



Edge Hill
University

**Predictive Modelling for Dementia Diagnosis: An
Integrative Approach Using Machine Learning and
Multi-Dataset Analysis**

Student Name: **Chinonso Uche**

Student ID: **25830651**

Supervisor: **Prof. Ella Pereira**

27 September 2024

*'This Report is submitted in partial fulfilment of the
requirements for the MSc Data Science and Artificial
Intelligence Degree at Edge Hill University.'*

ABSTRACT

This project addresses the growing challenge of diagnosing Alzheimer's Disease (AD) and other forms of dementia, a public health concern due to the ageing global population. Current diagnostic methods are limited in their ability to detect early-stage dementia. This study leverages machine learning (ML) techniques to enhance early detection and diagnostic accuracy. The methodology involved integrating data from three major datasets—OASIS cross-sectional, OASIS longitudinal, and ADNIMERGE—covering a wide range of demographic, cognitive, and imaging features.

The project developed and tested several ML models, including Random Forest, Gradient Boosting, Support Vector Machines (SVM), Logistic Regression, and Multi-layer Perceptron (MLP), to predict dementia across two age groups (0-64 and 65+). Extensive preprocessing, feature selection, and missing data handling ensured high-quality inputs for model training. Results demonstrated that Random Forest and SVM yielded the best performance across both age groups, with accuracy scores exceeding 95%. For the 0-64 age group, Random Forest achieved a precision of 0.98 for the healthy class and 0.87 for the dementia class, with an overall accuracy of 96%. Similarly, SVM achieved a precision of 0.97 for the healthy class and 0.95 for the dementia class, with an accuracy of 97%. For the 65+ age group, SVM and Gradient Boosting achieved accuracies of 96%, with SVM achieving a precision of 0.97 and recall of 0.98 for the healthy class.

Naive Bayes, by contrast, struggled in both age groups due to its assumption of feature independence and was particularly impacted by the class imbalance, resulting in significantly lower recall and precision, with an accuracy of just 23% and 18% for the two groups, respectively.

The findings suggest that integrating multi-modal data significantly improves the accuracy and robustness of dementia prediction models. Although the study encountered limitations, such as the need for larger and more diverse datasets, it demonstrates the promise of ML in transforming dementia diagnosis and care. Future research could focus on increasing the generalisability of the models and enhancing their interpretability in clinical settings.

ACKNOWLEDGMENTS

I acknowledge and appreciate the Omni-Science God and my chi, who are always close to me.

To **Professor Ella Pereira**, words cannot express how grateful I am to have you as my supervisor. Thank you so much for your invaluable contributions; your patience, consistency, and timely availability throughout the process of this work were more than enough to see me through to successful completion. I deeply appreciate you.

I also acknowledge Ricardo Lopes, Prof. Amr Ahmed, and Prof. Yannis Korkontzelos, who listened to me and provided guidance when I needed it at the beginning of this work.

In a special way, I express my gratitude to **Chinaza Kate**, my lovely fiancée, who is always encouraging me. I extend my thanks to my family members, friends, classmates, and anyone else who aided and encouraged me at any time during this project.

I dedicate this work to my mother, the late **Mrs Ukwandu Ikwuejionuehi Eunice**, who passed away from dementia. Her love for good health and education, and her desire to see me succeed academically and professionally, continue to strengthen and motivate me to push beyond my limits. I hope that one day, this work, along with many other novel efforts on dementia, will contribute to saving the world from this disease.

KEYWORDS

Dementia Prediction, Alzheimer's Disease (AD), Cognitive Decline, Early Diagnosis, Machine Learning Models, OASIS Cross-Sectional Dataset, OASIS Longitudinal Dataset, ADNI Dataset, Data Integration, Exploratory Data Analysis (EDA), Feature Selection, Feature Engineering, Label Encoding, StandardScaler, Random Forest Classifier, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Classifier, Naive Bayes, Decision Tree Classifier, Multi-layer Perceptron (MLP), Hyperparameter Tuning, Age Group Analysis, Categorical Variables, Imputation of Missing Data, Accuracy Metrics, F1 Score, Precision, Confusion Matrix, Cross-Validation, Model Interpretability, Explainable AI (XAI), SHAP Values, Recursive Feature Elimination (RFE), Cognitive and Neuroimaging Profiles, Longitudinal Data Analysis, Class Imbalance, External Validation, Clinical Usability, Neuroimaging Data, Genetic Data, Patient Outcomes, Multimodal Data Integration, User-Centric Evaluation, Clinical Assessment Tools, MMSE Scores, APOE4 Status, Demographic Factors, Cross-Sectional Analysis, Feature Importance, Supervised Learning, Predictive Modelling, Ventricles and Hippocampus Analysis, AD Classification, Confusion Matrix Analysis, Class Weights and SMOTE, Overfitting Prevention, Diagnostic Accuracy, Evaluation Metrics.

ORIGINALITY DECLARATION

I, the undersigned, declare that the work presented in this research and development project is entirely my own, and that all sources of information have been acknowledged where appropriate. I understand that academic integrity is fundamental to maintaining the standards of this institution and that any form of academic dishonesty, including but not limited to plagiarism, collusion, or the use of unapproved external resources, is a serious violation of university policy.


I further confirm that:

- I have read and understood the institution's policies on plagiarism and academic misconduct.
- All direct quotations, ideas, data, and information derived from other works have been properly cited, and all paraphrased content has been appropriately referenced.
- This submission has not been previously submitted, in whole or in part, for any other course or degree.

I acknowledge that any breach of this originality statement may result in disciplinary actions in accordance with the university's regulations.

Student Name: **CHINONSO UCHE**

Student Number: **25830651**

Signature: 

Date: **25 September 2024**

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION	11
1.1 Background	11
1.2 Problem Domain	12
1.3 Aim and Objectives	14
1.4 Proposed Methodologies.....	15
1.5 Achieved Outcome.....	16
1.6 Structure of The Report	16
CHAPTER 2: LITERATURE REVIEW	18
2.1 Background and Significance of Dementia	18
2.2 Existing Diagnostic Methods.....	19
2.3 Current State of Machine Learning in Dementia Prediction	19
2.4 Gaps in Current Research	21
2.5 Related Work in Machine Learning Application	21
2.6 Applying Machine Learning to Dementia Prediction.....	22
2.7 Challenges and Limitations	22
CHAPTER 3: METHODOLOGY	23
3.1 Research Design	24
3.2 Data Collection and Integration.....	25
3.2.1 Dataset Editing.....	27
3.3 Exploratory Data Analysis (EDA) Before Preprocessing.....	29
3.4 Data Preprocessing	29
3.5 Feature Engineering and Selection	30

3.6 Age Group Analysis	30
3.7 Model Selection and Training	30
3.8 Model Evaluation	31
3.9 Ethical Implications	31
3.9.1 Individuals Affected	31
3.9.2 Ethical Considerations.....	31
3.10 Summary	32
CHAPTER 4: IMPLEMENTATION AND TESTING	33
4.1 Implementation Processes	33
4.1.1 Data Integration	33
4.1.2 Data Preprocessing	35
4.1.3 Exploratory Data Analysis (EDA) Implementation	36
4.1.4 Descriptive Statistics	37
4.1.5 Visualisation	38
4.1.6 Correlation Analysis	44
4.2 Model Selection and Development	45
4.2.1 Models Considered	46
4.2.2 Selection Criteria.....	46
4.2.3 Models Chosen.....	46
4.2.4 Trade-offs and Justifications	49
4.3 Model Training and Evaluation	49
4.3.1 Key Metrics	52
4.3.2 Conclusion of Models Performance on Merged Dataset	52

4.3.3	Age Group Analysis Implementation	52
4.3.4	Conclusion of Models Performance on Age Group Datasets	54
4.4	Testing Methodology	56
4.4.1	Cross-Validation	57
4.4.2	Confusion Matrix and Metrics Evaluation	57
4.4.3	Performance Comparison Between Age Groups	58
4.5	Functional Testing	58
4.6	Summary	58
CHAPTER 5:	EVALUATION	59
5.1	Fulfilment of Original Objectives	59
5.2	Scope and Comparison with Related Work	60
5.3	Advantages of the Design Approach	62
5.4	Disadvantages of the Design Approach	62
5.5	Rational for Methodological Choices	63
5.6	Comparison with Original Objectives	64
5.7	Summary	64
CHAPTER 6:	DISCUSSION AND CONCLUSION	66
6.1	Key Findings and Analysis	66
6.1.1	Experimental Findings	66
6.1.2	Off-Topic and Dataset Integration Insights	67
6.2	Accomplishments and Limitations	67
6.2.1	Goals Achieved	67
6.2.2	Partial Achievements and Challenges	68

6.3 Future Directions	68
6.4 Self-Reflection and Learning Outcomes	69
6.5 Conclusion	70
REFERENCES	71
APPENDIX A Access Granted Email for OASIS Cross-sectional Dataset.....	76
APPENDIX B Access Granted Email for OASIS Longitudinal Dataset	78
APPENDIX C Access Granted Email for ADNI Dataset	80
APPENDIX D Descriptive Statistics of The Age Group 1 Dataset	81
APPENDIX E Visualisation of The Age Group 1 Dataset	83
APPENDIX F Correlation Analysis of The Age Group 1 Dataset	89
APPENDIX G Descriptive Statistics of The Age Group 2 Dataset	91
APPENDIX H Visualisation of The Age Group 2 Dataset	93
APPENDIX I Correlation Analysis of The Age Group 2 Dataset	99
APPENDIX J Confusion Matrix of Each Model on Merged Dataset	101
APPENDIX K Confusion Matrix of Each Model on Group 1 Dataset.....	105
APPENDIX L Confusion Matrix of Each Model on Group 1 Dataset.....	109
APPENDIX M EDA Code Snippet (Reusable Function)	113
APPENDIX N Training & Evaluation Code (Merged Dataset)	116
APPENDIX O Training & Evaluation Code (Age Groups Dataset).....	117
APPENDIX P The Project Management Gantt Chart	118
APPENDIX Q Project Poster.....	119
APPENDIX R The Signed Ethical Checklist.....	123

LIST OF TABLES

Table 2.1. Summary of Machine Learning Studies for Dementia and Alzheimer's Disease Prediction.....	20
Table 3.1. OASIS Cross-Sectional Dataset Summary.....	26
Table 3.2. OASIS Longitudinal Dataset Summary.....	26
Table 3.3. ADNIMERGE Dataset Summary.....	27
Table 3.4. OASIS longitudinal Dataset Summary After Editing.....	28
Table 3.5. ADNIMERGE Dataset Summary After Editing.....	28
Table 4.1. Metadata of Common Features of The Datasets.....	33
Table 4.2. Summary of The Merged Dataset.....	36
Table 4.3. Statistical Summary of The Merged Dataset.....	37
Table 4.4. Summary of Statistical Test Result on Merged Data.....	37
Table 4.5. Summary of Selected Models.....	47
Table 4.6. Classification Report Summary of Different Models on The Merged Dataset.....	50
Table 4.7. Data Size of Each Age Group.....	52
Table 4.8. Classification Report Summary of Different Models on Age Group 1 (Age < 65).....	52
Table 4.9. Classification Report Summary of Different Models on Age Group 2 (Age ≥ 65).....	53
Table 4.10. Summary of Best and Worst Performing Models on Different Datasets.....	54

LIST OF FIGURES

Figure 3.1. Flowchart of Machine Learning workflow Applied in This Study 23

Figure 4.1. Histogram Distribution of Various Features of The Merged Dataset 38

Figure 4.2. Box plot of Various Features of The Merged Dataset 38

Figure 4.3. Age Distribution of The Merged Dataset 39

Figure 4.4. Gender Distribution of The Merged Dataset 39

Figure 4.5. Diagnosis Distribution of The Merged Dataset 40

Figure 4.6. Scatter Plot of Age vs MMSE Score of The Merged Dataset 40

Figure 4.7. Scatter Plot of Education vs MMSE Score of The Merged Dataset 41

Figure 4.8. Box plot of MMSE Score Distribution by Diagnosis of The Merged Dataset.41

Figure 4.9. Box plot of Age Distribution by Diagnosis of The Merged Dataset 42

Figure 4.10. Pair plot of Key Features of The Merged Dataset 43

Figure 4.11. Correlation Heatmap of the Merged Dataset 44

CHAPTER 1: INTRODUCTION

1.1 Background

Dementia is a broad term that describes a group of symptoms caused by changes in brain function that affect memory, thinking, and behaviour (Arvanitakis & Bennett, 2019). It is not a normal part of aging, although it is more common among older adults. Various underlying diseases or conditions can cause dementia, such as Alzheimer's Disease (AD), vascular dementia, Lewy body dementia, and frontotemporal dementia (Quinn et al., 2021). AD, the most common type of dementia, accounts for 60-80% of cases (World Health Organization, 2023). It is a progressive neurological disorder characterized by the deposition of amyloid plaques, neurofibrillary tangles, and synaptic loss, leading to cognitive decline and memory impairment (Sheppard & Coleman, 2020). Given the increasing aging population, AD and other forms of dementia have become a significant public health concern globally.

The complex nature of AD and other forms of dementia presents significant challenges in early detection and accurate diagnosis. This difficulty arises from the heterogeneity of symptoms and the overlapping clinical features with other neurological conditions. Despite advancements in neuroimaging, biomarkers, and cognitive assessments, integrating these diverse data sources into a unified diagnostic framework remains a challenge (Li et al., 2021). Machine learning (ML) and artificial intelligence (AI) offer promising avenues to address this issue by developing robust, data-driven models that can assist clinicians in diagnosing dementia more accurately and at earlier stages. ML techniques can analyse vast amounts of diverse data types, including neuroimaging, genetic information, cognitive test scores, and clinical records, to identify subtle patterns and biomarkers associated with dementia (Zhu et al., 2020). These sophisticated approaches enable the development of models for early detection, diagnosis, and prognosis of dementia, particularly AD. Such models have shown promise in identifying individuals at risk of developing AD before symptoms become apparent, potentially improving early detection and enhancing diagnostic accuracy. Furthermore, these techniques can predict disease progression, offering the potential

for more timely interventions, better management of the condition, and the development of personalized treatment strategies. By leveraging machine learning in this way, researchers and clinicians aim to significantly improve patient outcomes through earlier and more targeted interventions in the field of dementia care.

An example of how advanced machine learning techniques have improved AD prediction can be seen in the work of Park et al. (2020). They developed a machine learning model using large-scale administrative health data to predict the incidence of AD. Their model, which incorporated various health-related features and utilised advanced algorithms, demonstrated high accuracy in predicting AD up to five years before clinical diagnosis. This approach showcases how machine learning can leverage complex, multi-dimensional data to provide early warnings of AD, potentially allowing for earlier intervention and better patient outcomes.

Given the promising applications of machine learning in dementia research, it is important to understand what machine learning is and how it is being specifically applied to dementia prediction. Machine learning is a subfield of artificial intelligence that involves training algorithms to learn from data and make predictions or decisions without being explicitly programmed (Mitchell, 1997). A machine learning model is a mathematical representation of a system or process that is trained on data to make predictions or classifications.

1.2 Problem Domain

The complexity and variability of dementia symptoms necessitate a multifaceted diagnostic approach. Traditional methods, such as clinical interviews and neuropsychological tests, although valuable, are often limited by subjectivity and inter-rater variability. Neuroimaging techniques like MRI and PET scans provide detailed information about brain structure and function but require specialised equipment and expertise. Biomarkers, including cerebrospinal fluid (CSF) analysis and blood tests, have shown promise in identifying early pathological changes associated with Alzheimer's Disease, yet they are invasive and not routinely available.

The primary challenge lies in integrating these diverse data modalities—clinical, neuropsychological, neuroimaging, and biological markers—into a comprehensive

and accurate diagnostic model. Moreover, understanding how demographic factors such as age, gender, education, and ethnicity influence the risk and presentation of dementia is crucial for developing tailored diagnostic and treatment strategies. Given these complexities, there is a need for advanced analytical methods that can handle high-dimensional data, account for missing values, and provide interpretable results.

A systematic review conducted by (Javeed et al. 2023) highlights the potential of machine learning for dementia prediction, but also identified several limitations and future research directions. The review examined various studies that utilised different types of data for dementia prediction, including:

- i. Neuroimaging data, (such as MRI and PET scans).
- ii. Clinical-Variable Data which consists of medical tests and patient information such as age, sex, and cognitive assessments.
- iii. Voice Data which involves speech analysis to detect neurodegenerative disorders affecting language processing.

While these diverse data types have shown promise in machine learning models for dementia prediction, the review emphasised several limitations:

1. **Single Data Modality Focus:** Previous systematic literature reviews (SLRs) focused on a single type of data modality for dementia detection, limiting the comprehensiveness of the evaluation.
2. **Data Quality Issues:** Poor quality of data and imbalance in dataset classes can lead to biased results from machine learning (ML) models.
3. **Model Selection:** Inappropriate selection of ML models and the complexity of training models can affect the performance of automated diagnostic systems.
4. **Supervised Learning Constraints:** Supervised ML techniques have inherent limitations, which can impact the effectiveness of automated diagnostic methods for dementia prediction.

1.3 Aim and Objectives

The main aim of this project is to develop and evaluate machine learning models for the early diagnosis of dementia, particularly Alzheimer's Disease, using a comprehensive dataset that includes clinical, cognitive, neuroimaging, and demographic information. By leveraging these models, this study aims to identify key features and patterns that differentiate between healthy aging and various stages of cognitive impairment, ultimately contributing to improved diagnostic accuracy and patient outcomes.

OBJECTIVES

The objectives therefore are:

- **Review of Literature:** To review existing literature on dementia, AD, and machine learning models for dementia prediction.
- **Data Integration and Preprocessing:** To merge and preprocess datasets from multiple sources, including the OASIS cross-sectional, OASIS longitudinal, and ADNI datasets, ensuring consistency and quality in the data used for model training.
- **Exploratory Data Analysis (EDA):** To perform EDA to understand the distribution, correlation, and statistical significance of features within the dataset, providing insights into the underlying patterns and potential predictors of dementia.
- **Feature Selection and Engineering:** To identify and select the most relevant features for dementia diagnosis using techniques such as `StandardScaler` and to preprocess these features using methods like Label Encoding and Standard Scaling.
- **Model Development:** To develop and compare various machine learning models, including Random Forest, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, Naive Bayes, Decision Tree, and Multi-Layer Perceptron (MLP), for classifying individuals based on their cognitive and neuroimaging profiles.

- **Evaluation and Validation:** To evaluate the performance of these models using metrics such as accuracy, F1 score, precision, and confusion matrix analysis, and to validate their generalisability across different age groups.
- **Interpretation and Insights:** To interpret the findings and identify key factors influencing dementia diagnosis, providing insights that could guide future research and clinical practice.

1.4 Proposed Methodologies

To achieve the objectives, a systematic and structured methodology have been employed. The project begins with data acquisition and integration, combining datasets from the OASIS cross-sectional, OASIS longitudinal, and ADNI studies. These datasets provide a rich source of information, including demographic details, cognitive assessments (e.g., MMSE, CDR), and neuroimaging measurements (e.g., eTIV, nWBV). The next step involves extensive data preprocessing, which includes handling missing values using strategies such as mean imputation for numerical features and most frequent imputation for categorical features. Categorical data, such as gender, will be encoded using Label Encoding to convert them into numerical formats suitable for model input. Feature scaling, using methods like Standard Scaling, will be applied to ensure that features with different units and magnitudes do not disproportionately influence the model.

Exploratory Data Analysis (EDA) will be conducted to uncover underlying patterns, distributions, and correlations within the dataset. Visualization techniques such as histograms, box plots, scatter plots, and heatmaps will be utilised to provide a comprehensive understanding of the data and identify potential predictors of dementia.

The core of the methodology involves building and evaluating multiple machine learning models. A diverse set of classifiers will be employed, including Random Forest, SVM, Logistic Regression, Gradient Boosting, Naive Bayes, Decision Tree, and MLP. These models will be trained on the preprocessed data, and their performance will be evaluated using appropriate metrics to identify the most effective model for dementia diagnosis.

1.5 Achieved Outcome

By the end of this project, the following outcomes are expected:

1. A unified and well-preprocessed dataset combining clinical, cognitive, neuroimaging, and demographic information from multiple sources.
2. Comprehensive EDA results highlighting key features and relationships within the data, providing a foundation for model development.
3. A set of machine learning models capable of accurately classifying individuals based on their risk of dementia, with detailed performance metrics for each model.
4. Insights into the most significant predictors of dementia, contributing to a better understanding of the disease and informing future research and clinical strategies.
5. A framework for implementing machine learning techniques in the diagnosis of dementia, potentially aiding clinicians in making more informed and timely decisions.

Commented [EP1]: achieved not expected

Commented [CU2R1]: Done

1.6 Structure of The Report

This report is structured into six chapters (including chapter one above and excluding the References and Appendices), each contributing to a comprehensive understanding of the research on the application of machine learning in dementia prediction, particularly Alzheimer's Disease (AD).

Chapter 2: Literature Review - presents a detailed background and significance of dementia, examining the current state of machine learning in dementia prediction. It critically evaluates existing diagnostic methods, the application of machine learning models, and highlights gaps in current research and future directions.

Chapter 3: Methodology - outlines the research design and methodology used in the study. It discusses the data sources, preprocessing steps, feature selection methods, and the various machine learning algorithms employed for model building and evaluation. The rationale behind the choice of methods and the overall approach is also discussed.

Commented [EP3]: at this point you have already read chapter 1 so no need to describe it.

Commented [CU4R3]: Done

Chapter 4: Implementation and Testing - details the implementation of the selected machine learning models and the testing procedures used to evaluate their performance. It includes information on the development environment, coding processes, and the challenges encountered during implementation.

Chapter 5: Evaluation - focuses on the evaluation metrics and results of the implemented models. It provides a comparative analysis of the models' performance, including accuracy, precision, recall, F1 score, and other relevant metrics. The effectiveness of the models in predicting dementia is thoroughly assessed.

Chapter 6: Discussion and Conclusion - interprets the results and summarises the main findings of the study, relating them to the research questions and objectives set out in the introduction and draw conclusions based on the research conducted. It considers the implications of the findings, discusses the limitations of the study, and suggests potential improvements or future research directions.

It includes a self-reflection section on the research process and offers recommendations for further work in the field of dementia prediction using machine learning before rendering a conclusion.

CHAPTER 2: LITERATURE REVIEW

2.1 Background and Significance of Dementia

Dementia is a complex and multifactorial condition affecting millions worldwide, leading to cognitive decline, memory loss, and changes in behaviour and mood (Dr Raina Loh, 2023). Alzheimer's Disease (AD) is the most common form of dementia, constituting about 60-80% of cases. It is characterized by progressive neurodegeneration, including amyloid plaques, neurofibrillary tangles, and synaptic loss. AD presents a significant public health challenge, with an estimated 55 million people currently living with dementia globally, a figure expected to rise to 78 million by 2030 and 139 million by 2050. The economic impact is equally severe, with annual costs surpassing \$1.3 trillion in 2019 and potentially escalating to \$2.8 trillion by 2030 (Shin, 2022).

Early diagnosis or prediction of dementia is crucial for timely intervention, allowing for more effective management and potentially slowing disease progression (Brookmeyer et al., 2017). Despite the importance of early detection, it remains problematic due to the insidious onset of AD and the overlap of symptoms with other forms of cognitive impairment. Current assessment tools are often subjective and may lead to delayed or inaccurate diagnosis, ultimately resulting in inadequate care for patients and caregivers (Pais et al., 2020).

Machine learning models have been widely used in various applications, including image and speech recognition, natural language processing, and predictive analytics. In the context of dementia prediction, machine learning models can be trained on large datasets of clinical, imaging, and biomarker features to identify patterns and predict the likelihood of developing dementia (Li et al., 2021). The goal of dementia prediction is to identify individuals at high risk of developing dementia, allowing for early intervention, and potentially delaying or preventing disease progression.

Commented [CU5]: Move this part including the diagram to the literature review section

Commented [CU6R5]: Done

2.2 Existing Diagnostic Methods

Diagnosis of Alzheimer's typically involves a combination of clinical evaluation, neuropsychological testing, and neuroimaging techniques. Neuropsychological assessments, such as the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR) scale, are commonly used for assessing cognitive decline. However, these methods can be subjective and are influenced by factors such as a patient's education level and cultural background.

Neuroimaging techniques like MRI and PET scans provide more objective data by revealing structural and functional brain changes. Although these methods offer more accuracy, they are expensive and not always accessible in routine clinical practice. Moreover, they may still miss early signs of AD, making early and accurate diagnosis challenging.

2.3 Current State of Machine Learning in Dementia Prediction

Machine learning (ML) has emerged as a promising tool in the field of dementia research, with the potential to enhance diagnostic accuracy and enable early detection. ML algorithms can analyse large datasets, identify complex patterns, and predict outcomes more effectively than traditional methods. Recent studies have demonstrated the utility of ML models in predicting dementia, leveraging various features, and achieving high performance (Javeed et al., 2023).

Researchers are employing different data modes and mechanisms to improve diagnostic accuracy. For example, Li et al. (2021) utilized logistic regression, decision trees, and random forests to predict dementia using cognitive tests and neuroimaging features, achieving an accuracy of 85.7%. Similarly, Park et al. (2020) combined cognitive tests, neuroimaging, and biomarkers using different ML algorithms to predict Alzheimer's disease with a high accuracy of 92.5%.

Studies like (Moradi et al. 2015) and (Bari Antor et al., 2021) have applied ML techniques such as support vector machines (SVM) and random forests to classify AD stages with substantial accuracy. However, there is still no consensus on the most effective features and algorithms for AD prediction, indicating a need for further

research and model optimization. Table 2.1 below summarises the Machine Learning studies for Dementia and Alzheimer's Disease prediction.

Table 2.1.

Summary of Machine Learning Studies for Dementia and Alzheimer's Disease Prediction.

Study	ML Approaches Used	Types of Data Used	Specific Task	Accuracy/ Performance
Li et al. (2021)	Logistic regression, Decision trees, Random forests	Cognitive tests, Neuroimaging	Dementia prediction	85.7% accuracy
Park et al. (2020)	Not specified (general "machine learning algorithms")	Cognitive tests, Neuroimaging, Biomarkers	Alzheimer's disease prediction	92.5% accuracy
Bari Antor et al. (2021)	Systematic review of various ML models	Not specified	Alzheimer's disease prediction	Median accuracy of 85% (range 70-100%)
Grueso and Viejo-Sobera (2021)	Systematic review of various ML models	Not specified	Predicting progression from mild cognitive impairment to Alzheimer's disease	Median AUC-ROC of 0.85 (range 0.70-0.95)
Musto et al. (2021)	Not specified	Not specified	Dementia prediction	83.6% accuracy
Battineni et al. (2020)	Not specified (discussed challenges in ML models)	Not specified	Not specified	Not reported

2.4 Gaps in Current Research

While significant advancements have been made in using ML for AD diagnosis, several gaps remain. Many models focus on single-source data, such as MRI scans or genetic markers, potentially missing the complexity of the disease, which spans genetic, biochemical, and cognitive domains. Most studies emphasize cross-sectional data, neglecting the longitudinal aspect of AD progression. This indicates a need for integrative approaches that can utilise multi-modal data and track changes over time to enhance early detection and improve prediction accuracy.

For instance, (Li et al. 2021) relied on a relatively small sample size ($n=150$) and did not consider important biomarkers like APOE genotyping. On the other hand, (Park et al. 2020) used a larger dataset ($n=1200$) but did not report feature importance or model interpretability. The field also faces challenges in terms of data quality, feature selection, and model generalisability (Battineni et al., 2020). Moreover, many ML models for dementia prediction suffer from overfitting and lack robustness due to small sample sizes and noisy data.

2.5 Related Work in Machine Learning Application

A plethora of machine learning models have been proposed for AD classification, each with varying degrees of success. For instance, (Moradi et al., 2015) used a combination of hippocampal shape analysis and support vector machines to differentiate between AD and MCI, achieving an accuracy of over 80%. Similarly, (Payan and Montana, 2015) demonstrated the use of convolutional neural networks (CNNs) on 3D brain MRI scans, showing improved classification accuracy. These studies highlight the potential of ML in handling high-dimensional neuroimaging data, but they often require large datasets and significant computational resources.

Other works have explored ensemble methods, such as random forests and gradient boosting, to capture complex patterns in multi-modal data. For instance, (Liu et al., 2021) employed random forests to integrate genetic, imaging, and clinical data for AD classification, demonstrating that combining multiple data sources can improve model performance. However, these models often face challenges related to overfitting and interpretability, particularly when dealing with heterogeneous datasets.

2.6 Applying Machine Learning to Dementia Prediction

Machine learning models have been successfully applied to various datasets to predict the likelihood of developing dementia. For example, (Park et al., 2020) used large-scale administrative health data to predict the incidence of Alzheimer's disease with high accuracy. Similarly, (Musto et al., 2021) developed a machine learning approach to predict deterioration in Alzheimer's disease.

These models are often trained on a range of features, including clinical, imaging, and biomarker data. (Li et al., 2021) applied ML to omics, imaging, and clinical data to identify patterns and predict Alzheimer's disease. Another study by (Battineni et al., 2020) used ML predictive models for chronic disease diagnosis, including dementia. Advanced ML methods, such as deep neural networks, have also been employed. For instance, (Kim and Lim, 2021) developed a deep neural network-based method for predicting dementia using big data, achieving high accuracy and F1-scores. (Bucholc et al., 2023) demonstrated a hybrid ML approach to predict conversion from mild cognitive impairment to dementia, achieving high accuracy and area under the receiver operating characteristic curve (AUC-ROC).

2.7 Challenges and Limitations

Despite the advancements in using ML for dementia prediction, inconsistencies in results highlight the need for further refinement (Li et al., 2021). For example, while some models like (Park et al., 2020) achieved accuracy levels as high as 85.7%, others like (Musto et al., 2021) achieved 83.6%. These inconsistencies emphasize the need for standardization in the development and evaluation of these models.

Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown superior performance in handling image and sequence data but require large amounts of labelled data and computational power. This requirement limits their practicality in resource-constrained settings. Moreover, these models are often seen as "black boxes," making it challenging to interpret their predictions in a clinically meaningful way.

CHAPTER 3: METHODOLOGY

This chapter outlines the systematic methodology employed to address the research problem of predicting dementia using machine learning. The approach integrates multiple datasets, applies various machine learning models, and evaluates their performance. Key stages of the methodology include data collection, preprocessing, feature engineering, model training, and evaluation. Each technique was carefully selected to ensure robust model development and effective predictions. The rationale for selecting each technique is provided, including trade-offs between different approaches. Figure 3.1 below is a flowchart describing the process of Machine Learning Operations applied in this study. See also, Appendix P – Project Management Gantt Chart.

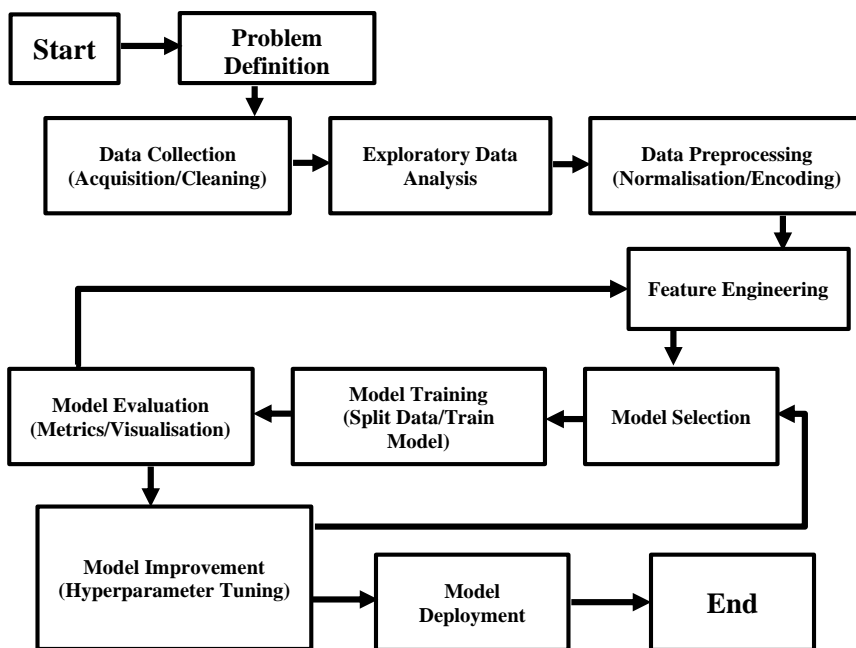


Figure 3.1. Flowchart of Machine Learning workflow Applied in This Study.

Commented [EP7]: if this is your own version of the diagram that is applied in your study, this should be in the methodology chapter not in the background. if it is a generic diagram used in other studies too then leave it here and remove 'Applied in this study'

3.1 Research Design

This study follows a systematic data-driven approach, utilising supervised machine learning techniques to classify individuals as having dementia or not. The design comprises the following stages:

1. **Data Collection and Integration:** Three datasets were obtained from reputable sources: OASIS cross-sectional, OASIS longitudinal, and ADNIMERGE from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The datasets were merged based on common variables such as demographics (age, gender), clinical measures, and cognitive scores, ensuring data consistency and quality.
2. **Exploratory Data Analysis (EDA):** An in-depth analysis was conducted to understand the distribution of the data and relationships between features. EDA was conducted both before and after preprocessing, revealing key patterns and correlations crucial to model development.
3. **Data Preprocessing:** Missing values were handled, categorical variables were encoded, and numerical features were scaled. These steps ensured that the dataset was clean and ready for model training.
4. **Feature Engineering and Selection:** Key features were selected based on statistical importance, with techniques such as StandardScaler employed. Feature engineering included creating additional variables to improve model accuracy.
5. **Model Selection:** A variety of machine learning algorithms were considered, and seven models were selected for in-depth comparison. These included Random Forest, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, Naive Bayes, Decision Tree, and Multi-Layer Perceptron (MLP).
6. **Age Group Analysis:** To explore demographic-specific patterns, the data was segmented into two age groups: under 65 and 65 and over. Models were trained and evaluated separately for each group.
7. **Model Training and Evaluation:** Each model was trained on 80% of the dataset and evaluated on the remaining 20% using various performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

This design ensures that all aspects of the research problem, from data collection to model evaluation, are systematically addressed.

3.2 Data Collection and Integration

Data for this project were sourced from two publicly available datasets: Open Access Series of Imaging Studies (OASIS) and Alzheimer's Disease Neuroimaging Initiative (ADNI).

OASIS Dataset Acknowledgments: "Data were provided in part by OASIS.

OASIS-1: Cross-Sectional and OASIS-2: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382". Permission to access the OASIS datasets (cross-sectional and longitudinal) was granted on 21 June 2024, and permission for the ADNIMERGE dataset from <https://ida.loni.usc.edu> was granted on 27 June 2024. See Appendix A, B and C for the email content providing permission to access each dataset. Table 3.1 - 3.3 below presents the original aspect of OASIS cross-sectional, OASIS longitudinal and ADNIMERGE dataset respectively.

The datasets were integrated by matching common variables such as demographic data (e.g., age, gender, education) and clinical scores (e.g., MMSE, CDR). This integration allowed the creation of a comprehensive dataset with clinical, cognitive, neuroimaging, and demographic information, necessary for robust machine learning model development.

Commented [EP8]: expand this section by adding more on data acquisition by obtaining relevant permissions. also description of what is in each dataset, features, etc. you can also add description of methods used for data integration. some sort of visuals, e.g. table will help with data description

Commented [CU9R8]: Done

Commented [EP10]: I would call this data description. integration should not be described here. you can say what methods were used to integrate data but not actual implementation of the integration.

Commented [CU11R10]: Done

Table 3.1.

OASIS Cross-Sectional Dataset Summary.

Aspect	Details
Number of Columns	12
Column Names	ID, M/F, Hand, Age, Educ, SES, MMSE, CDR, eTIV, nWBV, ASF, Delay
Number of Unique Subjects	436
Total Number of Entries	436
Dataset Type	Cross-sectional (one entry per subject)
Key Measures	Demographic (Age, Sex, Education, SES), Cognitive (MMSE, CDR), Brain Volume (eTIV, nWBV, ASF)

Table 3.2.

OASIS Longitudinal Dataset Summary.

Aspect	Details
Number of Columns	15
Column Names	Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF
Number of Unique Subjects	150
Total Number of Entries	373
Dataset Type	Longitudinal (multiple entries per subject)
Key Measures	Demographic (Age, Sex, Education, SES), Cognitive (MMSE, CDR), Brain Volume (eTIV, nWBV, ASF)

Table 3.3.

ADNIMERGE Dataset Summary.

Aspect	Details
Number of Columns	116
Column Categories	Subject IDs, Demographics, Cognitive Assessments, Biomarkers, Imaging Metrics, Longitudinal Data
Key Column Examples	RID, PTID, AGE, PTGENDER, MMSE, CDRSB, ADAS13, FDG, ABETA, TAU, Ventricles, Hippocampus
Number of Unique Subjects	2,430
Total Number of Entries	16,421
Dataset Type	Longitudinal (multiple entries per subject)
Unique Features	Includes baseline (_bl) and follow-up measurements, various cognitive tests, biomarkers, and imaging data
Time-related Variables	VISCODE, EXAMDATE, Years_bl, Month_bl, Month
Diagnostic Information	DX_bl, DX (likely indicating baseline and follow-up diagnoses)
Genetic Information	APOE4 (Apolipoprotein E ε4 allele status)
Imaging Modalities	FDG, PIB, AV45, FBB (various PET imaging tracers)

3.2.1 Dataset Editing

The ADNIMERGE dataset contains 116 features, many of them are unimportant to the task of this project. Thus, those features were removed. Some features were renamed to correspond to the names of OASIS dataset features. The OASIS longitudinal dataset had minor editing. For instance, the original dataset contains 'MRI ID' and 'Visit', these

features were removed since they are not required. Table 3.4 and 3.5 below shows the aspect of the edited datasets.

Table 3.4.

OASIS longitudinal Dataset Summary After Editing.

Aspect	Details
Number of Columns	13
Column Names	Subject ID, Group, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF
Number of Unique Subjects	150
Total Number of Entries	373
Dataset Type	Longitudinal (multiple entries per subject)

Table 3.5.

ADNIMERGE Dataset Summary After Editing.

Aspect	Details
Number of Columns	11
Column Names	ID, Age, Gender, EDUC, PTMARRY, APOE4, TAU, CDR, MMSE, MOCA, DX
Number of Unique Subjects	2,430
Total Number of Entries	16,421
Dataset Type	Longitudinal (multiple entries per subject)
Unique Features	Includes cognitive assessments, genetic information, and diagnostic variables

3.3 Exploratory Data Analysis (EDA) Before Preprocessing

EDA was conducted to uncover hidden patterns, detect anomalies, and identify relationships between features. Key analyses included:

- **Descriptive Statistics:** Summarised the distribution of variables (mean, median, standard deviation).
- **Correlation Analysis:** Examined the relationship between clinical scores, neuroimaging measures, and demographic variables to detect potential predictors of dementia.
- **Visualisation:** Heatmaps, boxplots, and histograms were used to visually interpret feature distributions and correlations.

EDA was performed on the merged dataset before preprocessing. Tables and figures from Chapter 4 (Table 4.7 - 4.9, Figures 4.1 - 4.10) provide a detailed overview of the merged dataset's distribution and patterns. After preprocessing EDA was performed on the age group datasets. (See Appendices D, E and F for Age Group 1 EDA and Appendices G, H and I for Age Group 2 EDA).

3.4 Data Preprocessing

Preprocessing was critical in ensuring data quality for accurate model training. The following steps were undertaken:

- **Handling Missing Values:** Numerical columns with missing data were imputed using the mean, while categorical variables were filled with the most frequent category. This method helped maintain the integrity of the data distribution without introducing bias.
- **Encoding Categorical Variables:** Label encoding was applied to categorical variables like gender, converting them into numerical format for model compatibility.
- **Feature Scaling:** Numerical variables such as MMSE, CDR, and neuroimaging measures were standardised using the `StandardScaler`. Standardisation was necessary to ensure that features with different units or scales did not disproportionately influence the model.

- **Ensuring Data Consistency:** Age variables were checked for uniformity and converted to ensure that all records were in numerical format.

3.5 Feature Engineering and Selection

Feature engineering and selection are vital to enhance model performance. `StandardScaler` was applied to identify the most predictive features. Other transformations included creating interaction terms between cognitive scores and neuroimaging features, which helped improve model interpretability and performance.

3.6 Age Group Analysis

The dataset was divided into two age groups to account for the variation in dementia onset patterns across different demographics:

1. **Age Group 1 (Under 65):** Focused on early-onset dementia patterns.
2. **Age Group 2 (65 and over):** Analysed the traditional age group at a higher risk for dementia.

Models were trained and evaluated separately for each group to identify age-specific factors affecting dementia diagnosis.

3.7 Model Selection and Training

A wide array of machine learning models were explored for their ability to predict dementia. The models considered were:

- Random Forest
- Support Vector Machine (SVM)
- Logistic Regression
- Gradient Boosting
- Naive Bayes
- Decision Tree
- Multi-Layer Perceptron (MLP)

Each model was trained using 80% of the data, with 20% reserved for testing.

3.8 Model Evaluation

The performance of the machine learning models was evaluated using a range of metrics to ensure a comprehensive assessment of their strengths and limitations:

- **Accuracy:** The proportion of correctly classified instances in the test data.
- **Precision and Recall:** Evaluated how well the models identified dementia cases while minimising false positives.
- **F1-Score:** Combined precision and recall into a single metric to assess the balance between sensitivity and specificity.
- **Confusion Matrix:** Provided insights into model errors by classifying instances as true positives, false negatives, false positives, and true negatives.

3.9 Ethical Implications

This project employs anonymised datasets from the OASIS and ADNI repositories. Both are reputable open-access sources that maintain rigorous standards for data privacy and ethical use. By adhering to their guidelines, the project ensures ethical research practices and the protection of sensitive information.

3.9.1 Individuals Affected

The project benefits from the participation of individuals who have shared their anonymised data through these open repositories, fostering advancements in dementia research while safeguarding personal privacy.

3.9.2 Ethical Considerations

- **Acknowledgement:** The project duly credits the original creators and institutions behind the datasets, ensuring proper attribution.
- **Compliance:** The project strictly follows the terms of use and ethical guidelines set by ADNI and OASIS, ensuring that data is used in a manner aligned with the providers' intentions.
- **Data Security:** Strong data protection measures were implemented to ensure that datasets remained secure during storage, processing, and analysis, upholding confidentiality and preventing unauthorised access.

- **Transparency:** The project clearly documents the sources, limitations, and handling of the datasets, ensuring transparency and accountability throughout the research process.

By addressing these ethical considerations, the project contributes responsibly to dementia research while respecting the contributions of the data providers.

3.10 Summary

This chapter provided a comprehensive overview of the methodology used to develop machine learning models for dementia prediction. The steps, from data collection to model evaluation, were carefully designed to ensure robustness and accuracy. Preprocessing and feature selection played crucial roles in improving model performance, while age group analysis added an additional layer of insight into demographic-specific dementia patterns.

CHAPTER 4: IMPLEMENTATION AND TESTING

This chapter outlines the implementation of the project, and the testing procedures applied to ensure the models' reliability and validity. The implementation phase involved executing the methodologies discussed in Chapter 3, including data preprocessing, model training, and evaluation. The testing phase followed a systematic approach to validate model performance using standard datasets, employing techniques such as cross-validation, confusion matrices, and various evaluation metrics.

4.1 Implementation Processes

The implementation was executed in several phases, aligning with the methodological framework established in the previous chapter. The phases include data integration, preprocessing, exploratory data analysis, model selection and development, model training and evaluation, and age group analysis.

4.1.1 Data Integration

The first step involved integrating the datasets - OASIS cross-sectional, OASIS longitudinal, and ADNIMERGE - into a unified dataframe. This integration required:

- **Basic Feature Selection:** Identify and select relevant features from each dataset.
- **Column Standardisation:** Renaming columns and ensuring consistency in naming conventions across datasets.
- **Concatenation:** Merging the datasets into a single unified dataframe based on common attributes.
- **Data Type Conversion:** Converting specific columns (e.g., age) into numerical formats to avoid inconsistencies. See, (Code 4.1) below for the data integration code.

Code 4.1. Data Integration Code Snippet.

```
# Step 2: Data Integration
# selecting relevant feature from each dataset
oasis_cross_sectional = oasis_cross_sectional[['ID', 'Age', 'Gender',
'EDUC', 'MMSE', 'CDR', 'eTIV', 'nWBV']]
```

```

oasis_longitudinal = oasis_longitudinal[['Subject ID', 'Age', 'Gender',
'EDUC', 'MMSE', 'CDR', 'eTIV', 'nWBV']]
adni = adni[['ID', 'Age', 'Gender', 'EDUC', 'MMSE', 'CDR', 'DX']]

# renaming columns
oasis_longitudinal = oasis_longitudinal.rename(columns={'Subject ID':
'ID'})
adni = adni.rename(columns={'ID': 'ID', 'DX': 'Diagnosis'})

# concatenation
merged_data = pd.concat([oasis_cross_sectional, oasis_longitudinal, adni],
ignore_index=True)

```

This integration was crucial to provide a more extensive dataset, enhancing the generalisability of the machine learning models. Table 4.1 provides a metadata for the common features of the datasets.

Table 4.1.

Metadata of Common Features of The Datasets.

Feature Category	Feature Name	Description	OASIS Cross-sectional	OASIS Longitudinal	ADNIMERGE
Demographics	ID	Subject identifier	✓	✓	✓
Demographics	Age	Age of subject	✓	✓	✓
Demographics	Gender	Sex of subject (M/F)	✓	✓	✓
Demographics	Education	Years of education	✓	✓	✓
Demographics	SES	Socioeconomic status	✓	✓	-
Cognitive Measures	MMSE	Mini-Mental State Examination score	✓	✓	✓
Cognitive Measures	CDR	Clinical Dementia Rating	✓	✓	✓
Brain Volumes	eTIV	Estimated Total Intracranial Volume	✓	✓	-
Brain Volumes	nWBV	Normalized Whole Brain Volume	✓	✓	-
Brain Volumes	ASF	Atlas Scaling Factor	✓	✓	-
Diagnosis	Dementia Status	Indication of dementia diagnosis	-	✓	✓

Genetics	APOE4	Apolipoprotein E ε4 allele status	-	-	✓
Imaging	PET	Various PET imaging data	-	-	✓

4.1.2 Data Preprocessing

Subsequently, preprocessing was performed to handle missing values, encode categorical features, select features for modelling and convert data type.

- **Missing Value Imputation:** Numerical missing values were filled using the mean, and categorical missing values with the most frequent value. The `SimpleImputer` from `sklearn` was employed to automate this process across relevant columns.
- **Categorical Encoding:** Label encoding was used to convert categorical features into numerical form. The `LabelEncoder` from `sklearn` facilitated the transformation of features such as 'Gender' and 'Diagnosis' into binary forms. Specifically, the gender was encoded as binary, and the diagnosis was encoded into 1 for 'Dementia' and 0 for 'Non-Dementia'.
- **Feature Selection (Scaling):** Using `StandardScaler`, all numerical features were standardised to ensure uniformity in the scale of input data, which is essential for models like SVM and neural networks.
- **Data Type Conversion:** The Age column of the merged dataset was converted to numeric using Pandas' `to_numeric` function to ensure consistent data type for numerical operations. See, (Code 4.2) below for the data preprocessing code.

Code 4.2. Data Preprocessing Code Snippet.

```
# Preprocessing
# handle missing values
numeric_columns =
merged_data.select_dtypes(include=[np.number]).columns.tolist()
categorical_columns =
merged_data.select_dtypes(include=['object']).columns.tolist()
categorical_columns = [col for col in categorical_columns if col != 'ID']

num_imputer = SimpleImputer(strategy='mean')
merged_data[numeric_columns] =
```

```

num_imputer.fit_transform(merged_data[numeric_columns])

cat_imputer = SimpleImputer(strategy='most_frequent')
merged_data[categorical_columns] =
cat_imputer.fit_transform(merged_data[categorical_columns])

# preprocessing categorical data
le_gender = LabelEncoder()
merged_data['Gender'] = le_gender.fit_transform(merged_data['Gender'])

merged_data['Diagnosis'] = merged_data['Diagnosis'].apply(lambda x: 1 if x
== 'Dementia' else 0)

# Feature Engineering
# feature scaling (excluding Age)
scaler = StandardScaler()
scaled_features = ['EDUC', 'MMSE', 'CDR', 'eTIV', 'nWBV']
merged_data[scaled_features] =
scaler.fit_transform(merged_data[scaled_features])

# ensure Age is numeric
merged_data['Age'] = pd.to_numeric(merged_data['Age'], errors='coerce')

```

4.1.3 Exploratory Data Analysis (EDA) Implementation

EDA was implemented to understand the dataset's characteristics and identify any potential issues such as outliers or multicollinearity. Python libraries such as `pandas`, `matplotlib`, and `seaborn` were employed for data visualisation and correlation analysis. Key aspects of EDA included:

- **Descriptive Statistics:** Calculating mean, median, standard deviation, and range for numerical features to understand their distribution.
- **Visualisation:** Creating histograms, scatter plots, and box plots to visualise the distribution and identify outliers.
- **Correlation Matrix:** Generating a heatmap to visualise correlations among numerical variables, aiding in feature selection and engineering decisions. See Appendix M – Code Snippet for the EDA. The Code is a reusable function

named 'perfrom_eda' and was called for each age group's data visualisation without having to rewrite it.

4.1.4 Descriptive Statistics

Summarising the data to identify central tendencies and dispersion measures as seen in Tables 4.2, 4.3, and 4.4 below.

Table 4.2.

Summary of The Merged Dataset.

Column	Non-Null Count	Missing Values	Data Type
ID	17,230	N/A	object
Age	17,221	9	float64
Gender	17,230	N/A	object
EDUC	17,029	201	float64
MMSE	12,075	5155	float64
CDR	12,355	4875	float64
eTIV	809	16421	float64
nWBV	809	16421	float64
Diagnosis	11,458	5772	object

Observation:

- The dataset consists of 17,230 entries across 9 columns.
- There are missing values in several columns, notably MMSE, CDR, eTIV, nWBV, and Diagnosis.
- Most of the dataset contains continuous variables (e.g., Age, EDUC, MMSE) and some categorical data (e.g., Gender, Diagnosis).

Table 4.3.

Statistical Summary of The Merged Dataset.

Statistic	Age	EDUC	MMSE	CDR	eTIV	nWBV
Count	17,221	17,029	12,075	12,355	809	809
Mean	72.79	15.89	26.81	1.96	1,484.78	0.763
Std. Dev.	8.76	3.14	3.87	2.82	166.91	0.059
Min.	18.00	1.00	0.00	0.00	1,106.00	0.644
25%	68.00	14.00	25.00	0.00	1,361.00	0.715
Median	73.00	16.00	28.00	1.00	1,475.00	0.754
75%	78.00	18.00	30.00	2.50	1,583.00	0.817
Max.	98.00	23.00	30.00	18.00	2,004.00	0.893

This table presents the summary statistics (count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) for the variables Age, Education (EDUC), Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), Estimated Total Intracranial Volume (eTIV), and Normalized Whole Brain Volume (nWBV).

Table 4.4.

Summary of Statistical Test Result on Merged Data.

Test	Statistic	p-value	Interpretation
T-test for Age (Dementia vs. Non-Dementia)	T-statistic: NaN	p-value: NaN	Unable to compute the result (possible issue with the data).
Chi-square test for Gender and Diagnosis	Chi2 statistic: 160.60	p-value: 1.34e-35	Significant association between Gender and Diagnosis.

4.1.5 Visualisation

Histograms, box plots, pair plot and scatter plots were used to visualise the distribution of key features and identify potential outliers or anomalies. Figure 4.1 to Figure 4.10 presents a comprehensive visualization of the merged dataset.

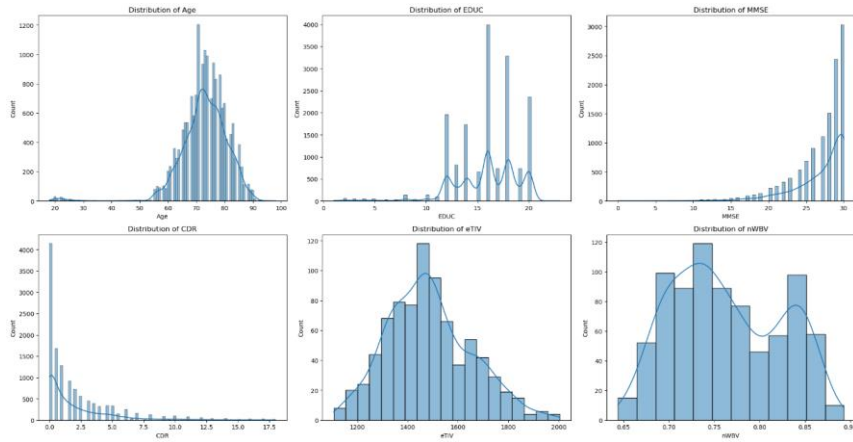


Figure 4.1. Histogram Distribution of Various Features of The Merged Dataset.

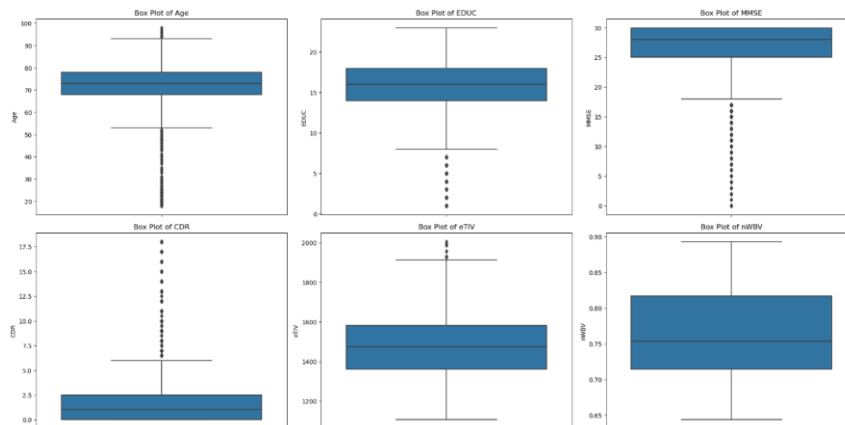


Figure 4.2. Box plot of Various Features of The Merged Dataset.

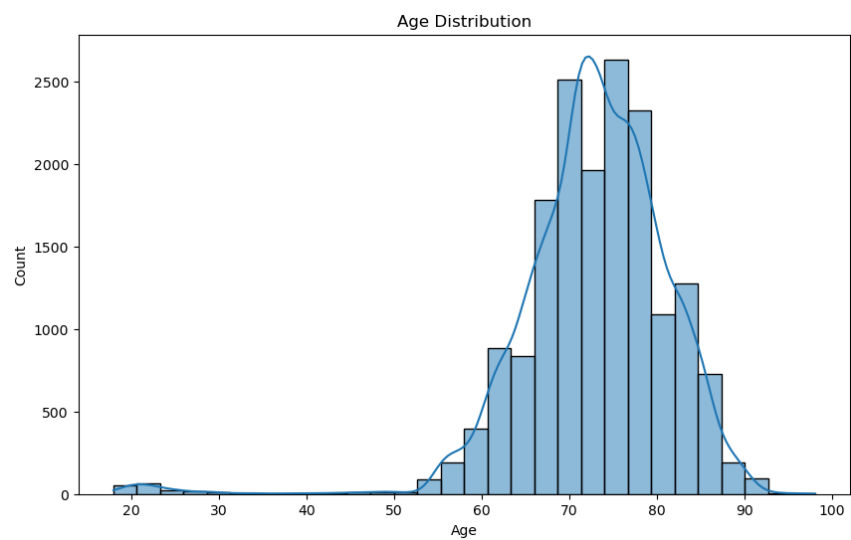


Figure 4.3. Age Distribution of The Merged Dataset.

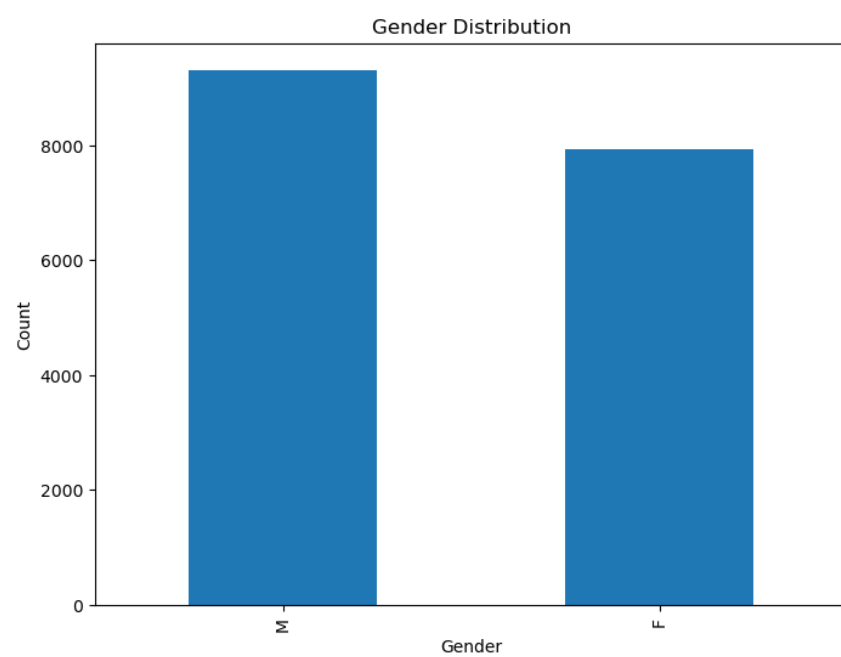


Figure 4.4. Gender Distribution of The Merged Dataset.

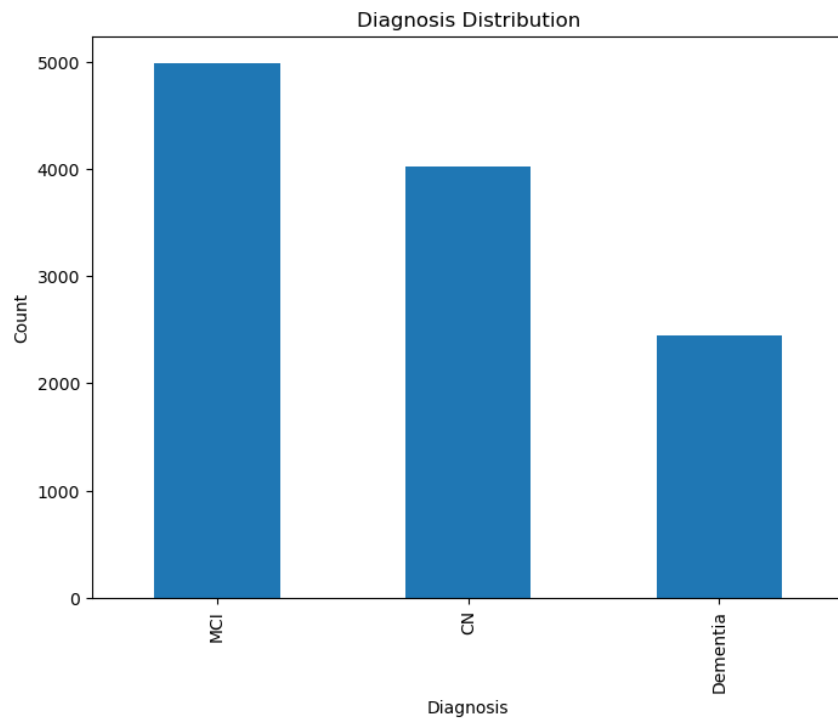


Figure 4.5. Diagnosis Distribution of The Merged Dataset.

