Leveraging Explainable AI for Dementia Classification: A Machine Learning Approach

RIJUL KUMAR, MANOJ T[a*], SUCHARITHA SHETTY, OMKAR PRABHU
Department of Computer Science and Engg., Manipal Institute of Technology
Manipal Academy of Higher Education, Manipal,
Karnataka, INDIA

*Corresponding Author
[a]ORCID : 0000-0002-5472-2418

Abstract: - In last two decades the growing life expectancy has spiralled the prevalence of neurogenerative disorder such as Dementia among elderly population posing a significant challenge to mental health globally. The dementia is generally characterized by progressive cognitive decline which it affects memory, thinking, behaviour, and the ability to perform everyday activities. Early diagnosis and intervention are crucial for improving patient outcomes and managing the societal burden of dementia. The primary aim of this study is to deploy various machine learning models such as Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest (RF). Extreme Gradient Boosting (XGBoost) for the classification of Dementia. The suitable features required for the classification is determined through efficient filter-based and wrapper-based feature selection techniques. The experimental evaluation of machine learning models is performed using standard metrics such as accuracy, precision, recall and F1 score. Furthermore, Explainable AI (XAI) techniques such as SHAP and LIME are employed to interpret the black-box nature of these models, offering transparency and insights into the contribution of individual features to the model predictions. The thorough evaluation of machine learning models exhibited that Random Forest outperforms other models with an accuracy of 96%, with CDR identified as a key predictor through XAI analysis.

Key-Words: - Dementia Classification, Explainable Artificial Intelligence, Health Informatics, Machine Learning, Neuro-Disorder, Random Forest

Received: March 11, 2024. Revised: November 13, 2024. Accepted: December 15, 2024. Published: January 29, 2025.

1. Introduction

Dementia is a sickness that can result from several illnesses that gradually harm the brain and kill nerve cells, impairing cognitive function (i.e., the capacity to think critically) more than would be predicted from the normal ageing process. The physical, psychological, social, and economic aspects of dementia affect not just the person with the disease but also their family, caretakers, and society at large. There is often a lack of awareness and understanding of dementia, resulting in stigmatization and barriers to diagnosis and care. Dementia, a condition characterized by cognitive decline, can be caused by various illnesses and injuries that affect the brain, either directly or indirectly. The most common form, accounting for 60–70% of cases, is Alzheimer's disease. Other types include dementia with Lewy bodies, marked by abnormal protein deposits in nerve cells, dementia caused by vascular disease, and conditions that exacerbate frontotemporal dementia, in volving the degeneration of the brain's frontal lobe. Clear differentiations among the different types of dementia are often lacking, and mixed forms are commonly found to coexist. Currently ranking as the seventh most common cause of death worldwide, dementia is also a major contributor to impairment and dependency in the elderly population [1]. Someone in the world develops dementia every 3 seconds. Currently dementia affects over 55 million people worldwide, with over 60% of them residing in low- and middle-income countries. The numbers will almost double every 20 years, reaching 78 million in 2030 and 139 million in 2050 [2]. There is a serious possibility that the number of dementia cases in India would rise. In India, the estimated prevalence of dementia among

persons 60 years of age and older is 7.4%. Approximately 8.8 million Indians who are older than 60 suffer from dementia. In both rural and urban regions, dementia is more common in women than in men. There is notable variance in the prevalence of dementia between states [3]. Clinical diagnosis of dementia is based on neurological tests, cognitive evaluations, and a thorough medical history collected from patients and their family members. Additional tests such as haematology, CT scans, and MRI scans are conducted to eliminate other potential causes of dementia. Neuropsychological tests are particularly important for identifying impairments in various cognitive functions. Despite the existence of several clinical measures for early dementia diagnosis, a significant degree of subjectivity persists. There is an urgent need to create more dependable diagnostic tools.

The precise identification of cognitive impairment is essential, not only for the individuals affected but also for advancing the medical field. The manual process of diagnosing cognitive impairment in clinical environments takes a lot of time and may involve a number of different pieces of evidence, such as reports from informed informants, laboratory study results, and data from neuropsychological tests. The practitioner's level of expertise influences the diagnostic' efficacy and precision. Classification and early dementia diagnosis will be substantially more challenging in a number of isolated locations with a shortage of trained workers. Machine learning embodies a sophisticated computational technology that enhances the analysis of medical data and autonomously derives diagnostic outcomes [4]. In the classification and diagnosis of dementia, machine learning has become a potent instrument that offers substantial improvements in early detection, accuracy, personalised medication, research insights, and resource optimisation. By analyzing large datasets comprising diverse sources of data such as medical records, imaging data (MRI, CT scans), and neuropsychological test results, machine learning algorithms can detect subtle patterns indicative of dementia at early stages. This early detection enables timely interventions and treatment planning, potentially slowing down disease progression and improving patient outcomes. Moreover, machine learning models integrate various data sources and learn complex patterns that might not be readily visible to human clinicians, leading to more accurate and reliable diagnostic predictions and reducing misdiagnosis rates.

While AI has made remarkable advancements and is increasingly integrated into daily life, many AI-powered systems function as "black boxes," providing limited transparency into how decisions are made. This opacity creates challenges, especially when understanding how AI systems arrive at critical conclusions, such as a medical diagnosis. In high-stakes applications like healthcare, it is crucial to understand the reasoning behind specific decisions. Explainable Artificial Intelligence (XAI) addresses this need by adding a layer of transparency to these models, promoting systems that clarify their internal workings and offering valuable insights into the factors influencing their outputs. By integrating explainability, AI becomes more trustworthy, accountable, and transparent, allowing clinicians and healthcare providers to engage with AI systems confidently. This is particularly important in dementia classification, where understanding how models weigh factors like cognitive test scores or brain volume metrics helps align AI decisions with medical expertise. XAI supports early interventions and improves patient outcomes by providing clear, comprehensible explanations, ensuring more ethical and informed clinical practice.

Therefore, in our study we have considered machine learning models like Random Forest (RF), Extreme Gradient Boost (XGBoost), K-Nearest Neighbours (KNN), Logistic Regression, Decision Tree and Support Vector Machine (SVM) to detect the presence of dementia as a binary classification problem. The models are trained on the features selected from the dataset using Recursive Feature Elimination (RFE) approach and assessed based on metrics like accuracy, recall, precision, and F1 score. Additionally, to ensure transparency and trust in these predictive models, we incorporated Explainable AI (XAI) techniques like SHAP

(Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). By utilizing XAI, we provide insights into how specific features, such as Clinical Dementia Rating (CDR) or brain volume measurements, contribute to the model's decisions, making the results more interpretable for clinicians and fostering greater confidence in AI-assisted healthcare. This interpretability is crucial in translating machine learning outcomes into actionable medical decisions.

