Jurnal Kejuruteraan 36(4) 2024: 1713–1725
https://doi.org/10.17576/jkukm-2024-36(4)-34

Transparent Insights into Alzheimer's Progression: A Time-Aware Approach with Explainable AI

Abraham Varghese[a], Ben George[a], Vinu Sherimon[a*] & Prashanth Gouda[c]

[a]*College of Computing and Information Sciences, University of Technology and Applied Sciences, Muscat, Sultanate of Oman.*

[b]*University of Technology and Applied Sciences, Muscat, Sultanate of Oman*

[c]*College of Medicine, National University of Science and Technology, Sultanate of Oman*

[*]*Corresponding author: vinu.sherimon@utas.edu.om*

Received 26 March 2024, Received in revised form 26 March 2024
Accepted 26 April 2024, Available online 30 July 2024

ABSTRACT

Alzheimer's disease, acknowledged for its intricate and degenerative characteristics, presents considerable challenges, particularly among the elderly population. The relentless nature of Alzheimer's, marked by the gradual deterioration of cognitive function, underscores the urgency to develop effective strategies for early diagnosis and intervention. Artificial intelligence has played a significant role in terms of disease diagnosis and treatments. However, it has got limited acceptance in the medical community due to its lack of transparency. This study aims to advance the understanding and prediction of Alzheimer's disease progression through the integration of time-aware modeling and Explainable AI techniques. It makes significant contributions by addressing two key objectives in the context of Alzheimer's disease. First, by including temporal aspects, it accurately depicts the pace at which relevant predictors change over time, thereby capturing the dynamic nature of Alzheimer's disease. Second, by giving interpretable insights into the algorithm's decision-making process, the study hopes to empower researchers and physicians. This approach not only enhances transparency but also builds trust in the model's outcomes. The ADNI dataset, comprising 2980 observations, was employed for developing a prediction model using various machine learning classifiers. Among these classifiers, the Random Forest model emerged as the top performer, exhibiting superior accuracy, a high Coefficient of Determination (R2), and an impressive F1 score. To enhance interpretability, subsequent analyses utilized LIME and SHAP techniques. By combining time-aware modeling with Explainable AI methods, we seek to unravel the dynamic relationships within the dataset, providing transparent insights into the temporal evolution of Alzheimer's disease. Thus, this paper contributes to the creation of a clinically relevant and practical model for monitoring Alzheimer's disease progression that holds the potential for a deeper understanding of the evolving nature of the disease and paving the way for personalized and timely interventions.

Keywords: Alzheimer's disease; explainable AI; time-aware modeling; machine learning; LIME; SHAP

INTRODUCTION

Alzheimer's disease, a complex and degenerative condition, presents formidable challenges, particularly among the elderly demographic. The gradual erosion of cognitive function and memory associated with Alzheimer's underscores an urgent need for innovative strategies in early diagnosis and intervention (Dubois et al. 2009). While artificial intelligence (AI) has emerged as a promising tool in disease diagnosis and treatment, its limited acceptance in the medical community, attributed to a lack of transparency, calls for transformative approaches (Chun et al. 2022; Fabrizio et al. 2021). This research aims to improve our knowledge and prediction of the course of Alzheimer's disease to tackle these issues. The focus is on the combination of Explainable AI methods and time-aware modelling to create a complex and comprehensible model that takes time into account in addition to present values. The study seeks to uncover the dynamic interactions within the dataset with the rate of change in significant predictors across time. This approach provides visible understanding of the temporal evolution of Alzheimer's disease.

Alzheimer's disease (AD), which is signified by a persistent decrease in memory and cognitive function, continues to be a major global health concern. To lessen the burden on people and healthcare systems, it is imperative to create efficient ways for early diagnosis and intervention. With the use of advanced computational methods, artificial intelligence (AI) has become a ray of hope, helping us make sense of complex statistics and improve our knowledge of AD. Recent studies highlight the multifaceted role of AI, encompassing early detection methods and the tailoring of personalized treatment plans for individuals navigating the complexities of AD.

