recall scores across all the diagnostic classes, specifically for the dementia and mild cognitive impairment categories. The finding is clinically significant as it indicates the model's ability to correctly identify individuals with Alzheimer's disease or mild cognitive impairment, which is useful for early intervention and disease management (Ritchie et al., 2017). The high recall scores by the transfer learned XGBoost-fine-tuned on best hyper-parameters, denote that the model is more sensitive in detecting true positive cases, therefore minimising the risk of missing individuals who may require further evaluation and treatment.

9.2 Results after Model Hyper-Parameter Tuning and Retraining

9.2.0 Bias and Fairness Assessment

The Fairlearn MetricFrame analysis reveals differences in model performance across groups defined by the gender-sensitive feature PTGENDER. The results, as presented in Appendix 7, show that the XGBoost baseline model registered close accuracies in both gender groups (Group 0 and Group 1). Nonetheless, the precision and recall results in the same appendices indicate that gender group 0 had a slightly higher precision and recall than gender group 1, which implies the model performs slightly better for Group 0 (male Group). The disparities are small and might not be statistically significant, but not negligible at scale, suggesting that there may be some bias in the model’s predictions based on gender. This could be a result of various factors such as imbalances in the training dataset, inherent biases in the model’s learning process, or other confounding variables altogether.

9.2.1 Handling Skewed Class Distribution

A strategy commonly referred to as the synthetic minority oversampling technique (SMOTE) was implemented as a further pre-processing step to handle the original imbalanced distribution in the sampled male and female categories. The SMOTE algorithmically generates new minority class samples based on nearest neighbours, therefore balancing the overall gender distribution across the model diagnostic class proportions.

Class=0, n=6843 (33.333%)
Class=1, n=6843 (33.333%)
Class=2, n=6843 (33.333%)

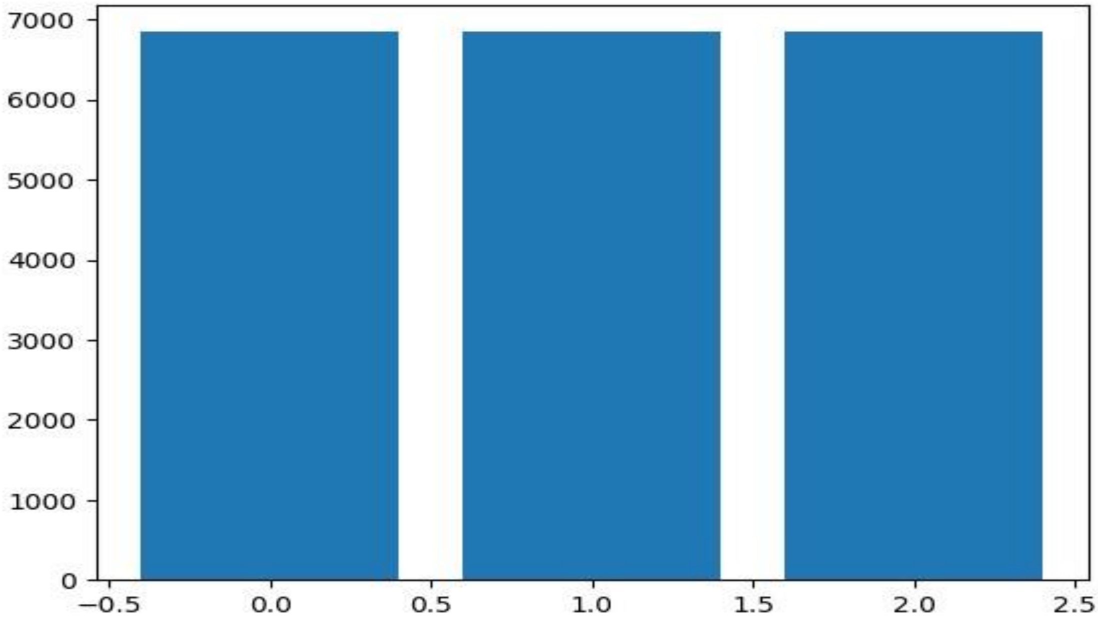


Figure 13: SMOTE gender balancing across diagnostic categories

After systematic oversampling, the training dataset attained an equal class distribution at 33% per diagnostic category presented in Figure 13 above. The balanced representation mitigates bias during model training, providing room for equivalent sensitivity across classes.

9.2.2 Retrained XGBoost Model Performance

After balancing the classes and retraining the XGBoost model on the best hyper-parameters on both the ADNI and the SHAREWave8 datasets, the XGBoost classifier registered varying performance improvement between the two datasets. Particularly, the retrained XGBoost model recorded a 3.59% improvement in the accuracy score for the ADNI dataset- from 88.64% accuracy on the baseline model to 92% when handled for imbalances and retrained with the best hyper-parameters (Table 1 and Table 2 in the results section). As for the weighted average precision, recall, and f1-scores, the XGBoost model had a 3% improvement, from 89% to 92% score on all three performance metrics. When the XGBoost model was retrained on the best hyper-parameters on the SHAREWave dataset, which was also handled for imbalances, the overall accuracy score improved from 97.94% to 99.02%- a 1.08% improvement (Tables 3 and Table 4). The same model had an improved weighted average precision, recall, and f1-score of 1%-having improved from 98% to 99%.

9.3 SHAP Explanations and Feature Importance

A SHAP summary plot generated in Figure 14 below quantifies the feature impact on model outputs for the diagnostic classification task. The visualisation presented in the stacked bar charts reveals that MRI-derived measurement MMSE wielded the most influential impact across all the diagnostic categories. Following close behind MMSE were the DX_bl and RAVLT_immediate features, which also emerged as informative inputs, especially the most influencing predictions for classes 0 and 2. Conversely, features including gender and RAVLT_forgetting exhibited minimal influence on model reasoning processes for assigning diagnoses. The SHAP summary plot confirms that MRI volumes, the Rey Auditory Verbal Learning Test (RAVLT) assessments and cognitive function exam results represent particularly pivotal data sources for Alzheimer’s disease or dementia diagnostic classification, while demographics confer little effect. The finding aligns with reported results on Alzheimer's disease pathophysiology, confirming that structural brain changes and cognitive impairments are hallmarks of the disease (Jessen et al., 2014; Hampel et al., 2018). The SHAP summary plot further insists on the importance of neuroimaging and cognitive assessments in the dementia diagnostic process, providing useful insights into the model's decision-making process and potentially guiding future feature selection strategies in Alzheimer’s risk assessment.