The main contributions of this research article are provided below:

- We performed relevant features selection for dementia classification using a combination of filter-based and wrapper-based techniques.
- We trained and evaluated multiple machine learning models such as Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest (RF), and Extreme Gradient Boosting (XGBoost) for dementia classification.
- We interpreted predictions obtained from best performing models utilizing XAI techniques like SHAP and LIME providing insights into feature importance and model's decision-making.

The remainder of the paper is organized as follows: Section 2 reviews the related work, while Section 3 outlines the proposed methodology. Section 4 presents the results and offers a comprehensive discussion. Finally, Section 5 concludes the paper and highlights potential directions for future research.

2. Related Work

Deep learning (DL) and machine learning (ML) have made significant strides in a number of fields lately, providing answers to today's problems. Machine learning methodologies have demonstrated considerable potential, particularly within the medical and healthcare domains, as evidenced by the utilization of machine learning techniques for the diagnosis of Alzheimer's disease [5] or utilizing machine learning and deep learning architectures to identify dementia through speech patterns [6]. However, amidst the plethora of applications, dementia remains a pressing and complex degenerative disorder demanding urgent attention. Recognizing its severity, numerous researchers have earnestly engaged with the problem, striving to uncover effective solutions and interventions. Herzog et al., [7] utilized data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to propose a novel approach for analyzing structural changes in brain asymmetries. By employing supervised machine learning algorithms and convolutional neural networks, the study yielded promising outcomes. Specifically, they achieved accuracies of 92.5% and 75.0% for distinguishing between normal cognition and early or progressive dementia stages, and 93.0% and 90.5% for differentiating between normal cognition and Alzheimer's Disease. This innovative pipeline presents an economical solution for dementia classification and holds promise for assisting in the diagnosis and monitoring of various brain degenerative disorders characterized by similar asymmetry changes. Castellazzi et al., [8] conducted research on distinguishing vascular dementia (VD) from Alzheimer's disease (AD) and predicting the primary disease in patients with mixed VD-AD dementia profiles. The study included 60 subjects (33 AD, 27 VD) and utilized various regional metrics from resting-state fMRI and diffusion tensor imaging as input features for three machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Adaptive Neuro-Fuzzy inference system (ANFIS). ANFIS emerged as the most effective algorithm, achieving a classification accuracy of over 84% using a limited feature set. When applied to the mixed VD-AD group, ANFIS accurately predicted the prevalent disease in 77.33% of cases. Overall, the research demonstrated the robust discriminative ability of the proposed method in distinguishing between AD and VD profiles and highlighted its potential in aiding physicians' diagnostic

assessments of dementia patients with ambiguous clinical presentations.

James et al., [9] investigated the efficacy of machine learning algorithms in forecasting incident dementia within a 2-year period using data from 15,307 memory clinic attendees without dementia. Their study aimed to compare the performance of these algorithms against established predictive models. Impressively, the machine learning algorithms outperformed the existing models, achieving a minimum accuracy of 90% with only 6 variables. Moreover, the assessment of the area under the receiver operating characteristic curve resulted in a value of 0.89, indicating strong predictive capability. These results highlight the capability of machine learning algorithms as valuable aids in clinical decision-making, particularly for accurately predicting dementia risk over a 2-year timeframe. Zhu et al., [10] conducted a study aimed at addressing the challenging issue of reliable diagnosis in the early stages of dementia. The study focused on creating and validating a novel machine learning-driven approach to aid in the initial diagnosis of normal cognitive function, mild cognitive impairment (MCI), very mild dementia (VMD), and dementia through the utilization of an informant-based questionnaire. A cohort of 5,272 participants took part in the research, completing a questionnaire comprising 37 items. Three distinct feature selection techniques were utilized to pinpoint the most significant features, with Information Gain emerging as the most effective method. Following this, the prominent features, in conjunction with six classification algorithms, were employed to construct diagnostic models. Among the various classification models examined, the Naive Bayes algorithm exhibited superior performance, attaining an accuracy of 0.81, precision of 0.82, recall of 0.81, and an F-measure of 0.81. The results indicate that the proposed diagnostic model serves as a robust tool for clinicians in diagnosing the early stages of dementia, providing valuable support for timely intervention and treatment. Salem et al., [11] in their study, investigated dementia diagnosis by employing the 10/66 one stage dementia diagnostic algorithm, utilizing data collected from three community-based surveys conducted in Lebanon. They tackled dataset imbalance by implementing oversampling and undersampling techniques, along with employing cost- sensitive methods to mitigate training bias. Utilizing three times repeated, 10-fold stratified cross-validation, they fine- tuned model hyperparameters, including incorporation cost and oversampling/undersampling percentages. Their results unveiled the balanced random forest as the most resilient probabilistic model, utilizing a mere 20 features and attaining an F2 score of 0.82, a G-Mean of 0.88, and a ROC AUC of 0.88. Furthermore, the Calibrated Weighted SVM emerged as the top classification model with similar features, obtaining an F2-score of 0.74 and a ROC AUC of 0.80. Mirzaei et al., [12] utilized the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, a publicly available resource designed to assess the feasibility of combining various imaging modalities, clinical evaluations, and neuropsychological assessments for measuring the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Their study focused on leveraging deep learning approaches, specifically convolutional neural networks (CNN) and recurrent neural networks (RNN), which have demonstrated remarkable accuracy without the need for extensive feature selection. Results indicated impressive accuracies of up to 96% for AD classification and 84.2% for MCI. Notably, these deep learning methods outperformed traditional machine learning techniques. The results emphasize the capacity of deep learning models to analyze neuroimaging data for diagnosing Alzheimer's disease (AD) and mild cognitive impairment (MCI), presenting a hopeful pathway for improving diagnostic precision in clinical environments.

Shahzad et al., [13] investigated the potential of instrumented gait assessment in home environments for the detection of Mild Cognitive Impairment (MCI), which serves as an early indicator of dementia. They gathered data from thirty individuals with mild cognitive impairment (MCI) and thirty cognitively normal (CN) sub- jects, utilizing shank-mounted inertial sensors in both standard and dual-task walking scenarios. The study evaluated various gait biomarkers derived from the