While AI-powered systems offer distinct competitive advantages, their inherent black-box nature raises concerns about transparency and the ability to elucidate decision-making processes. This challenge has prompted the emergence of explainable artificial intelligence (XAI), advocating for AI algorithms capable of revealing their internal processes and providing clarity on the rationale behind their decisions (Minh et al. 2022). Embedding explainable machines, particularly in healthcare, holds the potential to significantly streamline the repetitive tasks undertaken by medical professionals, allowing them to focus more on the interpretation of disease diagnoses. The challenge of black-box models, where predictions are highly accurate yet lack transparency in their internal mechanisms, has led to the development of various explainable artificial intelligence (XAI) frameworks (Adadi & Berrada, 2018). Notably, popular frameworks such as LIME (Local Model Agnostic Explanations) (Ribeiro et al. 2016), SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), and GradCAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017), among others, have found extensive use in addressing this challenge within the context of Alzheimer's disease (AD). Studies (Böhle et al. 2019; Gaur et al. 2022; Kamal et al. 2021; S. et al. 2021; Shad et al. 2021) provide a comprehensive list of works employing LIME, SHAP, LRP, and GradCAM methods and their combined approach.

The primary objective of this research is to contribute to the development of a more accurate and clinically relevant model for predicting Alzheimer's disease progression. This involves the integration of time-aware modeling and Explainable AI techniques to enhance the interpretability of the model. The study seeks to address critical questions: How can time-aware modeling and Explainable AI be synergistically employed to improve Alzheimer's disease progression prediction? What is the impact of incorporating both current values and the rate of change in key predictors over time on the development of an interpretable model?

The research intends to fill the existing gaps in our comprehension of Alzheimer's progression prediction and introduce a novel, temporally sensitive approach.

The goal of this research is to provide clinicians with better accurate and comprehensible tools by addressing the gaps in existing methods for predicting Alzheimer's disease. Time-aware modelling and Explainable AI are integrated with the goal of enhancing the clinical relevance of AI in Alzheimer's research. This approach may enhance our capacity to predict results and shed light on the intricate temporal dynamics of the disease. The significance of this research lies in its potential to aid in the creation of more effective, customized, and timely medications, ultimately improving the lives of Alzheimer's patients and their families.

The rest of the paper is organized as follows: Section 2 presents the methodology and the integration of Time-aware modelling with Explainable AI techniques. The results and analysis are presented in Section 3, and the implications of the findings are covered in Section 4. Section 5 provides a summary of the major contributions and future scope.

MATERIALS & METHODS

DATASET

The dataset includes 2980 observations and a wide range of predictors that were extracted from the ADNI dataset (*ADNI | ACCESS DATA*, n.d.). These predictors represent important clinical, genetic, demographic, and neuroimaging characteristics. A complete overview of the progression of a disease is provided by the representation of each predictor by both its current value and its rate of change over time. The full set includes genetic markers like APOE4 status, clinical assessments like CDRSB and ADAS scores, and demographic and educational data like age and education level. Brain structural volumes are included in neuroimaging measurements, and each variable has an associated rate of change. The target variable classifies the current disease stage into five categories: Normal (NL), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD), transition from NL to MCI, and transition from MCI to AD. This amalgamation of current values and dynamic change rates equips the dataset for advanced analyses, facilitating the prediction of Alzheimer's disease progression and the identification of potential transition patterns between different disease stages.

PROPOSED METHOD

An overview of the proposed method is given in Figure 1.