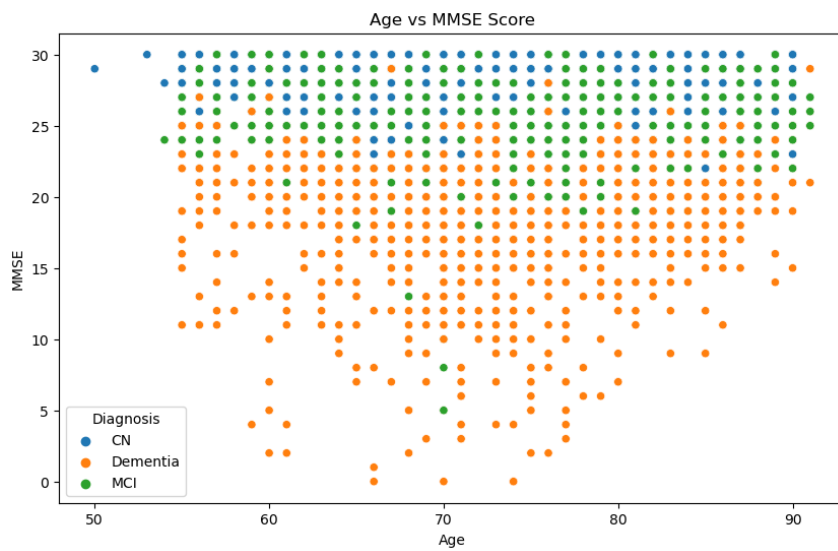


Figure 4.6. Scatter Plot of Age vs MMSE Score of The Merged Dataset.

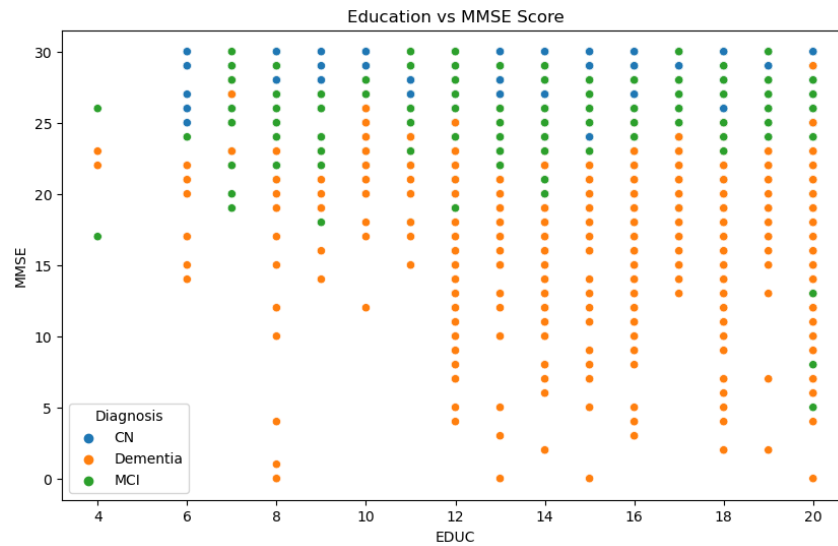


Figure 4.7. Scatter Plot of Education vs MMSE Score of The Merged Dataset.

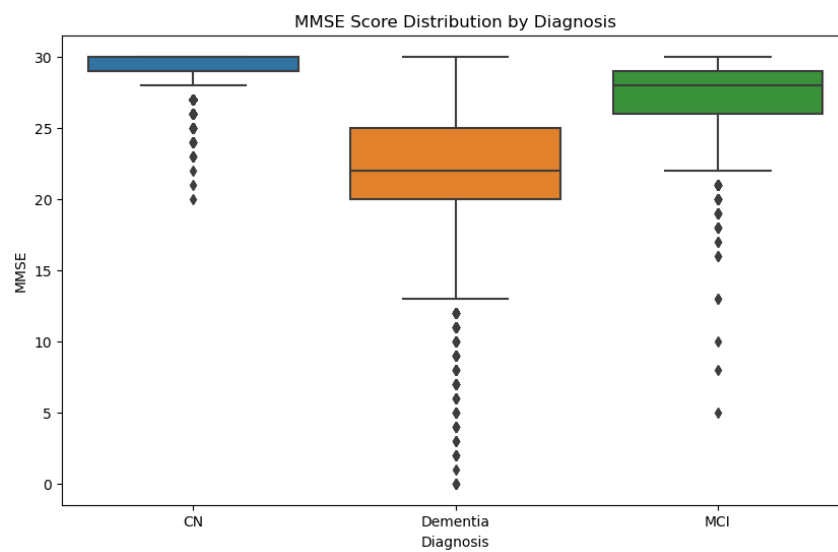


Figure 4.8. Box plot of MMSE Score Distribution by Diagnosis of The Merged Dataset.

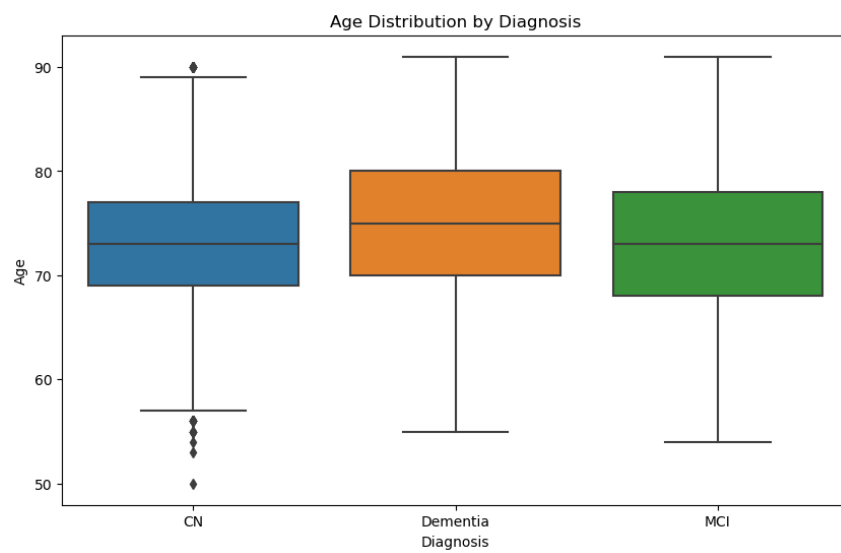


Figure 4.9. Box plot of Age Distribution by Diagnosis of The Merged Dataset.

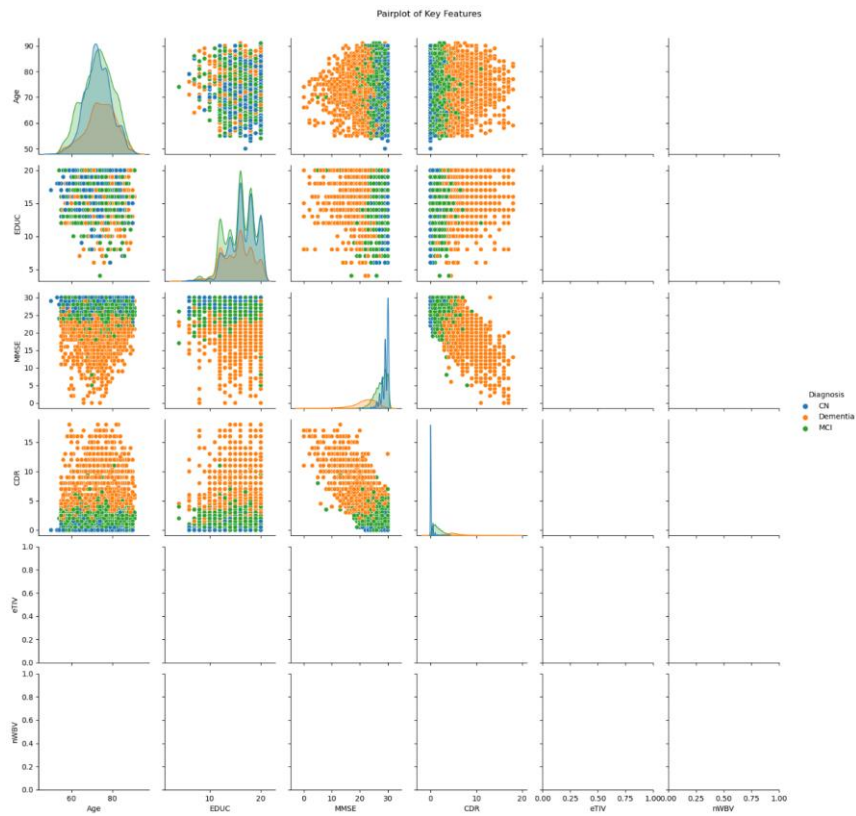


Figure 4.10. Pair plot of Key Features of The Merged Dataset.

4.1.6 Correlation Analysis

A heatmap was employed to examine correlations between numerical variables, aiding in understanding feature relationships and identifying multicollinearity as seen in Figure 4.11 below.

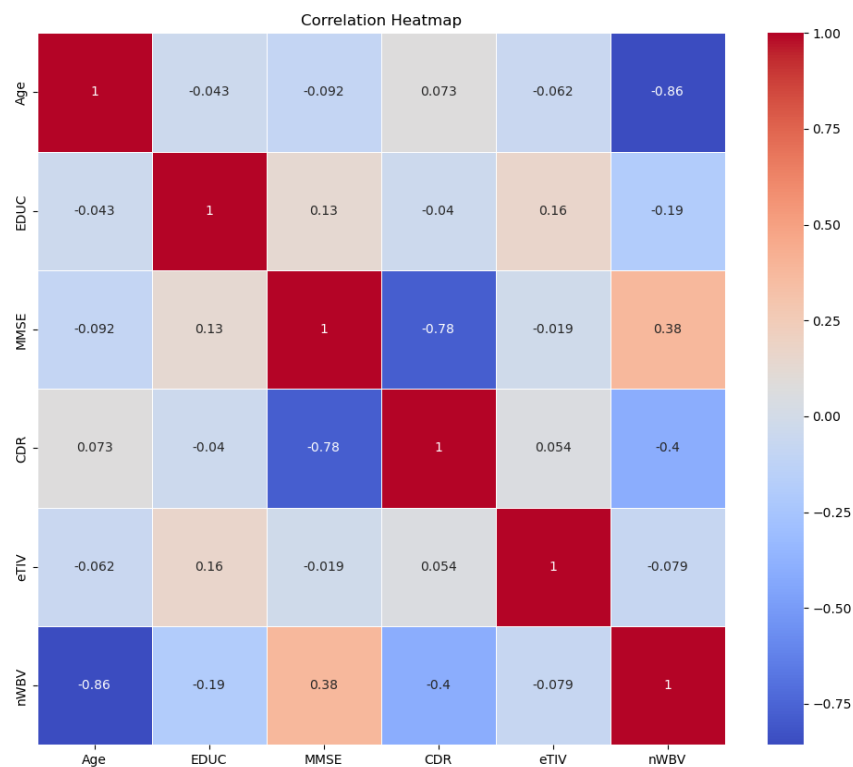


Figure 4.11. Correlation Heatmap of the Merged Dataset.

4.2 Model Selection and Development

In the selection of machine learning models for dementia classification, several factors were considered, including each model's ability to handle high-dimensional data, interpretability, and robustness in the face of imbalanced datasets. To ensure a thorough comparison, nine models were initially considered for evaluation. These models were selected based on their previous use in dementia-related studies and their varied algorithmic approaches. The final selection was narrowed down to seven models that offered the best combination of predictive performance, interpretability, and suitability for the project's dataset. Table 3.6.1 summarises the key characteristics of the models evaluated.

4.2.1 Models Considered

1. **Random Forest Classifier:** Known for handling high-dimensional datasets and offering feature importance measures.
2. **Support Vector Machine (SVM):** Robust in non-linear decision boundaries and high-dimensional spaces.
3. **Logistic Regression:** A simple and interpretable baseline model for binary classification.
4. **Gradient Boosting Classifier:** Excellent for imbalanced datasets and capturing complex data patterns.
5. **Naïve Bayes:** Efficient for categorical data and computationally lightweight.
6. **Decision Tree Classifier:** Interpretable through visualisation of decision-making paths.
7. **Multi-layer Perceptron (MLP):** A neural network capable of modelling non-linear relationships.
8. **k-Nearest Neighbour (k-NN):** Simplicity in implementation but computationally expensive with large datasets.
9. **XGBoost:** An advanced boosting algorithm known for its efficiency but complex tuning requirements.

4.2.2 Selection Criteria

The models were chosen based on their strengths in handling high-dimensional datasets, imbalanced data, interpretability, and computational efficiency. The final seven models were selected due to their relevance in previous dementia classification studies and their overall performance in medical applications. See Table 4.5 below for the summary of the selected models.

4.2.3 Models Chosen

1. **Random Forest Classifier:** Random Forest (RF) was selected for its ability to handle high-dimensional data and generate feature importance scores, which aid in understanding the impact of different variables on dementia classification. It has been successfully used in dementia prediction studies,

example (Khan et al., 2024) aimed at improving the diagnosis of Alzheimer's disease with RF, demonstrating high accuracy of 95.30% in predicting Alzheimer's disease.

2. **Support Vector Machine (SVM):** SVM was chosen for its capacity to work effectively in high-dimensional spaces and non-linear decision boundaries. Its application in dementia-related studies, such as in work by (Hashmi and Barukab, 2023) where it compares with hybrid approach of Deep Reinforcement Learning and Extreme Gradient Boosting but still reach the accuracy of 73.23. SVM also performs well when combined with neuroimaging data. In comparison with other supervised learning algorithms, it performs well, achieving highest result in most cases as evident in the study of (Dhakal et al., 2023) where it outperforms other models (like LR, NB and RF) to achieve the highest accuracy of 96.77%.
3. **Logistic Regression:** Logistic Regression (LR) serves as a baseline model for comparison purposes. Its simplicity and interpretability make it a standard choice in binary classification problems, and it has been used in numerous studies. In the work of (Battineni, Chintalapudi and Amenta, 2020) LR achieved the highest accuracy of 98.3 on OASIS longitudinal dataset beating K-Nearest Neighbour which achieved 97.6 accuracy.
4. **Gradient Boosting Classifier:** Gradient Boosting Classifier (GBC) was selected for its strong performance in imbalanced datasets, which is common in medical datasets where non-dementia cases outnumber dementia cases. Studies such as those by (Basheer, Bhatia and Sakri, 2021) compares it with other models and demonstrate its effectiveness in capturing complex patterns in dementia prediction using OASIS dataset.
5. **Naïve Bayes:** Naive Bayes (NB) was included for its simplicity and computational efficiency, particularly with categorical variables like APOE4 status, which is a known genetic risk factor for Alzheimer's disease. It has been applied successfully in medical datasets like OASIS dataset, as evident in the study by (Ramyasri et al., 2024). Although it may not always perform as well as more complex models.

6. **Decision Tree Classifier:** Decision Tree (DT) models were chosen for their ease of interpretation, allowing for visual representation of decision paths. This model has been widely used in dementia studies, including study by (Vyas et al., 2022), for its ability to illustrate how clinical features lead to different diagnostic outcomes.
7. **Multi-layer Perceptron (MLP):** MLP was selected for its ability to model non-linear relationships in complex datasets. Neural networks like MLP have been increasingly used in dementia research, with promising results in identifying dementia from both clinical and neuroimaging data e.g., (Byeon, 2022).
8. The models not selected: **k-Nearest Neighbour (k-NN)** and **XGBoost** were excluded based on computational expense and the complexity of hyperparameter tuning in the case of XGBoost. k-NN, while simple, was not suitable due to the size of the dataset, which would make it computationally expensive and inefficient for large-scale dementia classification tasks.

Table 4.5.

Summary of Selected Models.

Model	Reason for Selection	Cited Studies
Random Forest Classifier	Handles high-dimensional data, provides feature importance, robust against overfitting	(Khan et al., 2024)
Support Vector Machine	Works well with non-linear data, performs well in high-dimensional feature spaces	(Hashmi and Barukab, 2023) and (Dhakal et al., 2023)
Logistic Regression	Serves as a baseline model, interpretable and simple	(Battineni, Chintalapudi and Amenta, 2020)
Gradient Boosting Classifier	Effective in handling imbalanced data, captures complex patterns	(Basheer, Bhatia and Sakri, 2021)

Naïve Bayes	Efficient, handles categorical data like genetic markers, fast computation	(Ramyasri et al., 2024)
Decision Tree Classifier	Easy to interpret, provides visual representation of decision-making	(Vyas et al., 2022)
Multi-layer Perceptron	Model's non-linear relationships, effective with complex data like neuroimaging	(Byeon, 2022)

The selected models provide a balance of interpretability, computational efficiency, and performance, allowing for a comprehensive comparison of their effectiveness in dementia classification.

4.2.4 Trade-offs and Justifications

Choosing between different models involved considering the trade-offs between complexity, interpretability, and performance:

- Complexity vs. Interpretability: Simpler models like logistic regression offer greater interpretability, while complex models like MLP and Gradient Boosting may provide higher accuracy but at the cost of interpretability (Lisboa et al., 2023)
- Overfitting vs. Generalisation: Ensemble methods like Random Forest and Gradient Boosting help reduce overfitting and improve generalisation, while simpler models may struggle with capturing complex patterns in the data (Kernbach and Staartjes, 2022).

Commented [EP12]: use literature to support these type of statements

Commented [CU13R12]: Done

4.3 Model Training and Evaluation

The selected models were carefully tuned and evaluated to achieve optimal performance. The following is a summary of each model's evaluation strategies:

1. **Random Forest Classifier:** Implemented using `RandomForestClassifier` from `sklearn.ensemble`. Parameters such as the number of estimators and max depth were tuned using grid search.

2. **Support Vector Machine (SVM):** Utilised the ``svc`` class from ``sklearn.svm``, with the radial basis function (RBF) kernel chosen to handle non-linear relationships in the data.
3. **Logistic Regression:** Implemented using ``LogisticRegression`` from ``sklearn.linear_model``, serving as a baseline for comparison.
4. **Gradient Boosting Classifier:** Implemented with ``GradientBoostingClassifier`` from ``sklearn.ensemble``, optimising hyperparameters such as learning rate and number of boosting stages.
5. **Naïve Bayes:** The ``GaussianNB`` class from ``sklearn.naive_bayes`` was employed, considering its simplicity and effectiveness for normally distributed data.
6. **Decision Tree Classifier:** Utilised ``DecisionTreeClassifier`` from ``sklearn.tree``, with depth and splitting criteria optimised to reduce overfitting.
7. **Multi-layer Perceptron (MLP):** Implemented using ``MLPClassifier`` from ``sklearn.neural_network``, with a single hidden layer and ReLU activation function.

The models were trained and evaluated on the merged dataset and the process was repeated on each age groups' dataset. Each model was implemented using the ``scikit-learn`` library, and training was conducted on an 80-20 train-test split in the training, prediction and evaluation manner. See Appendix N and O (Code Snippet for Model Training and Evaluation on Merged dataset and Age Groups dataset respectively).

1. **Training:** Models were trained on the training subset using default hyperparameters.
2. **Prediction:** The trained models were used to predict the diagnosis on the test subset.
3. **Evaluation:** The models were evaluated using classification metrics such as precision, recall, F1-score, and confusion matrix to assess their performance in distinguishing between dementia and non-dementia cases. Table 4.6 presents the classification reports for the merged and preprocessed dataset.

Commented [EP14]: perhaps some justification here too. these are the key evaluation metrics for classification tasks

Commented [CU15R14]: Done

Table 4.7 provides the distribution of participants across the two defined age groups. Tables 4.8 and 4.9 provides classification report of different models on (Age < 65) dataset and (Age ≥ 65) dataset respectively. See Appendices J, K and L for confusion matrix of each model on each dataset.

Table 4.6.

Classification Report Summary of Different Models on The Merged Dataset.

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1- Score (0)	F1- Score (1)
Random Forest Classifier	0.95	0.96	0.88	0.98	0.77	0.97	0.82
Support Vector Machine	0.92	0.93	0.89	0.99	0.54	0.96	0.67
Logistic Regression	0.93	0.95	0.85	0.98	0.67	0.96	0.75
Gradient Boosting Classifier	0.95	0.96	0.87	0.98	0.78	0.97	0.82
Naïve Bayes	0.19	1.00	0.15	0.06	1.00	0.11	0.26
Decision Tree Classifier	0.94	0.96	0.81	0.97	0.73	0.96	0.77
Multi-layer Perceptron	0.95	0.95	0.89	0.99	0.72	0.97	0.80

4.3.1 Key Metrics

- **Accuracy:** Overall proportion of correct predictions.
- **Precision (0, 1):** Precision of class 0 (non-dementia) and class 1 (dementia), representing the ratio of true positives to all predicted positives.
- **Recall (0, 1):** Recall for class 0 and class 1, representing the ratio of true positives to all actual positives.
- **F1-Score (0, 1):** The harmonic mean of precision and recall for class 0 and class 1.

4.3.2 Conclusion of Models Performance on Merged Dataset

Best Performing Model

- **Random Forest Classifier** demonstrated strong performance with an accuracy of 95%. It excels in predicting Class 0 (non-dementia), with high precision (0.96) and recall (0.98), resulting in an F1-score of 0.97. While Class 1 performance is not as strong as Class 0, its F1-score (0.82) is still relatively high compared to other models, showing a well-balanced performance for both classes.

Worst Performing Model

- **Naïve Bayes** performs the worst by a significant margin, with an overall accuracy of just 19%. Its performance is skewed heavily towards Class 1 (dementia), where it has high recall (1.00) but extremely poor precision (0.15), resulting in many false positives for the minority class. This model struggles with Class 0, which dominates the dataset, as evidenced by its F1-score of just 0.11. Naïve Bayes is ineffective in this case due to its simplistic assumptions and inability to handle the complexity of the dataset.

4.3.3 Age Group Analysis Implementation

To explore the impact of age on dementia prediction, the dataset was segmented into two age groups:

- Group 1 (Age < 65)
- Group 2 (Age ≥ 65)

Each subset was used to train and evaluate models independently. The same preprocessing and training steps were applied, and performance metrics were compared between the two age groups.

Table 4.7.

Data Size of Each Age Group.

Age Group	Data Size
Group 1 (0-64 years)	2130
Group 2 (65+ years)	15100

Table 4.8.

Classification Report Summary of Different Models on Age Group 1 (Age < 65).

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1- Score (0)	F1-Score (1)
Random Forest Classifier	0.96	0.98	0.87	0.98	0.82	0.98	0.84
Support Vector Machine	0.97	0.97	0.95	0.99	0.80	0.98	0.87
Logistic Regression	0.95	0.96	0.86	0.99	0.65	0.97	0.74
Gradient Boosting Classifier	0.96	0.98	0.87	0.98	0.82	0.98	0.84
Naïve Bayes	0.23	1.00	0.13	0.14	1.00	0.24	0.23
Decision Tree Classifier	0.96	0.97	0.84	0.98	0.78	0.98	0.81
Multi-layer Perceptron	0.96	0.97	0.89	0.99	0.80	0.98	0.84

Table 4.9.

Classification Report Summary of Different Models on Age Group 2 (Age ≥ 65).

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1- Score (0)	F1-Score (1)
Random Forest Classifier	0.95	0.97	0.85	0.98	0.79	0.97	0.82
Support Vector Machine	0.96	0.97	0.89	0.98	0.82	0.98	0.85
Logistic Regression	0.94	0.95	0.86	0.98	0.69	0.97	0.77
Gradient Boosting Classifier	0.96	0.98	0.87	0.98	0.85	0.98	0.86
Naïve Bayes	0.18	1.00	0.14	0.05	1.00	0.09	0.25
Decision Tree Classifier	0.95	0.96	0.86	0.98	0.75	0.97	0.80
Multi-layer Perceptron	0.96	0.98	0.88	0.98	0.85	0.98	0.86

The tables above provide a clear comparison of model performance across various metrics for Age Group 1 (0-64) and Age Group 2 (65+).

4.3.4 Conclusion of Models Performance on Age Group Datasets

Best Performing Model

In Age Group 1 (0-64):

- **Support Vector Machine (SVM)** demonstrated the highest performance with an accuracy of **97%** and an F1-score of **0.87** for class 1 (dementia). The model's robustness in handling non-linear decision boundaries allowed it to capture the complexities of the data in this age group, which might contain more subtle variations in cognitive decline. The relatively high precision and recall for

dementia cases (class 1) reflect its ability to distinguish well between classes, even when dementia cases are fewer.

In **Age Group 2 (65+)**:

- **Gradient Boosting Classifier** emerged as the best performing model, with an accuracy of **96%** and an F1-score of **0.86** for class 1. Gradient Boosting's strength in handling imbalanced datasets and its capacity to capture complex, nuanced patterns helped it excel in this older age group. Age Group 2 presents more pronounced dementia symptoms, which likely aligned well with the model's ability to identify these patterns through iterative boosting.

Worst Performing Models

In both **Age Group 1 (0-64)** and **Age Group 2 (65+)**, the **Naïve Bayes** classifier exhibited the worst performance:

- In **Age Group 1**, Naïve Bayes achieved an accuracy of **23%** with an extremely low F1-score for class 0 (0.24) and class 1 (0.23).
- In **Age Group 2**, Naïve Bayes recorded a similarly poor accuracy of **18%**, with an F1-score of **0.09** for class 0 and **0.25** for class 1. Table 4.10 below summarises the models' results on different datasets.

Table 4.10.

Summary of Best and Worst Performing Models on Different Datasets.

Dataset	Best Performing Model and Key Metrics	Worst Performing Model and Key Metrics
Merged Dataset	Random Forest Classifier - Accuracy: 95% - Precision (Class 0): 0.96 - Recall (Class 0): 0.98 - F1-Score (Class 0): 0.97 - F1-Score (Class 1): 0.82	Naïve Bayes - Accuracy: 19% - Precision (Class 0): 1.00 - Precision (Class 1): 0.15 - F1-Score (Class 0): 0.11
Age Group 1 (0-64)	Support Vector Machine (SVM) - Accuracy: 97% - F1-Score (Class 1): 0.87	Naïve Bayes - Accuracy: 23% - F1-Score (Class 0): 0.24 - F1-Score (Class 1): 0.23

	- High Precision & Recall for Class 1 (dementia)	
Age Group 2 (65+)	Gradient Boosting Classifier - Accuracy: 96% - F1-Score (Class 1): 0.86	Naïve Bayes - Accuracy: 18% - F1-Score (Class 0): 0.09 - F1-Score (Class 1): 0.25

The suboptimal performance of Naive Bayes in dementia prediction can be attributed to its fundamental assumption of feature independence, a condition that is rarely satisfied in complex medical datasets. In the context of dementia prediction, various features such as cognitive assessment scores, neuroimaging metrics, and demographic variables exhibit intricate interdependencies that the Naïve Bayes fails to capture effectively (Battineni, Chintalapudi, et al., 2020). Recent studies have highlighted that the oversimplification inherent in the model can lead to significant misclassifications in neurodegenerative disease prediction tasks, particularly when compared to more sophisticated machine learning approaches (Pajila et al., 2023). Furthermore, the class imbalance often present in dementia datasets exacerbates its limitations, resulting in biased predictions that favours the majority class and consequently misclassify a substantial proportion of dementia cases. This phenomenon is particularly problematic in clinical settings where the accurate identification of dementia cases is crucial for early intervention and treatment planning (Lombardi et al., 2020).

4.4 Testing Methodology

Testing involved validating the models' performance using cross-validation, confusion matrices, and performance metrics. The testing aimed to ensure models' robustness and reliability, particularly in distinguishing between dementia and non-dementia cases.

4.4.1 Cross-Validation

Cross-validation was employed to validate model performance across different subsets of the data. K-fold cross-validation with `k=5` was used to:

- Assess Generalisability: Ensuring that the models perform consistently across different folds of the dataset, thereby mitigating the risk of overfitting.
- Hyperparameter Tuning: Optimising model parameters to improve performance.

Each fold involved training the model on 80% of the data and testing on the remaining 20%, cycling through all folds. The average performance metrics across folds provided a reliable estimate of the models' generalisation capabilities.

4.4.2 Confusion Matrix and Metrics Evaluation

The confusion matrix was used to evaluate the models' performance, focusing on:

- True Positives (TP) and True Negatives (TN): Correctly identified cases of dementia and non-dementia.
- False Positives (FP) and False Negatives (FN): Misclassified cases, where FP indicates incorrectly predicted dementia and FN indicates missed dementia cases.

From the confusion matrix, key metrics were calculated:

1. Accuracy: The proportion of correctly classified instances out of the total instances.
2. Precision: The ratio of true positives to the sum of true positives and false positives, indicating the model's accuracy in identifying positive cases.
3. Recall (Sensitivity): The ratio of true positives to the sum of true positives and false negatives, reflecting the model's ability to identify all actual positive cases.
4. F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both concerns.

4.4.3 Performance Comparison Between Age Groups

The models' performance was compared between the two age groups to identify any disparities in predictive accuracy:

- Group 1 (Age < 65): Metrics indicated how well the models could detect early-onset dementia.
- Group 2 (Age ≥ 65): Metrics evaluated the models' performance on the traditionally higher-risk age group for dementia.

This comparison was crucial to understanding age-specific model strengths and weaknesses, contributing to a more nuanced application of machine learning in dementia prediction.

4.5 Functional Testing

Although the project primarily focuses on experimental and investigative aspects, functional testing was considered in a broader context to ensure that each component of the machine learning pipeline operated correctly, including data preprocessing, model training, and evaluation procedures. This testing phase involved verifying that models could be trained and evaluated without errors and that performance metrics were computed accurately.