sensor signals and assessed their predictive power for MCI screening. Statistical analysis revealed significant differences in gait parameters between MCI and CN subjects, particularly under dual-task conditions. Through the utilization of machine learning models and feature selection techniques, the study developed a model to preliminarily screen for mild cognitive impairment (MCI) by utilizing gait biomarkers extracted from inertial sensors. The model achieved an accuracy of 71.67% and a sensitivity of 83.33%. These findings suggest the potential of gait assessment as a non-invasive and early screening tool for MCI, enabling timely intervention and treatment to mitigate dementia progression. Hane et al. [14] studied the utilization of deidentified clinical notes from different hospital systems collected over a decade to improve retrospective machine learning models focused on predicting the risk of Alzheimer's disease and related dementias (ADRD). They utilized two years of data to forecast the onset of ADRD and transformed clinical notes into a 100- dimensional vector space to identify clusters of associated terms and sentiments. The findings demonstrated that the inclusion of clinical notes substantially enhanced the area under the curve (AUC), increasing it from 0.85 to 0.94, and improved the Positive Predictive Value (PPV), which rose from 45.07% to 68.32% at the time of disease onset. Models incorporating clinical notes exhibited improved AUC and PPV in years 3-6, aligning with increased note volume, although outcomes in years 7 and 8 with smaller cohorts yielded mixed results. These findings highlight the potential of deidentified clinical notes, collected via natural language processing across multiple hospital systems, to enhance the accuracy of risk models for ADRD prediction, offering valuable insights into early detection and improved patient care. Aschwanden et al. [15] conducted a study utilizing data from the Health and Retirement Study, which included nearly 10,000 participants aged 50 to 98 years, to evaluate 52 possible indicators of cognitive decline and dementia. By integrating machine learning techniques with semi-parametric survival analysis, they discovered that African American individuals and those experiencing heightened emotional distress were at an increased risk for cognitive decline and dementia. Furthermore, sociodemographic variables such as lower educational levels and Hispanic ethnicity, along with health-related factors like declining subjective health and elevated BMI, were identified as significant indicators of cognitive decline. Surprisingly, cardiovascular factors and polygenic scores were found to have less predictive value than initially anticipated. The study concluded that broader factors like emotional distress and subjective health exerted greater influence than specific clinical and behavioral indicators. It emphasized the necessity of interdisciplinary collaborations and diverse methodological approaches to gain deeper insights into the intricate mechanisms underlying dementia and to effectively tackle this critical global health challenge. Jin et al., [16] developed a machine learning model using data from the Harmonized Diagnostic Assessment of Dementia for the Longitudinal Aging Study in India (LASI-DAD), a nationally representative study on late-life cognition and dementia in India involving 4,096 participants. From a subsample of 2,528 respondents, clinicians provided clinical consensus diagnoses of dementia. The study utilized comprehensive data from LASI-DAD, encompassing sociodemographic details, medical histories, cognitive screening outcomes, and informant interviews. Using a two-step process, numerous machine learning models underwent training and evaluation utilizing diverse metrics such as area under the receiver operating curve (AUROC), accuracy, sensitivity, specificity, precision, F1 score, and kappa statistic. Among these models, the support vector machine demonstrated the highest sensitivity (0.81), F1 score (0.72), and kappa (0.70), signifying substantial agreement and leading to its selection as the primary model. Upon application to individuals lacking a clinical consensus diagnosis, this model predicted a dementia prevalence of 7.4%. The research findings revealed that the chosen machine learning model exhibited exceptional discriminatory capability and substantial alignment with clinically accepted diagnoses, highlighting its potential as a decision support tool for dementia diagnosis.

Machine learning algorithms, especially deep learning models, often lack transparency, which can

significantly impede trust in healthcare applications like dementia diagnosis. In critical healthcare settings, the absence of explainability in AI-driven decisions can lead to skepticism and hinder the adoption of these systems. Additionally, biases embedded in these algorithms can exacerbate inequalities, particularly affecting vulnerable populations. The need for AI systems in healthcare to be both accurate and interpretable is paramount for ensuring trust and fairness in diagnosis and treatment recommendations (Rai, 2020 , [17]). Lombardi et al. [18] established a comprehensive framework to categorize individuals into three groups: healthy controls, those exhibiting cognitive impairment, and individuals diagnosed with dementia, utilizing a range of cognitive metrics. Their research further examined the variability of SHAP (SHapley Additive exPlanations) values associated with the choices made by predictive models. They illustrated that SHAP values could effectively depict the influence of each cognitive metric on a patient's cognitive condition. Additionally, longitudinal SHAP value analyses provided valuable insights into the progression of Alzheimer's disease, identifying cognitive indexes most relevant for tracking neurodegeneration. This work highlights the potential of explainability indexes as markers to describe changes in cognitive status during both normal and pathological aging. These indexes offer a means to quantify the contribution of individual cognitive domains to overall patient health, thus enabling personalized neurodegeneration patterns and optimizing predictors for Alzheimer's classification (Lombardi et al., [18]). In recent studies, the application of machine learning algorithms has shown promise in the early diagnosis of mental health conditions, including schizophrenia, which has parallels in dementia diagnostics. Shivaprasad et al., [19] highlight the effectiveness of various classifiers, including Logistic Regression, SVM with linear kernel, and Ridge, in predicting schizophrenia with high accuracy rates ranging from 83% to 86%. Their research emphasizes the importance of feature selection and correlation analysis, demonstrating that specific attributes, such as age and sex, significantly contribute to predictive outcomes. Moreover, they employed five Explainable AI techniques—SHAP, LIME, QLattice, ELI5, and Anchor—to elucidate the decision-making processes behind their models. For instance, SHAP values were used to quantify feature importance, revealing critical insights into patient characteristics influencing predictions. This approach can inform the development of XAI methodologies for dementia, where understanding model behavior is crucial for clinical decision-making. The integration of such explainable frameworks can enhance trust and transparency in AI systems utilized within healthcare settings, ultimately aiding in faster and more accurate diagnoses of cognitive disorders like dementia (Shivaprasad et al., [19]). Bogdanovic, B., Eftimov, T. & Simjanoska, M., [20] explored recent advancements in understanding Alzheimer's disease (AD) through the application of Explainable Machine Learning (ML) methods, particularly the SHAP (SHapley Additive exPlanations) framework, which has proven instrumental in enhancing model interpretability. SHAP enables both global and local interpretability, allowing researchers to ascertain how different features influence model predictions. A comprehensive dataset comprising 12,741 individuals was utilized to test hypotheses regarding the causes and indicators of AD. Findings revealed that cognitive assessments, especially the Clinical Dementia Rating Scale Sum of Boxes (CDRSB), have a significant impact on diagnosis predictions, with higher CDRSB values correlating with more severe diagnoses. Additionally, the study highlighted the complex interplay of various features, such as education levels and MRI indicators, in diagnosing AD. Notably, it was found that AD cannot be attributed solely to genetic factors, age, or gender, indicating its multifactorial nature. These insights underscore the potential of explainable ML methods to provide valuable guidance for future research and clinical practices in AD, ultimately aiding in timely diagnoses and interventions.

3. Methodology

3.1 Dataset Description

The dementia dataset used in this study is sourced from Kaggle and includes records from 373 patients, divided into three categories: demented, non-demented, and individuals who transition to a demented state due to delays or reluctance in seeking timely medical intervention. In the dataset, "demented" denotes individuals diagnosed with dementia, a degenerative condition marked by cognitive deterioration and memory impairment, while "non-demented" signifies subjects exhibiting standard cognitive function. The term "converted" pertains to patients initially identified as non-demented but later diagnosed with dementia. In this study, "converted" patients are classified as "demented." The dataset includes various attributes recorded for each patient, such as Subject ID, MRI ID, Group, Visit, MR Delay, gender (M/F), handedness (Hand), age, education level (EDUC), socioeconomic status (SES), Mini-Mental State Examination (MMSE) score, Clinical Dementia Rating (CDR), estimated total intracranial volume (eTIV), normalized whole-brain volume (nWBV), and atlas scaling factor (ASF). Each entry provides invaluable insights into the demographic and clinical profiles of the subjects under investigation.

