SHAP-based Feature Selection and Explainable Machine Learning Classification of Alzheimer's Disease

Archana Menon P[1], R. Gunasundari[2]

[1]Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India,
Dept. of Cyber Security and Applied Computing, St.Teresa's College (Autonomous), Ernakulam, India,
Email: archananirmal1414@gmail.com
[2]Dept. of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, India,
Email: gunasoundar04@gmail.com

Received: 16.04.2024 Revised : 13.05.2024 Accepted: 22.05.2024

ABSTRACT
Alzheimer's Disease (AD) is a progressive neurodegenerative disease with a significant impact on healthcare. This study suggests a machine learning (ML)-based Explainable Artificial Intelligence (XAI) framework for early AD prediction. The system incorporates SHAP (Shapley Additive explanations) for interpretability in order to overcome the "black box" aspect of existing machine learning models. Using the OASIS dataset, we quickly investigate different classifiers for AD prediction by utilizing PyCaret, a low-code tool. With a classification accuracy of 96%, the Naive Bayes classifier was the most successful of the assessed models. In order to comprehend feature importance and get insights into model thinking, SHAP analysis is applied. The approach also selects features, determining the most important variables for AD prediction by utilizing SHAP values. The model's transparency is increased by this combination method, which uses SHAP for interpretability and feature selection and PyCaret for effective exploration. Clinicians gain deeper understanding of the model's decision-making process and the factors most critical for AD prediction. This study breaks new ground by demonstrating the efficacy of PyCaret and SHAP in building an interpretable and accurate framework for early AD prediction.

Keywords: Explainable AI, Alzheimer's Disease, Machine Learning, PyCaret, SHAP

1. INTRODUCTION

Brain is an exceptionally complex organ which is in charge of numerous body functions. Neurological disorders are caused due to many factors such as age, genetics, lifestyle etc. There are more than 600 types of neurological disorders with various causes. Most common types of neurological disorders are Alzheimer's Disease (AD), Parkinson's disease and Stroke [1]. These diseases act on different facet of nervous system. Each disease has its own causes, symptoms, diagnosis methods and treatments. Symptoms of such diseases gradually worsen over time. Even though these diseases have no cure, they can be controlled or symptoms can be mitigated if diagnosed on time and provides adequate treatment [1].

AD is a progressive neurological disorder that causes the brain to shrink and brain cells to die [2]. It is caused by a combination of factors affecting the brain such as genetic, lifestyle and environment. It is the common cause of dementia. Dementia is a state of continuous decline in thinking, social and behavioral skills which affects the ability of a person to function independently [2]. Early signs of AD include forgetting recent conversations and events. When the disease progresses, the patient suffers from memory impairment and they no longer would have the ability to do day-to-day activities. No treatment could cure this disease. Instead, early detection of this disease and medications could improve the patient's condition or atleast slow down the progression of the disease to a certain extent. Brain Imaging (MRI) helps the doctors to find out the shrinkage of brain tissues. MRI findings says if a patient is affected with AD or not. Machine Learning (ML) techniques are the proved techniques which can accurately predict the progress of a patient from mild cognitive impairment to AD [3].

In healthcare, the ability to predict a disease at an early stage could revolutionize the treatment and prevention strategies. Even though ML models excel in analyzing vast datasets and predict the diseases with greater accuracy surpassing traditional statistical methods, their black-box nature always acts as a barrier for explaining the decision-making process. ML models which gives mere results without explaining how they arrived at that particular decision are called black box models [4]. This black box nature of the complex models makes it untrusty to use by the clinicians [5]. For the wide acceptability of

complex ML models, Explainable Artificial Intelligence (XAI) evolved. XAI method makes the ML methods understandable by human beings [6]. They are integrated into critical decision-making processes which makes the model more transparent and trustworthy to the end users.

Medical data are collected from different patients around the world in different formats using various sources. So, the data will be a mixture of noisy, irrelevant as well as important data. These high dimensional data, if not processed properly, will affect the computational efficiency of the model. Many studies on feature selection techniques for AD have taken place [7-9]. Different from existing studies, a novel XAI framework is proposed in this paper combining ML models and SHAP algorithm to predict AD at an early stage by using only the prime features. Apart from providing explanations to the results, SHAP performs feature selection and reduces the dimensionality of the original dataset. The data is then fed to the classifier which diagnosis AD. A good classification performance exhibited by the classifiers with SHAP based feature selection can be effectively used by the doctors as a reference for the diagnosis of AD. Here, Brain MRI dataset from OASIS project is implemented to test the feasibility of the framework.

The main contributions of this paper are:

1. This work leverages PyCaret's user-friendly approach to streamline the machine learning workflow.This low-code library offers a vast collection of pre-built algorithms, enabling rapid experimentation and comprehensive evaluation through a variety of performance metrics for the prediction of AD.
2. The model identifies the patients at high risk. OASIS dataset is used to assess the performance of the complex model.
3. SHAP identifies the features (variables) which are high indicators for the prediction. Utilizing SHAP values enhances the interpretability of our model, fostering trust in its predictions and facilitating communication with medical professionals. The SHAP insights can be used in positive or negative result decision-making process.
4. SHAP explanation is not just limited to the average impact of the model but also exploring interactions and individual prediction explanations. Local explanations improve the clinical understanding of prediction.
5. An effective feature selection technique using SHAP is employed, to reduce variables in the dataset, which in turn increases computational efficiency. By leveraging SHAP values, we provide a nuanced understanding of feature importance. Analyzing feature importance allows us to delve into the "why" behind the predictions, providing valuable insights beyond just performance metrics.

The rest of the paper is organized as follows. Section 2 briefs out recent literatures.In Section 3, the background of this research is detailed such as, the need for explanations, PyCaret Library, intuition behind SHAP and some of the feature selection approaches used in this experiment. Section 4details the proposed methodology with an overview, algorithms and explanations. Section 5 describes the experiments carried out and the results achieved. A detailed discussion and comparison of results has been carried out in Section 6.Section 7 concludes the paper with the scope for future improvements.

2. RELATED WORKS

There's a growing body of research demonstrating the potential of ML techniques for accurate AD diagnosis. This is because ML approaches are more accurate in diagnosis, rely on non-invasive techniques and have data analysis power.

[10] suggests a novel ML-based approach to distinguish between AD patients, people with MCI, and healthy controls using structural magnetic resonance imaging (sMRI) data. Three ML algorithms were put into practice and contrasted by the authors: Regularised Extreme Learning Machine (RELM), Import Vector Machine (IVM), and Support Vector Machine (SVM). According to their results, RELM is more accurate in differentiating between the AD, MCI, and healthy control groups.

The potential of ML to forecast the course of AD in individuals with mild cognitive impairment (MCI) is examined in [11]. Researchers examined 116 papers that used ML to analyse neuroimaging (MRI, PET) data from the ADNI database, adhering to PRISMA principles. The results show that Support Vector Machines and Convolutional Neural Networks were able to predict the development of AD in patients with MCI with a promising accuracy of 75.4% and 78.5%, respectively. Moreover, research integrating complicated models (deep learning) with multimodal data (MRI & PET) showed better results. These findings point to the possibility of ML as a useful tool for doctors in identifying individuals at risk for AD progression.

[12]highlights the value of early intervention and determines the best criteria for AD prediction by analysing data from the Open Access Series of Imaging Studies (OASIS). Using a variety of ML algorithms, such as Random Forests and Decision Trees, the authors outperform previous techniques, obtaining an

83% test accuracy. These results demonstrate the potential of ML for better AD diagnosis, which could result in earlier treatment and better patient outcomes.

The challenge of diagnosing Early Mild Cognitive Impairment (EMCI) in AD is addressed in [13]. The authors suggest a unique DL strategy in response to the problem that many ML models have in differentiating between EMCI and typical cognition. This approach uses a Convolutional Neural Network with Long Short-Term Memory (LSTM) architecture to integrate multimodal data, such as MRI scans, PET scans, and results from neuropsychological tests. The model distinguishes between EMCI patients and normal controls with an astounding 98.5% accuracy rate. These results point to the possibility of accurate early detection of AD by deep learning with multimodal data, which could open the door to earlier intervention and better patient outcomes.

The model proposed in [14] makes use of an extensive dataset that includes clinical evaluations, neuroimaging scans (MRI, fMRI, and PET), and demographic data from both AD patients and healthy individuals. It combines characteristics from multiple sources to produce a two-fold strategy: 1) identifying AD phases and setting AD apart from healthy controls, and 2) forecasting the course of the disease and the clinical outcomes that AD patients will experience in the future. The outcomes show good classification and prognostic accuracy, indicating the framework's potential for early diagnosis, customised patient stratification, and individualised therapy regimens in the management of AD.