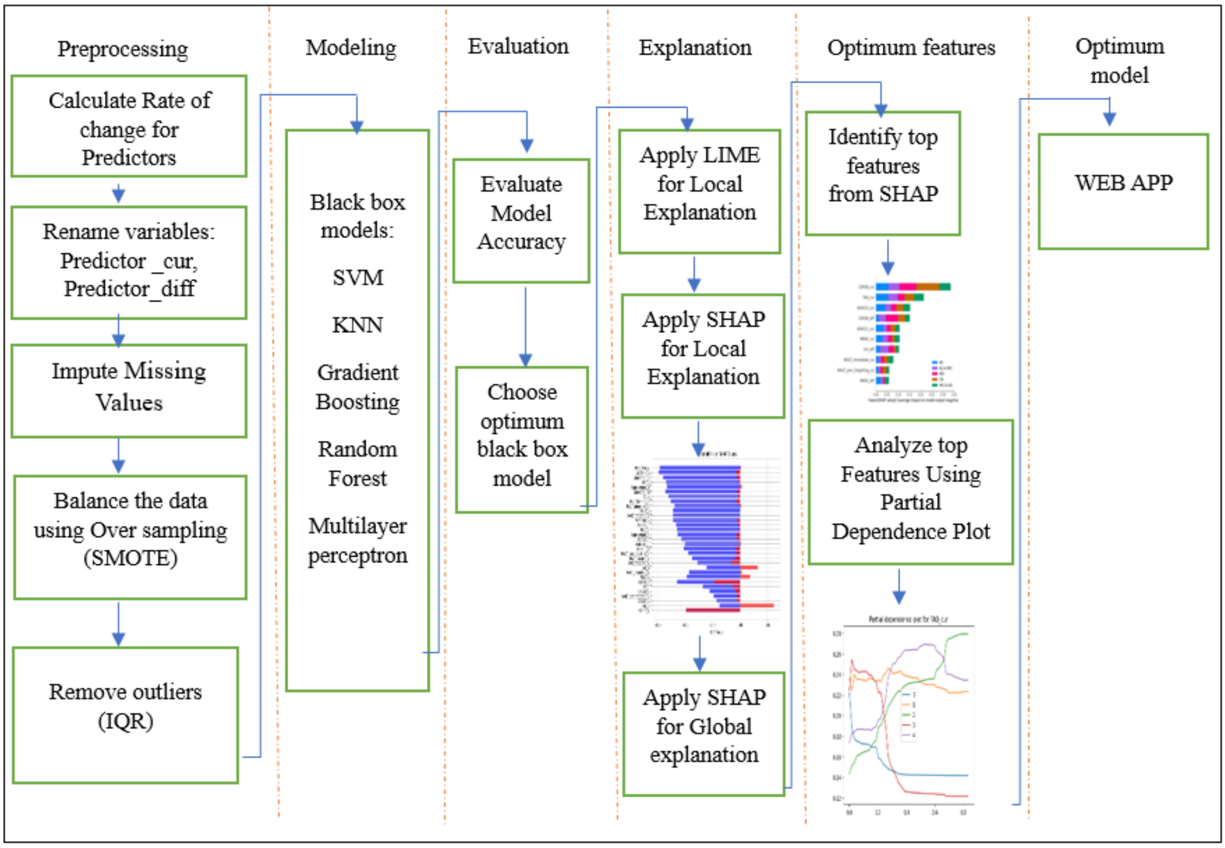

Preprocessing
Calculate Rate of change for Predictors
Rename variables: Predictor _cur, Predictor_diff
Impute Missing Values
Balance the data using Over sampling (SMOTE)
Remove outliers (IQR)
Modeling
Black box models:
SVM
KNN
Gradient Boosting
Random Forest
Multilayer perceptron
Evaluation
Evaluate Model Accuracy
Choose optimum black box model
Explanation
Apply LIME for Local Explanation
Apply SHAP for Local Explanation
Apply SHAP for Global explanation
Optimum features
Identify top features from SHAP
Analyze top Features Using Partial Dependence Plot
Optimum model
WEB APP

FIGURE 1. Overview of the proposed method

The methodology begins by calculating the rate of change for each predictor over different months, resulting in the creation of two versions for each variable: one denoting the current value (*predictor_cur*) and the other representing the rate of change (*predictor_diff*). Subsequently, missing values are imputed, and the dataset is balanced through oversampling, ensuring a representative distribution of classes. Outliers are then meticulously removed using the Interquartile Range (IQR) method to enhance the accuracy and reliability of subsequent analyses (Muslikh et al. 2023). To unravel the complex relationships within the dataset, various black-box models are employed, and their accuracies are systematically evaluated. The model exhibiting optimal accuracy is selected for further interpretability analysis. LIME, a tool for individual instance explanation, is applied to elucidate the predictions of the chosen model (Ribeiro et al. 2016). This process is replicated with SHAP (Lundberg & Lee, 2017), a method known for its global interpretability, to calculate feature importance for all instances. The most influential feature identified by SHAP is subjected to a thorough analysis using Partial Dependence Plots (PDP) (Jerome H., 2001), shedding light on its clinical relevance and contribution to disease progression. This comprehensive methodology integrates the strengths of black-box models with interpretable techniques, providing a multifaceted understanding of predictor dynamics and their impact on disease progression.