3.2 Data Preprocessing

This section focuses on describing the data preprocessing steps performed on the dementia dataset considered in the present study. Initially, the exploratory data analysis is conducted by identifying missing values, outliers and checking for skewness. Missing values were found in two features in the dementia dataset namely ‘SES’, ’MMSE’. The percentage of missing values in ‘MMSE’ feature is insignificant (0.54%) and hence the rows corresponding to missing values are dropped. For missing values in ‘SES’(5.09%),the K-Nearest Neighbours (KNN) imputation technique is used. KNN imputation operates on the premise that similar data points should exhibit similar values. Outliers were identified across all features in the dataset. IQR method, Z-Score method and Percentile method are used based on the distribution of the respected features as detailed in Table I. For the features in which outliers were detected, capping was performed. In process of capping we set a maximum and minimum threshold of a variable to limit outliers and the outliers detected are set to minimum or maximum limit based on their values. Subsequently, features having skewness are detected and transformed using Box Cox, Yeo Johnson and Log Transform to avoid bias in the further usage of features.

TABLE I: Outliers Detected in Features

Feature	Distribution	Tested Method
Visit	Skewed	IQR
MR Delay	Skewed	IQR
CDR	Skewed	IQR
EDUC	Neither normal nor skewed	Percentile
MMSE	Skewed	IQR

In this study, feature selection techniques such as supervised and unsupervised selection techniques are utilized. Features such as Subject ID, MRI ID and Hand are removed directly from the dataset due to their lack of relevance. Supervised feature selection techniques may be further divided into three categories, i.e. intrinsic, wrapper, filter methods. During the implementation, filter and wrapper-based methods are employed. For filter-based method of feature selection, the input variable had both Numerical and categorical features which are separated. The output is categorical. Chi-Squared and Mutual Information are applied on categorical features. Anova and Mutual Information are applied on numerical features. In wrapper-based method of feature selection, recursive feature elimination and iterative feature selection techniques are applied. The models Decision Tree, Random Forest, XGBoost and Logistic Regression are used as estimator for recursive feature elimination. The best results from all the feature selection techniques utilized are obtained from recursive feature selection with Random Forest as estimator and six features were chosen based on their relevance as shown in Figure 1.

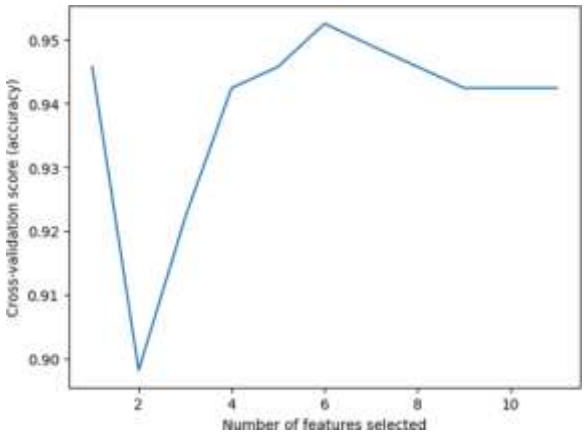


Fig. 1: Feature Selection Using Recursive Feature Elimination and Random Forest as estimator

3.3 Machine Learning Models

3.3.1 Random Forest

Random Forest is an ensemble learning technique that aggregates predictions from multiple decision trees to enhance both accuracy and robustness. During the training process, it constructs numerous trees and determines either the majority class for classification tasks or the mean prediction for regression tasks. Known for its effectiveness, scalability, and ability to handle complex, high-dimensional datasets, Random Forest emerges as a prominent method in the realm of machine learning.

3.3.2 XGBoost

XGBoost presents an optimized distributed gradient boosting library crafted for efficiency, adaptability, and scalability. Employing a gradient boosting framework, it sequentially trains a conglomerate of weak learners, typically decision trees, amalgamating their predictions to refine accuracy. Widely recognized for its exceptional performance and speed, XGBoost serves as a cornerstone in various machine learning competitions.

3.3.3 Decision Tree

A decision tree is a supervised learning algorithm that excels in both classification and regression endeavours. It partitions datasets recursively into subsets based on the most informative features at each node, thereby creating a tree-like structure. Each leaf node encapsulates a class label or numerical value, rendering decision trees interpretable, visually intuitive, and proficient in capturing intricate data relationships.

3.3.4 K-Nearest Neighbours

KNN is a simple instance-based learning algorithm designed for both classification and regression purposes. It assigns labels to new data points by considering the majority class among their closest neighbours in the feature space. Devoid of assumptions regarding the underlying data distribution, KNN proves particularly adept in handling datasets featuring complex decision boundaries.

3.3.5 Support Vector Machine

SVM, a formidable supervised learning algorithm, finds application in both classification and regression scenarios. By constructing hyperplanes in a high-dimensional feature space, SVM effectively segregates classes or predicts continuous outcomes. Its objective lies in maximizing the margin between classes while minimizing classification errors, thereby excelling in both linearly and non- linearly separable datasets.

3.3.6 Logistic Regression

Logistic Regression, categorized as a linear classification algorithm, is particularly useful in tasks involving binary classification. It works by modeling the probability of an input being assigned to a particular class through the logistic function. Despite its nomenclature, Logistic Regression operates as a classification rather than regression algorithm. Renowned for its simplicity, interpretability, and efficiency with linearly separable datasets, Logistic Regression stands as a foundational method in machine learning.

3.4 Explainable AI Techniques

3.4.1 SHAP (SHapley Additive exPlanations)

SHAP is a game-theory-based method utilized to interpret the outcomes of machine learning models. It attributes an importance score to each feature based on how much it contributes to the final prediction. By calculating Shapley values for all possible feature combinations, SHAP offers a

comprehensive view of how individual features affect the model's prediction. Its consistency and ability to handle complex models make SHAP one of the most widely adopted explainability methods in the field.

SHAP Summary Plot:
The SHAP summary plot provides a consolidated representation of feature importance and their effects on model predictions, derived from SHAP values. In this visualization, features are ranked by their mean absolute SHAP values, with the most important features at the top and the least important at the bottom. Each dot in the scatter plot represents an individual observation for a particular feature. The Y-axis lists the features, while the X-axis displays the corresponding SHAP values, indicating the direction and magnitude of each feature's effect on the predictions. This plot provides an in-depth analysis of how each feature influences the model's outputs.

SHAP Bar Plot:
A SHAP bar plot visualizes the contribution of each feature to a machine learning model's output by displaying the average magnitude of SHAP values across all predictions. The length of each bar represents how much influence a feature has on the model's decisions, with longer bars indicating greater impact. Ranked by importance, this plot offers a clear, interpretable summary of the most influential features, making it especially useful for understanding complex models and ensuring transparency in predictions.