[15] proposes a model that makes use of several algorithms, including GaussianNB and Random Forest, and highlights the significance of early diagnosis and treatment. Using a Voting Classifier method, the model trained on the OASIS dataset has a high validation accuracy of 96%. According to these results, ML has a great deal of potential to raise the detection rates of AD, which could improve patient outcomes and lower mortality.

[16] investigates the early diagnosis of AD and associated dementias (ADRD) using ML in real-world electronic health records (EHRs). Researchers contrasted knowledge-driven versus data-driven machine learning techniques after analysing data from more than a million patients. The best results were obtained by gradient boosting trees trained using the data-driven approach, with AUC scores above 0.9 for ADRD prediction up to five years prior to diagnosis. These results point to the possibility of using ML in conjunction with EHR data to identify high-risk patients in order to improve patient outcomes and initiate early intervention.

[17] looks into how well ML algorithms predict the conversion of MCI patients to AD. Using multimodal data from the ADNI database, researchers compared three algorithms: Random Forest, Gradient Boosting, and XGBoost. All algorithms performed to some extent, although biomarkers associated with AD and neuropsychology had the best accuracy (90%) in predicting conversion. The results of this study imply that a combination of several ML techniques and a range of data sources, such as biological and clinical measurements, can result in a prediction of AD progression in MCI patients that is more accurate and dependable.

Even though ML classifiers could produce good classification accuracy, those were not interpretable. The emerging Explainable AI (XAI) techniques aim to provide explanations for complex models, allowing doctors to understand the reasoning behind a prediction.

[18] addresses the discrepancy in the diagnosis and course of AD between research and clinical practice. Conventional research constraints encompass dependence on singular data categories, distinct diagnosis and progression models, and incomprehensibility in intricate models. The authors suggest a two-layer model that integrates 11 data modalities from more than 1000 ADNI individuals and uses Random Forest classification. With its excellent accuracy (94% for diagnosis and 87% for progression), this model provides interpretability using decision-tree based explanations and SHAP algorithms. The goal of these plain-spoken explanations is to increase doctor confidence and comprehension of the model's predictions. This method has the potential to close the gap between research and practice and enhance clinical decision-making in AD diagnosis and progression management by combining accuracy with interpretability.

[19] uses a large dataset of medical, cognitive, and lifestyle variables from over 12,000 people to explore the difficulties of early AD diagnosis. The study highlights the significance of a strong experimental design and data preparation in order to tackle imbalances, redundancies, and missing data. After training, an XGBoost model obtained a competitive F1-score of 0.84. That being said, interpretability is the main priority. Model predictions were explained by using SHAP values, which showed the positive and negative effects of important features. This interpretability scheme challenges preexisting hypotheses and provides clinicians with insightful information. The work emphasises the importance of explainable machine learning in revealing associations between characteristics and diagnoses, which may help in the early identification of AD.

The problem of early detection of mild cognitive impairment (MCI) employing cognitive tests at the bedside, such as clock drawing, is addressed in [20]. Even while DL models seem promising, clinical adoption is still hindered by interpretability. The authors present a unique framework that integrates soft labels, self-attention processes, and data from several drawing tasks (clock drawing, cube copying, and trail creating) into a DL model. According to medical professionals, this method increases interpretability and accuracy (from 75% to 81%) and may be a useful tool for early MCI detection with improved clinical integration.

Current methods for predicting amnestic mild cognitive impairment (aMCI) conversion to AD often lack complexity. [21] addresses this by employing interpretable machine learning (IML) to develop a more accurate and informative prediction model. They recruited a large cohort of aMCI patients and analyzed neuropsychological test results, APOE genotype, and clinical data. The findings demonstrate that an extreme gradient boosting model achieved the best performance, identifying factors like age, education, cognitive scores, and genetic risk as crucial predictors. Notably, the IML approach allows for individual patient analysis, pinpointing the specific factors most relevant to their risk of conversion to AD. This personalized approach holds promise for improved clinical decision-making and targeted interventions.

[22] proposes a promising approach for AD stage prediction that addresses the challenges of limited data and lack of interpretability. The proposed approach, EfficientNetB7 for AD stage prediction (CN, AD, EMCI, MCI, LMCI) utilizes data augmentation to address limited data and XAI for model interpretability. EfficientNetB7 with Score-CAM or Grad-CAM++ achieves high accuracy (96.34%) and interpretability for AD stage prediction. The use of data augmentation and Explainable Artificial Intelligence (XAI) techniques can potentially improve the accuracy and trustworthiness of machine learning models for AD diagnosis.

PyCaret is a relatively new framework. There are fewer research papers specifically focusing on AD prediction using PyCaret compared to other machine learning frameworks.

[23] examined hippocampus volume, APOE genotype, and cognitive function in MCI patients using longitudinal data from the ADNI database. Two separate ML models were able to predict which MCI patients would develop AD with a cross-validation accuracy of 70%. Impaired memory was found to be a consistent predictor of this accuracy, which was maintained across several models and validation techniques. These results, which are higher than chance predictions, point to the potential of ML in helping to create instruments that will help doctors detect patients who are at risk of acquiring AD.

[24] investigates the use of resting-state fMRI (rs-fMRI) data from five datasets to identify individuals with MCI from healthy controls using the XGBoost algorithm. Through SHAP analysis, the significance of interpretability is addressed while attaining a maximum accuracy of 65.14%. The results indicate that distinct brain regions have differing importance depending on the rs-fMRI analysis techniques used, pointing to a complex effect of MCI on brain function. The work shows the potential of XGBoost with interpretable models for MCI diagnosis using rs-fMRI data, despite limits in overall accuracy.

[25] uses wearable lifelog data to investigate the potential of Automated Machine Learning (AutoML) for detecting cognitive decline in senior populations. PyCaret in Google Colaboratory is used in the study to examine data from individuals with high-risk dementia. They determine that Gradient Boosting, Random Forest, and Voting Classifier are the best models by comparing them. The study also identifies "Average heart rate per minute during sleep" and "Average respiration per minute during sleep" as critical variables for precise prediction. These results imply that wearable data and ML together have potential for better management of cognitive decline in the elderly and prophylactic approaches.

Traditional ML and DL methods have achieved promising accuracy in predicting AD, their inherent "black box" nature limits our understanding of the decision-making process. This lack of interpretability hinders clinical adoption and trust in these models. Conversely, XAI techniques offer a compelling solution by providing insights into the model's reasoning. However, XAI techniques can often be complex and time-consuming to implement. PyCaret, a low-code machine learning library, streamlines the development process, allowing researchers to focus on the interpretability aspect. This paper presents a novel approach that leverages the strengths of both PyCaret's efficiency and XAI techniques' interpretability. By combining these advancements, we can achieve not only high accuracy in AD prediction but also gain valuable insights into the underlying factors that contribute to the disease. This interpretability can foster trust in the model's predictions, paving the way for its integration into clinical decision-making, ultimately leading to better patient outcomes.

3. Preliminaries

3.1 Need for Explanations

Zero knowledge about the reasoning behind predictions and lack of clarity in its decisions make the ML models unadoptable by the clinicians. Moreover, these models might perpetuate biases or hidden errors. XAI techniques tackles the black box nature of the ML models by making the results more understandable

and interpretable. XAI models provide interpretable results which could be analyzed and evaluated by the clinicians along with their expertise promoting accuracy and transparency in the decision-making process. Also, the insights from explanations in disease prediction can lead to novel discoveries and better patient outcomes. XAI enhances interpretability of the results thereby building trust. Establishing trust is crucial for the societal acceptance of algorithmic decision-making. In AD prediction, XAI techniques can show which factors the model deems important and how much each factor contributed to the final prediction.

3.2 PyCaret

The aim of the proposed method is to understand why a person is predicted with high chance of AD. Instead of using different ML models to train the dataset and get the prediction, an open-source, low code ML library in Python-PyCaret is availed here which is used for automating the workflows in ML. PyCaret replaces hundreds of lines of code with a few lines and speeds up the experiment cycle exponentially [26]. It can train and deploy both supervised and unsupervised ML models. It speeds up the experiment cycle thereby increasing productivity and efficiency [26].

3.3 SHAP Explanations

Now a days, AI makes most of the decisions for us. So, it is crucial, atleast in some scenarios, to understand why AI is making a particular decision. Complex black box ML models ignores why it made a particular decision. This makes the model untrusty to use by the clinicians. Although they can be quite effective at predicting predictions, machine learning models frequently lack transparency. It is difficult for us to comprehend how the model makes its decisions. In contrast to the black-box concepts in ML models, XAI explains the results of the solutions [3]. XAI ensures the social right of human beings for "explanation". The goal of XAI is to understand how exactly does the model work to make a particular prediction, to know the relationship between input and output and to recognize the impact of each feature on prediction [27-28]. Interpretable ML algorithms can be broadly categorized as model-specific and model-agnostic approaches [29]. Model-specific approaches are specific to certain models whose parameters as well as internal structure can be interpreted. As the internal working of the model is known, the decision taken by such models can be easily interpreted. Linear Regression, Decision Tree and other white-box models falls under this category [29]. For complicated black-box models, model agnostic approaches are applied as post-hoc methods. Here, the explanation is separated from the model. Model-agnostic methods generate explanations in human-understandable way as even non-experts can understand. LIME and SHAP are examples of such models [29].