PREPROCESSING

In the preprocessing phase, the Alzheimer's disease progression dataset underwent meticulous transformations to enhance its suitability for predictive modeling. The inclusion of diverse predictors, such as demographic variables, cognitive assessments, and neuroimaging metrics, necessitated careful handling of missing values and normalization. To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) (N. V. et al. 2002) was applied, generating synthetic samples to augment the representation of underrepresented classes. Additionally, the original class labels representing Alzheimer's disease stages were replaced with more interpretable identifiers, facilitating a clearer understanding of the disease progression. These preprocessing steps collectively laid the groundwork for subsequent analyses, ensuring a robust and balanced dataset ready for the development of predictive models aimed at uncovering patterns in Alzheimer's disease progression. Figure 2 shows the distribution of 5 classes before and after applying oversampling.

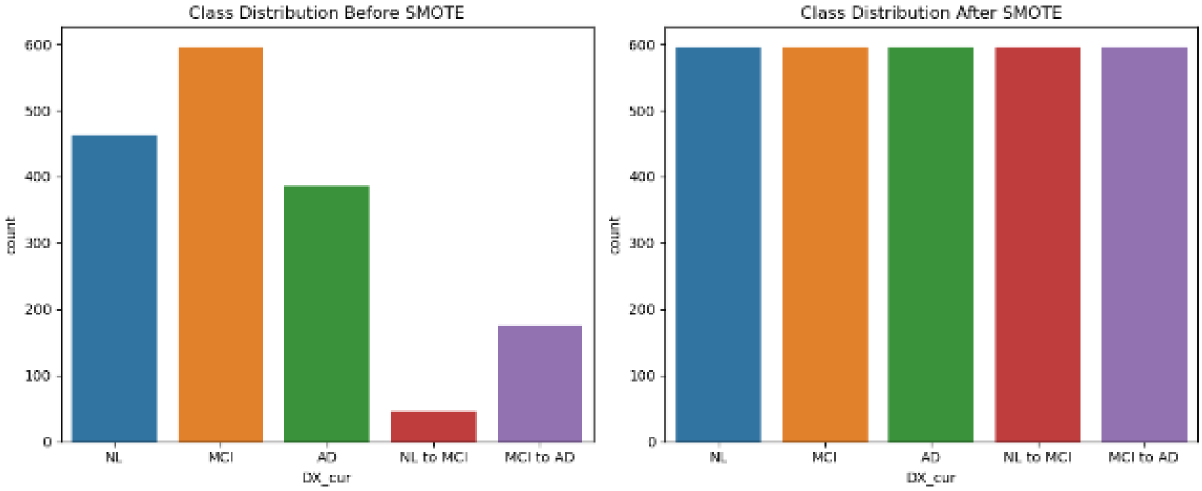


FIGURE 2. Class distribution before and after oversampling

The visualizations were created after employing a robust outlier removal technique based on the Interquartile Range (IQR) method. This method identifies outliers as data points that fall outside 1.5 times the interquartile range from the first and third quartiles. Subsequently, these outliers were systematically removed from each variable in the dataset. The resulting box plots exhibit a more accurate representation of variable distributions by excluding extreme values, providing a clearer insight into the central tendency, and spread. Similarly, the bar plots illustrating variable means have been refined, offering a more precise comparison among features after the removal of outliers. This rigorous outlier handling enhances the reliability of the visualizations, ensuring that the depicted patterns are not unduly influenced by extreme observations, and thereby fostering a more accurate interpretation of the dataset. Figure 3 gives the boxplot of the variables after removing the outliers.