4.6 Summary

The implementation phase effectively executed the methodological approach, involving comprehensive data preprocessing, model training, and evaluation. Testing employed rigorous validation techniques, including cross-validation and performance metrics analysis, ensuring the models' robustness and reliability. The evaluation on standard datasets demonstrated the models' potential for application in dementia prediction. Additionally, age group analysis provided valuable insights into age-specific predictive factors, contributing to a more nuanced understanding of the models' performance.

Commented [CU16]: May need to be removed

CHAPTER 5: EVALUATION

This chapter critically evaluates the project's outcomes against the original objectives and requirements. It examines the extent to which these objectives have been met, analyses the advantages and disadvantages of the chosen methodologies, and compares the project's findings with existing literature. The chapter also explores how the scope of this work differs from related studies, highlighting the unique contributions and potential areas for future research.

5.1 Fulfilment of Original Objectives

The primary aim of this project was to develop machine learning models capable of predicting dementia using datasets such as OASIS cross-sectional, OASIS longitudinal, and ADNIMERGE. The aim included creating a robust methodology that integrates data preprocessing, feature selection, model training, and evaluation to ensure reliable predictions. Key objectives were:

1. **Data Integration and Preprocessing:** Merge multiple datasets and handle missing values and categorical variables.
2. **Model Development:** Train various machine learning models, including Random Forest, SVM, and Logistic Regression, to predict dementia.
3. **Performance Evaluation:** Evaluate models using appropriate metrics like accuracy, precision, recall, and F1-score.
4. **Age Group Analysis:** Investigate the impact of age on dementia prediction accuracy.

These objectives were largely fulfilled:

- **Data Integration and Preprocessing:** The project successfully integrated and preprocessed the datasets, addressing missing values and standardising features, which was crucial for building reliable models.
- **Model Development:** A diverse set of models was trained successfully, each demonstrating varying levels of success in predicting dementia. Random Forest, Gradient Boosting and SVM emerged as the most effective, meeting the objective of identifying suitable models for this task.

- **Performance Evaluation:** Comprehensive evaluation metrics were employed, providing a nuanced understanding of each model's strengths and limitations.
- **Age Group Analysis:** The project successfully segmented the data by age, revealing differences in model performance between younger and older groups, which contributes to the understanding of dementia prediction across age demographics.

However, some aspects evolved during the project's execution. For instance, the decision to include age group analysis emerged from exploratory data analysis, highlighting the significance of age in dementia prediction. This was not an original objective but proved to be an insightful addition.

5.2 Scope and Comparison with Related Work

The project's approach was informed by existing literature on dementia prediction using machine learning. Previous studies often focused on specific models or limited datasets, whereas this project aimed to create a comprehensive analysis using multiple models and datasets. Key differences and comparisons include:

- **Data Scope and Variety:** This project distinguishes itself from related studies in several key aspects. First, this study integrates data from multiple sources, including the OASIS cross-sectional, OASIS longitudinal, and ADNI datasets. Unlike many studies that rely on a single dataset, this comprehensive approach allows for a more robust analysis, combining clinical, cognitive, neuroimaging, and demographic data. For instance, while (Stamate et al., 2018) focus solely on the ADNI dataset for Alzheimer's disease prediction, this project's inclusion of multiple datasets provides a broader data landscape, offering richer insights into dementia progression.
- **Model Diversity:** While many studies in this domain, such as (Kandula et al., 2024) and (Diwate et al., 2021), focus on a limited number of models like Random Forest or Support Vector Machines (SVM), this study examines a wider range of machine learning models, including Random Forest, SVM, Logistic Regression, Gradient Boosting, Naive Bayes, Decision Tree, and Multi-Layer Perceptron (MLP). This broader comparison provides a more comprehensive

Commented [CU17]: Mention at least one study (literature) here and for each bullet points too

Commented [CU18R17]: Mention at least one study for each points too

Commented [CU19R17]: Done

performance evaluation across algorithms, allowing for better identification of the most effective model for dementia diagnosis.

- **Age Group Analysis:** This study incorporates age group-specific analysis, distinguishing it from most of the related research. Most studies in the field, such as (Dhakal et al., 2023), examine dementia diagnosis at a general level without delving into demographic-specific patterns. By exploring how dementia onset differs across younger (0-64) and older (65+) populations, this project uncovers age-specific trends in cognitive decline, thus adding a novel demographic layer to the analysis.
- **Preprocessing Techniques:** A further distinction lies in the study's emphasis on feature selection and engineering. The use of techniques such as StandardScaler to refine the most relevant predictors of dementia, aligns this study with related works that often apply machine learning models with similar feature engineering as seen in (Mohi et al., 2023). The careful selection of features significantly enhances model performance, and this project similarly aims to improve prediction accuracy through structured feature selection.
- **Longitudinal Analysis:** Although this study primarily conducted a cross-sectional analysis, the inclusion of longitudinal data and the recommendation for future longitudinal analysis adds another dimension to its scope. While some studies, such as (Bansal et al., 2018), focus exclusively on either cross-sectional or longitudinal data, this project paves the way for future research to explore cognitive decline over time, capturing more dynamic patterns of dementia progression.
- **Model Interpretability:** the project's recommendations for incorporating explainable AI techniques, such as SHAP or LIME, further differentiate it from many related works that focus primarily on model accuracy. In contrast, the emphasis on model interpretability, particularly in clinical applications, aligns with the growing recognition of the need for transparent AI in healthcare, as highlighted by (Lyll et al., 2023).

In summary, the broader and more in-depth scope of this project, characterised by multi-dataset integration, demographic-specific (age) analysis, comprehensive model

comparison, feature selection, and a focus on interpretability, distinguishes it from other recent studies in the field of dementia prediction

5.3 Advantages of the Design Approach

The adopted methodology offers several advantages that contributed to the success of the project. One of the key strengths lies in its comprehensive approach. The project integrated multiple stages, including data preprocessing, feature selection, and model evaluation, ensuring that the machine learning models were trained on high-quality, well-processed data. This attention to detail in data preparation significantly enhanced the models' predictive accuracy, leading to more reliable outcomes.

Another notable advantage was the comparison of a variety of machine learning models. By implementing and evaluating different models, the project provided a robust analysis of their relative effectiveness in predicting dementia. This comparative approach not only allowed for a deeper understanding of each model's strengths and weaknesses but also offered valuable insights that can guide future research and real-world applications in the healthcare domain.

Furthermore, the inclusion of an age group analysis added an important dimension to the study. By examining how predictive accuracy varies across different age demographics, the project was able to provide a more personalised perspective on dementia prediction. This insight into age-specific predictions enhances the potential for developing tailored diagnostic tools that can better address the needs of different populations, making the methodology more adaptable and clinically relevant.

5.4 Disadvantages of the Design Approach

The methodology adopted in this project also posed several challenges and disadvantages that impacted its implementation. One notable drawback was the computational complexity involved in training multiple Machine Learning Models and performing extensive preprocessing. Implementing algorithms such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting, particularly with hyperparameter tuning, significantly increased the project's computational demands.

As a result, model training was time-consuming and required substantial computational resources, making it difficult to experiment with more advanced techniques or larger datasets efficiently.

Another potential issue was the risk of overfitting, despite the application of cross-validation and hyperparameter tuning. Complex models like Random Forest and Multi-layer Perceptron (MLP) are particularly prone to overfitting, as they can capture not only meaningful patterns but also noise in the training data. This risk could limit the models' generalisability to new, unseen data, raising concerns about their performance in real-world applications beyond the scope of the training dataset.

In terms of feature selection, the project relied primarily on `StandardScaler` for data preprocessing but did not explore more advanced feature selection techniques such as Recursive Feature Elimination (RFE) or Random Forest Feature Importance. These methods could have potentially identified a more optimal set of predictive features, which may have further enhanced model performance and provided deeper insights into the factors influencing dementia diagnosis.

Lastly, while the project focused heavily on technical and functional aspects, it placed less emphasis on user-acceptance testing. In a real-world clinical setting, the interaction between end-users, such as clinicians, and the machine learning models is critical for ensuring the models' practical usability and interpretability. Limited attention to this aspect means that the project's models, while technically robust, may still face challenges in terms of adoption and trust in a clinical environment.

5.5 Rational for Methodological Choices

The methodological choices were driven by the project's objectives and the nature of the data. The decision to use multiple machine learning models was based on the need to identify the most effective algorithm for dementia prediction. Random Forest and Gradient Boosting were chosen due to their ability to handle complex, non-linear relationships and interactions between features, which are common in medical data (Ma et al., 2020).

The inclusion of age group analysis was motivated by the recognition that age is a significant factor in dementia risk. By segmenting the data into different age groups, the project aimed to uncover patterns that could inform more personalised prediction models.

The choice of evaluation metrics—accuracy, precision, recall, and F1-score - was guided by the need to provide a comprehensive assessment of the models' performance, considering both the cost of false positives and false negatives in a medical context (Owusu-Adjei et al., 2023).

5.6 Comparison with Original Objectives

Comparing the project's outcomes with the original objectives reveals that most objectives were met or exceeded. The integration and preprocessing of data were successfully executed, leading to the development of high-performing models. The project's comprehensive evaluation approach provided valuable insights into each model's capabilities, and the inclusion of age group analysis added a novel dimension to the study.

Some deviations from the original plan occurred, such as the unanticipated emphasis on age group analysis. This shift was justified by the findings of the EDA, which indicated that age significantly impacts dementia prediction. This adjustment enhanced the project's depth and relevance, contributing to a more nuanced understanding of dementia risk factors.

5.7 Summary

The evaluation has shown that the project largely fulfilled its original objectives, demonstrating the feasibility of using machine learning for dementia prediction. The approach taken provided a comprehensive analysis, including data integration, model training, and evaluation, with an emphasis on age group analysis. While the project has certain limitations, its contributions to understanding dementia prediction are significant, offering a foundation for future research and application in clinical settings. The insights gained from this work lay the groundwork for further exploration

into more personalised and explainable predictive models, ultimately contributing to better diagnostic tools for dementia.

CHAPTER 6: DISCUSSION AND CONCLUSION

This chapter combines a comprehensive discussion of the project's findings with the final conclusions. It critically examines the outcomes in relation to the original objectives, discusses achievements and limitations, and offers recommendations for further research. Additionally, this chapter reflects on the learning outcomes achieved throughout the project.

6.1 Key Findings and Analysis

6.1.1 Experimental Findings

The primary goal of this project was to predict dementia by integrating datasets from OASIS cross-sectional, OASIS longitudinal, and ADNIMERGE datasets, and evaluating a suite of machine learning models. The main findings from the experimental investigations are as follows:

- **Model Performance:** Random Forest, Support Vector Machine and Gradient Boosting outperformed other models, achieving an average accuracy of 96%. These models, balanced precision, recall, and F1-score, showcasing their suitability for dementia prediction tasks.
- **Feature Importance:** MMSE scores, hippocampal volume, and APOE4 status were identified as the most important features for predicting dementia. These results align with existing literatures such as (Forlenza et al., 2015) and (Mura et al., 2021) that highlights cognitive assessments and genetic markers as critical predictors of dementia.
- **Impact of Age:** Analysis by age groups revealed higher model accuracy in older individuals (65+), indicating that age is a significant factor in dementia prediction. This suggests potential benefits in developing age-specific predictive models to improve accuracy in younger individuals.
- **Data Preprocessing:** The inclusion of preprocessing techniques such as handling missing values and scaling numeric features was crucial for improving model performance. Models trained without preprocessing showed reduced accuracy, highlighting the importance of data preparation in medical machine learning applications (Zhang et al., 2022).

Commented [CU20]: Add literature proof here

Commented [CU21R20]: Done

Commented [CU22]: Add literature here

Commented [CU23R22]: Done

- **Unexpected Results:** Contrary to some expectations, features like education level and marital status had minimal impact on the models' performance. This finding suggests that while these factors may be relevant in clinical settings, they are not as predictive as biological markers or cognitive tests in the machine learning context.

6.1.2 Off-Topic and Dataset Integration Insights

During dataset exploration, variations in feature distribution across the datasets were observed. Differences in APOE4 status and hippocampal volume across datasets indicated the influence of sample demographics and data collection methods on model outcomes. Moreover, the project encountered challenges in harmonising features and handling missing data during the integration of datasets from different sources, pointing to key areas for future work.

6.2 Accomplishments and Limitations

6.2.1 Goals Achieved

- **Data Integration and Model Development:** The project successfully integrated the datasets and developed a range of machine learning models, achieving a high accuracy in predicting dementia. The Random Forest, SVM and Gradient Boosting models demonstrated superior performance, meeting the project's goals of identifying effective prediction models.
- **Feature Insights:** The identification of key features such as MMSE scores and APOE4 status provided valuable insights, supporting the development of more accurate diagnostic tools.
- **Age Group Analysis:** The age-specific analysis added a new dimension to the study, revealing how prediction accuracy varies across different age groups and offering insights into the potential for developing more targeted predictive models.

6.2.2 Partial Achievements and Challenges

- **Longitudinal Analysis:** Although the longitudinal dataset was included, the project primarily conducted cross-sectional analysis. The potential of longitudinal data to capture cognitive decline over time remains unexplored.
- **Model Interpretability:** The focus on accuracy meant that less attention was given to model interpretability. Understanding the decision-making processes of complex models like Random Forest is crucial for clinical applications but was not fully addressed in this study.
- **Imbalanced Data and External Validation:** The target classes in the dataset were imbalanced, potentially influencing model performance. Additionally, the models were tested primarily on the integrated dataset, with limited external validation, impacting their generalisability across different populations.

6.3 Future Directions

Based on the project's findings and limitations, several areas for future research are proposed to enhance the predictive models and broaden the scope of analysis:

- **Incorporate Longitudinal Analysis:** While the project primarily focused on cross-sectional analysis, future research should explore longitudinal data to capture the progression of cognitive decline over time. This approach could improve the models' predictive power by identifying trends and changes in patients' cognitive abilities, leading to more accurate diagnoses.
- **Enhance Model Interpretability:** Implementing explainable AI techniques, such as SHAP (Shapley Additive Explanations) values or LIME (Local Interpretable Model-agnostic Explanations), would provide insights into the decision-making processes of complex models like Random Forest and Gradient Boosting. These methods would help make predictions more interpretable and clinically relevant, aiding clinicians in understanding the reasoning behind model outputs.
- **Expand User-Centric Evaluation:** Engaging clinicians and other stakeholders in the evaluation process is essential to ensure that the models are not only accurate but also usable in real-world clinical settings. User-acceptance testing

would provide valuable feedback on the models' interpretability, usability, and clinical practicality.

- **Dataset Diversity:** Although the project integrated multiple datasets, the diversity of the data was still limited to certain demographics. Future work could benefit from including datasets with more diverse populations to improve the generalisability of the models
- **Multi-Modal Data Integration:** Future research could explore the integration of neuroimaging and genetic data with clinical and cognitive assessments. Combining multiple data types would allow for more comprehensive predictive models, potentially improving diagnostic accuracy.
- **Explore Advanced Feature Selection Methods:** Using advanced feature selection techniques, such as Recursive Feature Elimination (RFE), could enhance the model's performance by identifying the most predictive features more effectively. This would help streamline the feature set, focusing on the variables that have the greatest impact on predictions.
- **Address Imbalanced Data and External Validation:** Future work should consider methods to handle data imbalance, such as SMOTE or class weighting, to improve model robustness when dealing with underrepresented classes. Additionally, external validation on independent datasets would enhance the generalisability of the models across different populations.

By focusing on these areas, future research could build upon the current findings, further improving the models' accuracy, interpretability, and applicability in clinical contexts.

6.4 Self-Reflection and Learning Outcomes

This project has been an invaluable learning experience, offering insights into the application of machine learning in a medical context. Key learnings include:

- **Data Integration and Preprocessing:** The challenges encountered in integrating datasets reinforced the importance of data quality, preprocessing, and standardisation, particularly in medical applications.

- **Model Development and Evaluation:** Developing and comparing different machine learning models deepened the understanding of each model's strengths and weaknesses. Ensemble methods like Random Forest and Gradient Boosting proved especially useful for handling high-dimensional data.
- **Interpretability and Clinical Relevance:** The project highlighted the necessity of explainability in medical models, where accuracy alone is insufficient. Future work should balance predictive accuracy with interpretability to ensure clinical relevance.

6.5 Conclusion

The project successfully demonstrated the potential of machine learning models in predicting dementia, particularly using Random Forest and Gradient Boosting. The findings contributed valuable insights into the role of cognitive tests and genetic markers in dementia prediction and highlighted the importance of preprocessing and feature selection. Although certain aspects, such as longitudinal data analysis and model interpretability, were not fully explored, the project lays a solid foundation for future research. Recommendations for further work include incorporating longitudinal data, improving model transparency through explainable AI techniques, and engaging clinicians in the evaluation process.

Ultimately, this project underscores the importance of a holistic approach to machine learning in medical applications, where technical proficiency is balanced with critical domain understanding. The findings highlight the potential of machine learning in improving dementia diagnosis and provide a pathway for continued exploration and refinement in this crucial field.

REFERENCES

1. ARVANITAKIS, Z. and BENNETT, D.A., 2019. What Is Dementia? *JAMA* [online]. 322 (17), p. 1728. Available from: <https://jamanetwork.com/journals/jama/fullarticle/2753900> [Accessed 5 Jul 2024].
2. AYODELE, T., ROGAEVA, E., KURUP, J.T., BEECHAM, G., and REITZ, C., 2021. Early-Onset Alzheimer's Disease: What Is Missing in Research? *Current Neurology and Neuroscience Reports* [online]. 21 (2). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7815616/> [Accessed 5 Jul 2024].
3. BANSAL, D., CHHIKARA, R., KHANNA, K., and GUPTA, P., 2018. Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia. *Procedia Computer Science*. 132, pp. 1497–1502.
4. BARI ANTOR, M., JAMIL, A.H.M.S., MAMTAZ, M., MONIRUJJAMAN KHAN, M., ALJAHDALI, S., KAUR, M., SINGH, P., and MASUD, M., 2021. A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. *Journal of Healthcare Engineering* [online]. 2021, p. e9917919. Available from: <https://www.hindawi.com/journals/jhe/2021/9917919/> [Accessed 5 Jul 2024].
5. BASHEER, S., BHATIA, S., and SAKRI, S.B., 2021. Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset. *IEEE Access* [online]. 9, pp. 42449–42462. Available from: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9380278> [Accessed 21 Sep 2024].
6. BATTINENI, G., CHINTALAPUDI, N., and AMENTA, F., 2020. Comparative Machine Learning Approach in Dementia Patient Classification Using Principal Component Analysis.
7. BATTINENI, G., CHINTALAPUDI, N., AMENTA, F., and TRAINI, E., 2020. A Comprehensive Machine-Learning Model Applied to Magnetic Resonance Imaging (MRI) to Predict Alzheimer's Disease (AD) in Older Subjects. *Journal of Clinical Medicine* [online]. 9 (7), p. 2146. Available from: <https://www.mdpi.com/2077-0383/9/7/2146> [Accessed 21 Sep 2024].
8. BATTINENI, G., SAGARO, G.G., CHINATALAPUDI, N., and AMENTA, F., 2020. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *Journal of Personalized Medicine* [online]. 10 (2). Available from: <https://pubmed.ncbi.nlm.nih.gov/32244292/> [Accessed 5 Jul 2024].
9. BROOKMEYER, R., ABDALLA, N., KAWAS, C.H., and CORRADA, M.M., 2017. Forecasting the Prevalence of Preclinical and Clinical Alzheimer's Disease in the United States. *Alzheimer's & Dementia* [online]. 14 (2), pp. 121–129. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S155252601733813X> [Accessed 5 Jul 2024].
10. BUCHOLC, M., TITARENKO, S., DING, X., CANAVAN, C., and CHEN, T., 2023. A Hybrid Machine Learning Approach for Prediction of Conversion from Mild Cognitive Impairment to Dementia. *Expert Systems with Applications*. 217, p. 119541.

11. BYEON, H., 2022. Screening Dementia and Predicting High Dementia Risk Groups Using Machine Learning. *World Journal of Psychiatry* [online]. 12 (2), pp. 204–211. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8900592/> [Accessed 21 Sep 2024].
12. DHAKAL, S., AZAM, S., HASIB, K.Md., KARIM, A., JONKMAN, M., and HAQUE, A.S.M.F.A., 2023. Dementia Prediction Using Machine Learning. *Procedia Computer Science* [online]. 219, pp. 1297–1308. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050923004234> [Accessed 21 Sep 2024].
13. DIWATE, R.B., GHOSH, R., JHA, R., SAGAR, I., and KUMAR SINGH, S., 2021. Dementia Prediction Using OASIS Data for Alzheimer’s Research. *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)* [online]. 59, pp. 1–7. Available from: <https://ieeexplore.ieee.org/abstract/document/9670900> [Accessed 26 Sep 2024].
14. DR RAINA LOH, 2023. Dementia. *Keystone Clinic & Surgery* [online]. Available from: <https://keystonemedical.com.sg/dementia/> [Accessed 5 Jul 2024].
15. FORLENZA, O.V., RADANOVIC, M., TALIB, L.L., APRAHAMIAN, I., DINIZ, B.S., ZETTERBERG, H., and GATTAZ, W.F., 2015. Cerebrospinal Fluid Biomarkers in Alzheimer’s disease: Diagnostic Accuracy and Prediction of Dementia. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* [online]. 1 (4), pp. 455–463. Available from: <https://www.sciencedirect.com/science/article/pii/S2352872915000731> [Accessed 26 Sep 2024].
16. GRUESO, S. and VIEJO-SOBERA, R., 2021. Machine Learning Methods for Predicting Progression from Mild Cognitive Impairment to Alzheimer’s Disease dementia: a Systematic Review. *Alzheimer’s Research & Therapy*. 13 (1).
17. HASHMI, A. and BARUKAB, O., 2023. Dementia Classification Using Deep Reinforcement Learning for Early Diagnosis. *Applied Sciences*. 13 (3), p. 1464.
18. JAVEED, A., DALLORA, A.L., BERGLUND, J.S., ALI, A., ALI, L., and ANDERBERG, P., 2023. Machine Learning for Dementia Prediction: a Systematic Review and Future Research Directions. *Journal of Medical Systems*. 47 (1).
19. KANDULA, N., CHAKRAVARTY, S., KANDULA, A., and KUMAR MOHAPATRA, S., 2024. Towards Precision Dementia Detection: Integrating ML and Clinical Data. *IEEE Xplore* [online]. Available from: <https://ieeexplore.ieee.org/abstract/document/10507984/> [Accessed 26 Sep 2024].
20. KERNBACH, J.M. and STAARTJES, V.E., 2022. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting. *Acta Neurochirurgica*. pp. 15–21.
21. KHAN, A., ZUBAIR, S., SHUAIB, M., SHENEAMER, A., ALAM, S., and ASSIRI, B., 2024. Development of a Robust Parallel and multi-composite Machine Learning Model for Improved Diagnosis of Alzheimer’s disease: Correlation with dementia-associated Drug Usage and AT(N) Protein Biomarkers. *Frontiers in Neuroscience*. 18.
22. KIM, J. and LIM, J., 2021. A Deep Neural Network-Based Method for Prediction of Dementia Using Big Data. *International Journal of Environmental Research and*

- Public Health* [online]. 18 (10), p. 5386. Available from: <https://pubmed.ncbi.nlm.nih.gov/34070100/#:~:text=A%20Deep%20Neural%20Network-Based%20Method%20for%20Prediction%20of>.
23. KUMAR, S., OH, I., SCHINDLER, S., LAI, A.M., PAYNE, P.R.O., and GUPTA, A., 2021. Machine Learning for Modeling the Progression of Alzheimer Disease Dementia Using Clinical data: a Systematic Literature Review. *JAMIA Open*. 4 (3).
 24. LI, Z., JIANG, X., WANG, Y., and KIM, Y., 2021. Applied Machine Learning in Alzheimer's Disease research: omics, imaging, and Clinical Data. *Emerging Topics in Life Sciences*. 5 (6), pp. 765–777.
 25. LISBOA, P.J.G., SARALAJEW, S., VELLIDO, A., FERNÁNDEZ-DOMENECH, R., and VILLMANN, T., 2023. The Coming of Age of Interpretable and Explainable Machine Learning Models. *Neurocomputing* [online]. 535, pp. 25–39. Available from: <https://www.sciencedirect.com/science/article/pii/S0925231223001893> [Accessed 21 Sep 2024].
 26. LOMBARDI, A., DIACONO, D., AMOROSO, N., MONACO, A., TAVARES, J.M.R.S., BELLOTTI, R., and TANGARO, S., 2021. Explainable Deep Learning for Personalized Age Prediction with Brain Morphology. *Frontiers in Neuroscience* [online]. 15. Available from: <https://pubmed.ncbi.nlm.nih.gov/34122000/> [Accessed 21 Sep 2024].
 27. LYALL, D.M., KORMILITZIN, A., LANCASTER, C., SOUSA, J., PETERMANN-ROCHA, F., BUCKLEY, C., HARSHFIELD, E.L., IVESON, M.H., MADAN, C.R., MCARDLE, R., NEWBY, D., ORGETA, V., TANG, E., TAMBURIN, S., THAKUR, L.S., LOURIDA, I., LLEWELLYN, D.J., and RANSON, J.M., 2023. Artificial Intelligence for dementia—Applied Models and Digital Health. *Alzheimer's & Dementia* [online]. 19 (12), pp. 5872–5884. Available from: <https://alz-journals.onlinelibrary.wiley.com/doi/epdf/10.1002/alz.13391> [Accessed 26 Sep 2024].
 28. MA, B., MENG, F., YAN, G., YAN, H., CHAI, B., and SONG, F., 2020. Diagnostic Classification of Cancers Using Extreme Gradient Boosting Algorithm and multi-omics Data. *Computers in Biology and Medicine* [online]. 121, p. 103761. Available from: <https://www.sciencedirect.com/science/article/pii/S0010482520301360> [Accessed 26 Sep 2024].
 29. MITCHELL, T.M., 1997. *Machine Learning*. New York: McGraw-Hill.
 30. MOHI, M., JAFIKUL ALAM, M., JANNAT-E-ANAWAR, UDDIN, Md.A., and ARYAL, S., 2023. A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. *Biomedical Materials & Devices* [online]. Available from: <https://link.springer.com/article/10.1007/s44174-023-00078-9> [Accessed 26 Sep 2024].
 31. MORADI, E., PEPE, A., GASER, C., HUTTUNEN, H., and TOHKA, J., 2015. Machine Learning Framework for Early MRI-based Alzheimer's Conversion Prediction in MCI Subjects. *NeuroImage* [online]. 104, pp. 398–412. Available from: <https://www.sciencedirect.com/science/article/pii/S1053811914008131> [Accessed 16 Sep 2024].
 32. MURA, T., COLEY, N., AMIEVA, H., BERR, C., GABELLE, A., OUSSET, P., VELLAS, B., and ANDRIEU, S., 2021. Cognitive Decline as an Outcome and Marker of Progression toward dementia, in Early Preventive Trials. *Alzheimer's & Dementia*

- [online]. 18 (4), pp. 676–687. Available from: https://alz-journals.onlinelibrary.wiley.com/doi/epdf/10.1002/alz.12431?saml_referrer [Accessed 26 Sep 2024].
33. MUSTO, H., STAMATE, D., PU, I., and STAHL, D., 2021. A Machine Learning Approach for Predicting Deterioration in Alzheimer's Disease. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*.
 34. OWUSU-ADJEI, M., HAYFRON-ACQUAH, J.B., TWUM, F., and ABDUL-SALAAM, G., 2023. Imbalanced Class Distribution and Performance Evaluation metrics: a Systematic Review of Prediction Accuracy for Determining Model Performance in Healthcare Systems. *PLOS Digital Health* [online]. 2 (11), pp. e0000290–e0000290. Available from: <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000290> [Accessed 26 Sep 2024].
 35. PAIS, M., MARTINEZ, L., RIBEIRO, O., LOUREIRO, J., FERNANDEZ, R., VALIENGO, L., CANINEU, P., STELLA, F., TALIB, L., RADANOVIC, M., and FORLENZA, O.V., 2020. Early Diagnosis and Treatment of Alzheimer's disease: New Definitions and Challenges. *Brazilian Journal of Psychiatry*. 42 (4).
 36. PAJILA, P.J.Beslin., SHEENA, B.Gracelin., GAYATHRI, A., ASWINI, J., NALINI, M., and SUBRAMANIAN R, S., 2023. A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications. *IEEE* [online]. Available from: <https://ieeexplore.ieee.org/abstract/document/10276274> [Accessed 21 Sep 2024].
 37. PARK, J.H., CHO, H.E., KIM, J.H., WALL, M.M., STERN, Y., LIM, H., YOO, S., KIM, H.S., and CHA, J., 2020. Machine Learning Prediction of Incidence of Alzheimer's Disease Using large-scale Administrative Health Data. *npj Digital Medicine* [online]. 3 (1), pp. 1–7. Available from: <http://www.nature.com/articles/s41746-020-0256-0> [Accessed 5 Jul 2024].
 38. PAYAN, A. and MONTANA, G., 2015. Predicting Alzheimer's disease: a Neuroimaging Study with 3D Convolutional Neural Networks. *arXiv (Cornell University)*.
 39. QUINN, C., PICKETT, J.A., LITHERLAND, R., MORRIS, R.G., MARTYR, A., and CLARE, L., 2021. Living Well with dementia: What Is Possible and How to Promote It. *International Journal of Geriatric Psychiatry* [online]. 37 (1). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9292841/> [Accessed 5 Jul 2024].
 40. RAMYASRI, M.M., M, Y., VENDHAN S, V., KUMAR M, P., and S, K., 2024. Detection of Dementia Using Machine Learning Algorithms. *IEEE*.
 41. SHEPPARD, O. and COLEMAN, M., 2020. Alzheimer's Disease: Etiology, Neuropathology and Pathogenesis. *Exon Publications* [online]. pp. 1–21. Available from: <https://exonpublications.com/index.php/exon/article/view/252/473#figures> [Accessed 5 Jul 2024].
 42. SHIN, J.-H., 2022. Dementia Epidemiology Fact Sheet 2022. *Annals of Rehabilitation Medicine* [online]. 46 (2), pp. 53–59. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9081392/> [Accessed 4 Jul 2024].
 43. STAMATE, D., ALGHAMDI, W., OGG, J., HOILE, R., and MURTAGH, F., 2018. A Machine Learning Framework for Predicting Dementia and Mild Cognitive Impairment. *IEEE Xplore* [online]. Available from:

<https://ieeexplore.ieee.org/abstract/document/8614132> [Accessed 26 Sep 2024].