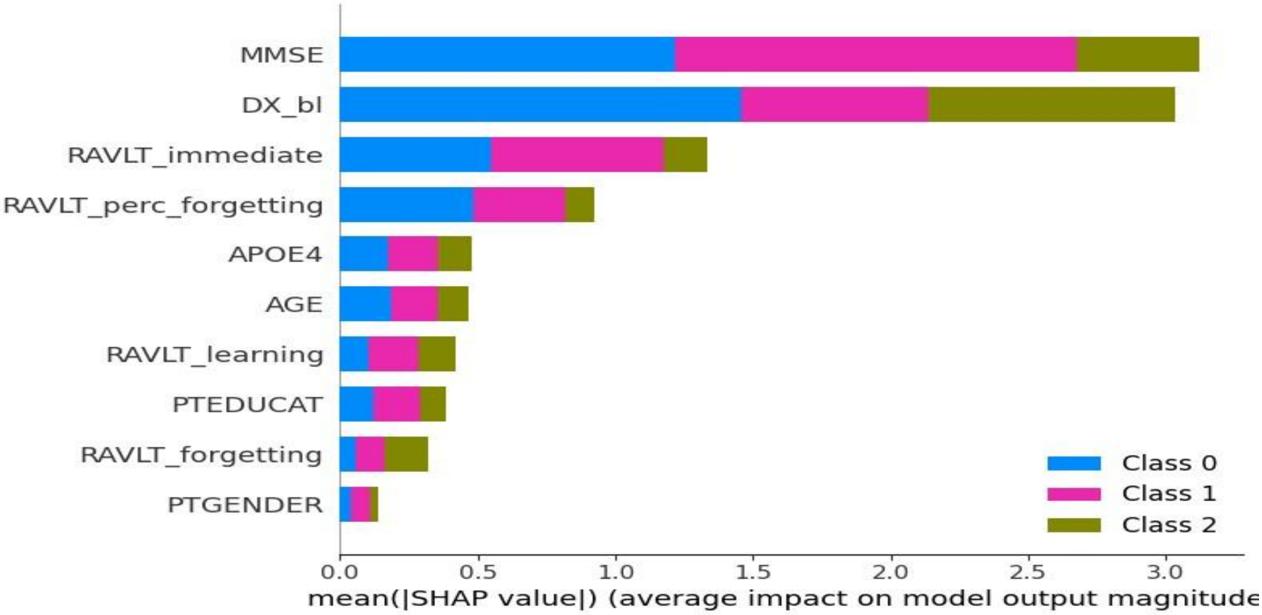


Figure 14: XGBoost SHAP Summary Plot

9.4 Survival analysis using Kaplan-Meier and Cox Proportional Hazards models

The Kaplan-Meier survival plots estimate the survival function for Alzheimer's disease progression based on age, with the first plot presented in Appendix 5, providing an overall survival estimate over time. While the second plot- in Appendix 6, differentiates between cohorts with and without positive axillary nodes detected in the ADNI dataset; though the cohort without positive axillary nodes showed no significant difference in survival probabilities between the two groups. The Cox Proportional Hazards Model, a semi-parametric survival analysis technique, indicates that input features such as APOE4 allele, initial diagnosis, and gender had statistically significant effects on survival times of the sampled population in the ADNI Dataset, with the APOE4 allele leading to a 47% increase in hazard for every one-unit increase (Appendix 3). Notably, the model's moderate predictive power is reflected in a 0.60 concordance index (Appendix 4). The Kaplan-Meier plots exhibit the expected decrease in survival probability over age, while the Cox model feature importance plot highlights the log (Hazard Ratio) for each feature, presenting insights into their relative impacts on survival.

The Cox proportional hazards model results, as presented in Appendices 3 and 4 suggest multiple feature variables had a statistically significant association with time to diagnosis of Alzheimer's disease. Specifically, the presence of the APOE4 genetic variant was associated with a 47% higher hazard of earlier diagnosis (hazard ratio 1.47), a diagnosis of mild cognitive impairment at baseline increased the hazard by 2.23 times compared to cognitively normal whereas being male decreased the hazard by 25% (hazard ratio 0.75) compared to females. Variables related to memory and learning, such as MMSE score and RAVLT forgetting rate, also showed significant associations. As additionally reported in the Table, in Appendix 3, the Cox proportional hazard model has a concordance or predictive accuracy value of 0.60, which suggests that the model has just above-random predictive power and is far from perfect. The log-likelihood ratio test suggests that the model is significantly better than a null model with no predictors.

The Cox Results Tables in Appendices 3 and 4 suggest that both genetic factors like APOE4 and clinical baseline factors like cognitive tests and gender appear to influence dementia disease prognosis in a statistically significant way.