SHAP Waterfall Plot:
The waterfall plot visually represents the impact of individual features on the model's prediction for a specific instance. Each row illustrates how a feature either positively (red) or negatively (blue) contributes to shifting the predicted value from the expected output, E[f(x)], which is the model's prediction based on the background data distribution, to the final prediction, f(x), for the given instance. The SHAP values, representing each feature's contribution, are accumulated to show how the model output transitions from a baseline value to the final prediction. This format effectively clearly illustrates how the model integrates evidence from all features to arrive at the predicted value.

3.4.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a model-agnostic explanation technique designed to interpret predictions of any black-box model. It operates by approximating the original model locally around the instance to be explained, using simpler interpretable models like linear regression or decision trees. By modifying the input data and analyzing the resulting changes in predictions, LIME provides a local surrogate model that helps understand the reasoning behind specific predictions.

4. Results and Discussion

The dataset undergoes partitioning into training and testing subsets, with an allocation ratio of 80:20 for the train-test split. In this study, cross-validation is also utilized for evaluation purposes, enhancing the reliability of the findings. Cross-validation stands as a statistical methodology vital for assessing the efficacy and resilience of machine learning models. This technique entails partitioning the dataset into distinct subsets, or folds, which are sequentially employed as both training and validation sets. By iteratively training and validating on various subsets, cross-validation provides a more reliable assessment of the model's performance compared to a single train-test split. Figure 2 depicts the confusion matrix obtained for different machine learning models.

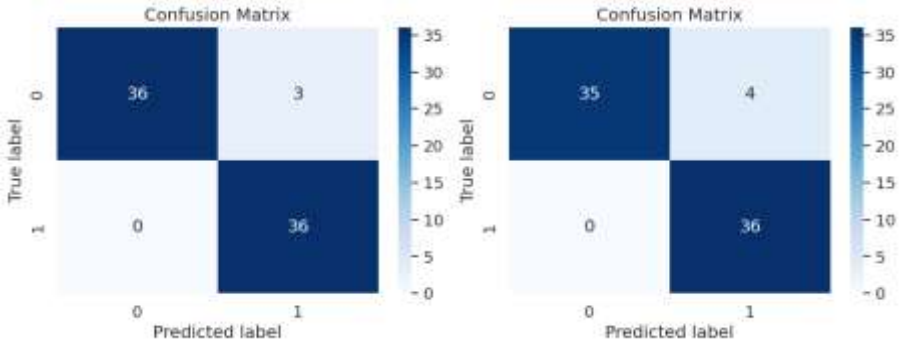
The selection of the optimal classifier model involves considering several performance metrics. The formulas for these metrics in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are presented in Equations (1) through (4), which are used in Table II. These equations are instrumental in evaluating the outcomes of different models.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

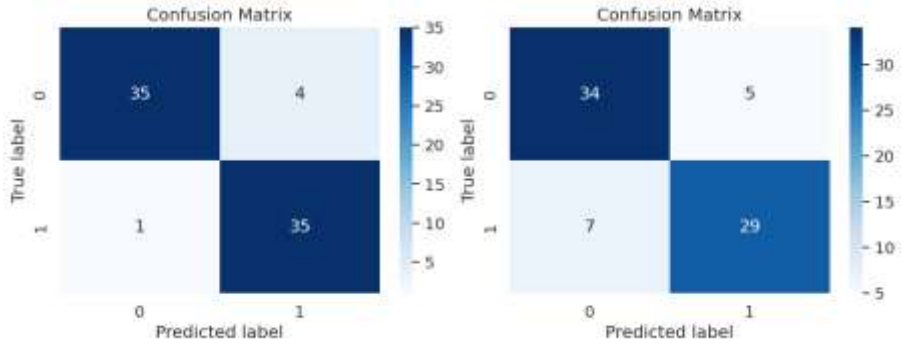
$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

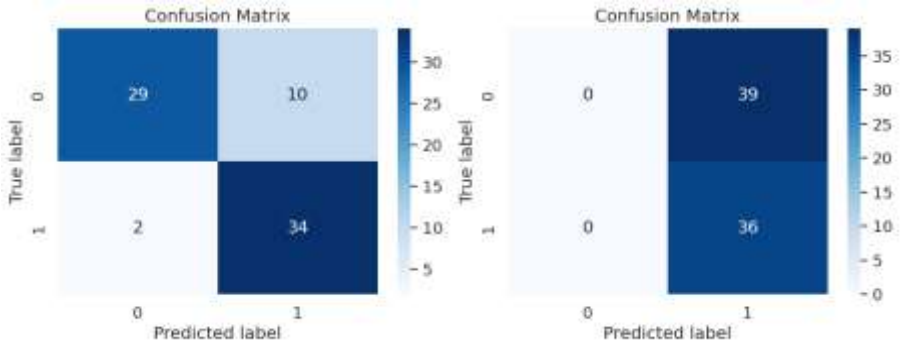
$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$



(a) Confusion Matrix for Random Forest (b) Confusion Matrix for XgBoost



(c) Confusion Matrix for Decision Tree (d) Confusion Matrix for kNN



(e) Confusion Matrix for SVM (f) Confusion Matrix for Logistic Regression

Fig. 2: Confusion Matrix for Different Machine Learning Models for Dementia Classification

TABLE II: Evaluation Metrics for Dementia Classification

Model	Cross-Validation Score				Test Data			
	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
Random Forest	0.95	0.93	0.95	0.98	0.96	0.92	0.96	1
XGBoost	0.95	0.92	0.95	0.98	0.95	0.90	0.95	1
Decision Tree	0.89	0.92	0.89	0.88	0.93	0.90	0.93	0.97
KNN	0.75	0.74	0.77	0.82	0.84	0.85	0.83	0.81
SVM	0.77	0.71	0.81	0.95	0.84	0.77	0.85	0.94
Logistic Regression	0.52	0.74	0.77	0.82	0.48	0.48	0.65	1

4.1 Implementation and Evaluation of ML Models

In this study, several machine learning models are evaluated for their effectiveness in classifying dementia. Random Forest is employed as an estimator in Recursive Feature Elimination (RFE) for feature selection, successfully identifying six key features and achieving high accuracy, precision, and recall. Its robust performance and interpretability, enhanced through SHAP and LIME analyses, lead to its selection as a final model. XGBoost, applied after feature selection, also demonstrates exceptional performance and strong predictive accuracy, supported by the same interpretability tools.

Decision Trees are initially used for the classification of dementia on the dataset but are ultimately outperformed by Random Forest and XGBoost in predictive performance. K-Nearest Neighbours (KNN) is considered but not used due to limitations with high-dimensional data. Similarly, Support Vector Machines (SVM) are briefly explored but do not achieve the accuracy and scalability of the ensemble methods. Logistic Regression provides useful insights into linear relationships but is less effective in capturing the dataset's complexities. Ultimately, Random Forest and XGBoost outperform the other models, making them the preferred choices for this analysis.