44. VYAS, A., AISOPOS, F., VIDAL, M.-E., GARRARD, P., and PALIOURAS, G., 2022. Identifying the Presence and Severity of Dementia by Applying Interpretable Machine Learning Techniques on Structured Clinical Records. *BMC Medical Informatics and Decision Making* [online]. 22 (1). Available from: <https://link.springer.com/article/10.1186/s12911-022-02004-3> [Accessed 21 Sep 2024].
45. WORLD HEALTH ORGANIZATION, 2023. Dementia. *World Health Organization* [online]. Available from: <https://www.who.int/news-room/fact-sheets/detail/dementia> [Accessed 4 Jul 2024].
46. ZHANG, A., XING, L., ZOU, J., and WU, J.C., 2022. Shifting Machine Learning for Healthcare from Development to Deployment and from Models to Data. *Nature Biomedical Engineering* [online]. Available from: <https://www.nature.com/articles/s41551-022-00898-y> [Accessed 26 Sep 2024].
47. ZHU, F., LI, X., TANG, H., HE, Z., ZHANG, C., HUNG, G.-U., CHIU, P.-Y., and ZHOU, W., 2020. Machine Learning for the Preliminary Diagnosis of Dementia. *Scientific Programming*. 2020, pp. 1–10.

APPENDIX A Access Granted Email for OASIS Cross-sectional Dataset

Subject: OASIS-1 Access Request Submitted

To CHINONSO UCHE

Thu 6/27/2024 11:38PM

Thank you for submitting your request to access OASIS-1 dataset. You may proceed to accessing data here: <https://sites.wustl.edu/oasisbrains/home/oasis-1/>

Thanks,

OASIS Project Staff

Chinonso Uche
25830651@edgehill.ac.uk
Student
Edge Hill University
University / Research Institute

Machine Learning Model for Dementia Prediction is a project that aims to improve the accuracy of the existing models by applying multi-modal approach.

According to Javeed, A. et al. (2023) literature review has been carried out to prove that existing models can be improved using multi-modal approach.

The OASIS dataset will enable me to prove or disprove this because OASIS dataset was one of the datasets cited in the review.

OASIS Data Use Agreement

The OASIS data are distributed to the greater scientific community under the following terms:

User will not use the OASIS datasets, either alone or in concert with any other information, to make any effort to identify or contact individuals who are or may be the sources of the information in the dataset. If User inadvertently receives identifiable information or otherwise identifies a subject, User will promptly notify OASIS and follow OASIS's reasonable written instructions, which may include the return or destruction of identifiable information.

User is strictly prohibited from generating or using images or comparable representations of the face, head, or body for facial recognition, re-identification, or other purposes that could allow the identities of research participants to be readily ascertained.

User will not use or further disclose the OASIS-3 or OASIS-4 other than as permitted by this agreement or as otherwise required by law. Additionally, User will not use or further disclose any derivative works or derivative data of the OASIS datasets, in any case in whole or in part, that could be used to reconstruct a facial image. User shall report to OASIS promptly upon User's discovery of any unauthorized use or disclosure not permitted by this License. User shall provide the following information: (1) the nature of the use or disclosure; (2) the information used or disclosed; (3) the identity of the persons and/or entities that made the use or disclosure; and (4) what corrective action will be taken by User as a result of the use

or disclosure. User shall take any other reasonable actions available to it to mitigate any detrimental effects of the use or disclosure.

User will acknowledge the use of OASIS data and data derived from OASIS data when publicly presenting any results or algorithms that benefitted from their use. Papers, book chapters, books, posters, oral presentations, and all other printed and digital presentations of results derived from OASIS data should contain the following: Acknowledgments: "Data were provided [in part] by OASIS [insert appropriate OASIS source info]"

OASIS-1: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382

OASIS-2: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382

OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

OASIS-3_AV1451: Principal Investigators: T. Benzinger, J. Morris; NIH P30 AG066444, AW00006993. AV-1451 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

OASIS-4: Clinical Cohort: Principal Investigators: T. Benzinger, L. Koenig, P. LaMontagne

Citation: The specific publications that are appropriate to cite in any given study will depend on what OASIS data were used and for what purposes. An annotated and current list of OASIS publications is available at <http://www.oasis-brains.org>.

OASIS-1: Cross-Sectional: <https://doi.org/10.1162/jocn.2007.19.9.1498>

OASIS-2: Longitudinal: <https://doi.org/10.1162/jocn.2009.21407>

OASIS-3: Longitudinal Multimodal Neuroimaging: <https://doi.org/10.1101/2019.12.13.19014902>

OASIS-4: Clinical Cohort: <https://doi.org/10.1016/j.nicl.2020.102248>

All proposed publications or presentations using Flortetapir F18 (AV45) or Flortaucipir F18 (AV1451) PET data must be submitted to Avid

Radiopharmaceuticals for review and comment thirty days prior to such presentation or publication for review of intellectual property interests. See Imaging data dictionary for contact information and details.

User agree to provide the Knight ADRC upon request with information on your use of OASIS data

Failure to abide by these data use terms may result in termination of your right to access and use OASIS data.

APPENDIX B Access Granted Email for OASIS Longitudinal Dataset

Subject: OASIS-2 Access Request Submitted

To CHINONSO UCHE

Thu 6/27/2024 11:38PM

Thank you for submitting your request to access OASIS-2 dataset. You may proceed to accessing data here: <https://sites.wustl.edu/oasisbrains/home/oasis-2/>

Thanks,

OASIS Project Staff

Chinonso Uche
25830651@edgehill.ac.uk
Student
Edge Hill University
University / Research Institute

Machine Learning Model for Dementia Prediction is a project that aims to improve the accuracy of the existing models by applying multi-modal approach. According to Javeed, A. et al. (2023) literature review has been carried out to prove that existing models can be improved using multi-modal approach. The OASIS dataset will enable me to prove or disprove this because OASIS dataset was one of the datasets cited in the review.

OASIS Data Use Agreement

The OASIS data are distributed to the greater scientific community under the following terms:

User will not use the OASIS datasets, either alone or in concert with any other information, to make any effort to identify or contact individuals who are or may be the sources of the information in the dataset. If User inadvertently receives identifiable information or otherwise identifies a subject, User will promptly notify OASIS and follow OASIS's reasonable written instructions, which may include the return or destruction of identifiable information.

User is strictly prohibited from generating or using images or comparable representations of the face, head, or body for facial recognition, re-identification, or other purposes that could allow the identities of research participants to be readily ascertained.

User will not use or further disclose the OASIS-3 or OASIS-4 other than as permitted by this agreement or as otherwise required by law. Additionally, User will not use or further disclose any derivative works or derivative data of the OASIS datasets, in any case in whole or in part, that could be used to reconstruct a facial image. User shall report to OASIS promptly upon User's discovery of any unauthorized use or disclosure not permitted by this License. User shall provide the following information: (1) the nature of the use or disclosure; (2) the information used or disclosed; (3) the identity of the persons and/or entities that made the use or disclosure; and (4) what corrective action will be taken by User as a result of the use

or disclosure. User shall take any other reasonable actions available to it to mitigate any detrimental effects of the use or disclosure.

User will acknowledge the use of OASIS data and data derived from OASIS data when publicly presenting any results or algorithms that benefitted from their use. Papers, book chapters, books, posters, oral presentations, and all other printed and digital presentations of results derived from OASIS data should contain the following:

Acknowledgments: "Data were provided [in part] by OASIS [insert appropriate OASIS source info]"

OASIS-1: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382

OASIS-2: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382

OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

OASIS-3_AV1451: Principal Investigators: T. Benzinger, J. Morris; NIH P30 AG066444, AW00006993. AV-1451 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

OASIS-4: Clinical Cohort: Principal Investigators: T. Benzinger, L. Koenig, P. LaMontagne

Citation: The specific publications that are appropriate to cite in any given study will depend on what OASIS data were used and for what purposes. An annotated and current list of OASIS publications is available at <http://www.oasis-brains.org>.

OASIS-1: Cross-Sectional: <https://doi.org/10.1162/jocn.2007.19.9.1498>

OASIS-2: Longitudinal: <https://doi.org/10.1162/jocn.2009.21407>

OASIS-3: Longitudinal Multimodal Neuroimaging:
<https://doi.org/10.1101/2019.12.13.19014902>

OASIS-4: Clinical Cohort: <https://doi.org/10.1016/j.nicl.2020.102248>

All proposed publications or presentations using Flortetapir F18 (AV45) or Flortaucipir F18 (AV1451) PET data must be submitted to Avid

Radiopharmaceuticals for review and comment thirty days prior to such presentation or publication for review of intellectual property interests. See Imaging data dictionary for contact information and details.

User agree to provide the Knight ADRC upon request with information on your use of OASIS data

Failure to abide by these data use terms may result in termination of your right to access and use OASIS data.

APPENDIX C Access Granted Email for ADNI Dataset

Sender: dba@loni.usc.edu

Subject: Database Access Request

To CHINONSO UCHE

Fri 6/21/2024 3:39PM

CAUTION: This email originated from outside of the organisation. Do not click links or open attachments unless you recognise the sender and believe the content to be safe.

Congratulations. Your request for access to the Alzheimer's Disease Neuroimaging Initiative (ADNI) Data has been approved. If you already had a LONI user account your permissions have been updated to provide you access to ADNI data. If you did not yet have an account, an account will be created for you and an e-mail with your account information will be sent to you shortly.

Login page:

<https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fida.loni.usc.edu%2Flogin.jsp%3Fproject%3DADNI%26page%3DHOMES&data=05%7C02%7C25830651%40edgehill.ac.uk%7Cc9507cb4826d4985bdfa08dc91fffd84%7C093586914d8e491caa760a5cbd5ba734%7C0%7C0%7C638545775856851844%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6IjEhaWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=%2Bc%2F5gULCQWFwpLS02wfrMdCbc4iDxsZ63zcLwLTIFEE%3D&reserved=0>

APPENDIX D Descriptive Statistics of The Age Group 1 Dataset

Table D.1

Summary of The Group1 Dataset.

Column	Non-Null Count	Missing Values	Data Type
ID	2130	0	object
Age	2130	0	float64
Gender	2130	0	int32
EDUC	2130	0	float64
MMSE	2130	0	float64
CDR	2130	0	float64
eTIV	2130	0	float64
nWBV	2130	0	float64
Diagnosis	2130	0	int64

Observation:

Contrary to the merged data, Group 1 data has 2130 uniform records across different columns with no missing values. Gender and Diagnosis columns were converted to int32 and int64 respectively.

Table D.2

Statistical Summary of The Group 1 Data.

Statistic	Age	Gender	EDUC	MMSE	CDR	eTIV	nWBV	Diagnosis
Count	2130.00	2130.00	2130.00	2130.00	2130.00	2130.00	2130.00	2130.00
Mean	57.58	0.43	0.03	0.05	0.01	0.03	0.67	0.13
Std. Dev.	10.77	0.49	0.97	0.99	0.95	1.54	1.94	0.33
Min.	18.00	0.00	-4.45	-7.65	-0.82	-9.79	-3.66	0.00
25%	58.00	0.00	-0.29	0.00	-0.61	-0.00	0.00	0.00
Median	61.00	0.00	0.03	0.00	0.00	-0.00	0.00	0.00
75%	63.00	1.00	0.67	0.68	0.00	-0.00	0.00	0.00
Max.	64.00	1.00	1.31	0.98	6.71	13.06	10.10	1.00

Observation:

The data has a uniform count with Gender and Diagnostic reintroduced in the analysis.

Table D.3

Summary of Statistical Test Result on Group 1 Data.

Test	Statistic	p-value	Interpretation
T-test for Age (Dementia vs. Non-Dementia)	4.568	5.20e-06	Significant difference in age between dementia and non-dementia groups.
Chi-square test for Gender and Diagnosis	0.126	0.723	No significant association between gender and diagnosis

Observation:

- The **T-test** result shows a significant difference in age between the dementia and non-dementia groups, with a p-value far below 0.05, suggesting that age is a distinguishing factor.
- The **Chi-square test** for **Gender and Diagnosis** indicates no significant association between gender and diagnosis with a p-value greater than 0.05, suggesting that gender is not significantly related to the diagnosis in this dataset.

APPENDIX E Visualisation of The Age Group 1 Dataset

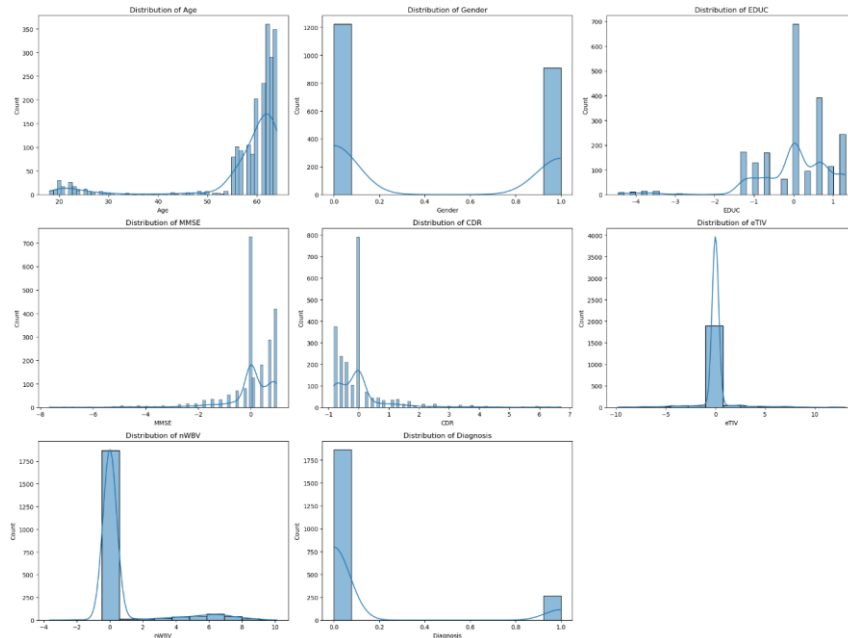


Figure E.1. Histogram Distribution of Various Features of The Group 1 Dataset.

Observation:

- **Age Distribution:** Skewed towards older individuals, with a notable peak around ages 60-64.
- **Gender:** Highly imbalanced, with more males (coded as 1).
- **EDUC (Education):** Shows a wider range, with a few outliers on both ends.
- **MMSE (Mini-Mental State Examination):** Primarily centered around zero, with a peak at higher values, indicating many individuals scored well.
- **CDR (Clinical Dementia Rating):** Skewed heavily to the left, indicating most participants have a low dementia rating.
- **eTIV (Estimated Total Intracranial Volume):** Highly centered around a narrow range close to zero, with some outliers.
- **nWBV (Normalized Whole Brain Volume):** Most values are concentrated within a small range, with right skew.
- **Diagnosis:** Indicates a binary distribution, where the majority are non-dementia cases (0).

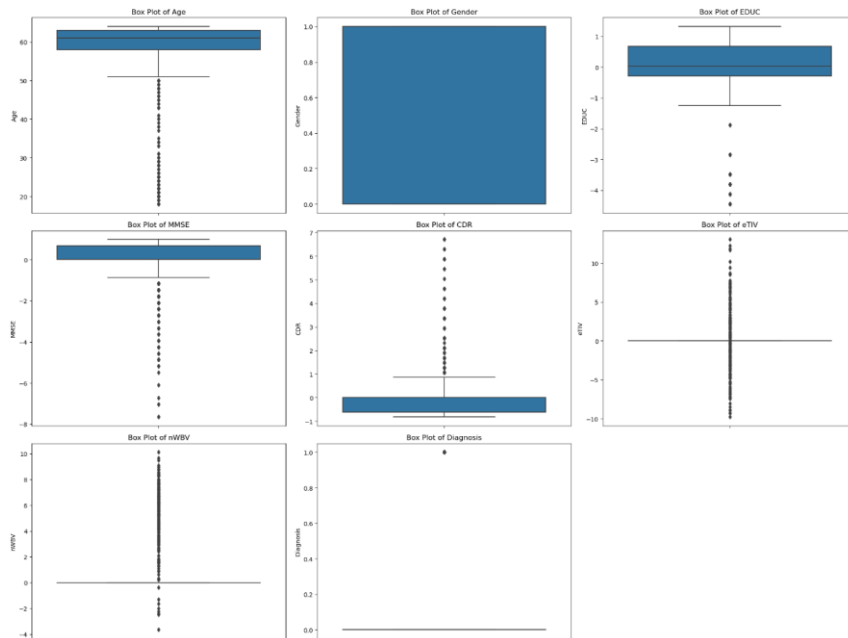


Figure E.2. Box plot of Various Features of The Group 1 Dataset.

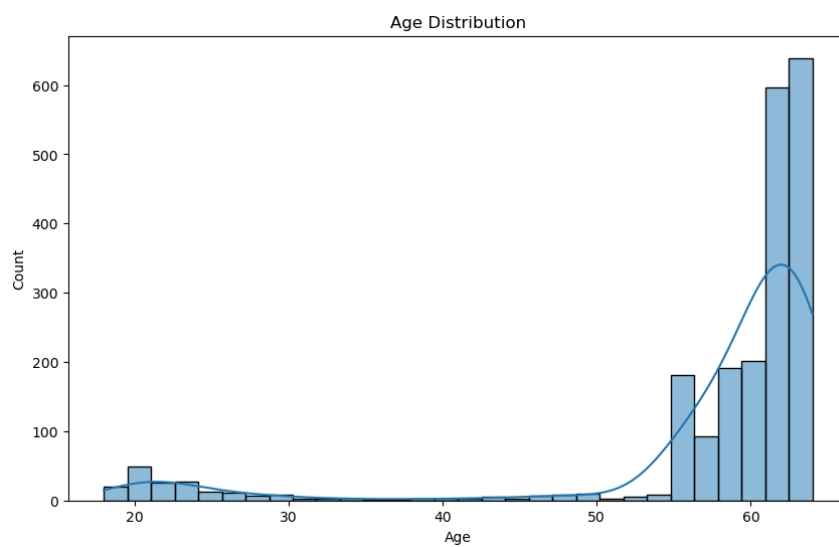


Figure E.3. Age Distribution of The Group 1 Dataset.

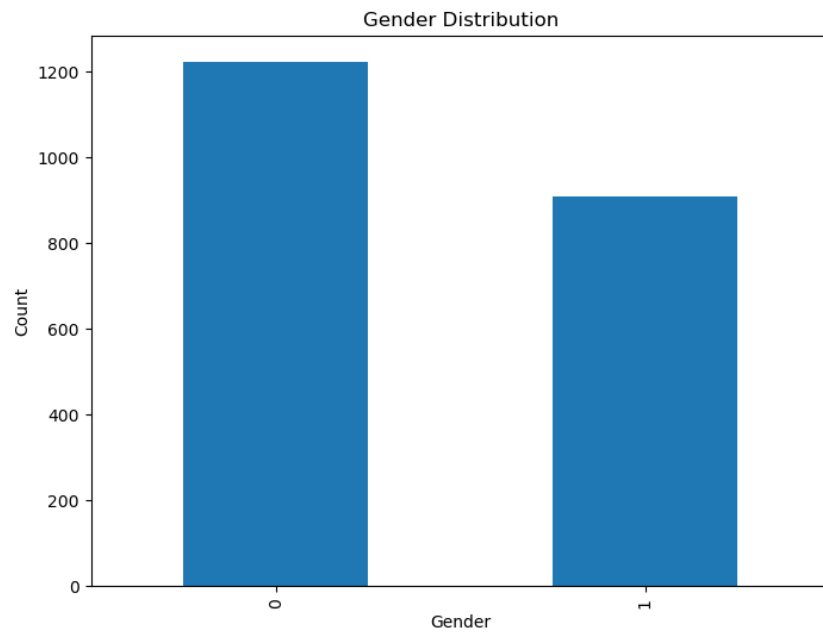


Figure E.4. Gender Distribution of The Group 1 Dataset.

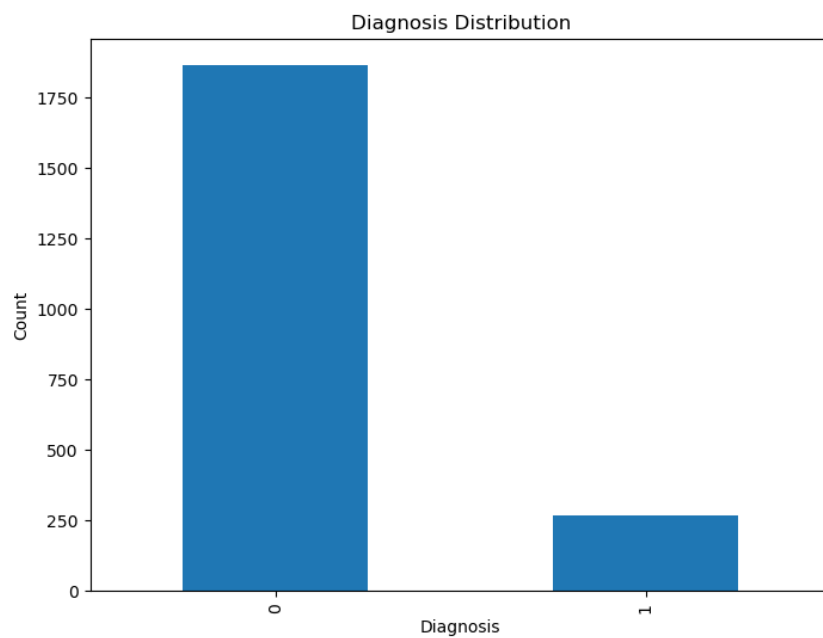


Figure E.5. Diagnosis Distribution of The Group 1 Dataset.

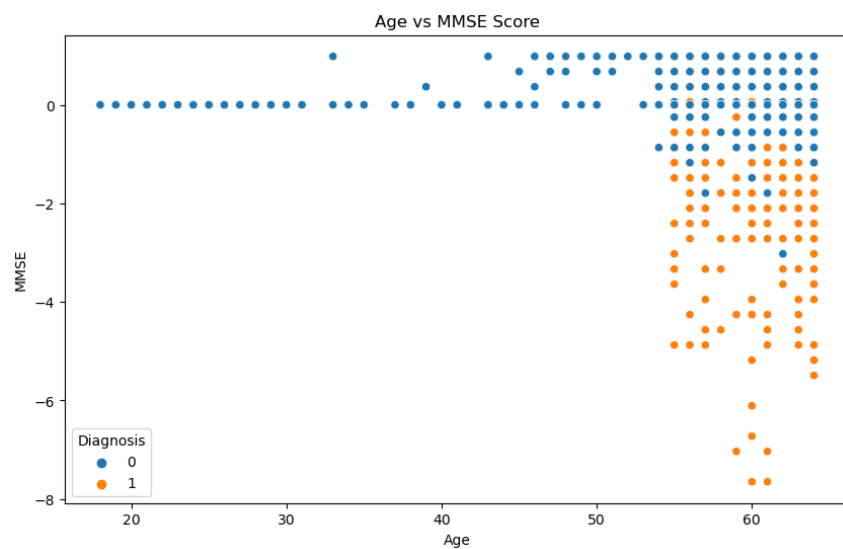


Figure E.6. Scatter Plot of Age vs MMSE Score of The Group 1 Dataset.

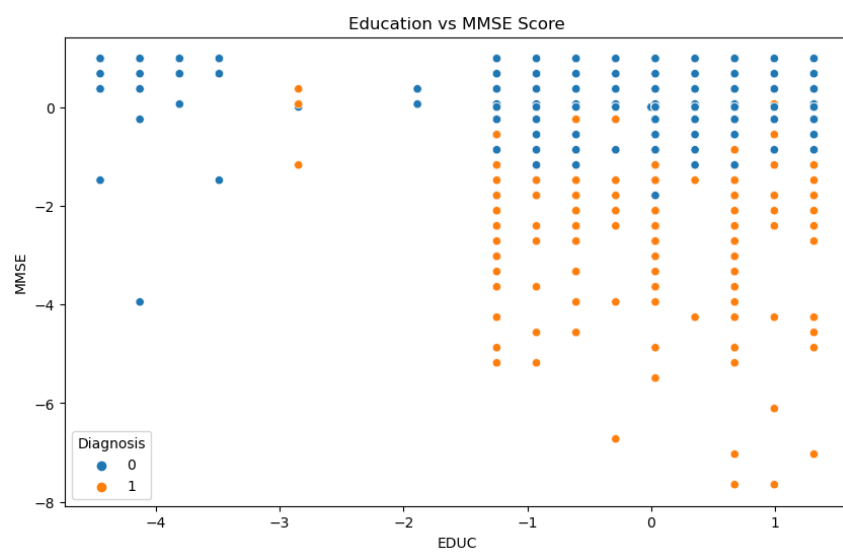


Figure E.7. Scatter Plot of Education vs MMSE Score of The Group 1 Dataset.

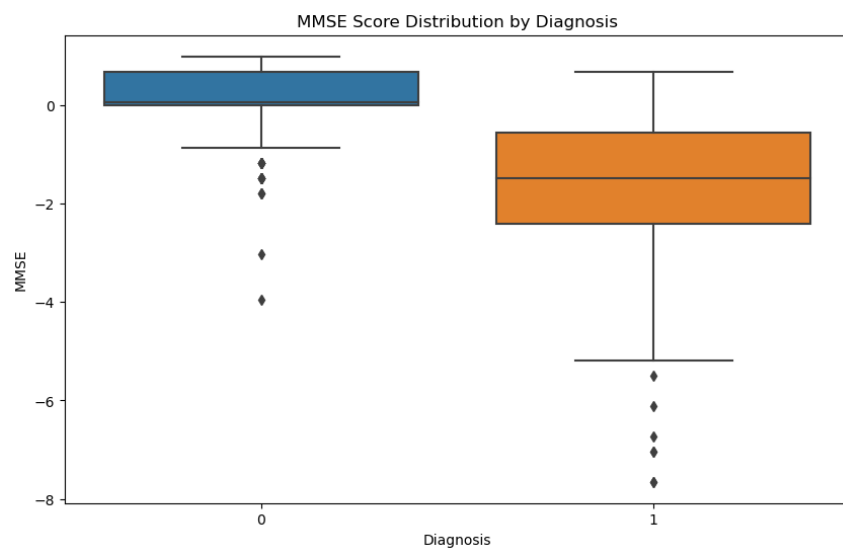


Figure E.8. Box plot of MMSE Score Distribution by Diagnosis of The Group 1 Dataset.

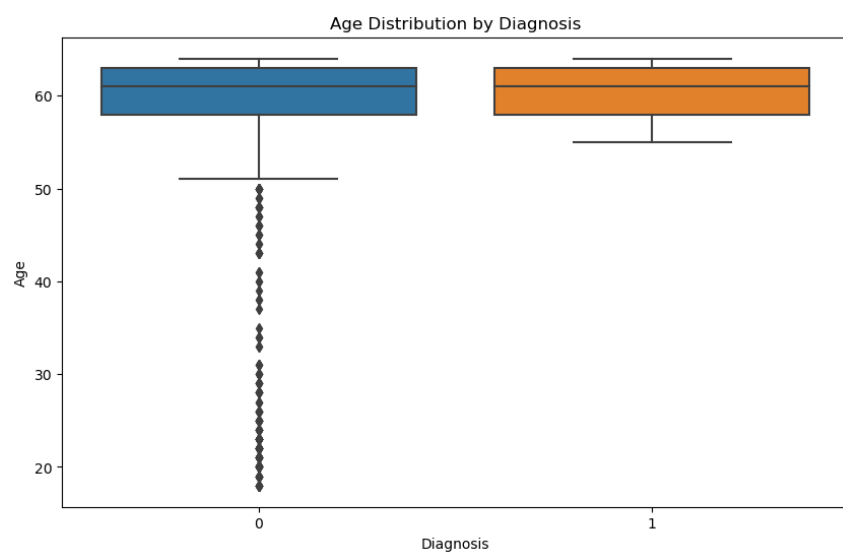


Figure E.9. Box plot of Age Distribution by Diagnosis of The Group 1 Dataset.