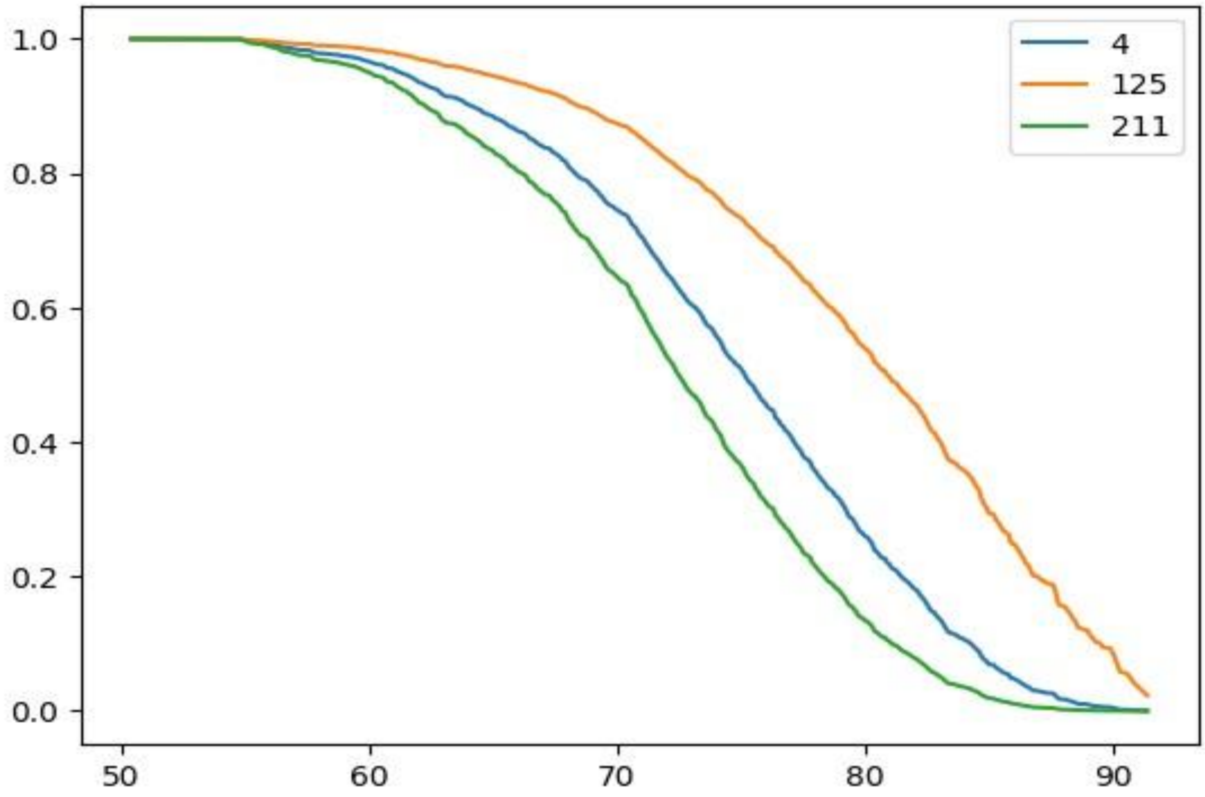


Figure 15: Predicted Survival Function for Three Selected Patients Based on Cox Proportional Hazards Model

Figure 15, above, presents the predicted survival function for three selected patients (with IDs 4, 125, 211) based on the Cox Proportional Hazards Model. The survival function is the probability of surviving beyond a given time point, in this case, age. In the plot, the model estimates the survival probability for each of the randomly selected patients at different ages, based on their features and the model coefficients. In that context, the patient with ID 4 had a survival probability of approximately 0.8 at age 70, signifying that there was an 80% chance that they would survive beyond 70 years old without being diagnosed with Alzheimer’s disease. The patient with ID 125 had a lower survival probability of approximately 0.6 at age 70, signifying that there is a 60% chance that they will survive beyond 70 years old without being diagnosed with Alzheimer’s disease. The patient with ID 211 had the lowest survival probability of approximately 0.4 at age 70, signifying that there was a 40% chance that they would survive beyond 70 years old without being diagnosed with Alzheimer’s disease.

The plot also demonstrates how the survival probabilities change over time for each of the sampled patients. For example, the patient with ID 4 had a relatively flat curve, signifying that their survival probability did not decrease much with age. The patient with ID 125 has a steeper curve, signifying that their survival probability decreases faster with age. The patient with ID 211 had the steepest curve, signifying that their survival probability decreased the fastest with age, which suggests that the patient with ID 211 had the highest risk of developing Alzheimer’s disease, while the patient with ID 4 had the lowest risk. The patient with ID 125 had a moderate risk. The observed differences may be due to the different features of each patient, such as their gender, education, initial diagnosis, APOE4 allele, MMSE score, and RAVLT scores.

The survival analysis assessment using the Cox Proportional Hazards model was included, not for comparative analysis, but for cross-validation of the fitted modelling features. The analysis results shed light on the prognostic factors associated with Alzheimer's disease progression. According to results, the presence of the APOE4 genetic variant, a well-established risk factor for Alzheimer's disease (Corder et al., 1993), was found to be associated with a 47% higher hazard of earlier diagnosis. The finding is consistent with previous studies that have highlighted the influence of the APOE4 allele on the age of onset and disease progression (Liu et al., 2013; Cosentino et al., 2008). Additionally, the baseline diagnosis (DX-bl) of mild cognitive impairment increased the hazard of Alzheimer's disease diagnosis by 2.23 times compared to cognitively normal individuals, further emphasising the importance of early detection and intervention in the dementia-prone population. The Kaplan-Meier survival plots provided visual representations of the survival probabilities over time, which demonstrated the expected decrease in survival probability with increasing age.

9.5 Implications for Clinical Practice

The study findings infer multiple implications that could enhance clinical practice in the diagnosis and management of Alzheimer's disease and related Dementia. On one hand, the high accuracy and precision achieved by the XGBoost model suggest its potential for integration into clinical decision support systems. By taking advantage of the transfer learned model's predictive capabilities, clinicians can enhance diagnostic accuracy, reducing the risk of misdiagnosis and enabling timely intervention as well as treatment initiation, which is a critical factor, particularly in the early stages of the disease when small cognitive changes may be challenging to detect through traditional diagnostic methods. The transfer learned XGBoost model's ability to accurately classify individuals into mild cognitive impairment and dementia classes signifies plausible application in the early detection of cognitive decline, in real-world clinical settings. On the other hand, the integration of the Cox Proportional Hazards model can provide clinicians with valuable insights into individual patient’s risk factors and prognostic indicators for Alzheimer's disease progression. Clinicians can develop personalised risk assessments and tailor treatment strategies accordingly by considering factors such as APOE4 genotype Baseline cognitive status, and other relevant features as informed by the SHAP assessment, which will eventually align with the growing emphasis on precision medicine and patient-centred care in Alzheimer's disease management. Finally, the machine learning models assessed in this study can serve as valuable clinical decision support tools and therefore complement expertise diagnoses in the healthcare sector. That would be possible by integrating the model's predictions with clinical judgment and multidisciplinary input from neurologists, radiologists, and other specialists forming a more comprehensive and informed decision-making process that can lead to improved patient outcomes.

10.0 Directions for Future Research

While the current assessment study implies promising application for the transfer learned XGBoost ML model, there is still more validation and cross-assessment to be done for definitive application in clinical settings. Notably, the study utilised the ADNI and SHAREwave8 datasets for model training and testing, which may not fully represent the diverse demographics and clinical characteristics of the general population in whichever clinical setting. As a result, future research should focus on validating the models’ performance on larger and more diverse datasets to ensure their generalizability and robustness across different population groups and clinical data attributes. For example, future research could explore the integration of multimodal data including genetic information, neuroimaging data (such as functional MRI, PET scans among others), and lifestyle factors, to further enhance the models' predictive power and provide a more holistic understanding of Alzheimer’s disease. Finally, while the current study incorporated survival analysis input feature cross-validation strategy, a more comprehensive analysis of longitudinal data could provide deeper insights into the temporal dynamics of Alzheimer's disease progression. Including longitudinal data from multiple time points could enable the development of predictive models that capture the disease trajectory along a diverse timeline and inform personalised treatment plans.