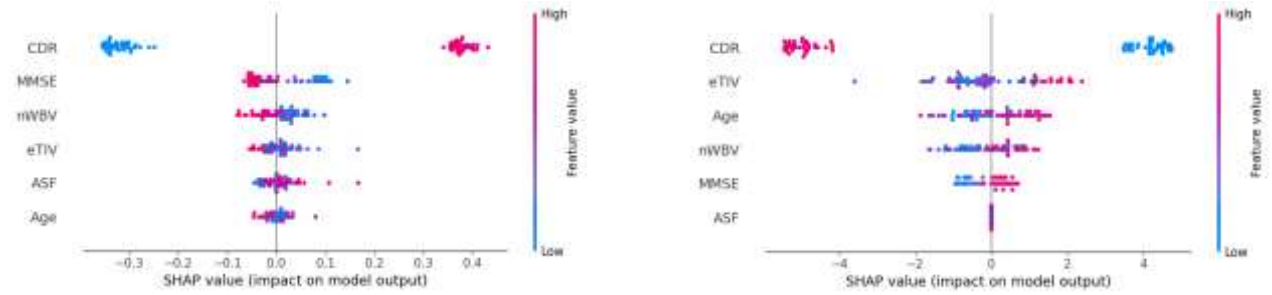
4.2 Implementation of XAI Techniques

4.2.1 SHAP

SHAP (SHapley Additive ExPlanations) is an interpretative framework grounded in game theory, designed to elucidate the outputs of machine learning models. It calculates SHAP values, which represent the contribution of each feature to a specific prediction, providing an intuitive and detailed understanding of feature influence. Features with larger SHAP values are deemed more significant, and their importance is visualized in descending order. As a model-agnostic approach, SHAP can be applied to any machine learning algorithm, offering consistent and interpretable insights into feature relevance. Its key innovation lies in defining a new class of additive feature attribution methods, with theoretical results establishing a unique solution that possesses desirable properties within this class.

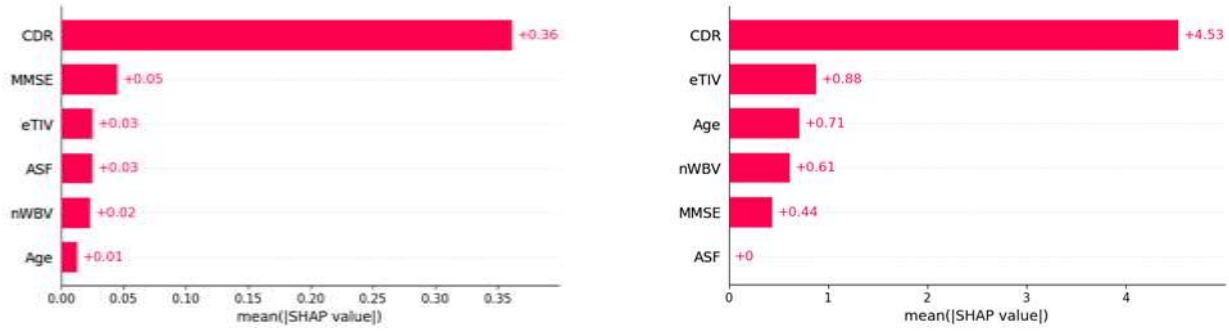
The SHAP summary plot for the Random Forest model regarding Class 0 (dementia) shown in Figure 3(a) demonstrates the predominance of Cognitive Decline Rating (CDR) as the key predictor in the classification process. The plot reveals a pronounced clustering of blue dots at -0.4, indicating that lower CDR values are strongly associated with non-demented classifications. Conversely, the aggregation of red dots at 0.4 illustrates that higher CDR scores correlate with an increased likelihood of dementia diagnosis. This clear delineation emphasizes the critical role of CDR in influencing model outcomes. Other features, such as Age and nWBV, present varied contributions, indicating relevance but being overshadowed by the dominant effect of CDR.

In the context of the XGBoost model, the SHAP summary plot depicted in Figure 3(b) further emphasizes the significance of CDR in dementia classification. The blue dots positioned near 4 suggest that lower CDR scores correspond to a reduced likelihood of the condition, while the red dots clustered around -4 (indicating higher CDR scores) suggest a greater probability of being classified as having dementia. This consistent finding across both models reinforces the conclusion that CDR serves as a vital predictor in identifying dementia.



(a) Dementia in Random Forest (b) Non dementia in XGBoost

Fig. 3: SHAP Summary Plot



(a) Random Forest (b) XGBoost

Fig. 4: SHAP Bar Plot

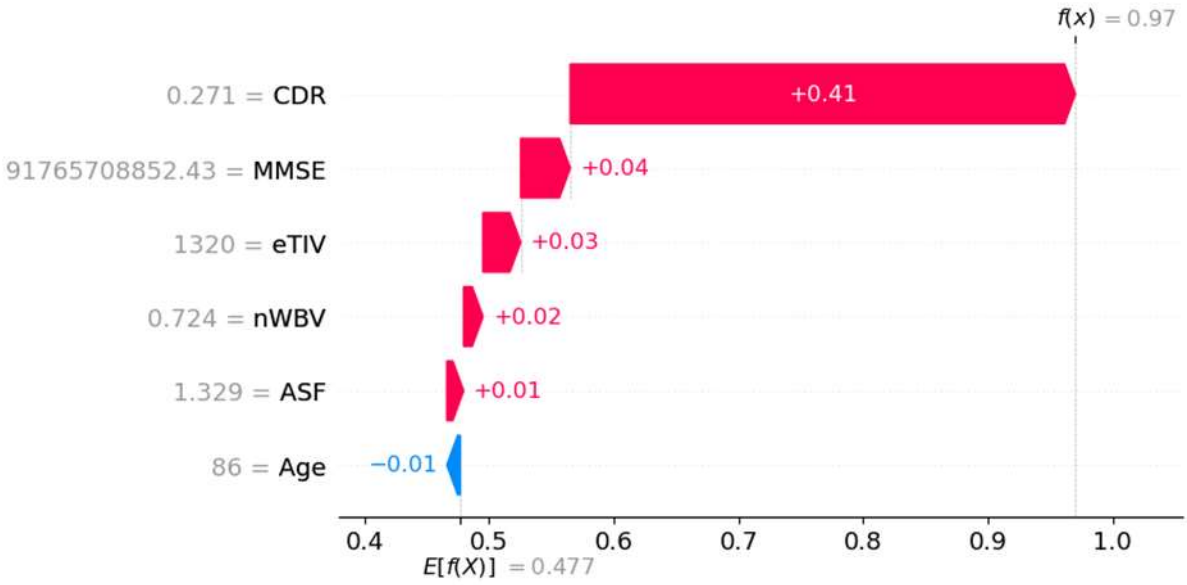


Fig. 5 SHAP waterfall plot of an instance of class 0 (dementia) using Random Forest