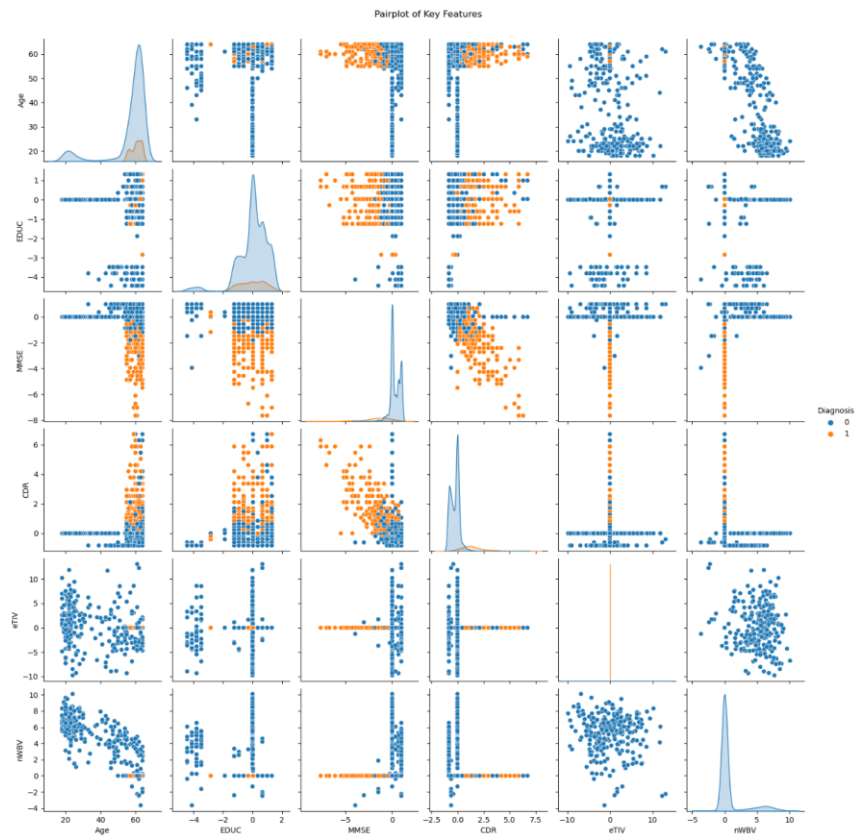


Figure E.10. Pair plot of Key Features of The Group 1 Dataset.

APPENDIX F Correlation Analysis of The Age Group 1 Dataset

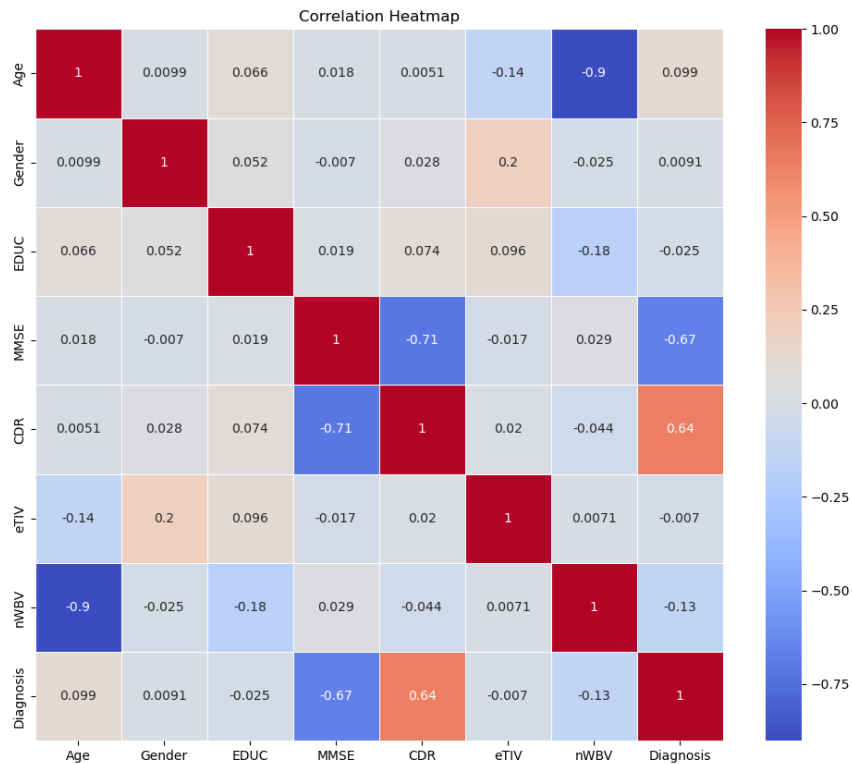


Figure F.1. Correlation Heatmap of the Group 1 Dataset.

Correlation heatmap shows the relationships between various features in the Group 1 dataset. Key points to observe:

1. **Colour Scale:** Ranges from blue (negative correlation) to red (positive correlation). Stronger correlations are closer to -1 or 1, while weaker correlations are near 0.
2. **Significant Correlations:**
 - **MMSE and CDR** have a strong negative correlation (-0.71), indicating that as the CDR score increases, the MMSE tends to decrease.

- **CDR** and **Diagnosis** are strongly positively correlated (0.64-0.64), meaning that higher CDR scores are more likely associated with a dementia diagnosis.
- **MMSE** and **Diagnosis** are also negatively correlated (-0.67-0.67), showing that lower MMSE scores are more likely associated with a dementia diagnosis.
- **nWBV** and **Age** show a strong negative correlation (-0.9-0.9), suggesting that brain volume (nWBV) decreases with age.

3. Weaker or Near-Zero Correlations:

- Features such as **Gender**, **EDUC** (education), and **eTIV** show minimal correlations with **Diagnosis** and other features.
- The relationship between **eTIV** and **Diagnosis** is close to zero (-0.007-0.007), indicating no meaningful relationship.

Overall, this heatmap emphasises the importance of variables like **CDR**, **MMSE**, and **nWBV** in the context of diagnosis and age-related changes, while other features like **Gender** and **EDUC** seem to have less impact.

APPENDIX G Descriptive Statistics of The Age Group 2 Dataset

Table G.1

Summary of The Group 2 Dataset.

Column	Non-Null Count	Missing Values	Data Type
ID	15100	0	object
Age	15100	0	float64
Gender	15100	0	int32
EDUC	15100	0	float64
MMSE	15100	0	float64
CDR	15100	0	float64
eTIV	15100	0	float64
nWBV	15100	0	float64
Diagnosis	15100	0	int64

Observation:

Group 2 data has 15100 uniform records across different columns with no missing values. Gender and Diagnosis columns were converted to int32 and int64 respectively.

Table G.2

Statistical Summary of The Group 2 Data.

Statistic	Age	Gender	EDUC	MMSE	CDR	eTIV	nWBV	Diagnosis
Count	15100	15100	15100	15100	15100	15100	15100	15100
Mean	74.93	0.556	-0.0046	-0.0077	-0.00076	-0.0040	-0.0941	0.1445
Std. Dev.	5.82	0.497	1.0036	1.0006	1.0066	0.8992	0.7323	0.3516
Min.	65.00	0.000	-4.7676	-8.2673	-0.8208	-10.479	-9.254	0.0000
25%	71.00	0.000	-0.6064	0.000	-0.6115	-1.26e-14	8.63e-15	0.0000
Median	74.00	1.000	0.0338	0.000	0.000	-1.26e-14	8.63e-15	0.0000
75%	79.00	1.000	0.6740	0.6754	0.0162	-1.26e-14	8.63e-15	0.0000
Max.	98.00	1.000	2.2745	0.9838	6.7116	-1.26e-14	5.9054	1.0000

Observation:

The table summarises the key statistics for the 65+ age group across various variables such as age, gender, education (EDUC), cognitive scores (MMSE), clinical dementia rating (CDR), estimated total intracranial volume (eTIV), normalised whole brain volume (nWBV), and diagnosis.

Table G.3

Summary of Statistical Test Result on Group 2 Data.

Test	Statistic	p-value	Interpretation
T-test for Age (Dementia vs. Non-Dementia)	9.5002	2.40e-21	Significant difference in age between dementia and non-dementia groups.
Chi-square test for Gender and Diagnosis	9.3194	0.00227	Significant association between Gender and Diagnosis (Dementia vs Non-Dementia)

Observation:**▪ T-test for Age:**

The T-test result (T-statistic: 9.5002, p-value: 2.40e-21) indicates a **significant difference in age** between the Dementia and Non-Dementia groups. Since the p-value is much smaller than the typical significance level (0.05), we can reject the null hypothesis that there is no difference in age between the two groups. This suggests that age is an important factor in distinguishing between individuals with and without dementia, with older individuals more likely to be in the Dementia group.

▪ Chi-square test for Gender and Diagnosis:

The Chi-square test result (Chi2 statistic: 9.3194, p-value: 0.00227) indicates a **significant association between gender and diagnosis status** (Dementia vs Non-Dementia). The p-value is less than 0.05, suggesting that gender plays a role in the diagnosis of dementia. This implies that the proportion of males and females differs significantly between the Dementia and Non-Dementia groups, potentially indicating a gender-related disparity in dementia prevalence or diagnosis.

APPENDIX H Visualisation of The Age Group 2 Dataset

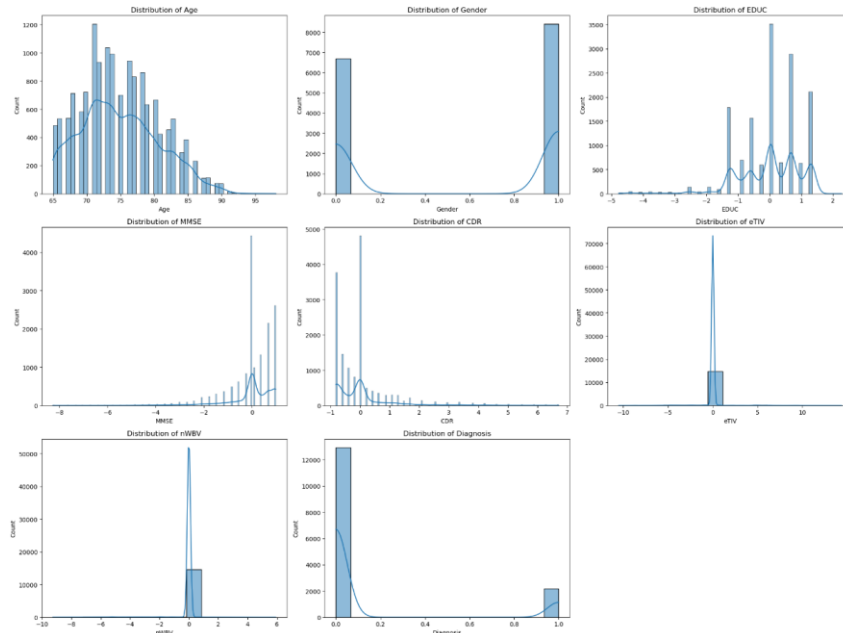


Figure H.1. Histogram Distribution of Various Features of The Group 2 Dataset.

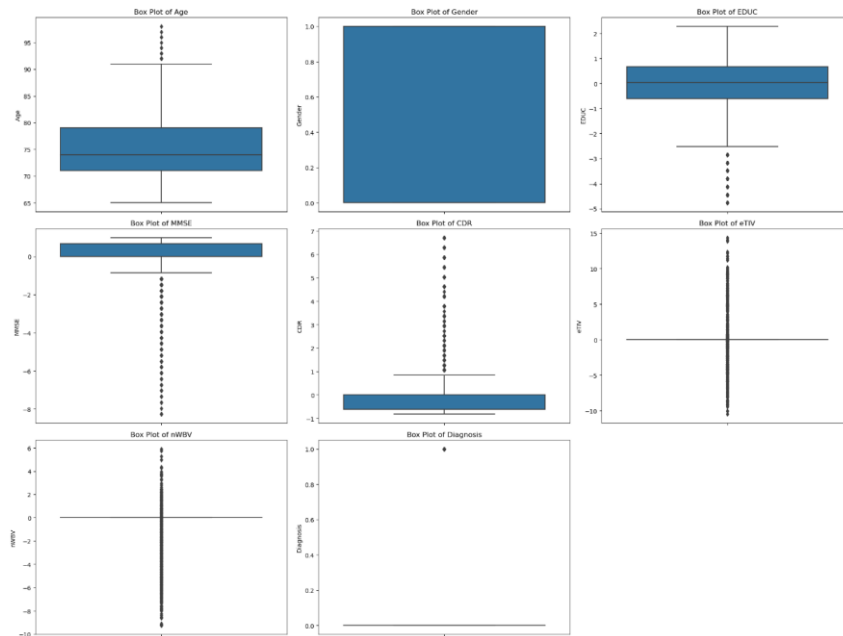


Figure H.2. Box plot of Various Features of The Group 2 Dataset.

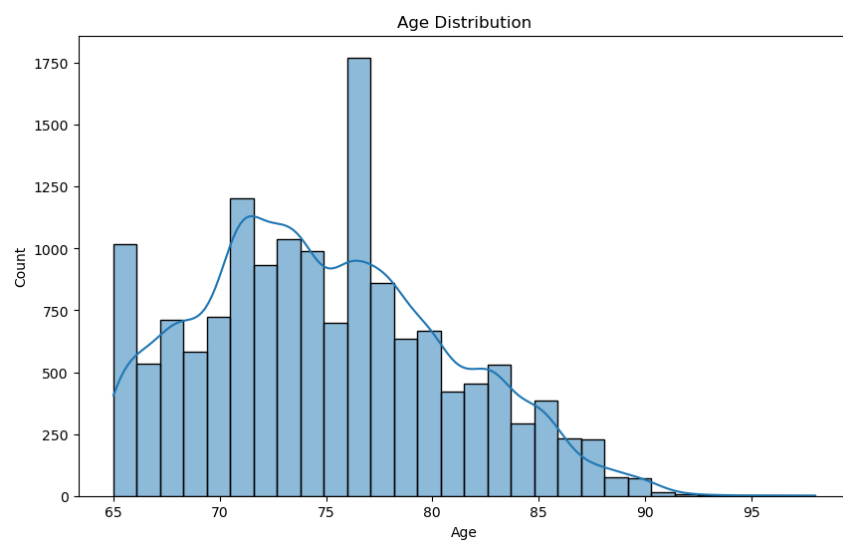


Figure H.3. Age Distribution of The Group 2 Dataset.

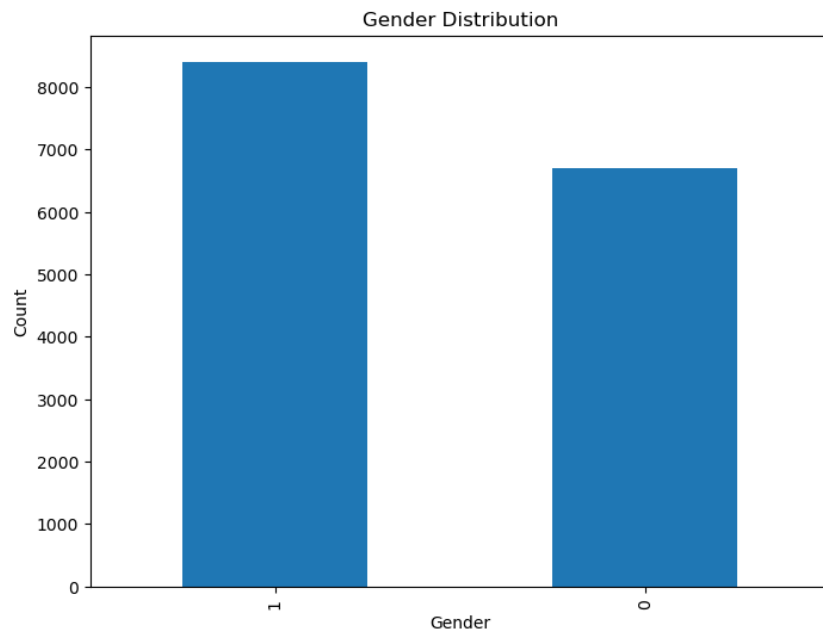


Figure H.4. Gender Distribution of The Group 2 Dataset.

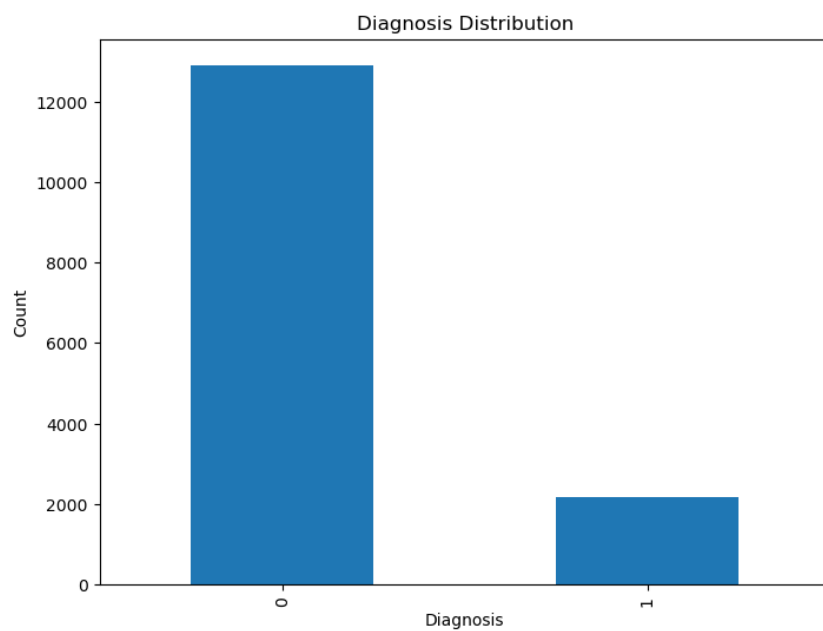


Figure H.5. Diagnosis Distribution of The Group 2 Dataset.

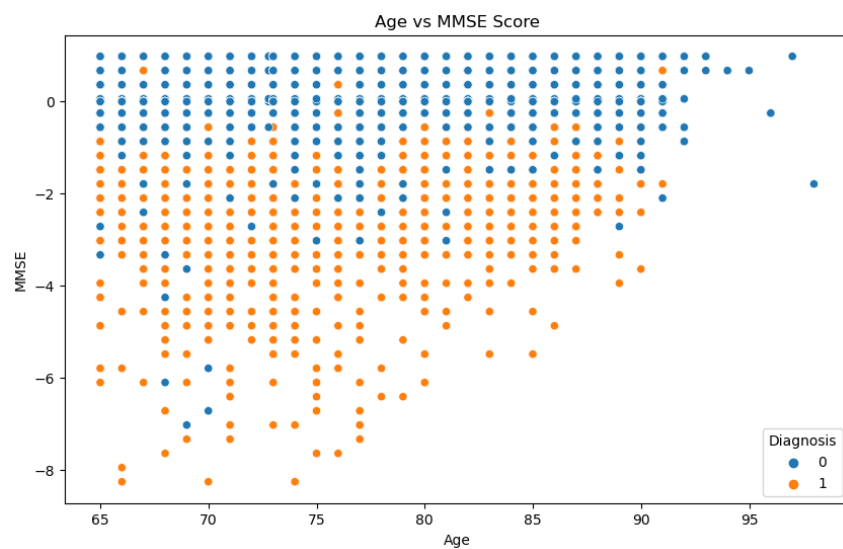


Figure H.6. Scatter Plot of Age vs MMSE Score of The Group 2 Dataset.

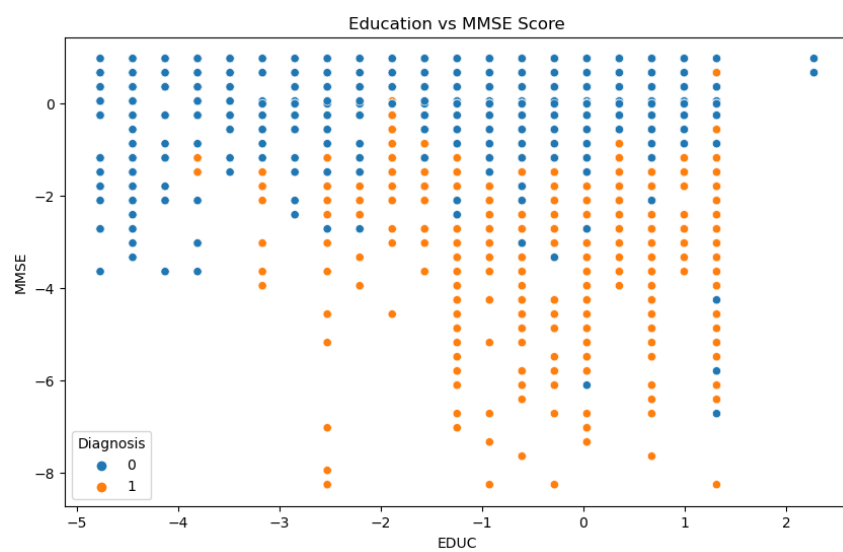


Figure H.7. Scatter Plot of Education vs MMSE Score of The Group 2 Dataset.

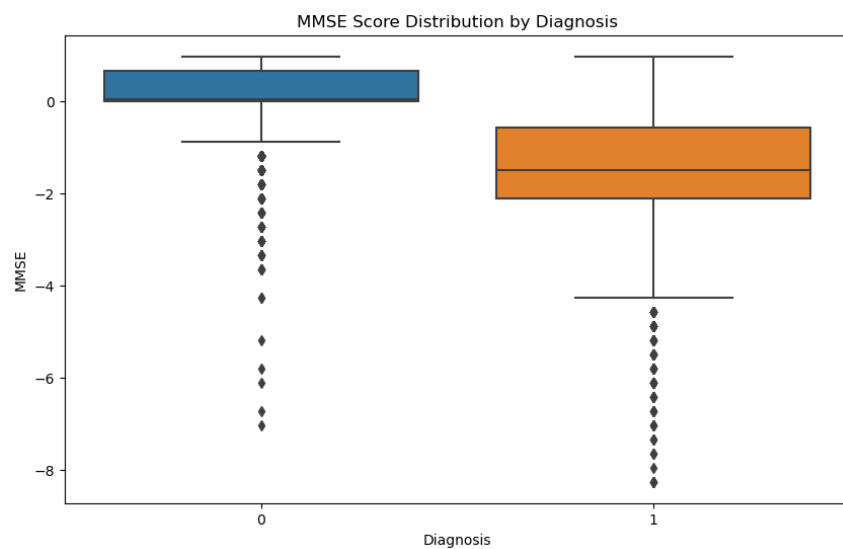


Figure H.8. Box plot of MMSE Score Distribution by Diagnosis of The Group 2 Dataset.

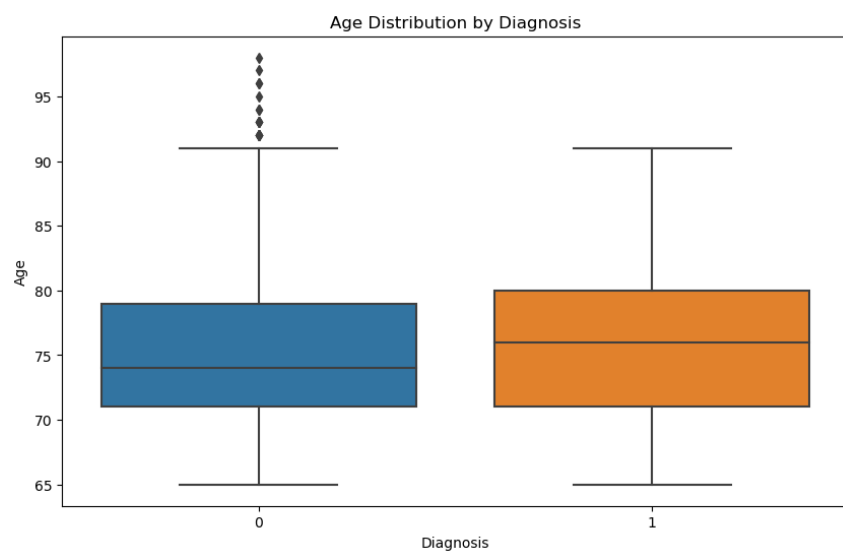


Figure H.9. Box plot of Age Distribution by Diagnosis of The Group 2 Dataset.

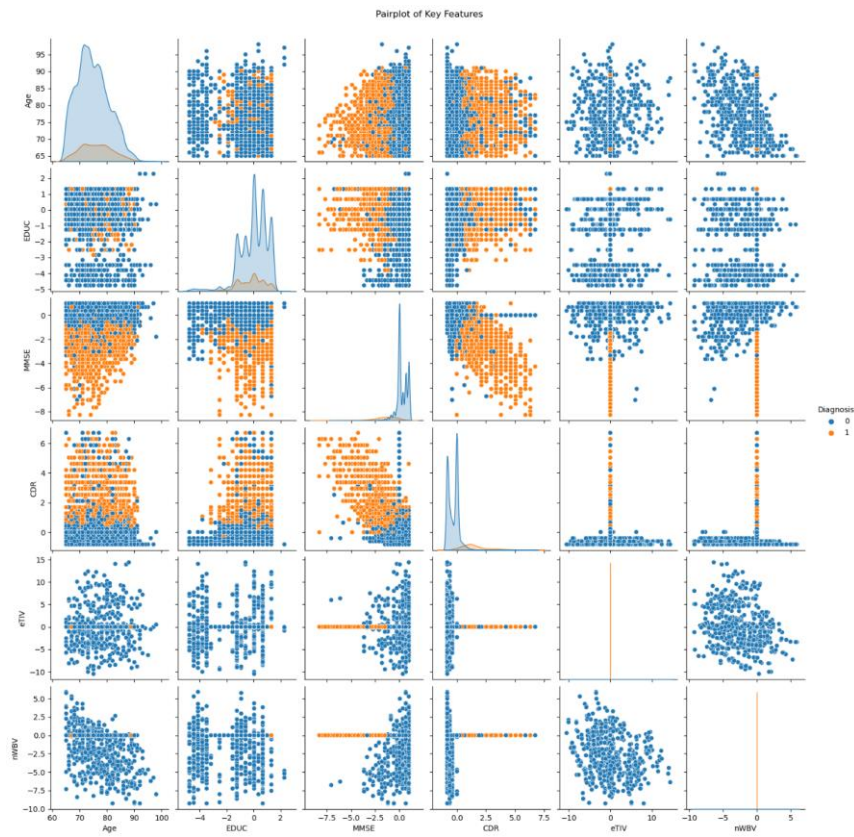


Figure H.10. Pair plot of Key Features of The Group 2 Dataset.

APPENDIX I Correlation Analysis of The Age Group 2 Dataset

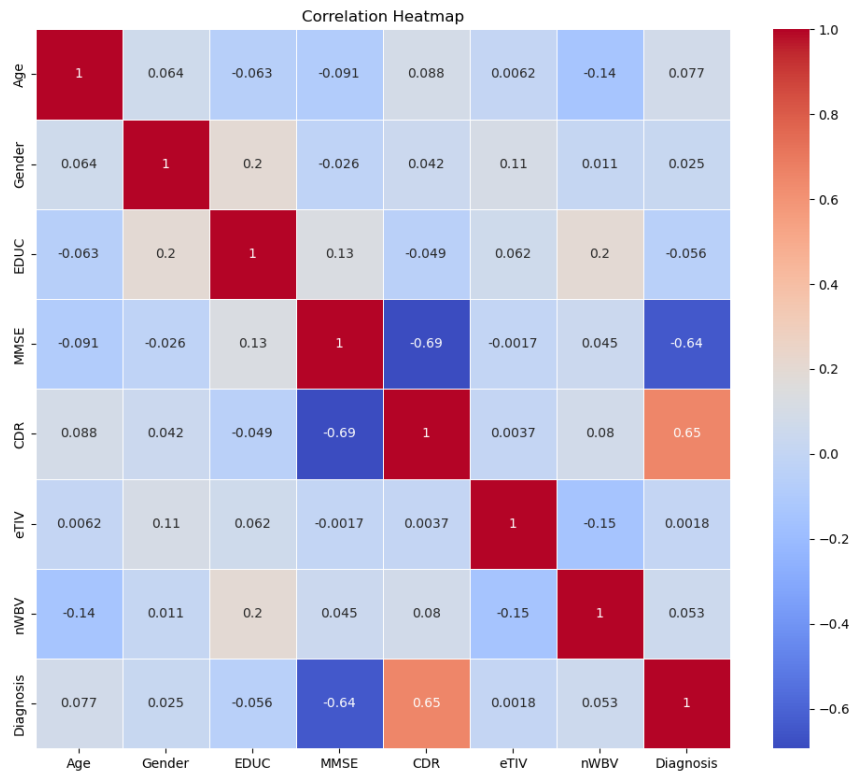


Figure I.1. Correlation Heatmap of the Group 2 Dataset.

Correlation heatmap represents the relationship between various features of the Group 2 dataset. Key points to observe:

1. **Colour Scale:** The heatmap uses a colour gradient from blue (negative correlation) to red (positive correlation), with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation).
2. **Key Findings:**
 - **MMSE and CDR** show a strong negative correlation ($-0.69 \sim -0.69$), indicating that as cognitive impairment increases (higher CDR), MMSE scores tend to decrease.

- **CDR** and **Diagnosis** have a strong positive correlation (0.65~0.65), suggesting that higher CDR scores are strongly associated with a dementia diagnosis.
 - **MMSE** and **Diagnosis** show a significant negative correlation (-0.64~ -0.64 -0.64), implying that lower MMSE scores are linked to a dementia diagnosis.
 - Other features, such as **Age**, **Gender**, **EDUC**, **eTIV**, and **nWBV**, show weaker correlations with **Diagnosis** and among themselves.
3. **Weak or No Correlations:** Many features, such as **Gender** and **Diagnosis** or **eTIV** and **Diagnosis**, show near-zero correlations, indicating little to no linear relationship.

This heatmap helps identify which features have the strongest relationships and may be critical for predicting outcomes, like dementia diagnosis, in the dataset.

APPENDIX J Confusion Matrix of Each Model on Merged Dataset

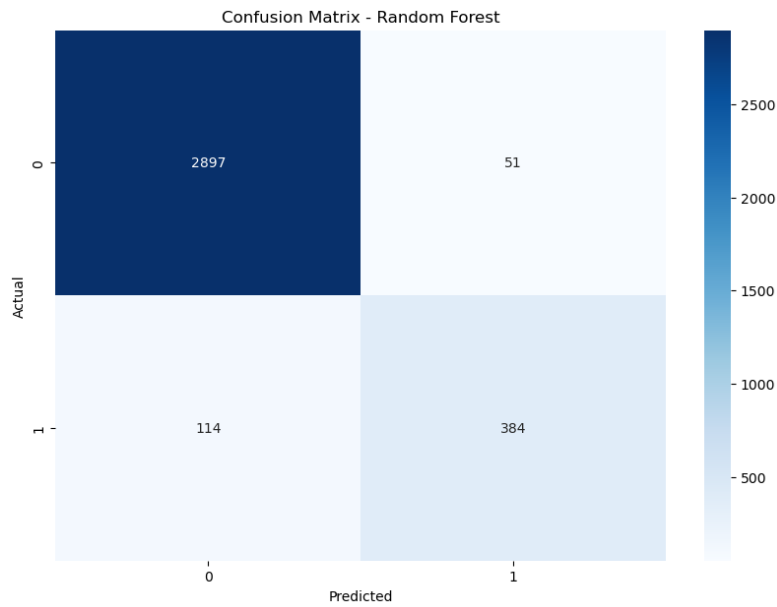


Figure J.1. Random Forest Confusion Matrix on Merged Dataset.