11.0 Conclusion

From the results, it is evident that interpretable machine learning models, particularly the transfer learned XGBoost demonstrate potentially superior performance in personalised dementia risk prediction over classical ML models without the transfer learning-Logistic Regression model. The incorporation of hyper-parameter tuning significantly improved performance, with the XGBoost model consistently outperforming the logistic regression model across both the SHAREWave8 and the ADNI datasets. The Kaplan-Meier/Cox Proportional Hazards model and the SHAP framework were resourceful in validating the model input features and interpretability of the models, hence, contributing to understanding the models' classification process. As discussed in the SHAP results, the MRI-derived measurements including the Mini-Mental State Examination (MMSE) score and cognitive assessments like the Rey Auditory Verbal Learning Test (RAVLT) were among the most influential features for the model's predictions, which confirmed the importance of neuroimaging and cognitive assessments in the Alzheimer's disease diagnostic process.

The interpretable gradient boosting framework with transfer learning, implemented through the XGBoost algorithm, significantly outperformed the Logistic Regression Model in the context of both the ADNI and the SHAREWave8 datasets. Additionally, the XGBoost model demonstrated superior discrimination capabilities, with an average AUC-ROC of 0.97 on the ADNI dataset, indicating remarkable sensitivity and specificity in distinguishing between diagnostic categories. The XGBoost model's calibration also, as assessed by the Brier score, recorded superior, with the best Brier score of 0.017 on the SHAREwave8 dataset. The results prove that the XGBoost model demonstrated substantial performance improvements after class imbalance handling and hyperparameter fine-tuning. On one hand, in the ADNI dataset, for example, the adjustments led to a 3.59% improvement in overall accuracy and a 3% improvement in weighted average precision, recall, and F1 scores. On the other hand, on the SHAREwave8 dataset, the adjustments enhanced the XGBoost model to an exceptional overall accuracy of 99.02% and a near-perfect weighted average precision, recall, and F1-score of 0.99, showcasing a 1.08% improvement compared to the baseline model.

Notably, transfer learning enabled the effective adaptation of the XGBoost model to heterogeneous real-world datasets, improving the model's ability to generalise and make accurate predictions, especially after hyperparametric tuning. The findings show that the transfer learned XGBoost model achieved higher performance metrics (approximately 1.38% higher overall accuracy, 7% higher weighted average precision, recall, and F1 scores) when trained on the larger SHAREwave8 dataset (20 input features, 79,144 instances) compared to the smaller ADNI dataset

(10 input features, 16,421 instances). The same improvement is also evident in the case of the Logistic Regression model, which recorded a 19%+ improvement across the performance metrics. In that context, it can be concluded that the characteristics of the dataset, particularly the size and number of input features, can influence the performance of machine learning models. The improved performance on the larger SHAREwave8 dataset presents room for speculation that having access to a larger and more diverse set of instances and input features during the training process allowed the transfer learned XGBoost model to better capture the underlying patterns and relationships within the data, leading to more accurate and generalizable predictions. Future research should focus on validating transfer ML model performance using different clinical datasets and assessing for the quantitative influence of data variability or diversity on transfer learned model performance.

Word Count: 9, 772

References

Abbas, A., Abdelsamea, M.M. and Gaber, M.M., 2020. Detrac: Transfer learning of class decom posed medical images in convolutional neural networks. *IEEE Access, 8,* pp.74901-74913.

Alzheimer's Disease Neuroimaging Initiative (ADNI), 2021. ADNI3 clinical data [Data set]. ADNI. https://adni.loni.usc.edu/adni-3/

Alzheimer's Disease International [ADI] 2022. World Alzheimer's Report 2022. *ADI.* Available at https://www.alzint.org/resource/world-alzheimer-report-2022/#:~:text=The%20World%20Alzheimer%20Report%202022%20is%20dedicated%20to%20the%20vast,their%20carers%20after%20a%20diagnosis. Retrieved on 31st Oct 2023.

Battineni, G., Hossain, M.A., Chintalapudi, N., Traini, E., Dhulipalla, V.R., Ramasamy, M. and Amenta, F., 2021. Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms. *Diagnostics*, *11*(11), p.2103.

Battista, P., Salvatore, C., Berlingeri, M., Cerasa, A. and Castiglioni, I., 2020. Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. *Neuroscience & Biobehavioral Reviews*, *114*, pp.211-228.

Bogdanovic, B., Eftimov, T. and Simjanoska, M., 2022. In-depth insights into Alzheimer's disease by using explainable machine learning approach. *Scientific Reports*, *12*(1), p.6508.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. and Zuber, S., 2013. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *International journal of epidemiology*, *42*(4), pp.992-1001.

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J., 2021. Explainable machine learning in credit risk management. *Computational Economics*, *57*, pp.203-216.

Chhetri, T.R., Dehury, C.K., Lind, A., Srirama, S.N. and Fensel, A., 2022. A Combined System Metrics Approach to Cloud Service Reliability Using Artificial Intelligence. *Big Data and Cognitive Computing*, 6(1), p.26.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G., ... & Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, *261*(5123), 921-923.

Cosentino, S., Scarmeas, N., Helzner, E., Glymour, M. M., Brandt, J., Albert, M., ... & Stern, Y. (2008). APOE ε4 allele predicts faster cognitive decline in mild Alzheimer dis ease. *Neurology*, *70*(19_part_2), 1842-1849.

Dara, O. A., Lopez-Guede, J. M., Raheem, H. I., Rahebi, J., Zulueta, E., & Fernandez-Gamiz, U. (2023). Alzheimer’s Disease Diagnosis Using Machine Learning: A Survey. *Applied Sciences*, *13*(14), 8298.

Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of animal ecology*, *77*(4), pp.802-813.

Hampel, H., O’Bryant, S. E., Molinuevo, J. L., Zetterberg, H., Masters, C. L., Lista, S., ... & Blennow, K. (2018). Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, *14*(11), 639-652.

Hampel, H., Vergallo, A., Aguilar, L. F., Benda, N., Broich, K., Cuello, A. C., ... & Lista, S. (2018). Precision pharmacology for Alzheimer’s disease. *Pharmacological research*, *130*, 331-365.

Iddi, S., Li, D., Aisen, P.S., Rafii, M.S., Thompson, W.K., Donohue, M.C. and Alzheimer’s Disease Neuroimaging Initiative, 2019. Predicting the course of Alzheimer’s progression. *Brain informatics*, *6*, pp.1-18.

Jessen, F., Wiese, B., Bickel, H., Eifflaender-Gorfer, S., Fuchs, A., Kaduszkiewicz, H., ... & Age CoDe Study Group. (2011). Prediction of dementia in primary care patients. *PloS one*, *6*(2), e16852.

Jessen, F., Wiese, B., Bickel, H., Eifflaender-Gorfer, S., Fuchs, A., Kaduszkiewicz, H., ... & Age CoDe Study Group. (2011). Prediction of dementia in primary care patients. *PloS one*, *6*(2), e16852.