A strong foundation for analysing machine learning models is offered by SHAP (SHapley Additive exPlanations). It enables us to comprehend the role that every feature (data point) plays in a model's finalprediction.Through a variety of charts, we may use SHAP to obtain important insights on how models' function. Comprehending the significance of features is crucial in determining which features exert the greatest impact on the model's aggregate forecasts.We can see how different features interact with one another to influence the prediction with the aid of SHAP. For example, a loan acceptance prediction model may consider both income and credit score; SHAP can show how these factors interact to affect the outcome. SHAP can be a valuable tool for identifying potential biases in the model by analysing the consistent effects of particular feature values on the predictions. As SHAP is a model agnostic approach, it can be applied on any ML model.SHAP, in short, enables us to transcend the "black-box" approach to machine learning. We can increase the reliability of our models' outputs and, eventually, boost their efficacy by deciphering their internal mechanisms.

SHAP is based on game theory [30] and is helpful to explain individual predictions [31]. An instance x is predicted by computing the contribution of individual features to the prediction. SHAP computes shapley values from coalitional game theory where the feature values of an instance act as players [29]. Shapley value explains the marginal contribution of each feature in a dataset. Apart from individual contribution, their interaction among others (i.e., subsets) must also be considered [31]. Shapley values for each feature is calculated by including as well as excluding the feature in the subset and finding out the difference. Hence, shapley value calculates a feature's contribution for each subset and then averages these contributions. This gives the marginal contribution of a feature to the entire dataset and hence it is also called as marginal value. And these marginal contributions are used to find out the actual contributions of each feature [30]. We can find out the shapley values of each feature using the equation

$$\emptyset_i(f,x) = \sum_{Z' \subseteq x'} \frac{|Z'|!\,(F-|Z'|-1)!}{F!}\,[f_x(Z') - f_x(Z' \backslash i)] \quad (1)$$

In eqn. (1) [10], $\emptyset_i(f,x)$ represents the shapley value for feature i of the given instance x in the black box model f. Let us see what each variable in the eqn. (1) represents.

z'- a subset of attributes (i.e., different possible combination of inputs),

x'- the simplified input data (for image data)
F- total number of features in the dataset
$f_x(Z')$- Output of black box model with the subset attributes z' (Prediction)
$f_x(Z'\backslash i)$- Output of black box model with the subset attributes z' excluding feature i.
$[f_x(Z') - f_x(Z'\backslash i)]$- this finds out the contribution of feature i for the prediction
$\frac{|Z'|!\,(F-|Z'|-1)!}{F!}$ – calculates the weight. Contribution of each feature is multiplied by this weight.
If, by adding a new feature in the subset increases the weight, then that feature is considered to have strong contributions to the prediction. SHAP provides both global and local interpretability to the model.

3.3.1 Global Explanability

When feature importance scores are averaged across all instances, we get global interpretation. Global explanation intimates the contribution of each feature to the output either positively or negatively [29]. Feature Imporatnce Plot, Summary Plot, SHAP Dependence Plot etc. are commonly used to deliver global explanations.

3.3.2 Local Explanability

Local explanation is achieved through individual SHAP values of predictors. It shows the contribution of each feature for that particular observation and explains the reason for that decision/prediction [29]. Force plot, Waterfall Plot, Local bar Plot, Individual SHAP value plot etc. delivers local explanation to the model.

3.4 Feature Selection

Data redundancy is a major problem in high-dimensional datasets. Feature selection is a crucial step in an ML pipeline that involves choosing a subset of relevant features from the original set of input features [32]. It improves the performance of a model by reducing the dimensions of the dataset. It focuses on the most informative and important features. Different standard techniques exist for feature selection and the main 3 categories are: Filter, Wrapper and Embedded [32].
Filter techniques considers the inherent properties of features and select the most important features as a subset from the original dataset [33]. Information gain, Chi-square test, Fisher's score, correlation coefficient, variance threshold etc. are example of filter techniques [33]. Wrapper methods follows greedy search technique and they act like surrogate models. They train the algorithm in an iterative manner using a specific set of features and the features will be added or removed in each cycle based on previous output [33]. Forward feature selection, backward feature elimination, exhaustive feature selection, etc. falls under wrapper method [33]. Embedded methods include interactions of features and the selection is performed during training [34]. These methods are faster and combines the advantages of filter and wrapper methods. L1 and L2 Regularization, Random forest importance etc. are embedded methods [34].
Feature selection process improves the model performance as well as enhances the model interpretability. Here, SHAP is designated to select important features from the given datasets. The need for model interpretation motivated to pick SHAP as feature selection mechanism. As a preprocessing technique, SHAP helps to explain the decisions while constructing the ML models.

4. METHODOLOGY

4.1 Overview

The proposed system is designed in 2-phases. Fig. 1 shows Phase I and Fig. 2 shows Phase II of the proposed system. In Phase I, longitudinal brain MRI dataset from OASIS (Open Access Serie of Imaging Studies) [35] is considered for the prediction of AD. They generated longitudinal MRI data from 150 individuals with or without AD. The imaging was done on older adults aged between 60 and 96 years. We employed PyCaret to analyze the dataset containing relevant features for AD prediction. The dataset is preprocessed and prepared for analysis within the PyCaret environment. Various ML classifiers are trained and evaluated with PyCaret's built-in functions and the instances are classified as either Demented or Non-Demented. Later, the model performance is assessed using common metrices such as accuracy, precision, recall and F1-score. The analysis identifies the most effective model for AD classification in our dataset. The results are explained using SHAP. The variables which influenced high and low for the prediction are found out. SHAP provides both global and local explanations.

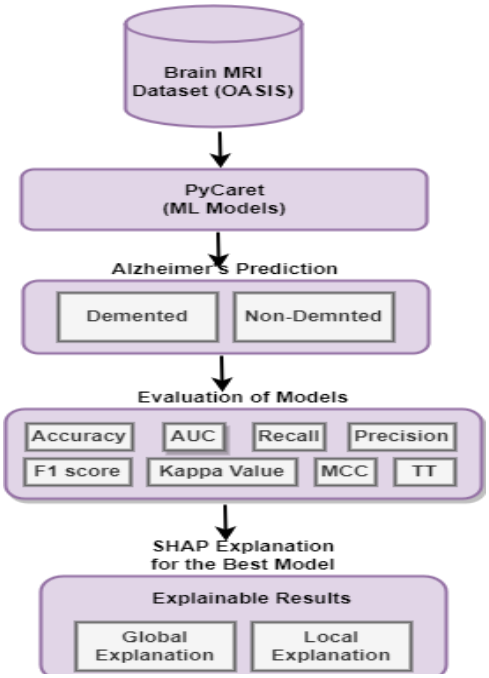

Fig.1 Architecture of Proposed Model- Phase I

In Phase II, the explainable results by SHAP are analysed to find out the important features responsible for the prediction. The impact of each feature on the prediction can be found out from the SHAP- values. Thus, prime features are extracted and the features with very low or no impact on the predictions are discarded. This reduces the dimensionality of the dataset and serves as the new dataset for Phase II. The new reduced dataset is used for training the classifiers in Phase II and then the classification accuracy of each model is calculated. Here again PyCaret library is employed for AD prediction. The classification accuracy of each model is reported and compared with those models in Phase I. The classifiers in Phase II, predicted diseases with greater accuracy and performed better in all the ways than in Phase I.