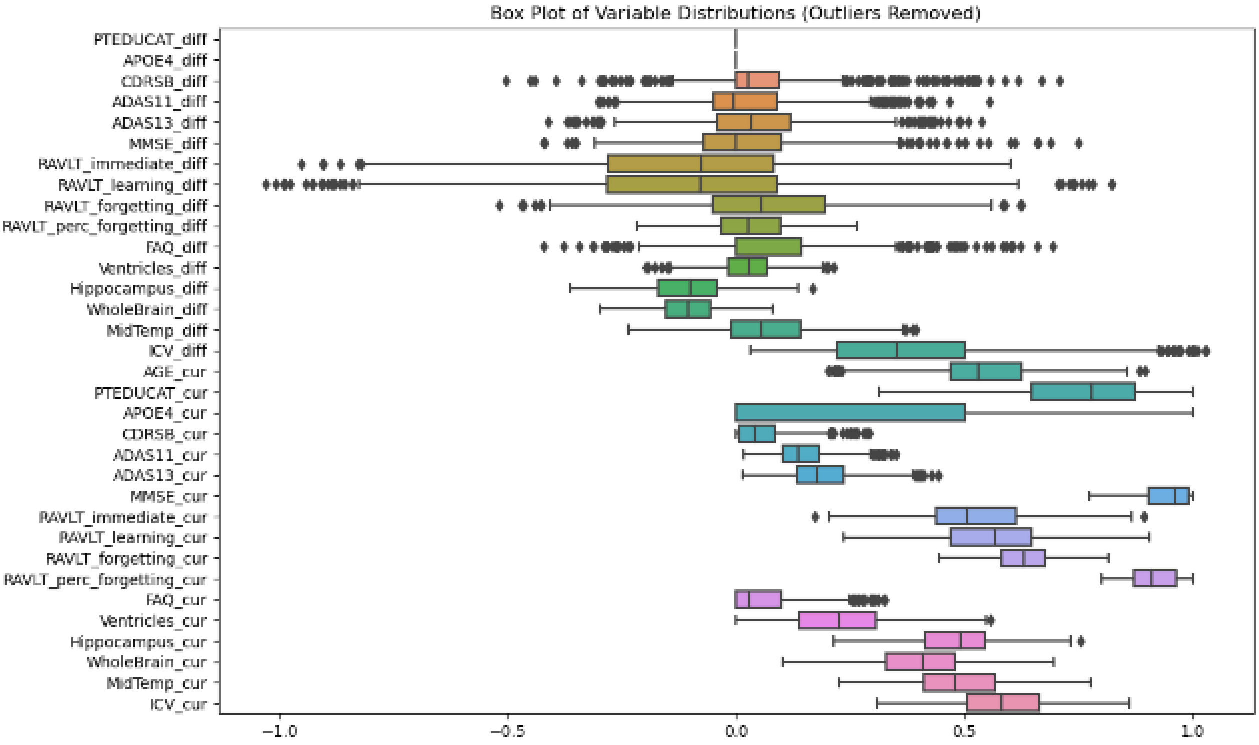


FIGURE 3. Boxplot of the predictors

MODELING AND EVALUATION

In this comprehensive modeling approach, an array of machine learning classifiers has been harnessed to tackle the intricacies of a multiclass classification task. The machine learning classifiers have been employed on three distinct sets of predictors: one associated with the current features (Predictor_cur), another linked to the rate of change features (Predictor_diff), and a combined set that incorporates both types of predictors. The process kicks off with essential data preprocessing steps, including the standardization of features to ensure uniform scaling. The classifiers, ranging from K-Nearest Neighbors (KNN) to Support Vector Machines (SVM), and encompassing Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP), undergo a meticulous optimization process.

GridSearchCV is a strategy that systematically investigates a range of hyperparameter combinations to achieve the most successful model configurations through hyperparameter tweaking. The GridSearchCV method is part of the popular Python machine learning toolbox, scikit-learn. This method, which is a component of scikit-learn's model selection module, is frequently used to tune hyperparameters by thoroughly searching a given hyperparameter grid. The KN and SVM classifiers are then trained using the optimal hyperparameters that have been chosen. Evaluation criteria, like accuracy, R-squared (R2) scores, and confusion matrices, are used to evaluate how well these models perform on a specific test dataset.