Jessen, F., Wolfsgruber, S., Wiese, B., Bickel, H., Mösch, E., Kaduszkiewicz, H., ... & on Aging, G. S. (2014). AD dementia risk in late MCI, in early MCI, and in subjective memory impairment. *Alzheimer's & Dementia*, *10*(1), 76-83.

Jessen, F., Wolfsgruber, S., Wiese, B., Bickel, H., Mösch, E., Kaduszkiewicz, H., ... & on Aging, G. S. (2014). AD dementia risk in late MCI, in early MCI, and in subjective memory impairment. *Alzheimer's & Dementia*, *10*(1), 76-83.

Jo, T., Nho, K., Risacher, S.L., Saykin, A.J. and Alzheimer’s Neuroimaging Initiative, 2020. Deep learning detection of informative features in tau PET for Alzheimer’s disease classification. *BMC bioinformatics*, *21*, pp.1-13.

Kanani, A., Vaidya, S. and Agarwal, H., 2021, September. LightFPGA: Scalable and Automated FPGA Acceleration of LightGBM for Machine Learning Applications. In *2021 25th International Symposium on VLSI Design and Test (VDAT)* (pp. 1-6). IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information pro cessing systems, 30.*

Kumar, S., Oh, I., Schindler, S., Lai, A.M., Payne, P.R. and Gupta, A., 2021. Machine learning for modelling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA open*, *4*(3), p.ooab052.

Lee, G., Nho, K., Kang, B., Sohn, K. A., & Kim, D. (2019). Predicting Alzheimer's disease pro gression using multi-modal deep learning approach. *Scientific reports, 9*(1), 1952.

Lee, G., Nho, K., Kang, B., Sohn, K.A. and Kim, D., 2019. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, *9*(1), p.1952.

Liu, C. C., Kanekiyo, T., Xu, H., & Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology, 9*(2), 106-118.

Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., Pantaleo, E., Logroscino, G., De Blasi, R., Tangaro, S. and Bellotti, R., 2022. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. *Brain informatics*, *9*(1), pp.1-17.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), pp.56-67..

Marcílio, W.E. and Eler, D.M., 2020, November. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)* (pp. 340-347). Ieee.

Masud, M., Hossain, M.S., Alhumyani, H., Alshamrani, S.S., Cheikhrouhou, O., Ibrahim, S., Muhammad, G., Rashed, A.E.E. and Gupta, B.B., 2021. Pre-trained convolutional neural networks for breast cancer detection using ultrasound images. *ACM Transactions on Internet Technology (TOIT)*, *21*(4), pp.1-17.

Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobot ics, 7,* p.21.

Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M. and Aydin, N., 2022. Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. *Computational biology and chemistry*, *97*, p.107619.

Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T. and Alzheimer's Disease Neuroimaging Initiative, 2020. Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, *222*, p.117203.

Petersen, R. C., Lopez, O., Armstrong, M. J., Getchius, T. S., Ganguli, M., Gloss, D., ... & Rae-Grant, A. (2018). Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology*, *90*(3), 126-135.

Ritchie, C. W., Russ, T. C., Banerjee, S., Barber, B., Boaden, A., Fox, N. C., ... & Burns, A. (2017). The Edinburgh Consensus: preparing for the advent of disease-modifying therapies for Alzheimer's disease. *Alzheimer's research & therapy*, *9*, 1-7.

Ritchie, C. W., Russ, T. C., Banerjee, S., Barber, B., Boaden, A., Fox, N. C., ... & Burns, A. (2017). The Edinburgh Consensus: preparing for the advent of disease-modifying therapies for Alzheimer's disease. *Alzheimer's research & therapy*, *9*, 1-7.

SHARE-Eric.Eu 2022. Wave 1. Retieved on 17th Nov 2023 from https://share-eric.eu/data/data-documentation/waves-overview/wave-1

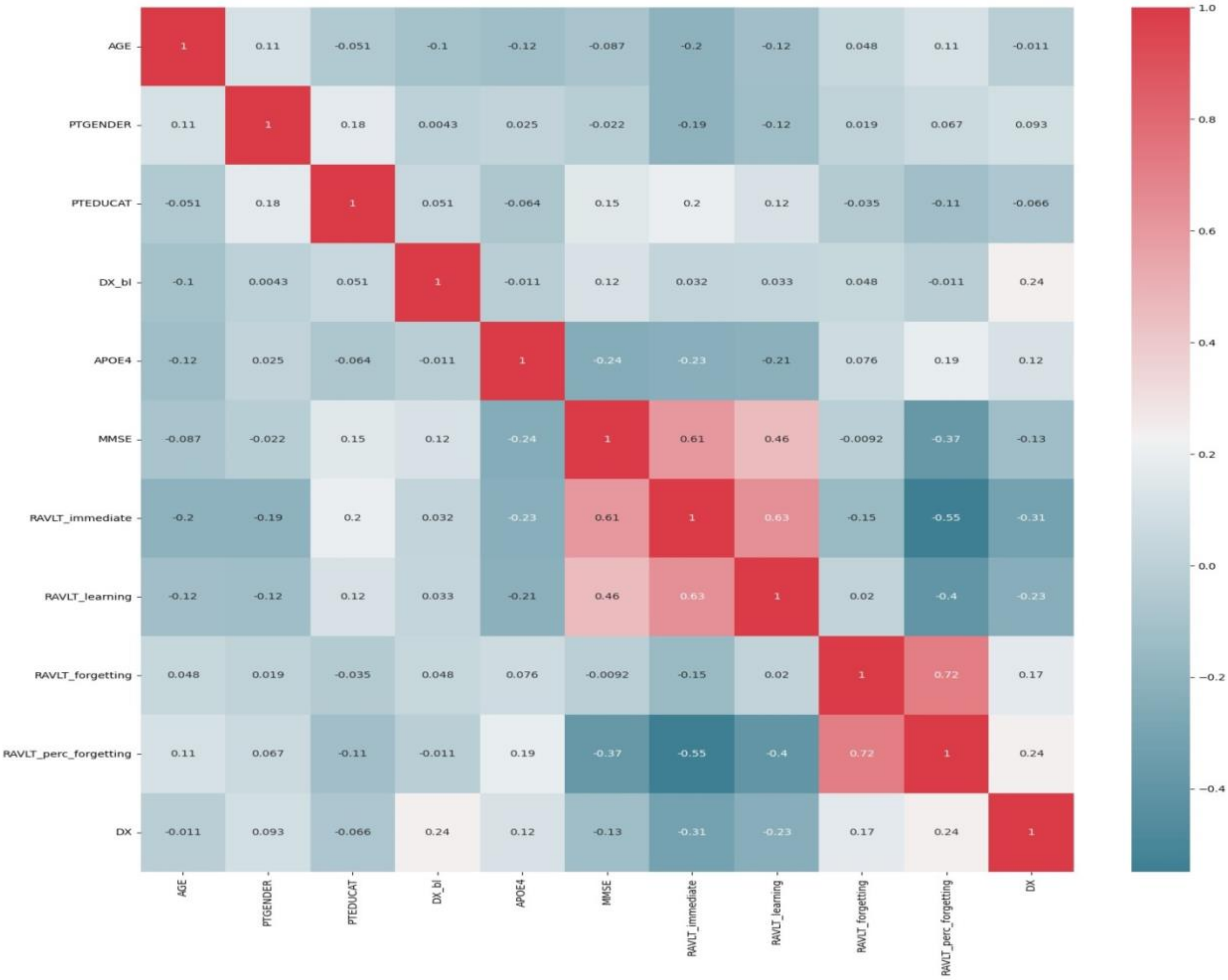
Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N.A., Trollor, J. and Brodaty, H., 2020. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, *10*(1), p.20410.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C., 2018. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27* (pp. 270-279). Springer International Publishing.