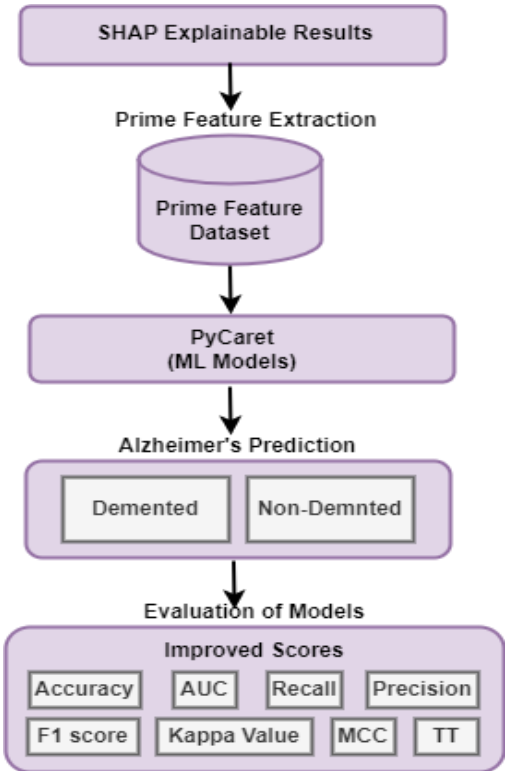

Fig 2. Architecture of Proposed Model- Phase II

4.2 Algorithm

Algorithm 1 details the working of Phase I. The original OASIS Brain MRI dataset, D_o, with n features, $f(x_1, x_2,\ldots,x_n)$,is considered for the experiment. The dataset containing relevant features is preprocessed and analyzed in the PyCaret environment. Several classifiers in the PyCaret library is trained using the OASIS dataset and each instance is predicted either as Demented or Non-Demented. Then the performance of these classifiers is evaluated and compared to select the best model in Phase I, BM_1. Predictions for new inputs are obtained and C_{p1} represents the predicted class (demented or non-demented) for Phase I. The predictions are explained using SHAP plots. Feature Importance plot, G_I, and SHAP Summary Plot, G_S, gives global explanation to the model. From these plots, most relevant features contributed for the prediction are extracted and builds a new dataset called Prime feature dataset D_p, with features$f(x_1, x_2,\ldots,x_m)$ where m<n. As the irrelevant features are discarded in D_p, it contains lesser number of features than in the original dataset, D_o. SHAP Force Plot, G_F, interprets individual instances thus giving local explanation to the model.

Algorithm 1: Phase I- Explainable AD prediction Using PyCaret and Prime feature selection
Input: OASIS MRI Dataset (D_o), Input features $f(x_1, x_2,\ldots,x_n)$
Output: Best Model (BM_1),
Output Class (C_{p1} -Demented or Non-demented),
Prime feature Dataset (D_p),
Output features $f(x_1, x_2,\ldots,x_m)$,
SHAP Global Explanations: (G_I)- SHAP Feature Importance plot and (G_S)- SHAP Summary Plot,
SHAP Local Explanations: (G_F)- SHAP Force plot
Step 1: Input D_o to PyCaret models for preprocessing and training
Step 2: Obtain the predictions C_{p1}, for the input $f(x_1, x_2,\ldots,x_n)$
Step 3: Evaluate and Compare the performance accuracy of the ML models and select the Best Model BM_1
Step 4: Obtain the predictions C_{p1}, for the input $f(x_1, x_2,\ldots,x_n)$
Step 5: Obtain SHAP Global explanation for the best ML model and Create Prime feature dataset
Step 5.1: Plot Feature Importance Graph G_I
Step 5.2: Draw SHAP Summary Plot G_S
Step 5.3: Select the Prime features from these graphs and create a new dataset D_p with only the prime features $f(x_1, x_2,\ldots,x_m)$ where m<n
Step 6: Obtain SHAP Local explanations for the best ML model
Step 6.1: Draw Force Plot for individual instances G_F
ReturnC_{p1}, BM_1, D_p, $f(x_1, x_2,\ldots,x_m)$, G_I, G_S,and G_F
Phase II is demonstrated in Algorithm 2. Here, the input dataset is the Prime feature Dataset, D_p, with the features $f(x_1, x_2,\ldots,x_m)$. The reduced feature dataset is then used to train the classifiers in the PyCaret library. The predictions are made and the performance of the classifiers are evaluated and compared in Phase II also and the best model, BM_2, is chosen. Predictions for new inputs are obtained from BM_2, and C_{p2} represents the predicted class (demented or non-demented) for Phase II. Again, a performance comparison is carried out between classifiers in Phase I and Phase II to show that all the ML models in Phase II outperformed Phase I models because of the positive impact of the reduced and relevant dataset, $D_p$in Phase II. A model which performed uniformly good in both Phase I and Phase II is selected as a consistent Model, CM.

Algorithm 2: Phase II- AD prediction from Prime feature dataset Using PyCaret with improved performance accuracy
Input: Prime feature Dataset (D_p), Input features $f(x_1, x_2,\ldots,x_m)$
Output: Best Model (BM_2),
Output Class (C_{p2}- Demented or Non-demented),
Consistent Model (CM)
Step 1: Input D_p to PyCaret models for training
Step 2: Obtain the predictions C_{p2}, for the input $f(x_1, x_2,\ldots,x_m)$
Step 3: Evaluate and Compare the performance accuracy of the ML models and select the Best Model BM_2 of Phase II
Step 4: Obtain the prediction C_{p2}, for new input $f(x_1, x_2,\ldots,x_m)$
Step 5: Compare the performance accuracy of the ML models in Phase I and Phase II and select a Consistent Model CM
ReturnC_{p2}, BM_2, and CM

4.3 Classification Using PyCaret

Initially the PyCaret environment has been setup to transform the pipeline for modeling and deployment. The data is pre-processed in the next step by encoding categorical values and imputing missing values. The PyCaret library simplifies model training by including functions which automates data pre-processing, data preparation and many other functions. The classification module (pycaret.classification) in PyCaret is a supervised ML module responsible for binary classification. PyCaret trains all the models in the library and rank them using stratified cross-validation for metric evaluation [26]. To evaluate the performance, we can analyze and compare the scores of the models from the scoring grid. PyCaret is used in Phase I and Phase II and analyzes the performance differences of all the models.

4.4 Explanations Using SHAP

Once the model prediction and evaluation are over, the results are explained using SHAP. SHAP, being a model agnostic interpretation method, helps to achieve the explanations of ML models employed. SHAP draws inspiration from game theory, specifically Shapley values, to explain how much each feature contributes to a specific prediction.It is more flexible to use as we can separate the model from explanations. SHAP draws insights from each ML models in predicting AD. It measures the positive or negative impact of each features along with its magnitude on a prediction. SHAP finds out the importance of all features in the OASIS MRI dataset for predicting AD risk by calculating the absolute Shapley values. Shapley value is calculated by taking the average of marginal contributions of each feature across all permutations. The intuition is that, the higher the Shapley value, the more important the feature is for the prediction. By this way, SHAP extracts relevant features from ML models.

SHAP gives both global and local explanation to the models. The average of absolute Shapley values of each feature across the dataset gives the global importance of features. These explanations are consistent with local explanations as the global value is contributed by local values from each instance. The SHAP value exposes contribution of each predictor to the output either positively or negatively. The most powerful aspect of SHAP is its ability to explain individual predictions. Every instance has its own SHAP value which gives local interpretation to that observation. We can take a single instance and say why is it predicted demented or non-demented and the contribution of each feature for this prediction. SHAP also offers several ways to visualize these explanations. In Phase I, the results are explained both locally and globally using SHAP.

4.5 Prime Feature Selection

SHAP explanations help to choose necessary and sufficient features to make the prediction. SHAP also aids in removing irrelevant features thus reducing the dimensionality of the data. Other feature selection algorithms have explainability issues. Especially in medical dataset, we cannot remove a few features simply without giving proper explanation. The weakness of feature selection techniques discussed in Section 3.4 is that they can't give appropriate explanation on why specific features are picked or removed. Filter methods does not consider the characteristics of a model for filtering the features whereas wrapper methods support a model's prediction.

In Phase II, a new dataset is built – Prime Feature Dataset- which consists of only the relevant features which is necessary for AD prediction. As SHAP is grounded by strong mathematical formulae, it helps to eliminate irrelevant features more sensibly. The contributions of each feature to the prediction are considered and sorted in descending order of their importance and applied it as feature selection approach. Now the prime feature dataset is a dimensionality reduced dataset with only the useful and needful features for AD prediction. With fewer features, the model performs more accurately with less complexity and computational cost.

4.6 Disease Prediction with Improved Performance

The prime feature dataset is trained and tested using PyCaret library.The classifier predicts each new observation as demented or non-demented. When the models are evaluated, all of them have exhibited a very good performance. Compared with the results of Phase I, all the models in Phase II enhanced their performance in terms of accuracy, precision, recall and F1-score. Hence, SHAP can be appraised as a preprocessing tool for feature selection in disease predictions. In the future, it can be applied in other domains too where model transparency is required.

5. Experiments and Results

The framework developed is mainly meant for classifying AD into demented and non-demented classes where the results are explained using SHAP. Other than explanation, SHAP shows the relevance of each feature input on the output which helps us to eliminate irrelevant features. Experimental results show that

the new models developed up on the prime features are more accurate than the previous models. The experiment is performed using brain MRI dataset from OASIS [35].