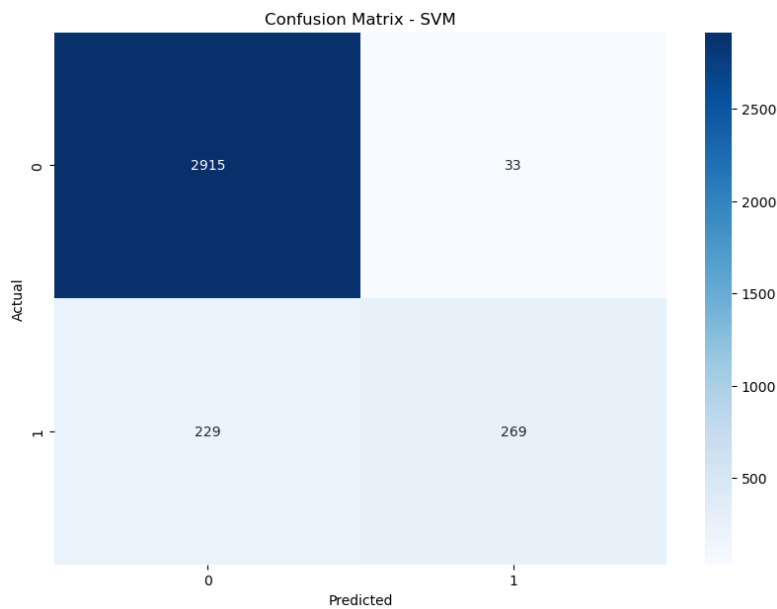


Figure J.2. SVM Confusion Matrix on Merged Dataset.

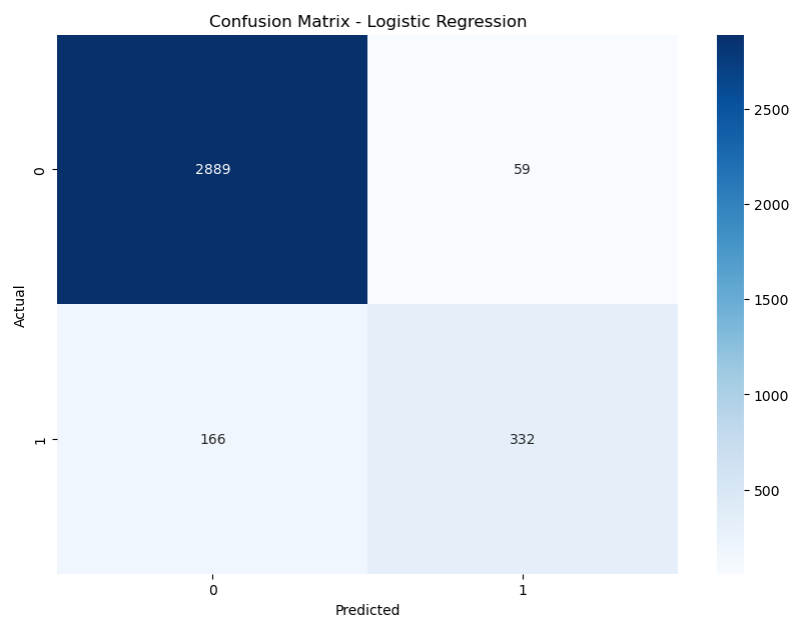


Figure J.3. Logistic Regression Confusion Matrix on Merged Dataset.

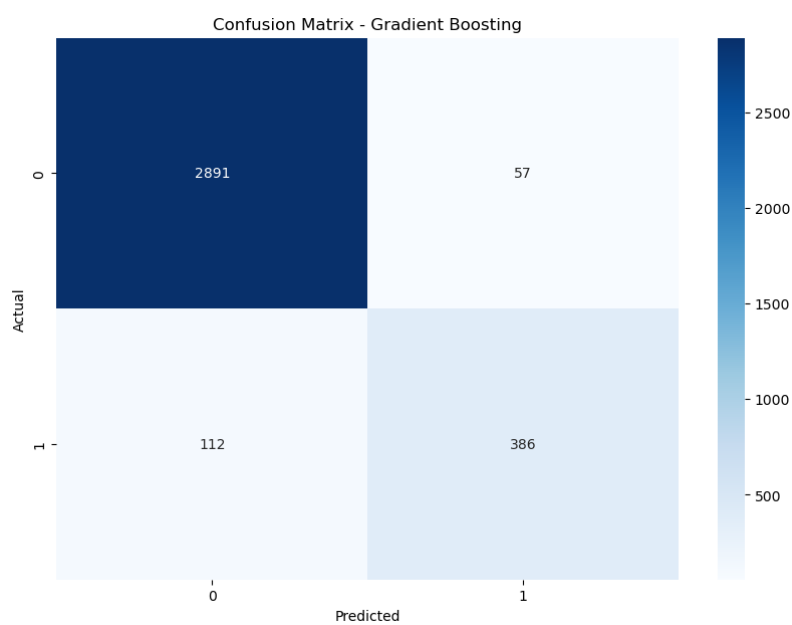


Figure J.4. Gradient Boosting Confusion Matrix on Merged Dataset.

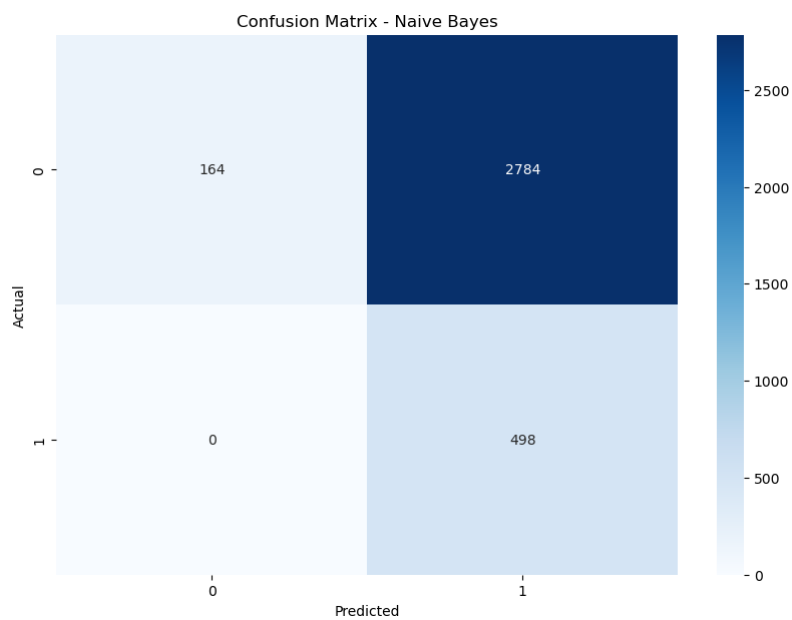


Figure J.5. Naïve Bayes Confusion Matrix on Merged Dataset.

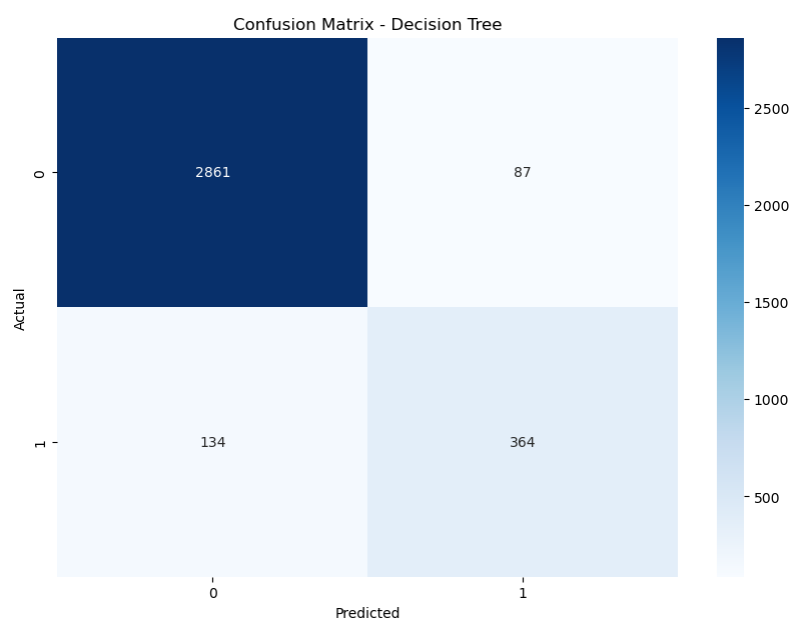


Figure J.6. Decision Tree Confusion Matrix on Merged Dataset.

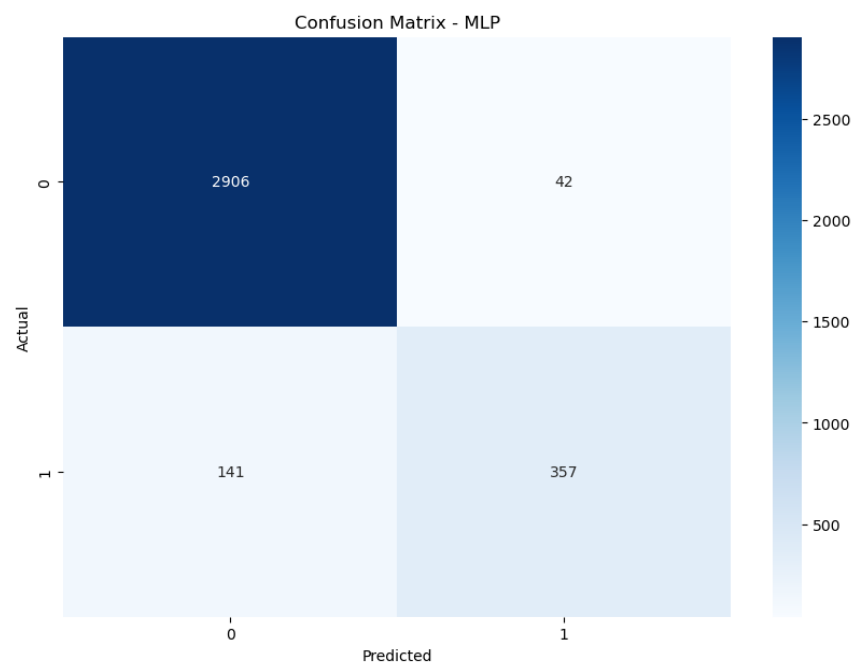


Figure J.7. MLP Confusion Matrix on Merged Dataset.

APPENDIX K Confusion Matrix of Each Model on Group 1 Dataset

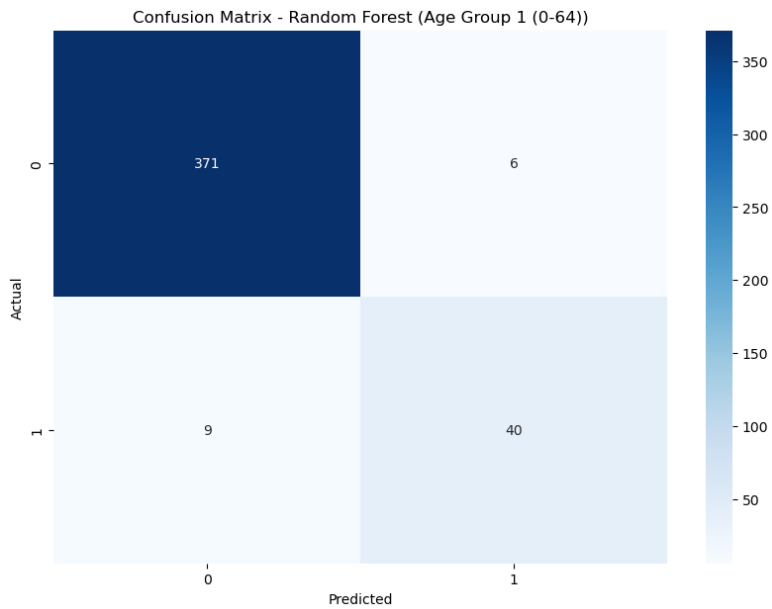


Figure K.1. Random Forest Confusion Matrix on Group 1 Dataset.

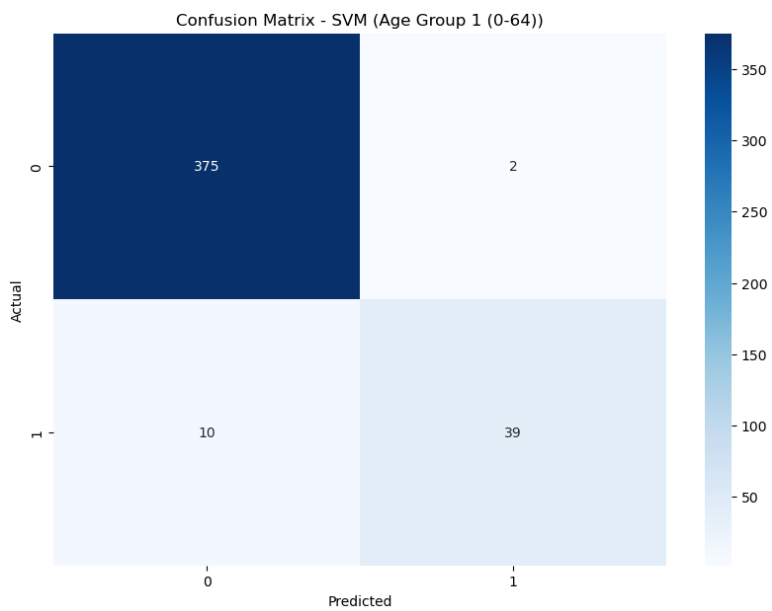


Figure K.2. SVM Confusion Matrix on Group 1 Dataset.

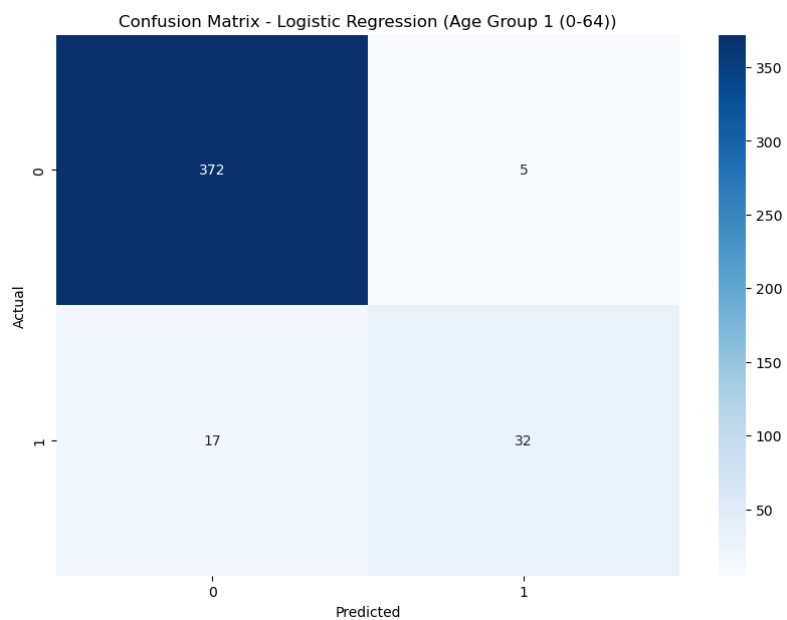


Figure K.3. Logistic Regression Confusion Matrix on Group 1 Dataset.

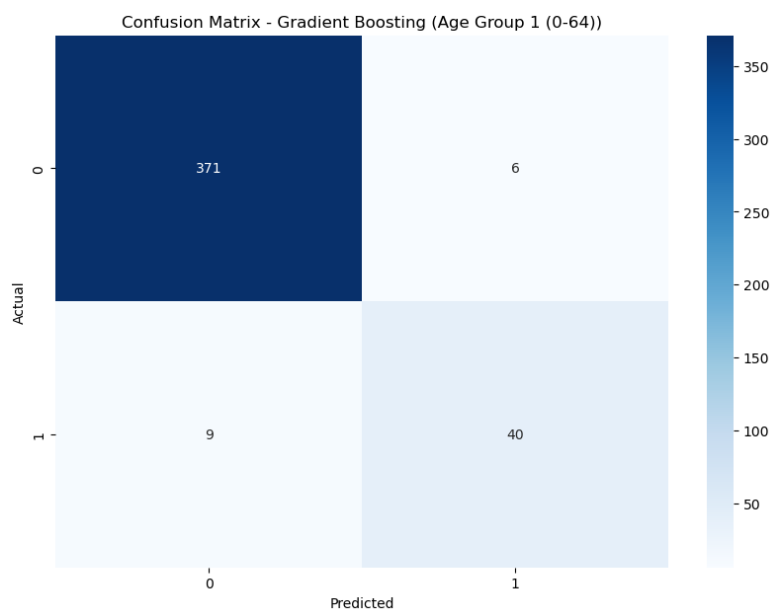


Figure K.4. Gradient Boosting Confusion Matrix on Group 1 Dataset.

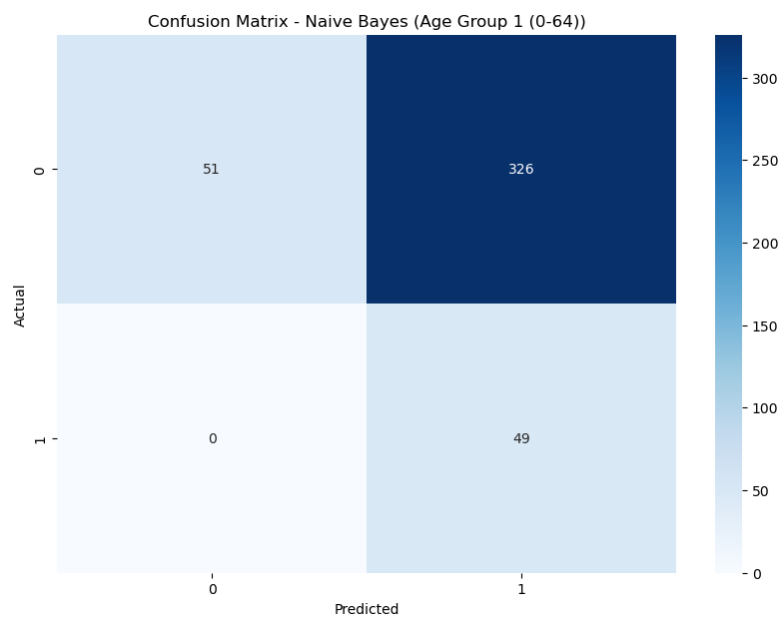


Figure K.5. Naïve Bayes Confusion Matrix on Group 1 Dataset.

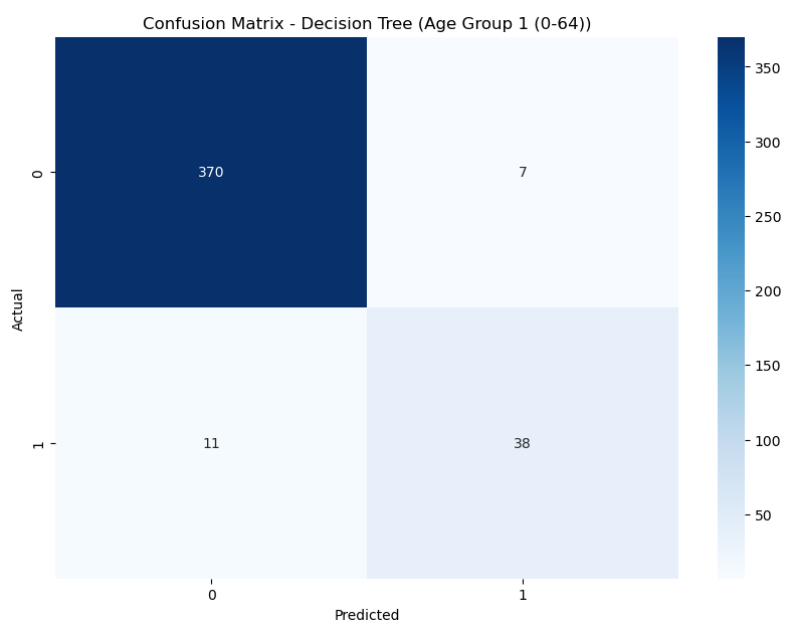


Figure K.6. Decision Tree Confusion Matrix on Group 1 Dataset.

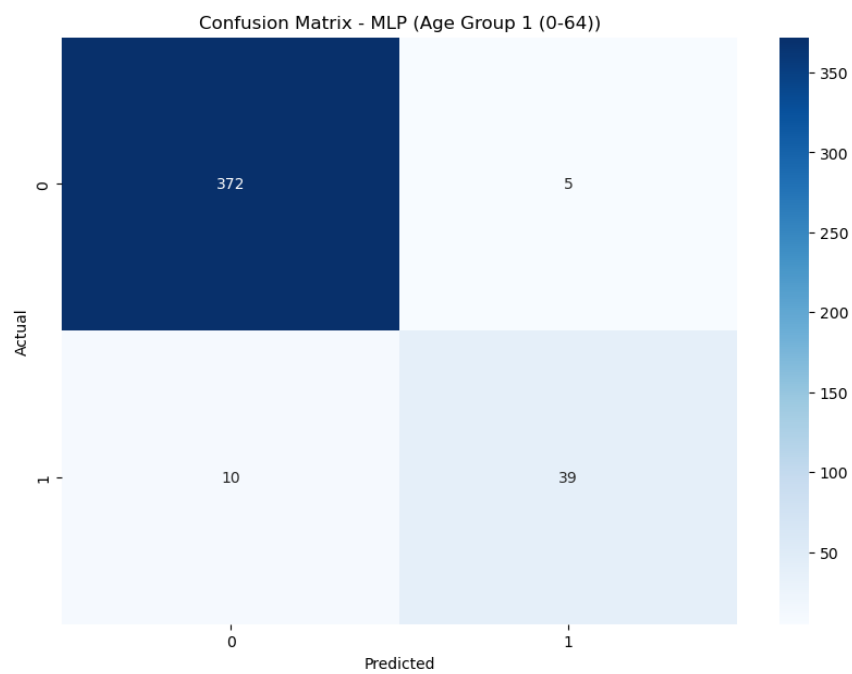


Figure K.7. MLP Confusion Matrix on Group 1 Dataset.

APPENDIX L Confusion Matrix of Each Model on Group 1 Dataset

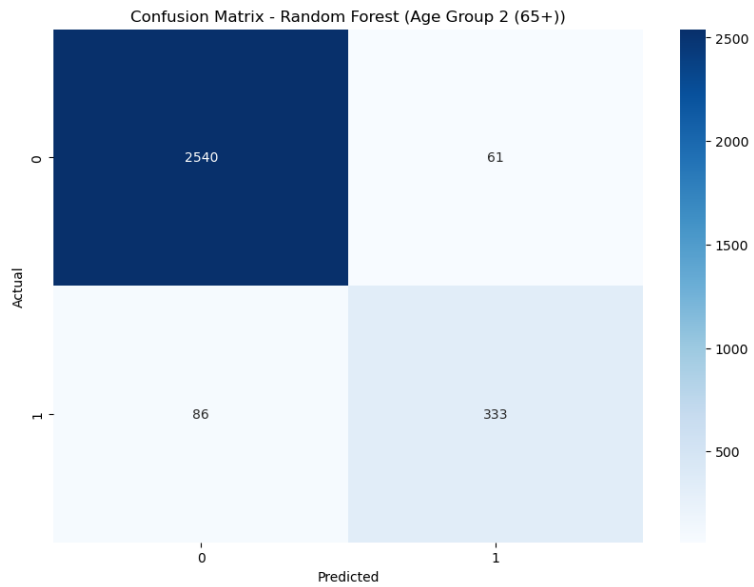


Figure L.1. Random Forest Confusion Matrix on Group 2 Dataset.

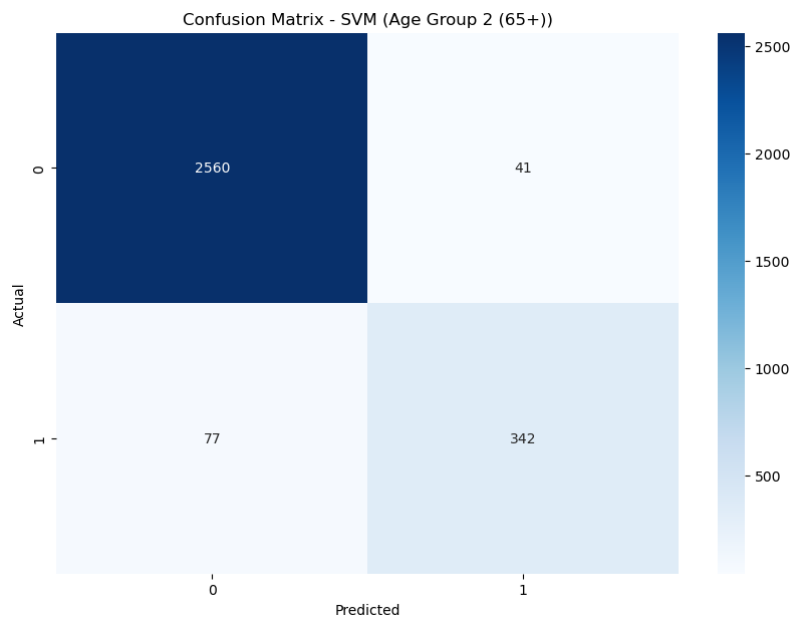


Figure L.2. SVM Confusion Matrix on Group 2 Dataset.

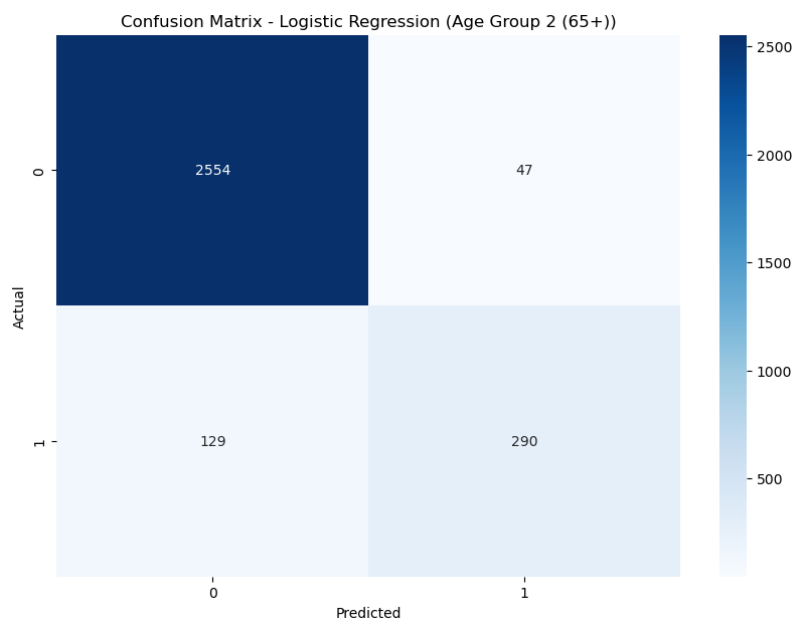


Figure L.3. Logistic Regression Confusion Matrix on Group 2 Dataset.

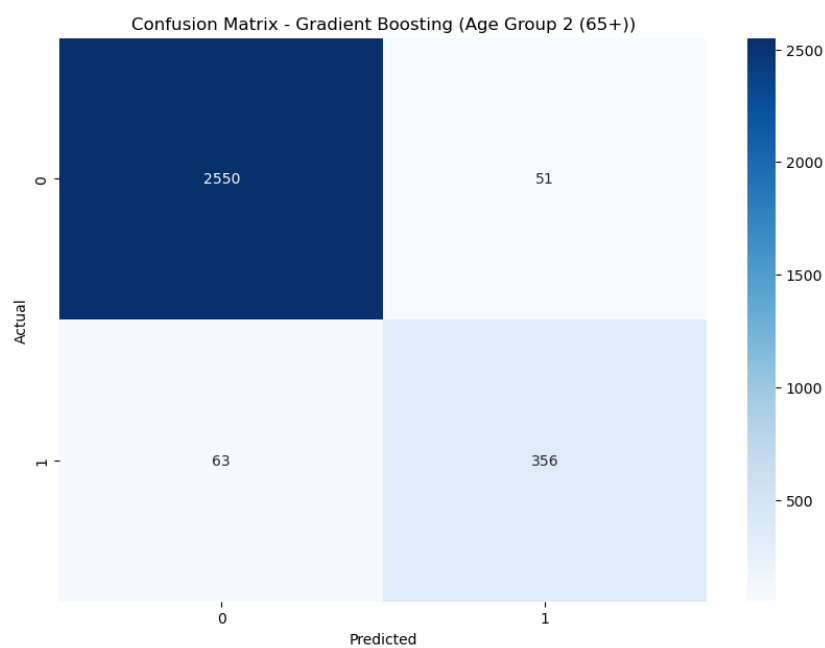


Figure L.4. Gradient Boosting Confusion Matrix on Group 2 Dataset.