Wang, M., Greenberg, M., Forkert, N.D., Chekouo, T., Afriyie, G., Ismail, Z., Smith, E.E. and Sajobi, T.T., 2022. Dementia risk prediction in individuals with mild cognitive impairment: a comparison of Cox regression and machine learning models. *BMC Medical Research Methodology*, *22*(1), p.284.

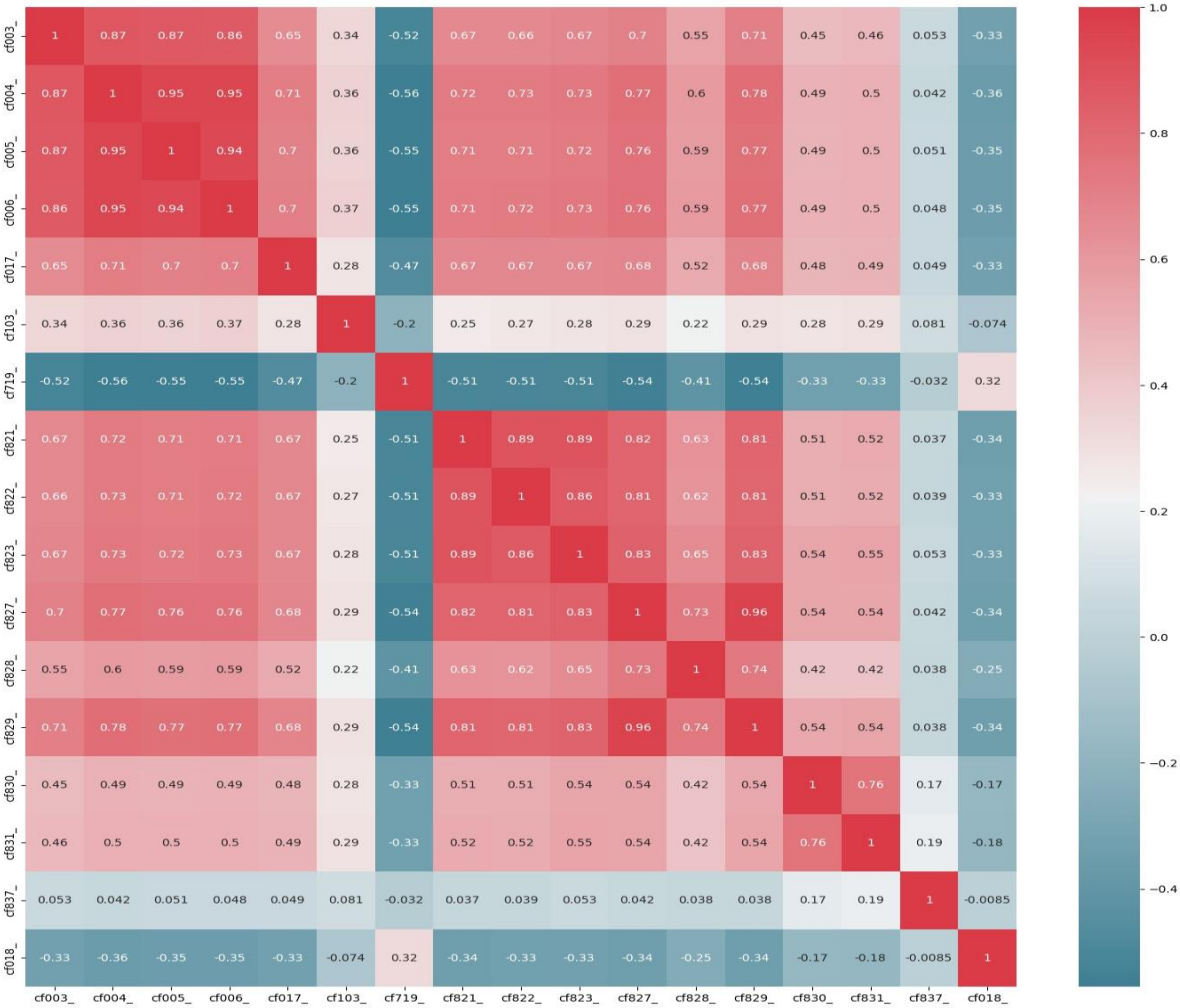
Appendices

Appendix 1: ADNI Input Feature Correlation Heat Map



	AGE	PTGENDER	PTEDUCAT	DX_bl	APOE4	MMSE	RAVLT_immediate	RAVLT_learning	RAVLT_forgetting	RAVLT_perc_forgetting	DX
AGE	1	0.11	-0.051	-0.1	-0.12	-0.087	-0.2	-0.12	0.048	0.11	-0.011
PTGENDER	0.11	1	0.18	0.0043	0.025	-0.022	-0.19	-0.12	0.019	0.067	0.093
PTEDUCAT	-0.051	0.18	1	0.051	-0.064	0.15	0.2	0.12	-0.035	-0.11	-0.066
DX_bl	-0.1	0.0043	0.051	1	-0.011	0.12	0.032	0.033	0.048	-0.011	0.24
APOE4	-0.12	0.025	-0.064	-0.011	1	-0.24	-0.23	-0.21	0.076	0.19	0.12
MMSE	-0.087	-0.022	0.15	0.12	-0.24	1	0.61	0.46	-0.0092	-0.37	-0.13
RAVLT_immediate	-0.2	-0.19	0.2	0.032	-0.23	0.61	1	0.63	-0.15	-0.55	-0.31
RAVLT_learning	-0.12	-0.12	0.12	0.033	-0.21	0.46	0.63	1	0.02	-0.4	-0.23
RAVLT_forgetting	0.048	0.019	-0.035	0.048	0.076	-0.0092	-0.15	0.02	1	0.72	0.17
RAVLT_perc_forgetting	0.11	0.067	-0.11	-0.011	0.19	-0.37	-0.55	-0.4	0.72	1	0.24
DX	-0.011	0.093	-0.066	0.24	0.12	-0.13	-0.31	-0.23	0.17	0.24	1