Other models, such as Random Forest, Gradient Boosting and MLP, are also used in addition to the basic classifiers to provide a thorough understanding of the relative efficacy of various techniques in addressing the multiclass classification problem. To observe and comprehend the discriminatory performance of the models, Receiver Operating Characteristic (ROC) curves are constructed. These curves highlight the trade-off between true positive and false positive rates, offering insights into each model's classification process. This multimodal method seeks to provide a thorough understanding of each classifier's relative strength and weaknesses in managing the intricacies present in multiclass classification scenarios, in addition to optimizing each classifier's performance. The resulting visualizations and metrics serve as invaluable tools for assessing the models' overall effectiveness and guiding further refinement for enhanced predictive capabilities.

EXPLAINABLE AI

Explainable AI (XAI) refers to the set of techniques and methods used to make the decisions and outputs of artificial intelligence (AI) systems understandable and interpretable by humans (Ribeiro et al. 2016). The goal of Explainable AI is to enhance transparency, trust, and accountability in AI systems, particularly in situations where the decision-making process might otherwise be considered a "black box."

LIME (Ribeiro et al. 2016): LIME is a model-agnostic technique employed in machine learning for rendering individual prediction interpretations from complex models. Mathematically, LIME's goal is to find a surrogate model g(z) (often linear) that approximates the behavior of the black box model f(x) in the vicinity of the instance x. π_x (z) are the sampling weights that measure the proximity of z to x. The weighted loss function is calculated as L(f,g,π_x)=∑▒〖π_x (z) [f(x)-g(z)]^2 〗, where π_x (z) is a weight assigned to each perturbed instance z.

In this implementation, local explanations for a user-selected instance are generated using LIME after a RandomForestClassifier classifier has been trained on a dataset. LIME creates an interpretable surrogate model by varying the input characteristics and tracking the resulting model responses. This provides insight into the significant variables and how they affect the model's choice. To improve the interpretability of black box models and to provide an improved comprehension of the local decision-making process, the results—actual class, predicted class, and LIME-generated explanations are provided.

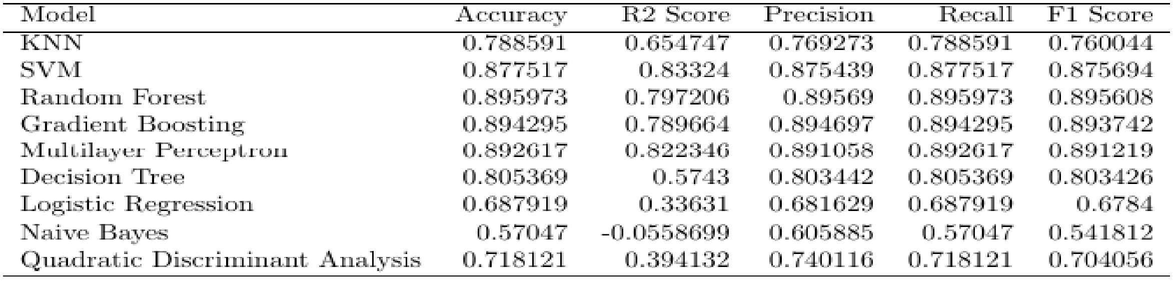
SHAP (Lundberg & Lee, 2017): A-n effective interpretability method called SHAP is intended to clarify the role of every feature in a model's output. Here, a RandomForestClassifier is trained on a dataset, and a TreeExplainer is used to calculate SHAP values. After the user chooses an example from the test set, SHAP values are produced to show how each characteristic affects the prediction made by the model. To provide a thorough understanding of feature relevance, the calculated SHAP values for individual features are displayed alongside the actual and predicted class labels. To further improve interpretability, a summary graphic is also created to show the overall influence of the features on the prediction.