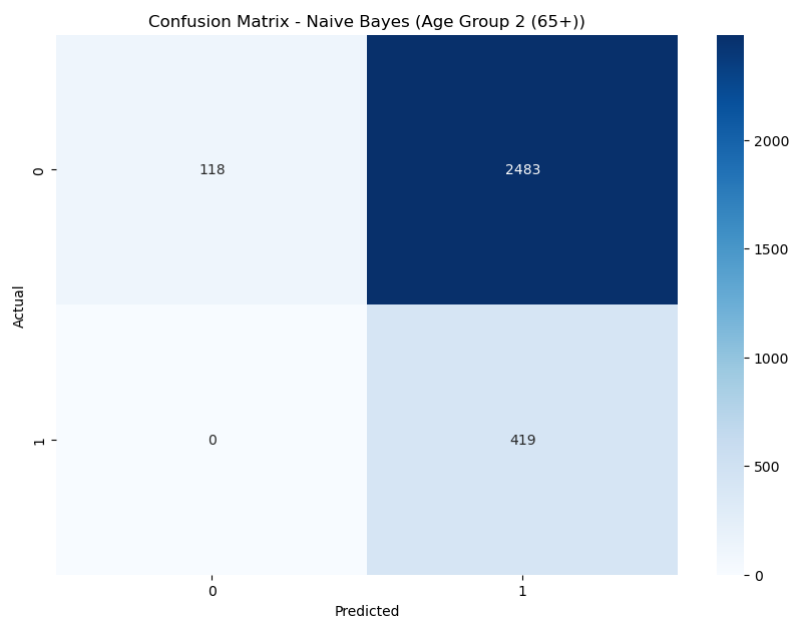


Figure L.5. Naïve Bayes Confusion Matrix on Group 2 Dataset.

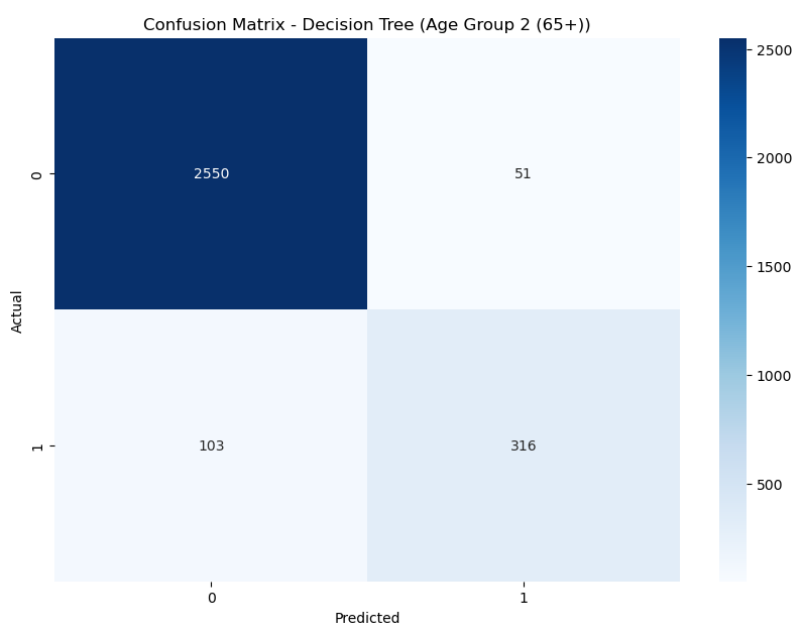


Figure L.6. Decision Tree Confusion Matrix on Group 2 Dataset.

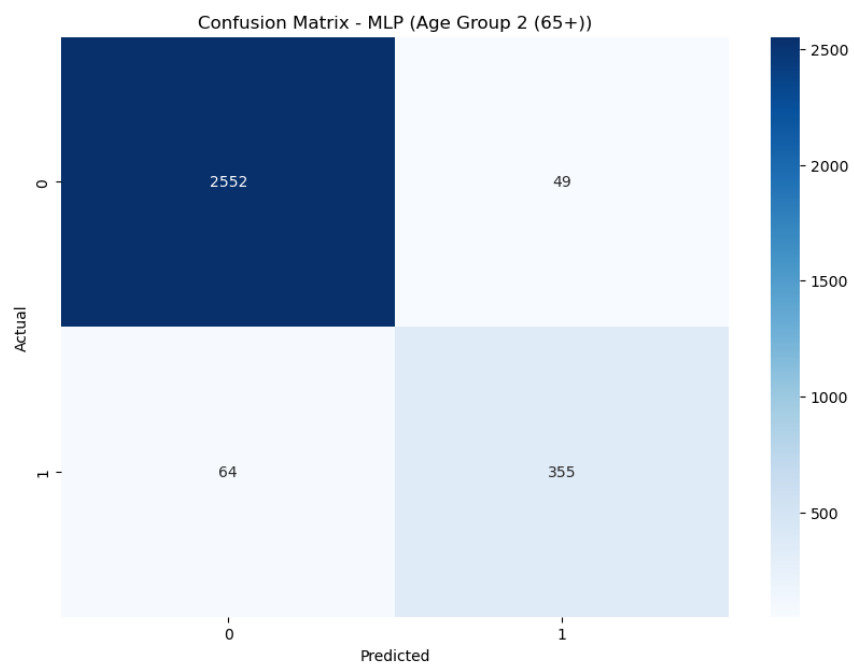


Figure L.7. MLP Confusion Matrix on Group 2 Dataset.

APPENDIX M EDA Code Snippet (Reusable Function)

```
# Exploratory Data Analysis Before Preprocessing
def perform_eda(data):
    print("Exploratory Data Analysis")
    print("=====")

    # Basic information about the dataset
    print("\n1. Dataset Information:")
    print(data.info())

    # Summary statistics
    print("\n2. Summary Statistics:")
    print(data.describe())

    # Missing values
    print("\n3. Missing Values:")
    missing_values = data.isnull().sum()
    print(missing_values[missing_values > 0])

    # Unique values in categorical columns
    print("\n4. Unique Values in Categorical Columns:")
    categorical_columns = data.select_dtypes(include=['object']).columns
    for col in categorical_columns:
        print(f"{col}: {data[col].nunique()} unique values")
        print(data[col].value_counts())
        print()

    # Distribution of numerical features
    numerical_columns = data.select_dtypes(include=[np.number]).columns
    plt.figure(figsize=(20, 15))
    for i, col in enumerate(numerical_columns, 1):
        plt.subplot(3, 3, i)
        sns.histplot(data[col], kde=True)
        plt.title(f'Distribution of {col}')
    plt.tight_layout()
    plt.show()

    # Box plots for numerical features
    plt.figure(figsize=(20, 15))
    for i, col in enumerate(numerical_columns, 1):
        plt.subplot(3, 3, i)
        sns.boxplot(y=data[col])
        plt.title(f'Box Plot of {col}')
    plt.tight_layout()
    plt.show()

    # Correlation heatmap
    plt.figure(figsize=(12, 10))
    sns.heatmap(data[numerical_columns].corr(), annot=True,
cmap='coolwarm', linewidths=0.5)
    plt.title('Correlation Heatmap')
    plt.show()

    # Age distribution
    plt.figure(figsize=(10, 6))
    sns.histplot(data['Age'], kde=True, bins=30)
    plt.title('Age Distribution')
```

```

plt.show()

# Gender distribution
plt.figure(figsize=(8, 6))
data['Gender'].value_counts().plot(kind='bar')
plt.title('Gender Distribution')
plt.ylabel('Count')
plt.show()

# Diagnosis distribution
plt.figure(figsize=(8, 6))
data['Diagnosis'].value_counts().plot(kind='bar')
plt.title('Diagnosis Distribution')
plt.ylabel('Count')
plt.show()

# Age vs MMSE scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='MMSE', hue='Diagnosis', data=data)
plt.title('Age vs MMSE Score')
plt.show()

# Education vs MMSE scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='EDUC', y='MMSE', hue='Diagnosis', data=data)
plt.title('Education vs MMSE Score')
plt.show()

# MMSE distribution by Diagnosis
plt.figure(figsize=(10, 6))
sns.boxplot(x='Diagnosis', y='MMSE', data=data)
plt.title('MMSE Score Distribution by Diagnosis')
plt.show()

# Age distribution by Diagnosis
plt.figure(figsize=(10, 6))
sns.boxplot(x='Diagnosis', y='Age', data=data)
plt.title('Age Distribution by Diagnosis')
plt.show()

# Pairplot for key features
sns.pairplot(data[['Age', 'EDUC', 'MMSE', 'CDR', 'eTIV', 'nWBV',
'Diagnosis']], hue='Diagnosis')
plt.suptitle('Pairplot of Key Features', y=1.02)
plt.show()

# Statistical tests
print("\n5. Statistical Tests:")
# T-test for Age between Dementia and Non-Dementia groups
dementia = data[data['Diagnosis'] == 1]['Age']
non_dementia = data[data['Diagnosis'] == 0]['Age']
t_stat, p_value = stats.ttest_ind(dementia, non_dementia)
print(f"T-test for Age between Dementia and Non-Dementia groups:")
print(f"T-statistic: {t_stat}, p-value: {p_value}")

# Chi-square test for Gender and Diagnosis
gender_diagnosis = pd.crosstab(data['Gender'], data['Diagnosis'])
chi2, p_value, dof, expected = stats.chi2_contingency(gender_diagnosis)
print(f"\nChi-square test for Gender and Diagnosis:")

```

```
print(f"Chi2 statistic: {chi2}, p-value: {p_value}")  
  
# Call the EDA function  
perform_eda(merged_data)
```

APPENDIX N Training & Evaluation Code (Merged Dataset)

```
# Function to train and evaluate models
def train_and_evaluate_merged_data(X, y):
    if len(X) == 0 or len(y) == 0:
        print("No data available for classification")
        return
    if len(y.unique()) < 2:
        print("Only one class present. Cannot perform classification.")
        return

    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

    classifiers = {
        'Random Forest': RandomForestClassifier(n_estimators=100,
random_state=42),
        'SVM': SVC(kernel='rbf', random_state=42),
        'Logistic Regression': LogisticRegression(random_state=42),
        'Gradient Boosting': GradientBoostingClassifier(random_state=42),
        'Naive Bayes': GaussianNB(),
        'Decision Tree': DecisionTreeClassifier(random_state=42),
        'MLP': MLPClassifier(random_state=42, max_iter=1000)
    }

    for name, clf in classifiers.items():
        print(f"\nTraining {name}...")
        clf.fit(X_train, y_train)
        y_pred = clf.predict(X_test)

        print(f"{name} Classification Report:")
        print(classification_report(y_test, y_pred))

        cm = confusion_matrix(y_test, y_pred)
        plt.figure(figsize=(10,7))
        sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
        plt.xlabel('Predicted')
        plt.ylabel('Actual')
        plt.title(f'Confusion Matrix - {name}')
        plt.show()

# Prepare data for classification
X = merged_data.drop(columns=['ID', 'Diagnosis'])
y = merged_data['Diagnosis']

# Perform classification on the entire merged dataset
print("\n--- Classification Results for Merged Dataset ---")
train_and_evaluate_merged_data(X, y)
```

APPENDIX O Training & Evaluation Code (Age Groups Dataset)

```
# Function to train and evaluate models
def train_and_evaluate(X, y, group_name):
    if len(X) == 0 or len(y) == 0:
        print(f"No data available for {group_name}")
        return
    if len(y.unique()) < 2:
        print(f"Only one class present in {group_name}. Cannot perform
classification.")
        return

    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

    classifiers = {
        'Random Forest': RandomForestClassifier(n_estimators=100,
random_state=42),
        'SVM': SVC(kernel='rbf', random_state=42),
        'Logistic Regression': LogisticRegression(random_state=42),
        'Gradient Boosting': GradientBoostingClassifier(random_state=42),
        'Naive Bayes': GaussianNB(),
        'Decision Tree': DecisionTreeClassifier(random_state=42),
        'MLP': MLPClassifier(random_state=42, max_iter=1000)
    }

    for name, clf in classifiers.items():
        print(f"\nTraining {name} for {group_name}...")
        clf.fit(X_train, y_train)
        y_pred = clf.predict(X_test)

        print(f"{name} Classification Report:")
        print(classification_report(y_test, y_pred))

        cm = confusion_matrix(y_test, y_pred)
        plt.figure(figsize=(10,7))
        sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
        plt.xlabel('Predicted')
        plt.ylabel('Actual')
        plt.title(f'Confusion Matrix - {name} ({group_name})')
        plt.show()

# Train and evaluate for each age group
for group, name in [(group1, "Age Group 1 (0-64)"), (group2, "Age Group 2
(65+)")]:
    print(f"\n\n--- Results for {name} ---")
    X = group.drop(columns=['ID', 'Diagnosis', 'Age'])
    y = group['Diagnosis']
    train_and_evaluate(X, y, name)
```

APPENDIX P The Project Management Gantt Chart

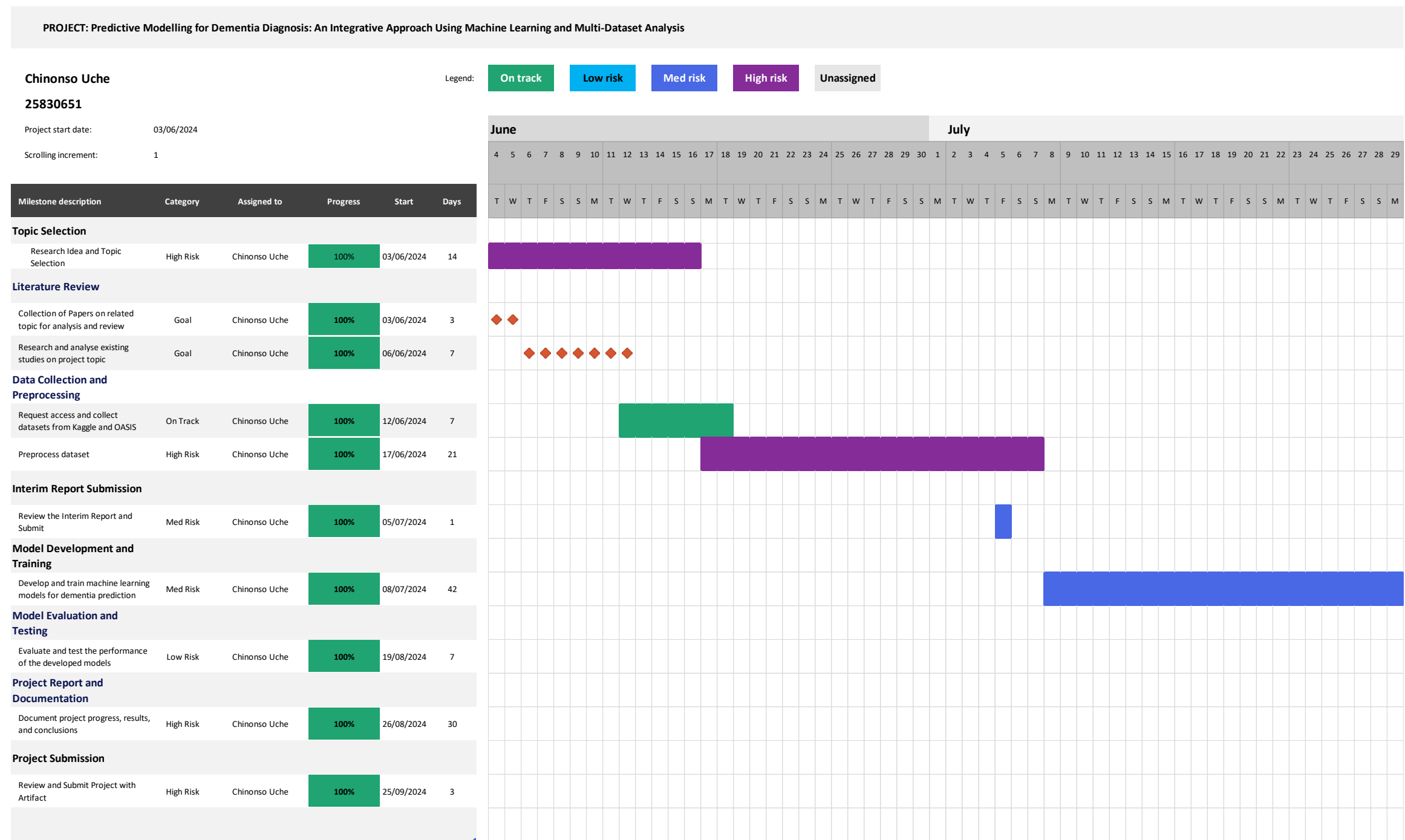


Figure P.1. Project Management Gantt Chart.

APPENDIX Q Project Poster

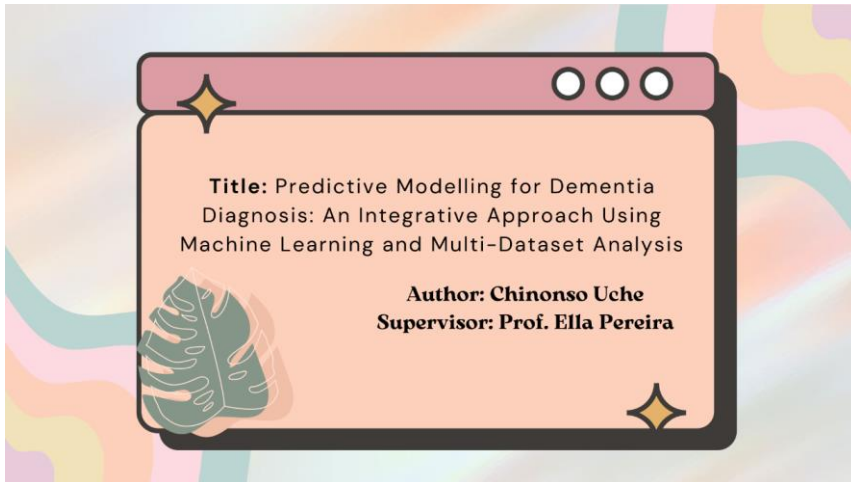


Figure Q.1. Poster Title.

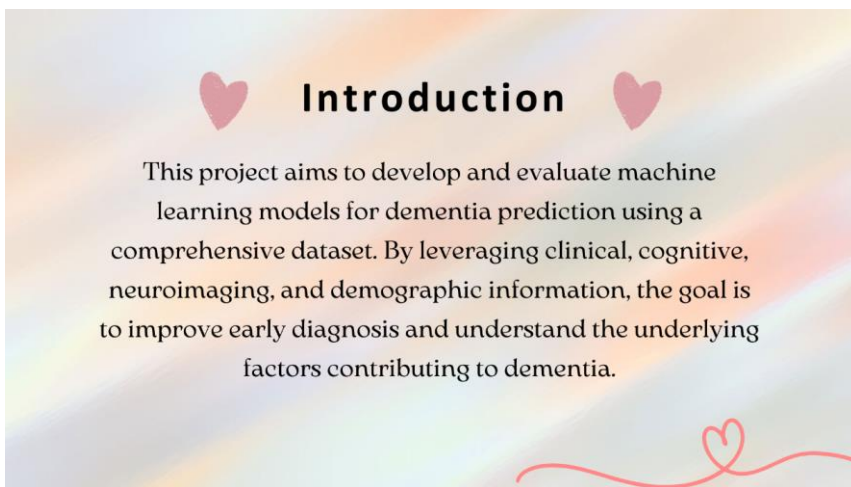


Figure Q.2. Poster Introduction.

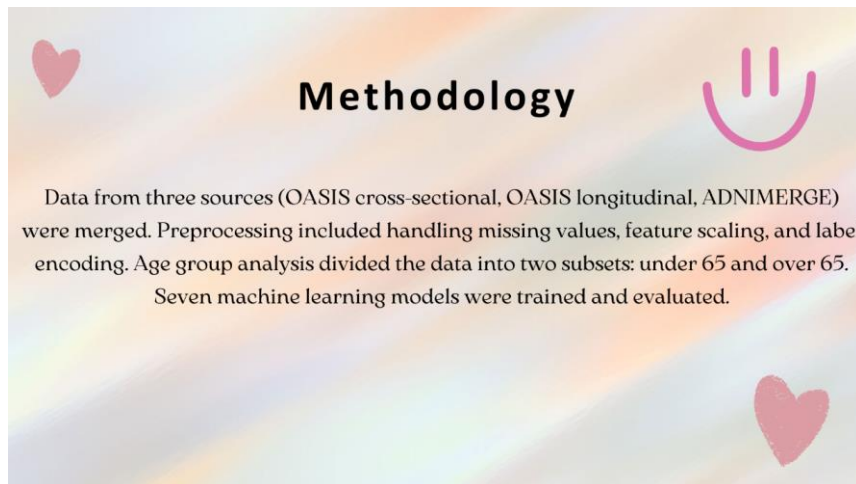


Figure Q.3. Poster Methodology.

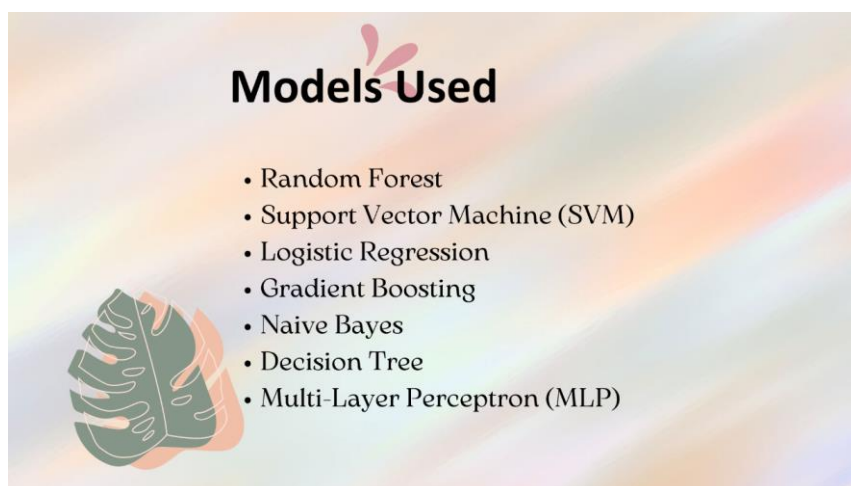


Figure Q.4. Poster Models Used.



Figure Q.5. Poster Result.

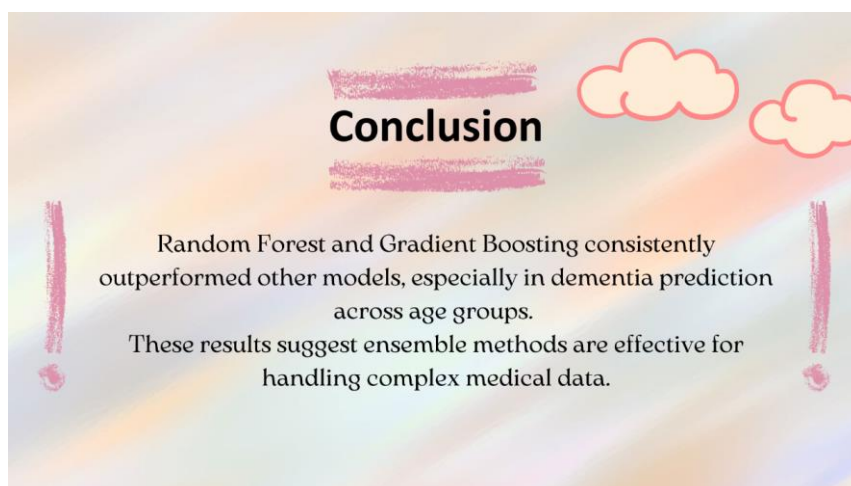


Figure Q.6. Poster Conclusion.

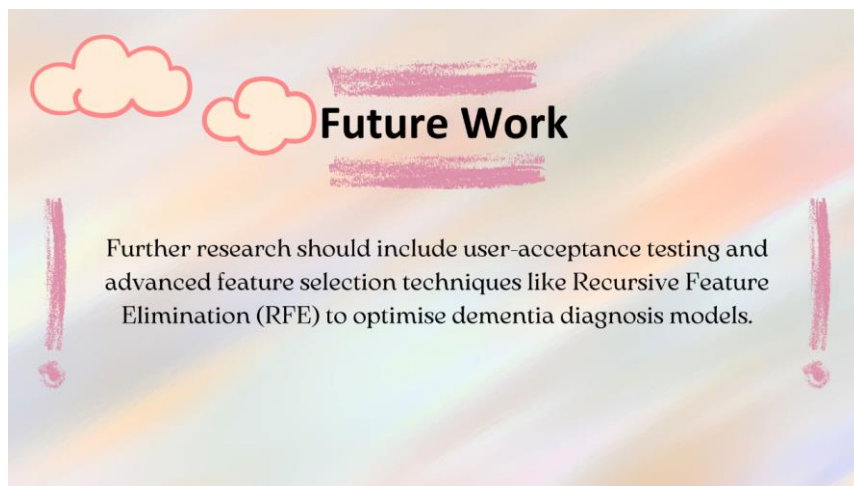


Figure Q.7. Poster Future Work.



Figure Q.8. Poster Greetings.

APPENDIX R The Signed Ethical Checklist

Ethical considerations

Your research project is unlikely to result in any major ethical issues, but please read what follows in order to acquaint yourself with some of the current thinking about the ethics of doing research. Not all of what follows may be applicable to your research.

This document **MUST** be read in conjunction with the Universities [Research Guidance](#)

Ethical considerations are paramount. They need to be fully discussed with your tutor before setting out on your project. Any concerns which are relevant to your investigation should be noted and will constitute an important aspect of your final discussion. For example, particular issues concerning the involvement of children or vulnerable adults need to be thought through carefully; likewise, issues around discussion of sensitive topics; issues around intrusion, if observation strategies are employed, and so on. All these matters need to be addressed before you set about collecting data. Evidence of in-depth reflection on ethical issues should be clear in your final report.

There are certain generally accepted guidelines governing research practice of which you must be aware:

1. Consent

As much information as possible should be provided to participants so that they can give - or withhold - their agreement to participate. Establishing consent is not always a straightforward business and requires careful and perceptive handling.

2. Deception

Intentional deception of participants about the purpose and general nature of the investigation should normally be avoided. If your project involves withholding any information from your participants, you must discuss this in detail with your tutor before going ahead.

3. Debriefing

At the end of the study you should give participants any further information needed to complete their understanding of the nature of the research, what you hope to do with it, and how it might affect them personally at any later date.

4. Rights to withdraw from the investigation

Participants have the right to drop out of the study at any time and this must be made clear to them from the outset. Remember that participants also have the right to withdraw consent retrospectively and require that their data be destroyed. They should be informed about this and enabled to take appropriate action.

5. Confidentiality and/or anonymity.

These issues should be fully discussed with prospective participants. You need to be very clear about how 'confidentiality' and 'anonymity' are different, and what guarantees you will - and conversely will not - be able to give to participants in terms of respecting either, or both, of these.

6. Protection of participants from physical and mental harm during the investigation.

This is essential. It may appear to be a rather extreme consideration in relation to the small-scale research project, but you must think carefully about any levels of stress or distress which participation might cause for your participants either during or after the research.

Sensitivity to Multicultural Issues

When constructing interview schedules, questionnaires or other data-gathering instruments, it is important to be sensitive to different cultural perspectives. Depending upon the nature and location of your research, the following issues should be considered:

1. Your own prejudices and biases.
2. Though we may strive to be objective, we all have our own, often unacknowledged, prejudices. They may relate to a person's age, ethnicity, gender, sexual orientation, religion, disability, or marital, parental or socio-economic status.
3. Sensitivity to the language used in the research process to describe the different groups involved: e.g. Are explicitly derogatory terms used in describing children or adults from 'other' groups? Are subtly derogatory terms used, such as 'these people' when describing participants in a project?
4. Is the language used in data-collection instruments accessible and understandable to all participants?
5. Are members/representatives of all groups of participants involved in planning, implementing, and reviewing results from the research?
6. Have multicultural issues been addressed openly at all stages of the research?
7. Does the group of participants represent the cultural diversity of the institution/area? What implications may this have for the research findings?
8. Could the results of the research be viewed differently by different cultural groups? What has been done to ensure that their perspectives have been included?

Stage 1 Self-Assessment

Part A


If your research involves human participants, are any of the following concerns relevant?	
Yes /No	1. The involvement of vulnerable participants or groups, such as children (under the age of 16), people with a learning disability or cognitive impairment, or persons in a dependent relationship?
Yes /No	2. The sensitivity of the research topic, e.g. the participants' sexual, political, or legal behaviour, or their experience of violence, abuse or exploitation?
Yes /No	3. The gender, ethnicity, language, or cultural status of participants?
Yes /No	4. The use of deception, trickery, or other procedures that may contravene participants' full or informed consent, without timely and appropriate debriefing, or activities that cause stress, humiliation, or anxiety, or the infliction of more than minimal pain?
Yes /No	5. Access to records of personal or other confidential information, including genetic or other biological information, concerning identifiable individuals without their knowledge or consent?
Yes /No	6. The use of intrusive interventions, such as the administration of drugs or other treatments, excessive physical exertion, or techniques such as hypnotherapy without the participants' knowledge or consent?
Yes/No	6. Research related to the NHS is strongly advised to seek advice from their supervisor before commencing the project

If you have answered 'Yes' to any of the questions then the project is considered to be of **high ethical risk** and may need to be approved by the Departments Ethics Committee.
Please discuss your project with your supervisor.

Otherwise, your project may be considered **low ethical risk**. Please sign below and submit your self-assessment document to your supervisor and upload it to BB.

Approval for Low Risk Research Projects

Part A

If your research involves human participants, are any of the following concerns relevant? I can confirm that : <i>(confirm you have read these)</i>	
<ul style="list-style-type: none"> - I have read the Edge Hill University Framework for Research Ethics https://www.edgehill.ac.uk/document/research-ethics-policy/ - I have read the Computing Departments Ethics Policy Document (see BB). - I agree to abide by their principles 	YES / NO YES / NO
Your Signature ⁱ	
Your name	CHINONSO UCHE
Date:	5 July 2024
Supervisor's Signature ⁱⁱ	
Supervisor's name	Prof. Ella Pereira
Date	