Appendix 2: SHAREWave8 Input Feature Correlation Heatmap



	cf003_	cf004_	cf005_	cf006_	cf017_	cf103_	cf719_	cf821_	cf822_	cf823_	cf827_	cf828_	cf829_	cf830_	cf831_	cf837_	cf018_
cf003_	1	0.87	0.87	0.86	0.65	0.34	-0.52	0.67	0.66	0.67	0.7	0.55	0.71	0.45	0.46	0.053	-0.33
cf004_	0.87	1	0.95	0.95	0.71	0.36	-0.56	0.72	0.73	0.73	0.77	0.6	0.78	0.49	0.5	0.042	-0.36
cf005_	0.87	0.95	1	0.94	0.7	0.36	-0.55	0.71	0.71	0.72	0.76	0.59	0.77	0.49	0.5	0.051	-0.35
cf006_	0.86	0.95	0.94	1	0.7	0.37	-0.55	0.71	0.72	0.73	0.76	0.59	0.77	0.49	0.5	0.048	-0.35
cf017_	0.65	0.71	0.7	0.7	1	0.28	-0.47	0.67	0.67	0.67	0.68	0.52	0.68	0.48	0.49	0.049	-0.33
cf103_	0.34	0.36	0.36	0.37	0.28	1	-0.2	0.25	0.27	0.28	0.29	0.22	0.29	0.28	0.29	0.081	-0.074
cf719_	-0.52	-0.56	-0.55	-0.55	-0.47	-0.2	1	-0.51	-0.51	-0.51	-0.54	-0.41	-0.54	-0.33	-0.33	-0.032	0.32
cf821_	0.67	0.72	0.71	0.71	0.67	0.25	-0.51	1	0.89	0.89	0.82	0.63	0.81	0.51	0.52	0.037	-0.34
cf822_	0.66	0.73	0.71	0.72	0.67	0.27	-0.51	0.89	1	0.86	0.81	0.62	0.81	0.51	0.52	0.039	-0.33
cf823_	0.67	0.73	0.72	0.73	0.67	0.28	-0.51	0.89	0.86	1	0.83	0.65	0.83	0.54	0.55	0.053	-0.33
cf827_	0.7	0.77	0.76	0.76	0.68	0.29	-0.54	0.82	0.81	0.83	1	0.73	0.96	0.54	0.54	0.042	-0.34
cf828_	0.55	0.6	0.59	0.59	0.52	0.22	-0.41	0.63	0.62	0.65	0.73	1	0.74	0.42	0.42	0.038	-0.25
cf829_	0.71	0.78	0.77	0.77	0.68	0.29	-0.54	0.81	0.81	0.83	0.96	0.74	1	0.54	0.54	0.038	-0.34
cf830_	0.45	0.49	0.49	0.49	0.48	0.28	-0.33	0.51	0.51	0.54	0.54	0.42	0.54	1	0.76	0.17	-0.17
cf831_	0.46	0.5	0.5	0.5	0.49	0.29	-0.33	0.52	0.52	0.55	0.54	0.42	0.54	0.76	1	0.19	-0.18
cf837_	0.053	0.042	0.051	0.048	0.049	0.081	-0.032	0.037	0.039	0.053	0.042	0.038	0.038	0.17	0.19	1	-0.0085
cf018_	-0.33	-0.36	-0.35	-0.35	-0.33	-0.074	0.32	-0.34	-0.33	-0.33	-0.34	-0.25	-0.34	-0.17	-0.18	-0.0085	1

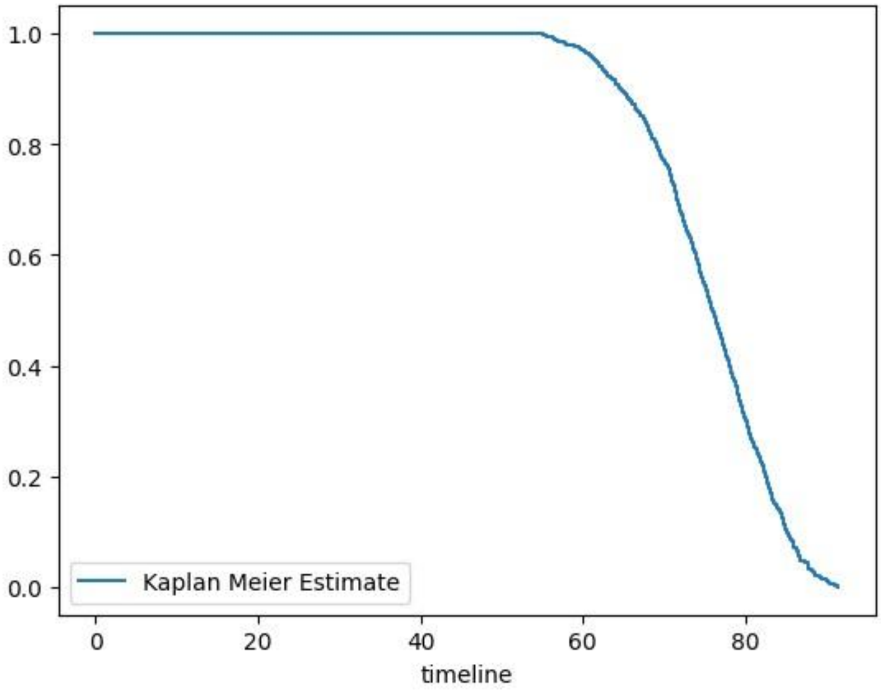
Appendix 3: Cox Proportional Hazards Model Results for Alzheimer's Disease Progression-ADNI Dataset