Partial Dependence Plot (Jerome H., 2001): Visualization tools called Partial Dependence Plots (PDPs) are utilized to investigate the link between a particular feature and a machine learning model's anticipated result. The process is choosing one interesting feature, holding other variables constant, and examining how changes in its values affect the model's predictions. This is achieved by systematically varying the chosen feature and recording the average prediction across all instances. The resulting plot illustrates the impact of the selected feature on the model's output, providing valuable insights into its behavior and aiding in the interpretation of complex machine learning models.

RESULTS

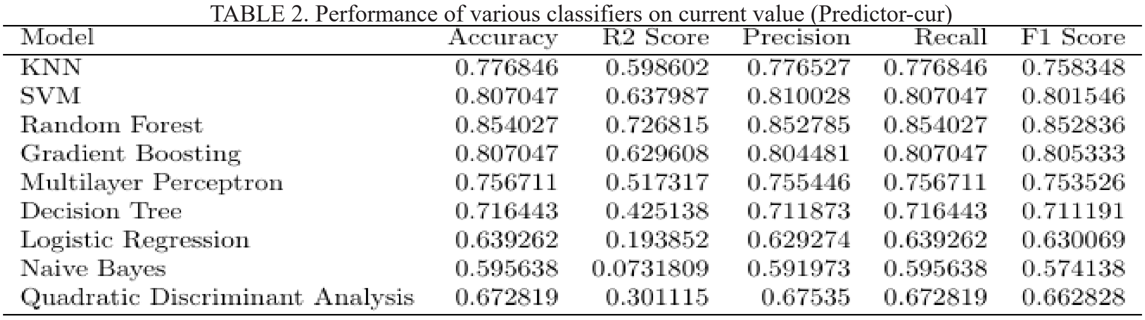
The presented results focus on evaluating the influence of incorporating temporal elements into the dataset. In the initial dataset, the rate of each predictor across distinct time periods is also integrated. We have assessed various classifiers on both the datasets, one containing temporal information and the other without. Tables 1 and 2 delineate the performance results of the classifiers under these conditions.

TABLE 1. Performance of various classifiers on entire dataset (Predictor-cur and Predictor-diff)



Model	Accuracy	R2 Score	Precision	Recall	F1 Score
KNN	0.788591	0.654747	0.769273	0.788591	0.760044
SVM	0.877517	0.83324	0.875439	0.877517	0.875694
Random Forest	0.895973	0.797206	0.89569	0.895973	0.895608
Gradient Boosting	0.894295	0.789664	0.894697	0.894295	0.893742
Multilayer Perceptron	0.892617	0.822346	0.891058	0.892617	0.891219
Decision Tree	0.805369	0.5743	0.803442	0.805369	0.803426
Logistic Regression	0.687919	0.33631	0.681629	0.687919	0.6784
Naive Bayes	0.57047	-0.0558699	0.605885	0.57047	0.541812
Quadratic Discriminant Analysis	0.718121	0.394132	0.740116	0.718121	0.704056

TABLE 2. Performance of various classifiers on current value (Predictor-cur)