5.1 Performance Analysis

There are 18 different ML classifiers and different plots provided by the PyCaret module to analyze the performance of the models. PyCaret trains all these models using stratified cross-validation and ranks the models. The performance of the supervised ML models for AD prediction are quantitatively evaluated using evaluation metrics such as average Accuracy, AUC, Recall, F1 Score, Kappa, MCC and TT (Training Time). Table I shows the performance score for each trained model. From the scoring grid, it can be concluded that the best results are achieved by the Naïve Bayes (NB) algorithm. Eventhough, Ridge Classifier (Ridge), Random Forest Classifier (RF), Linear Discriminant analysis (LDA) and ExtraTrees Classifier (ET) also have produced same accuracy that of NB, considering all other metrics together, NB came in the top list.

Accuracy measures the proportion of correct predictions made by the model and helps to understand and compare the models in a simpler way. Since the dataset is imbalanced, AUC (Area Under the ROC Curve) is useful as it is independent of class distribution. Recall identifies all relevant cases especially when missing important positives are costly in our problem. F1 score combines and balances the strengths of precision and recall. Kappa values are more robust than accuracy in imbalanced datasets. MCC is also a very important metric for imbalanced dataset which considers four types of predictions- True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN)- thereby providing a more comprehensive evaluation. Both F1 score and MCC gives a balanced view of precision and recall. A combination of these metrics provided by the PyCaret library gives a more complete picture of each model's performance. Phase I classifiers made use of all the features provided by the dataset for the prediction of AD.

Table 1. Analysis of Classifiers in Phase I

Model	Accuracy	AUC	Recall	Precision	F1 Score	Kappa	MCC	TT (Sec)
Naive Bayes (nb)	0.9089	0.9350	0.8150	1.0000	0.8885	0.8169	0.8388	0.0140
Ridge Classifier (ridge)	0.9089	0.9200	0.8150	1.0000	0.8885	0.8169	0.8388	0.0220
Random Forest Classifier (rf)	0.9089	0.9420	0.8350	0.9800	0.8935	0.8169	0.8333	0.1080
Linear Discriminant Analysis (lda)	0.9089	0.9160	0.8150	1.0000	0.8885	0.8169	0.8388	0.0270
Extra Trees Classifier (et)	0.9089	0.9395	0.8350	0.9800	0.8935	0.8169	0.8333	0.1000
Gradient Boosting Classifier (gbc)	0.9000	0.9200	0.8800	0.9348	0.8959	0.8000	0.8175	0.090
Logistic Regression (lr)	0.8889	0.8730	0.8150	0.9550	0.8713	0.7769	0.7925	1.4120
Light Gradient Boosting Machine (lightgbm)	0.8789	0.9120	0.8350	0.9267	0.8692	0.7569	0.7728	0.1810
Extreme Gradient Boosting (xgboost)	0.8689	0.9120	0.8350	0.9067	0.8603	0.7369	0.7511	0.0550
Ada Boost Classifier (ada)	0.8589	0.9200	0.8350	0.8900	0.8512	0.7169	0.7328	0.0650
Decision Tree Classifier (dt)	0.8367	0.8375	0.8550	0.8331	0.8363	0.6741	0.6879	0.0140
Quadratic Discriminant Analysis (qda)	0.7589	0.7830	0.8750	0.7687	0.7925	0.5169	0.5474	0.0220

Model	Accuracy	AUC	Recall	Precision	F1 Score	Kappa	MCC	TT (Sec)
K Neighbors Classifier (knn)	0.6344	0.6990	0.6500	0.6716	0.6445	0.2702	0.2673	0.0250
SVM - Linear Kernel (svm)	0.5056	0.5000	0.1000	0.0500	0.0667	0.0000	0.0000	0.0170
Dummy Classifier (dummy)	0.5056	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0180

Table 2. Analysis of Classifiers in Phase II

Model	Accuracy	AUC	Recall	Precision	F1 Score	Kappa	MCC	TT (Sec)
Naive Bayes (nb)	0.9575	0.9932	0.9458	1.0000	0.9721	0.9014	0.9017	0.0040
Ridge Classifier (ridge)	0.9471	0.9411	0.9397	1.0000	0.9689	0.8785	0.8785	0.0050
Random Forest Classifier (rf)	0.9357	0.9871	0.9257	1.0000	0.9614	0.9143	0.9152	0.0870
Extra Trees Classifier (et)	0.9214	0.9872	0.9016	1.0000	0.9482	0.8952	0.8954	0.0980
Linear Discriminant Analysis (lda)	0.9281	0.9660	0.8980	1.0000	0.9462	0.8649	0.8482	0.0120
Gradient Boosting Classifier (gbc)	0.9176	0.9783	0.8786	0.9689	0.9215	0.8354	0.8522	0.0210
Logistic Regression (lr)	0.9105	0.9668	0.8786	0.9624	0.9185	0.8206	0.8319	1.000
Light Gradient Boosting Machine (lightgbm)	0.9024	0.9143	0.8365	0.9532	0.8910	0.8091	0.8117	0.0820
Ada Boost Classifier (ada)	0.8871	0.8914	0.8333	0.9441	0.8852	0.7602	0.7512	0.0150
Extreme Gradient Boosting (xgboost)	0.8754	0.9479	0.8459	0.9246	0.8835	0.7932	0.7864	0.0050
Decision Tree Classifier (dt)	0.8501	0.9707	0.8120	0.8696	0.8398	0.8037	0.8098	0.0040
Quadratic Discriminant Analysis (qda)	0.8097	0.8787	0.8495	0.8640	0.8566	0.6115	0.6254	0.0010
KNeighbors Classifier (knn)	0.7157	0.7143	0.7714	0.8230	0.7963	0.4298	0.4471	0.0050
SVM - Linear Kernel (svm)	0.5229	0.5205	0.5839	0.7064	0.6393	0.0406	0.0366	0.0070
Dummy Classifier (dummy)	0.5133	0.5000	0.1000	0.5133	0.1673	0.0000	0.0000	0.0080

After getting explanations from SHAP for the Phase I results, certain irrelevant features are removed from the original dataset and created a new dataset called prime feature dataset. Now, this acts as the input features for Phase II classifiers. Table II shows the performance of classifiers in Phase II. It is

experimentally proved that all the classifiers in Phase II has shown an improved performance along all the metrics.

Table 3. Performance Accuracy Comparison of Top Ten Classifierswith All features and Only Prime Features Included

Model	Accuracy in % with all features included (Phase I)	Accuracy in % with only prime features included (Phase II)
Naive Bayes (nb)	91	96
Ridge Classifier (ridge)	91	95
Random Forest Classifier (rf)	91	94
Linear Discriminant Analysis (lda)	91	93
Extra Trees Classifier (et)	91	92
Gradient Boosting Classifier (gbc)	90	92
Logistic Regression (lr)	89	91
Light Gradient Boosting Machine (lightgbm)	88	90
Extreme Gradient Boosting (xgboost)	87	88
Ada Boost Classifier (ada)	86	89

Because accuracy is easy to interpret, calculate, understand and gives the general picture of the model, we use the performance accuracies of top 10 models from each phase for a quick comparison and is depicted in Table III. The formula to calculate Accuracy= (Number of correct predictions)/(Total number of predictions). i.e.,

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (2)$$

It can be observed that all the models in Phase II have shown an improved performance accuracy than in Phase I.

5.2 SHAP Explanations

Model agnostic explanations for the classification models are provided by SHAP. We can see important patterns in the behaviour of the model by utilising SHAP to create several plots. The best classifier among the PyCaret classifiers in both Phase I and Phase II is NB and is considered as the Consistent Model. So, SHAP explainability plots for the NB classifier is depicted here, though plots for other models can also be drawn. SHAP is used for both global and local explanation of the decisions taken by NB in AD.

5.2.1 Global Explaination

The impact of each feature on the model output is shown in Fig. 3 using SHAP Feature Importance plot.The shapley values of features across all the observations are aggregated and then sorted. Feature Importance plot lists the features in the descending order of their importance. As top features have higher shapley values, they contribute more to the prediction compared to the bottom ones and hence they have high predictive power.

Here, the feature CDR has more impact on the model than any other features whereas Hand and Visit has the least importance. It is also obvious that the impact of features on the classes Demented and Non-Demented are equal.

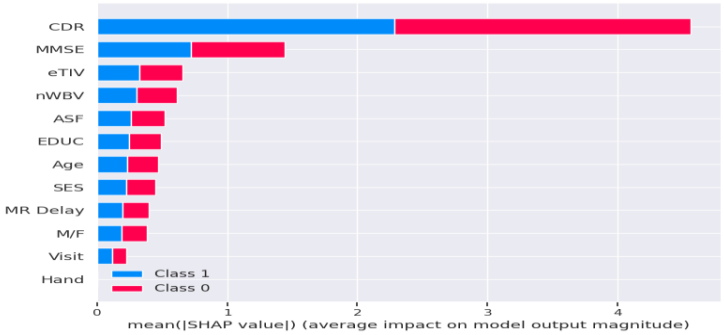

Fig 3. Feature Importance SHAP plot for AD prediction

The SHAP summary plot in Fig. 4 reveals the users both the positive as well as negative relationships of all the input features with the target variable. It displays the average impact of each feature on the model's prediction. The Feature Names (X-axis) lists all features used by the model. Dot Position (Y-axis) represents the feature's average impact on the model's prediction. All the observations in the AD training dataset are shown as dots in this plot. Features with higher absolute values (further from zero) have a stronger influence. This plot allows us to quickly identify the most influential features for the model's overall predictions.Here, the features are ranked in descending order of their importance. Color Bar indicates the feature effect. Red represents a positive influence (increasing the prediction value), and blue represents a negative influence (decreasing the prediction value). i.e., red color indicates high feature value and blue color indicates a low value of the feature for that observation. From the plot we can interpret that a high value of CDR has a positive as well as high impact on AD prediction (Red color indicates a high value and the position in X-axis shows its positive impact).