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	*-log 2(p)
PTGENDER	-0.15	0.86	0.02	-0.2	-0.11	0.82	0.9	0	-6.92	<0.005	37.66
PTEDUCAT	0	1	0	0	0.01	1	1.01	0	0.63	0.53	0.92
DX_bl	0.15	1.16	0.01	0.13	0.17	1.14	1.18	0	17.49	<0.005	225.02
APOE4	0.39	1.47	0.02	0.35	0.42	1.42	1.52	0	23.49	<0.005	402.94
MMSE	-0.03	0.97	0	-0.03	-0.02	0.97	0.98	0	-7.47	<0.005	43.52
RAVLT_immediate	0	1	0	0	0	1	1	0	-1.11	0.27	1.91
RAVLT_learning	-0.04	0.97	0.01	-0.05	-0.02	0.95	0.98	0	-5.93	<0.005	28.27
RAVLT_forgetting	0.04	1.04	0.01	0.02	0.06	1.02	1.06	0	5.04	<0.005	21.01
RAVLT_perc_forgetting	0	1	0	0	0	1	1	0	-0.82	0.41	1.27

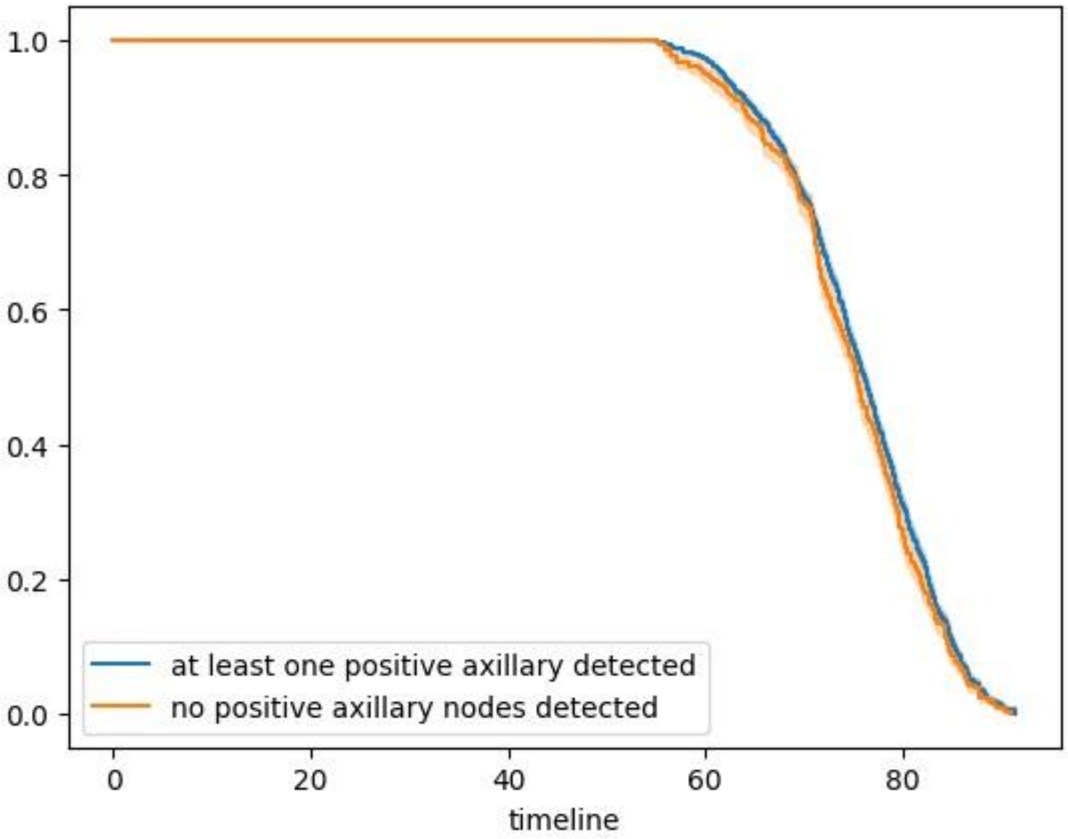
Appendix 4: Concordance Score-ADNI Dataset

Concordance	0.6
Partial AIC	155751.8
log-likelihood ratio test	1278.20 on 9 df
***-log2(p) of ll-ratio test**	892.94

Appendix 5: First Kaplan-Meier plot-ADNI Dataset



Appendix 6: Second Kaplan-Meier plot- ADNI Dataset



Appendix 7: Fairlearn accuracy, Precision, and Recall scores by the ADNI Baseline model

0.886380737
PTGENDER
0   0.889535
1   0.883769
Name: accuracy_score, dtype: float64

```
true=0  pred=0    0.901282
true=1  pred=0    0.870824
true=2  pred=0    0.883135
dtype: float64
PTGENDER
0    true=0  pred=0    0.919315
true=1  pred=0    0...
1    true=0  pred=0    0.881402
true=1  pred=0    0...
Name: precision_score_pd, dtype: object

true=0  pred=0    0.911803
```

```
true=1  pred=0    0.789899
true=2  pred=0    0.906609
dtype: float64
PTGENDER
0    true=0  pred=0    0.930693
true=1  pred=0    0...
1    true=0  pred=0    0.891008
true=1  pred=0    0...
Name: recall_score_pd, dtype: object
```

Appendix 8: Work plan

The project is expected to span 5 months, including 6 weeks for setup and approvals- including project proposal approval, 8 weeks for methodology, model development and evaluation, and 6 weeks for analysis, dissemination, presentation video and final report. A detailed work plan is included in Appendix 1.

Activities in the work plan shall include:

Phase 1: SHARE Dataset

- Preprocessing SHARE dataset for missing data, encodings, normalisation
- Visualising feature distributions and correlation analysis between features, to guide further investigation and modelling decisions
- Feature engineering, to optimise the data representation for model learning, and contribute to enhanced model performance and interpretability
- Implementing gradient boosting models, such as XGBoost

Phase 2: Second Dataset

- Preprocessing the second dataset for missing data, encodings, normalisation
- Visualising feature distributions and correlation analysis between features, to guide further investigation and modelling decisions
- Feature engineering, to optimise the data representation for model learning, and contributing to enhanced model performance and interpretability

Phase 3: Transfer Learning

- Applying transfer learning
- Tuning hyperparameters via cross-validation
- Evaluate final model on unseen data/ held-out test data

Months 2 and 3 will focus on methodology, model development and evaluation:

- Finalising model based on validation results
- Evaluating predictive performance on held-out test data
- Comparing to baseline models, such as logistic regression
- Generating SHAP explanations and analysing feature importance
- Evaluating explanation consistency and accuracy
- Analysing model biases and fairness
- Identifying areas for model improvement

Months 4 and 5 will involve final analysis, dissemination, and future directions:

- Compare models and explanations across both datasets
- Summarise key results and limitations
- Draft findings for publication in conferences and journals
- Suggest extensions like multi-modal frameworks
- Plan presentation video and final report