In the SHAP bar plot for Random Forest model illustrated in Figure 4(a), the Clinical Dementia Rating (CDR) stands out with a SHAP value of +0.36, indicating a strong association between higher CDR scores and dementia presence. Other features, such as the Mini-Mental State Examination (MMSE) and normalized Whole Brain Volume (nWBV), have lower contributions of +0.05 and +0.03, respectively, suggesting that cognitive performance and brain volume are also relevant but less impactful. These are followed by Effective Total Intracranial Volume (eTIV) with a contribution of +0.03. Alzheimer's Severity Factor (ASF) and Age exhibit minimal influence with values of +0.02 and +0.01 respectively. In contrast,

the SHAP bar plot for XGBoost model shown in Figure 4(b) assigns an even greater importance to CDR, with a SHAP value of +4.53, highlighting its critical role in dementia prediction. The eTIV, Age, and nWBV follow with SHAP values of +0.88, +0.71, and +0.61, respectively, indicating their meaningful contributions to the diagnosis. The MMSE retains significance (+0.44), while ASF shows no impact. Overall, both models affirm the pivotal role of CDR in diagnosing dementia, with variations in feature influence suggesting a multifaceted approach is essential for accurate predictions.

The SHAP waterfall plot outlined in Figure 5 deconstructs the classifier's prediction for an individual patient, elucidating the specific contributions of various features to the overall dementia assessment. This plot visually represents the shift from the expected value to the predicted outcome, highlighting how individual feature impacts align with the dementia diagnosis. In this instance, the model outputs a high prediction value of $f(x)=0.97f(x) = 0.97f(x)=0.97$, indicating a strong likelihood of dementia. The Clinical Dementia Rating (CDR) significantly influences the prediction, with a notable positive contribution of +0.41, demonstrating its critical role in identifying dementia. The contributions from other features, such as the Mini-Mental State Examination (MMSE), Effective Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), and Alzheimer's Severity Factor (ASF), further support this classification. Each of these features contributes positively to the prediction, with values of +0.04, +0.03, +0.02, and +0.01, respectively, indicating their relevance in the dementia context.

Conversely, the Age feature exhibits a minor negative contribution of -0.01, suggesting that older age has a slight diminishing effect on the prediction, although it does not overshadow the significant positive influences of other features. This combined effect reinforces the classifier's identification of this patient as being dementia-positive. The insights drawn from the waterfall plot align with existing dementia research, highlighting the importance of these features in predicting dementia and confirming the patient's condition as consistent with clinical expectations.

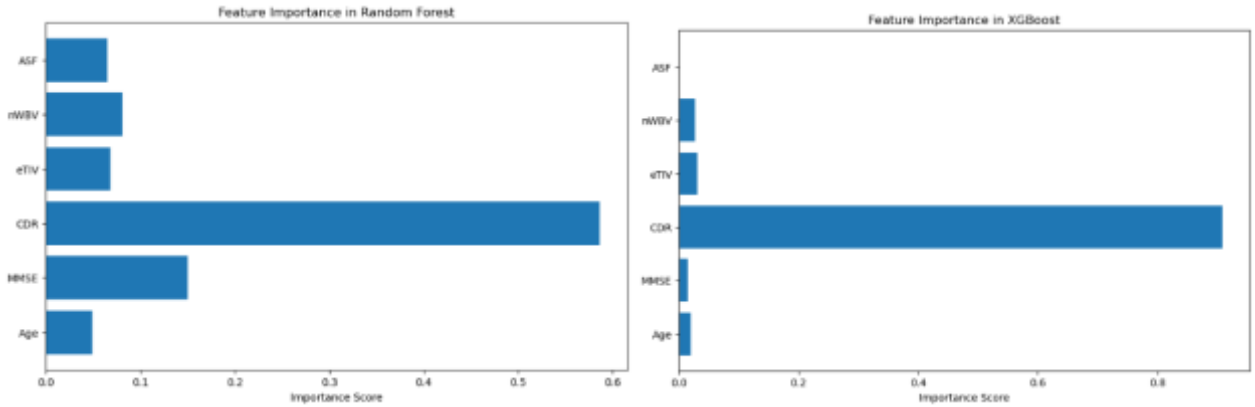
4.2.2 LIME

LIME is a model-agnostic technique in Explainable AI that generates locally interpretable models to explain how individual features influence the predictions of complex machine learning systems. The LIME framework provides insight into individual model outputs by generating a local surrogate model that approximates the behaviour of the original model in the neighbourhood of a given prediction. Given its focus on localized explanations, LIME does not endeavour to elucidate all potential decisions a model may render across the entire input space. Rather, it selectively examines the features that contribute to the classification of a specific instance, providing insight into the model's decision-making process for that particular prediction.

The Local Interpretable Model-agnostic Explanations (LIME) analysis of Figure 6 elucidates the contributions of various features influencing the prediction probabilities for a specific patient regarding dementia classification. The model predicts a high probability of class 0 (dementia) at 0.99, while the probability of class 1 (non-dementia) is minimal at 0.01. The Clinical Dementia Rating (CDR) of 0.27, indicating borderline status, contributes a significant probability of 0.58 to the prediction of class 0. This value falls within the range of $-0.00 < \text{CDR} \leq 0.27$, suggesting a tendency towards dementia. The Mini-Mental State Examination (MMSE) score of 91765708852.43 is notably high and reinforces the classification as demented, with a threshold of MMSE $\leq$ 91765708852.43 adding an additional 0.08 probability to class 0.

Fig. 6 LIME plot of an instance of class 0 (dementia) using Random Forest

Fig. 7 LIME plot of an instance of class 0 (dementia) using XGBoost



(a) Random Forest (b) XGBoost

Fig. 8 Feature contribution plot

Conversely, the Alzheimer's Severity Factor (ASF) of 1.19 contributes positively to class 1, as indicated by the threshold $1.11 < ASF \leq 1.19$, adding a probability of 0.02. The Effective Total Intracranial Volume (eTIV) of 1477.00 also supports class 1, falling within the range of $1473.00 < eTIV \leq 1583.75$, contributing 0.00 to the prediction for class 0. Additionally, the Normalized Whole Brain Volume (nWBV) of 0.73 and the patient's Age of 82.00 contributes 0.00 probabilities to classes 0 and 1, respectively. Notably, similar results were obtained using the XGBoost model, as shown in Figure 7, along with LIME analysis, further reinforcing the reliability of the findings across various model interpretations.

4.2.3 Feature Contribution

Feature contribution, often referred to as feature importance, quantifies the influence of each feature on the predictions made by the model. This measurement can be derived for either the entire dataset or specific individual instances. Understanding feature importance not only aids in

interpreting complex models but also enhances model performance by identifying relevant features for inclusion and potentially eliminating irrelevant or redundant ones. Moreover, understanding feature importance is especially critical in high-stakes fields such as healthcare and finance, where comprehending the reasoning behind predictions is paramount for informed decision-making.