Fig 4. SHAP Summary Plot for AD Prediction

5.2.2 Local Explanation

When a model makes a prediction, it is for the entire dataset. To make the predictions more sensible, each individual observation needs to be explained. This makes the model more adoptable. Individual SHAP value plots for observations aids us to explain each observation in detail. The mean value of all the features in the training dataset of AD is shown in the Table IV. Shapley values of each feature will be calculated and compared against their mean values before the model takes a decision.

Table 4. Mean Values of the Features of AD Training Dataset

Sl. No.	Features	Mean Values
1	Visit	1.870445
2	MR Delay	591.445344
3	M/F	0.408907
4	Age	76.744939
5	EDUC	14.619433
6	SES	2.554656
7	MMSE	27.303644
8	CDR	0.273279
9	eTIV	1481.214575
10	nWBV	0.731158
11	ASF	1.200911
12	Hand	1.000000

SHAP value plots are a visual representation of feature contributions to a model's prediction for a single data point. They depict features as forces pulling the prediction in a specific direction (positive or negative) based on their individual impact. The strength of the force corresponds to the feature's importance in influencing the prediction. This plot uses arrows to visualize features as forces, with the

length and direction of the arrow indicating the feature's influence.Two observations from the AD training dataset – one from demented category and the other from non-demented category- are shown below and interpreted using individual SHAP value plots.

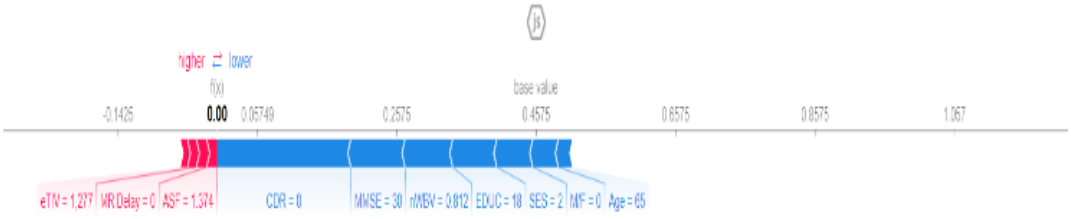

Fig 5. SHAP Value Plot for AD Prediction- Non-Demented Class

Fig. 5 shows individual SHAP value plot for obseravtion 1 in the training dataset. The ouput value 0 in the plot represents the prediction for this observation, which means that this person is classified as non-demented. The base value represents the mean of the model output. Red color shows the variables which drives the prediction higher and blue color indicates the fetaures which drives the prediction lower. Here, the features CDR, MMSE, nBW, etc. in blue color have contributed for classifying the observation as a negative class (non-demented). The feature CDR correlates positively to the prediction and its value 0 which is lesser than the mean value 0.273279 drives the prediction negatively. The variable EDUC negatively corelates to prediction and hence its value 18 which is greater than the mean value 14 pushed it towards the class non-demented.

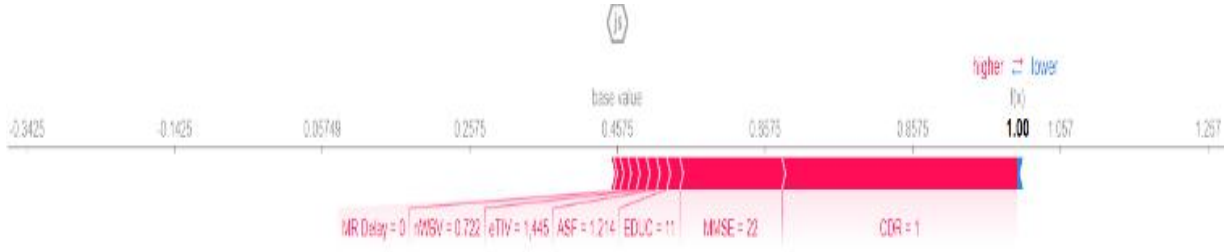

Fig 6. SHAP Value Plot for AD Prediction- Demented Class

Consider the individual SHAP value plot for another observation -Observation 2 which is depicted in Fig. 6. It has an output of 1 which means that the patient is classified as demented. Variable CDR is having value 1 which is greater than 0.273279, its mean value. CDR has a positive impact on dementia classification and pushes prediction to the right in this case. The variables shown in red color, MMSE, EDUC, nWBV, etc. have low values compared to their mean values. As these features are negatively corelated to the label, their low values have pushed the prediction to the right. Similarly, the individual SHAP value plot for all observations can be produced and hence local interpretability can be achieved.

DISCUSSIONS

In this section, the performance metrics and SHAP interpretations are discussed and the performance accuracies in Phase I and Phase II are evaluated.

While our PyCaret model successfully predicts AD, understanding the rationale behind these predictions is crucial. This is where SHAP comes in. SHAP provides a framework for interpreting the impact of individual features on the model's predictions. The performance score of classifiers of PyCaret library in training and classification of AD in Phase I and Phase II is given in Table I and II respectively. In Phase I, the classification is done on the original dataset, whereas in Phase II, the classification is made on prime feature dataset, which is obtained by removing irrelevant features from the original dataset with the help of SHAP values. So, by using SHAP as a feature selection approach, it not only gives explanation but also shows important features required for the prediction task. Among the 18 classifiers, NB exhibited a

consistent performance in both the Phases. In Phase I, NB exhibited 91% accuracy in AD classification whereas in Phase II, NB could produce an accuracy of 96%. The same model predicted AD with different accuracies when the number of features differed. This indicates that with reduced number of features and including only the relevant features, a model could improve the prediction accuracy. From the score grid it can also be observed that all the classifiers could improve their performance in terms of all the metrics.

The prime thing to be noticed from Table III is that the classification accuracy of all the ML models in Phase II has been improved and shown a supreme performance than in Phase I. NB improved its accuracy score from 91% to 96%, Ridge from 91% to 95%, RF from 91% to 94%, LDA from 91% to 93% and so on. High accuracy is definitely a good metric for evaluating a model which indicates that both individuals with and without AD is identified effectively. Considering the imbalanced nature of the dataset, we also analyze other metrics. AUC measures the ability of a model to discriminate between demented and non-demented cases. Models with AUC values closer to 1 can be considered as effective models. Recall identifies the proportion of TP identified by the model. High Recall avoids false negative cases and signifies that the model could capture all the patients with AD. High Precision enhances the accurate identification of AD minimizing false positives. F1-Score gives a balanced metric between precision and recall and the models with a score closer to 1 indicates that they are the best ones. Kappa considers the agreement occurring by chance and is more robust than accuracy for imbalanced datasets. MCC is a balanced measure as it considers all 4 cases- TP, TN, FP and FN. It is a valuable addition to imbalanced datasets for comprehensive evaluation. TT (Sec) represents the total time taken by the ML model to train on the dataset. This column allows us to compare the computational efficiency of different classifiers. By analyzing Table I and II, it is vivid that all the scores mentioned here has got an improved value for all the classifiers in Phase II compared to Phase I. Hence, we can say that all the classifiers performed well in SHAP reduced feature dataset than in the original dataset and hence SHAP can be considered as a feature selection technique.

Furthermore, the classification results by the classifiers in Phase I are explained by computing the SHAP values of each predictors. In section 4, the benefits of SHAP to interpret models in AD is demonstrated. Both global and local interpretation to the NB model is achieved by SHAP plots. SHAP explains predictions of ML modelsbased on shapley values. These values appraise the contribution of each variable in the model. Global explanation for the model is achieved through feature importance plots and summary plots. When we analyze feature importance SHAP plot, Fig. 3, it can be noticed that the features in the SHAP plot are ranked in descending order and hence the top features in the SHAP plot contribute more to the model and have high predictive power compared to the bottom features. This is done using shapley values which indicates the marginal contribution of a given feature for the prediction. This plot summarizes the average SHAP value for each feature across all patients. Positive values indicate the feature generally pushes the prediction towards AD (higher probability), while negative values suggest it pushes the prediction towards healthy (lower probability). Features with higher absolute SHAP values (further from zero) have a stronger influence on the model's overall predictions. So, the feature position in horizontal location shows if the value results in a lower or higher prediction. From Fig. 3, we can conclude that CDR, MMSE, eTIV, nWVB and ASF are the features which have contributed more to the model prediction.

SHAP Summary plot in Fig. 4 also shows variable importance. SHAP Summary plot is made of all instances in the train data. These plots take the mean absolute value of each feature over all observations of the dataset to create a global measure of feature importance. SHAP plot shows the positive and negative relationship of the features with the label. Red color indicates a high value and blue color indicates a low value of the feature for a particular observation. Feature importance plot and SHAP summary plot gives global interpretation to the model. These plots are helpful for explaining the output of any ML model. Fig. 4 explains the feature impact in AD prediction as that CDR, eTIV, nWBV, EDUC and MMSE are the top 5 features which contributes more for the prediction whereas the feature Hand contributes very less for the prediction.

SHAP value discovers patterns in data and explains model decisions. Feature importance deduced from SHAP values can be inferred for feature selection. By using SHAP, a major preprocessing technique- feature selection- benefits from explainability. SHAP explanations are supported by mathematical formulae. It is also established that when a model classifies a dataset after SHAP feature selection, the model could improve its performance accuracy. Also, finding out the most important features from feature importance plots and removing irrelevant features from the dataset, reduces computational time and ends in faster results so that the patients doesn't need to wait for a long time. To provide local explanation to the model, individual SHAP value plots are employed. This plot explores the relationship between a specific feature and the model's prediction for a single patient. It reveals how changes in that feature value affect the predicted probability of AD. Each individual observation's prediction can be explained to make sense of it.They are particularly useful for analyzing and explaining a specific

prediction you're interested in, providing a deeper dive into its reasoning.Fig. 5 and Fig. 6 explains two observations among which one is in demented category and the other one is non-demented. The horizontal line in these plots marks the base value of the model. The feature values are also displayed next to the line. Features that cause increase in the prediction are represented in red color, while those that decrease in prediction are depicted in blue. When a feature exhibits a positive correlation with the label, having a value greater than its mean will make a positive contribution to the prediction. Likewise, when a feature demonstrates a negative correlation with the label, having a value greater than its mean will result in a negative contribution to the prediction.

Let us interpret the individual SHAP plot, Fig.5, drawn for an observation which is classified as non-demented. The features such as CDR, MMSE, nWBV, ASF, EDUC, SES, Age and eTIV contributes more for the prediction. Among these features, CDR, Age, MMSE, nWBV, ASF and eTIV are positively corelated to the prediction, i.e., as the value of these features increases it push the prediction towards the class demented. Rest of the features are negatively corelated to the model output, which means that when the value of these features decreases, it drives the prediction towards the class demented. The absolute shap values of these features will be compared against their mean values which is shown in Table 2. In the Fig. 5, it can be seen that the values for CDR (=0), MMSE (=30), nWBV (=0.812), Age (=65) are all lesser than their mean value, and EDUC (=18) is greater than its mean value. That means these variables contributes to make the prediction as non-demented. For eg, a high value in EDUC contributes negatively to AD prediction, which states that as the number of years of education increases, the chance for AD decreases. Similarly, Fig. 6 provides the explanation for a Demented class. As EDUC is negatively corelated to dementia and has a value 11, which is lesser than the mean value, this feature pushed the observation to the class demented. Likewise, the values of CDR (=1), MMSE (=22), ASF (=1.214), eTIV (=1,445), nWBV(=0.722) are higher or very closer to their mean values and hence pushed the prediction towards the class demented. Thus, SHAP helps explain why the model predicts AD for a specific patient by revealing which features contributed most significantly to that prediction.Shap value plots are beneficial toShow a large number of feature effects clearly, display the cumulative effect of interactions, explore feature effects for a range of feature values, identify outliers and to identify typical prediction paths.

Both the accuracies and inaccuracies from the model disseminates some explanation which makes the users or clinicians understand more about the model. The reason behind NB model classifying a person into demented or non-demented category is clearly understood from the plots. Like so, SHAP transforms any black-box model into glass-box model. By incorporating feature selection method in model classification and explaining the predictions improves the prognosis of AD. This approach can be effectively used to predict any diseases and explain the results. The proposed framework can be extended to use in other domains also where model transparency is vital such as loan approval, detection of spam emails, etc.

Comparison of the proposed method with SOTA methods

Our model, with a broader focus on overall disease prediction, might prioritize different features compared to the State-of-the-Art (SOTA) methods discussed in the Literature Review. The SHAP analysis in our framework revealed that features like CDR, MMSE, eTIV, nWVB and ASF emerged as the most influential features for predicting AD. While there is some overlap with existing research, some key differences are noteworthy.Previous research on Alzheimer's disease prediction has primarily focused on feature selection techniques like filter, wrapper, and embedded methods. This paper presents a novel approach that utilizes SHAP for identifying the most significant features.Additionally, while numerous ML algorithms exist, implementing and comparing each one can be time-consuming. This work leverages PyCaret, a low-code library that streamlines the process. PyCaret offers a comprehensive suite of pre-built ML algorithms, enabling rapid experimentation and providing a variety of performance metrics for evaluation.

Out of all the reviewed articles, six articles used ML or ensembled techniques[10], [12], [15-17], [19], [21] and five articles employed DL or transfer learning methods[11], [14], [18], [20], [22]for AD prediction. It is worth noting that five articles [18-22] leveraged various explainable methods alongside DL or ensemble approaches and only three articles [23-25] utilized PyCaret for their analysis. No prior research has combined PyCaret's efficient exploration of ML algorithms with SHAP for feature selection and interpretability in the context of AD prediction. This work fills this gap by proposing a novel and insightful methodology.

Table V summarizes the performance of previous studies on AD prediction. The proposed model outperforms these methods in terms of accuracy, precision, recall, F1-score, and interpretability. These

findings suggest that the proposed XAI framework, combining PyCaret's efficient exploration and SHAP's interpretability, has the potential to offer a reliable and explainable approach to AD prediction.

Table 5. Comparison of the Proposed Method with SOTA Methods

Research Study	Year	Proposed Model	Accuracy (%)	Dataset	Explainability method
Proposed Pycaret method	2024	NB (Pycaret)	96	OASIS	SHAP
[22]	2024	EfficientNetB7	96	ADNI	CAMs
[15]	2023	Voting Classifier	96	OASIS	NO
[16]	2023	Gradient Boosting Tree	93.9 (AUC)	EHR data from OneFlorida+ Clinical Research Consortium	SHAP
[14]	2023	Ensemble Method	91	ADNI	SHAP
[17]	2023	Voting Classifier	90	ADNI	NO
[19]	2022	XGBoost	84.2	ADNI	SHAP
[12]	2022	RF	83	OASIS	NO
[20]	2022	CNN+ multi-input+ self-attention mechanism+ soft labelling	81	Live Patients from King Chulalongkorn Memorial Hospital	GradCAM
[21]	2022	XGBoost	80.7	Live Patients at Samsung Medical Center	ICE+ SHAP
[23]	2022	Ensemble	74.9	ADNI	NO
[24]	2022	XGBoost	65.14	From the Second Affiliated Hospital of Hangzhou Normal Uni-versity	SHAP
[18]	2021	RF	93.95	ADNI	SHAP
[11]	2021	CNN SVM	78.5 75.4	ADNI	NO
[10]	2021	RELM	77.62 (for binary classification)	ADNI	NO

The major accomplishments of this paper are:

- Achieves high classification accuracy (96%) for AD prediction using Naive Bayes classifier in the OASIS dataset.
- Integrates SHAP for interpretability, addressing the "black-box" challenge of ML models.
- Utilizes SHAP values for feature selection, identifying the most significant features for AD prediction.
- Offers a transparent and explainable framework for early AD diagnosis, aiding clinicians in decision-making.

CONCLUSION

Early detection of AD slows down its progression and the patients can lead a healthy life. ML models have been adopted in AD prediction and is increasingly common in predicting many other diseases. But the major challenge lies in the interpretability of the model which limits the adoption of the model.

Advancements in XAI techniques helps us to peep into the black-box models. In the proposed XAI framework, different ML models are analyzed using PyCaretfor predicting AD at an early stage and a popular interpretability technique, SHAP is employed for the model explanation. Instead of using individual ML models separately PyCaret gives a list of important and mostly used classifiers under the umbrella which makes the users convenient to use with less code and time. The models are evaluated on Brain MRI OASIS dataset using metrics such as Auuracy, Precision, Recall, AUC, F1-Score etc. In the experiment, best and consistent results are provided by NB among the other classifiers in the PyCaret library. SHAP, a model agnostic approach, provides robust explanation for the model on the basis of solid algorithms by integrating game theory. ML and SHAP is an exceptional duo which identifies important features and explore the relationships among data.