In both the Random Forest and XGBoost models as shown in Figure 8, the Clinical Dementia Rating (CDR) emerges as the most significant predictor of dementia, with a stark contrast in importance compared to other features. In the Random Forest model, CDR has an importance score close to 0.6, while all other features have scores below 0.1. Similarly, in the XGBoost model, CDR's importance exceeds 0.8, with negligible contributions from other features. This consistent dominance of CDR across both models highlights its critical role in dementia classification, underscoring its relevance as a key determinant in predicting the condition.

5. Conclusion

Dementia represents a significant global health issue, with a shifting focus towards risk mitigation, early intervention, and prompt identification in elderly individuals, rather than solely seeking a cure. With the aging of society, the prevalence of dementia rises, impacting not only older adults but also a growing number of younger individuals afflicted by the condition. This study demonstrated the effectiveness of machine learning models for binary classification, accurately predicting the presence or absence of dementia among patients and categorizing them as either “Demented” or “Non-Demented”. We utilized a range of machine learning models, encompassing Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and XGBoost. After comparison, Random Forest demonstrated superior performance relative to the other models, attaining an impressive accuracy rate of 96%. From the Table II it can be concluded that XGBoost is the close competitor followed by Decision Tree.

Additionally, to ensure transparency and interpretability, Explainable AI (XAI) techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were employed to analyze feature contributions. These techniques provided valuable insights into how specific features and clinical parameters contributed to the model's predictions. This interpretability enhances the trust and usability of machine learning models in clinical practice. As future work, these efficient models can be integrated with multimodal data sources such as neuroimaging, genetic, and clinical data to further enhance accuracy and understanding of disease progression. Investigating the development of interpretable machine learning models specific to clinical settings could also facilitate real-time adaptability by healthcare professionals.

References:

[1] World Health Organization (2023). Dementia, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia/ (Accessed Date: February 14, 2024).

[2] Alzheimer’s Disease International (2022). Dementia statistics, [Online]. Available: https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ (Accessed Date: February 14, 2024).

[3] Lee, J., Meijer, E., Langa, K. M., Ganguli, M., Varghese, M., Banerjee, J., et al. "Prevalence of dementia in India: National and state estimates from a nationwide study." *Alzheimer’s and Dementia*, vol. 19, no. 7, pp. 2898–2912, 2023. doi: 10.1002/alz.12928.

[4] Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F., and Li, J. "Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease." *Translational Research*, vol. 194, pp. 56–67, 2018. doi: 10.1016/j.trsl.2018.01.001.

[5] Tanveer, M., Richhariya, B., Khan, R., Rashid, A., Khanna, P., Prasad, M., and Lin, C. "Machine learning techniques for the diagnosis of Alzheimer’s disease: A review." *ACM Transactions on Multimedia Computing*, Communications and Applications, vol. 16, no. 1s, 2020. doi: 10.1145/3344998.

[6] Kumar, M. R., Vekkot, S., Lalitha, S., Gupta, D., Govindraj, V. J., Shaukat, K., et al. "Dementia detection from speech using machine learning and

deep learning architectures." *Sensors*, vol. 22, no. 23, 2022. doi: 10.3390/s22239311.

[7] Herzog, N. J., and Magoulas, G. D. "Brain asymmetry detection and machine learning classification for diagnosis of early dementia." *Sensors*, vol. 21, no. 3, pp. 1–17, 2021. doi: 10.3390/s21030778.

[8] Castellazzi, G., Cuzzoni, M. G., Cotta Ramusino, M., Martinelli, D., Denaro, F., Ricciardi, A., et al. "A machine learning approach for the differential diagnosis of Alzheimer and vascular dementia fed by MRI selected features." *Frontiers in Neuroinformatics*, vol. 14, 2020. doi: 10.3389/fninf.2020.00025.

[9] James, C., Ranson, J. M., Everson, R., and Llewellyn, D. J. "Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients." *JAMA Network Open*, vol. 4, no. 12, 2021. doi: 10.1001/jamanetworkopen.2021.36553.

[10] Zhu, F., Li, X., Tang, H., He, Z., Zhang, C., Hung, G.-U., et al. "Machine learning for the preliminary diagnosis of dementia." *Scientific Programming*, vol. 2020, 2020. doi: 10.1155/2020/5629090.

[11] Salem, F. A., Chaaya, M., Ghannam, H., Al Feel, R. E., and El Asmar, K. "Regression based machine learning model for dementia diagnosis in a community setting." *Alzheimer's & Dementia*, vol. 17, p. e053839, 2021. doi: 10.1002/alz.053839.

[12] Mirzaei, G., and Adeli, H. "Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia." *Biomedical Signal Processing and Control*, vol. 72, p. 103293, 2022. doi: 10.1016/j.bspc.2021.103293.

[13] Shahzad, A., Dadlani, A., Lee, H., and Kim, K. "Automated prescreening of mild cognitive impairment using shank-mounted inertial sensors based gait biomarkers." *IEEE Access*, vol. 10, pp. 15,835–15,844, 2022. doi: 10.1109/ACCESS.2022.3149100.

[14] Hane, C. A., Nori, V. S., Crown, W. H., Sanghavi, D. M., and Bleicher, P. "Predicting onset of dementia using clinical notes and machine learning: case-control study." *JMIR Medical Informatics*, vol. 8, no. 6, p. e17819, 2020. doi: 10.2196/17819.

[15] Aschwanden, D., Aichele, S., Ghisletta, P., Terracciano, A., Kliegel, M., Sutin, A. R., et al. "Predicting cognitive impairment and dementia: A machine learning approach." *Journal of Alzheimer's Disease*, vol. 75, no. 3, pp. 717–728, 2020. doi: 10.3233/JAD-190967.

[16] Jin, H., Chien, S., Meijer, E., Khobragade, P., Lee, J., et al. "Learning from clinical consensus diagnosis in India to facilitate automatic classification of dementia: machine learning study." *JMIR Mental Health*, vol. 8, no. 5, p. e27113, 2021. doi: 10.2196/27113.

[17] Rai, A. "Explainable AI: From black box to glass box." *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2020. doi: 10.1007/s11747-019-00710-5.

[18] Lombardi, A., Diacono, D., Amoroso, N., et al. "A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease." *Brain Informatics*, vol. 9, no. 17, 2022. doi: 10.1186/s40708-022-00165-5.

[19] Shivaprasad, S., Chadaga, K., Dias, C. C., Sampathila, N., and Prabhu, S. "An interpretable schizophrenia diagnosis framework using machine learning and explainable artificial intelligence." *Systems Science & Control Engineering*, vol. 12, no. 1, pp. 2364033, 2024. doi: 10.1080/21642583.2024.2364033.

[20] Bogdanovic, B., Eftimov, T., and Simjanoska, M. "In-depth insights into Alzheimer's disease by using explainable machine learning approach." Scientific Reports, vol. 12, no. 6508, 2022. doi: 10.1038/s41598-022-10202-2.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.