In this work, SHAP is also analyzed as a feature selection method which prioritize and picks features based on their contribution to model output. The experimental results show that, models which used datasets after feature selection using SHAP performed superior. Hence, SHAP can be considered in the future as a preprocessing approach in other domains too where model interpretability is crucial. SHAP also helps to build trust in the ML models by making the model decisions more transparent to the clinicians. It is a step towards a future where AI empowers healthcare professionals in the fight against AD.

We also note the directions for future work. First, Although the PyCaret library is promising in terms of performance, flexibility and ease of use, it is no way perfect. Second, the experiments for AD prediction should also be conducted in image datasets with multiclass classification. Future research can explore even more sophisticated XAI techniques to gain a deeper understanding of the disease and identify potential targets for intervention. Finally, instead of post-hoc filtering, incorporating causal knowledge helps to determine the cause and effect relationships.

Author Contributions

Archana Menon P conceived the original idea, implemented the frame work, performed analytical calculations and wrote the manuscript. Dr. R. Gunasundari was involved in planning, verifying the results, reviewing the manuscript and supervising the entire work. Both the authors discussed the results and contributed to the final manuscript.

REFERENCES

[1] Masters, C., Bateman, R., Blennow, K. et al. Alzheimer's disease. Nat Rev Dis Primers 1, 15056 (2015). https://doi.org/10.1038/nrdp.2015.56

[2] Alzheimer’s and dementia,Alzheimer’s Disease and Dementia. Available at: https://www.alz.org/alzheimer_s_dementia (Accessed: 17 July 2023).

[3] Alzheimer's Association National Plan Care and Support Milestone Workgroup, et al. "Report on milestones for care and support under the US National Plan to Address Alzheimer's Disease." Alzheimer's & Dementia 12.3 (2016): 334-369.

[4] Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019, 9, e1312.

[5] Watson, David S., et al. "Clinical applications of machine learning algorithms: beyond the black box." Bmj 364 (2019).

[6] Sheu RK, Pardeshi MS. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors (Basel). 2022 Oct 21;22(20):8068. doi: 10.3390/s22208068. PMID: 36298417; PMCID: PMC9609212.

[7] K. Tejeswinee, Gracia Jacob Shomona, R. Athilakshmi, Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer’s And Parkinson’s Disease, Procedia Computer Science, Volume 115, 2017, Pages 188-194, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.09.125.

[8] Muhammed Niyas K.P., Thiyagarajan P., Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer’s classification, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 8, Part A, 2022, Pages 4993-5006, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.12.009.

[9] Alshamlan H, Omar S, Aljurayyad R, Alabduljabbar R. Identifying Effective Feature Selection Methods for Alzheimer's Disease Biomarker Gene Detection Using Machine Learning. Diagnostics (Basel). 2023 May 17;13(10):1771. doi: 10.3390/diagnostics13101771. PMID: 37238255; PMCID: PMC10217314.

[10] M. Sudharsan, G. Thailambal, Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA), Materials Today: Proceedings, Volume 81, Part 2, 2023, Pages 182-190, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.03.061

[11] Grueso, S., Viejo-Sobera, R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alz Res Therapy 13, 162 (2021). https://doi.org/10.1186/s13195-021-00900-w

[12] Kavitha C, Mani V, Srividhya SR, Khalaf OI and Tavera Romero CA (2022) Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Front. Public Health 10:853294. doi: 10.3389/fpubh.2022.853294

[13] Bogdanovic, B., Eftimov, T. & Simjanoska, M. In-depth insights into Alzheimer's disease by using explainable machine learning approach. Sci Rep 12, 6508 (2022). https://doi.org/10.1038/s41598-022-10202-2

[14] Kasula, Balaram Yadav. "A Machine Learning Approach for Differential Diagnosis and Prognostic Prediction in Alzheimer's Disease." International Journal of Sustainable Development in Computing Science [Online], 5.4 (2023): 1-8. Web. 10 Jun. 2024.

[15] Uddin, K.M.M., Alam, M.J., Jannat-E-Anawar et al. A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. Biomedical Materials & Devices 1, 882–898 (2023). https://doi.org/10.1007/s44174-023-00078-9

[16] Li Q, Yang X, Xu J, Guo Y, He X, Hu H, Lyu T, Marra D, Miller A, Smith G, DeKosky S, Boyce RD, Schliep K, Shenkman E, Maraganore D, Wu Y, Bian J. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimers Dement. 2023 Aug;19(8):3506-3518. doi: 10.1002/alz.12967. Epub 2023 Feb 23. PMID: 36815661; PMCID: PMC10976442.

[17] Raffaella Franciotti, Davide Nardini, Mirella Russo, Marco Onofrj, Stefano L. Sensi, Comparison of Machine Learning-based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment, Neuroscience, Volume 514, 2023, Pages 143-152, ISSN 0306-4522, https://doi.org/10.1016/j.neuroscience.2023.01.029 .

[18] El-Sappagh, S., Alonso, J.M., Islam, S.M.R. et al. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep 11, 2660 (2021). https://doi.org/10.1038/s41598-021-82098-3

[19] Bogdanovic, B., Eftimov, T. & Simjanoska, M. In-depth insights into Alzheimer's disease by using explainable machine learning approach. Sci Rep 12, 6508 (2022). https://doi.org/10.1038/s41598-022-10202-2

[20] Ruengchaijatuporn, N., Chatnuntawech, I., Teerapittayanon, S. et al. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. Alz Res Therapy 14, 111 (2022). https://doi.org/10.1186/s13195-022-01043-2

[21] Chun MY, Park CJ, Kim J, Jeong JH, Jang H, Kim K, Seo SW. Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. Front Aging Neurosci. 2022 Aug 5;14:898940. doi: 10.3389/fnagi.2022.898940. PMID: 35992586; PMCID: PMC9389270.

[22] Jahan, S., Saif Adib, M.R., Mahmud, M., Kaiser, M.S. (2023). Comparison Between Explainable AI Algorithms for Alzheimer's Disease Prediction Using EfficientNet Models. In: Liu, F., Zhang, Y., Kuai, H., Stephen, E.P., Wang, H. (eds) Brain Informatics. BI 2023. Lecture Notes in Computer Science(), vol 13974. Springer, Cham. https://doi.org/10.1007/978-3-031-43075-6_31

[23] Rye, I., Vik, A., Kocinski, M. et al. Predicting conversion to Alzheimer's disease in individuals with Mild Cognitive Impairment using clinically transferable features. Sci Rep 12, 15566 (2022). https://doi.org/10.1038/s41598-022-18805-5

[24] Hu M, Yu Y, He F, Su Y, Zhang K, Liu X, Liu P, Liu Y, Peng G, Luo B. Classification and Interpretability of Mild Cognitive Impairment Based on Resting-State Functional Magnetic Resonance and Ensemble Learning. Comput Intell Neurosci. 2022 Aug 19;2022:2535954. doi: 10.1155/2022/2535954. PMID: 36035823; PMCID: PMC9417789

[25] Choi, Hyunchul, et al. "Cognitive Impairment Prediction Model Using AutoML and Lifelog." Journal of the Korea Society of Computer and Information, vol. 28, no. 11, 한국컴퓨터정보학회, Nov. 2023, pp. 53–63, doi:10.9708/JKSCI.2023.28.11.053

[26] Moez Ali, PyCaret: An open source, low-code machine learning library in Python, PyCaret 3.0 - Docs. (n.d.), 2020. https://pycaret.gitbook.io/docs/#citation.

[27] Antoniadi, Anna Markella, et al. "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review." Applied Sciences 11.11 (2021): 5088.

[28] Opening the blackbox:- Azodi, Christina B., Jiliang Tang, and Shin-Han Shiu. "Opening the black box: interpretable machine learning for geneticists." Trends in genetics 36.6 (2020): 442-455.
[29] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/
[30] Shapley, L. S.. "17. A Value for n-Person Games". Contributions to the Theory of Games (AM-28), Volume II, edited by Harold William Kuhn and Albert William Tucker, Princeton: Princeton University Press, 1953, pp. 307-318. https://doi.org/10.1515/9781400881970-018
[31] Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774
[32] Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021. https://doi.org/10.1155/2021/5581806
[33] Shardlow, Matthew. "An analysis of feature selection techniques." The University of Manchester 1.2016 (2016): 1-7.
[34] Ang JC, Mirzal A, Haron H, Hamed HN. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. IEEE/ACM Trans Comput Biol Bioinform. 2016 Sep-Oct;13(5):971-989. doi: 10.1109/TCBB.2015.2478454. Epub 2015 Sep 14. PMID: 26390495.
[35] Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci. 2010 Dec;22(12):2677-84. doi: 10.1162/jocn.2009.21407. PMID: 19929323; PMCID: PMC2895005.