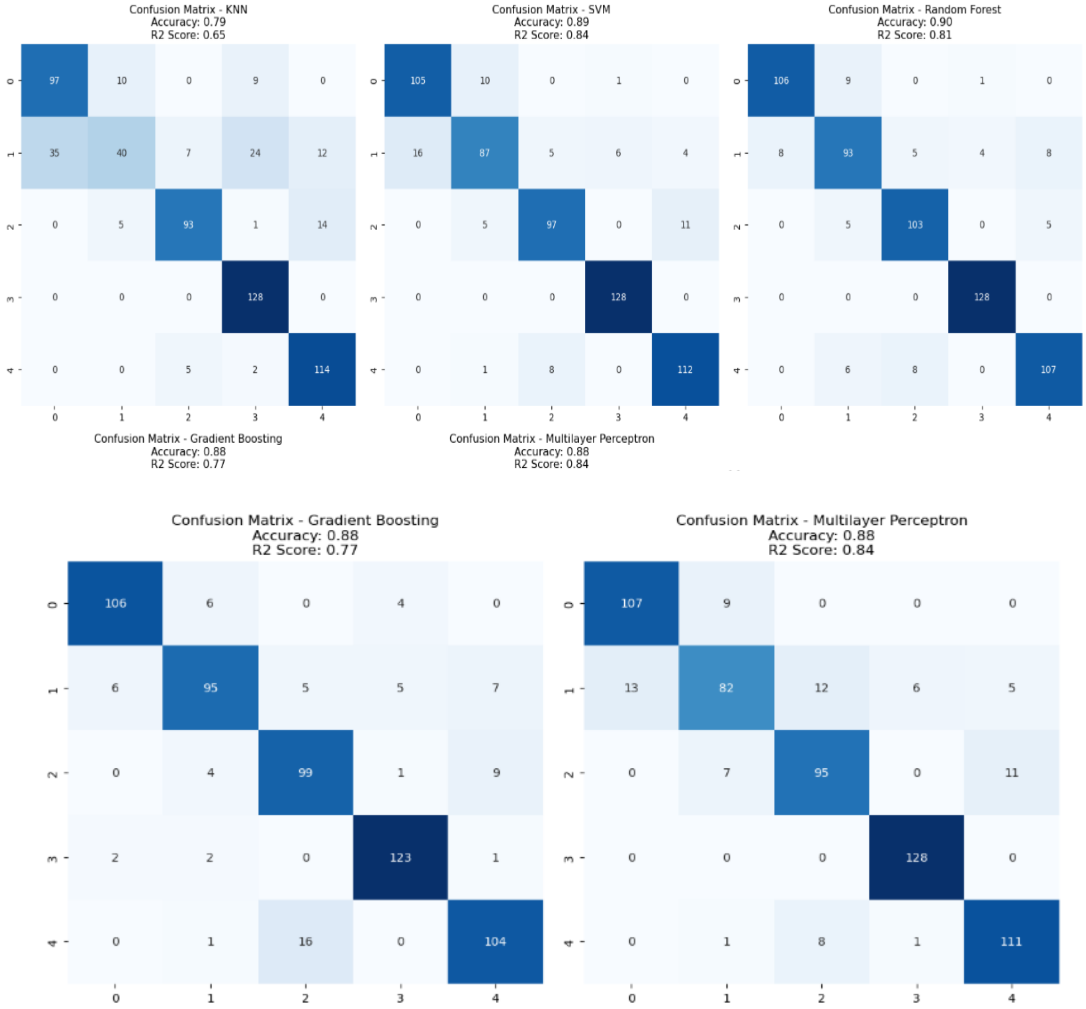
Model	Accuracy	R2 Score	Precision	Recall	F1 Score
KNN	0.776846	0.598602	0.776527	0.776846	0.758348
SVM	0.807047	0.637987	0.810028	0.807047	0.801546
Random Forest	0.854027	0.726815	0.852785	0.854027	0.852836
Gradient Boosting	0.807047	0.629608	0.804481	0.807047	0.805333
Multilayer Perceptron	0.756711	0.517317	0.755446	0.756711	0.753526
Decision Tree	0.716443	0.425138	0.711873	0.716443	0.711191
Logistic Regression	0.639262	0.193852	0.629274	0.639262	0.630069
Naive Bayes	0.595638	0.0731809	0.591973	0.595638	0.574138
Quadratic Discriminant Analysis	0.672819	0.301115	0.67535	0.672819	0.662828

The comparison between the two datasets, one with temporal information and the other without, reveals that incorporating temporal information generally enhances the performance across various classifiers. Across most classifiers, the dataset with temporal information consistently exhibits higher accuracy, R2 score, precision, recall, and F1 score compared to the dataset without temporal information. This suggests that the inclusion of temporal aspects contributes positively to the predictive capabilities of the classifiers. Random Forest stands out in both datasets, demonstrating its robustness and adaptability. In the dataset with temporal information, Random Forest achieves superior performance, reinforcing its effectiveness in capturing temporal patterns. Other classifiers, such as SVM and Gradient Boosting, also benefit from the inclusion of temporal features. Random Forest's superior performance across various tasks can be attributed to its ensemble learning approach, where multiple decision trees are constructed during training to form a collective model. This ensemble strategy enhances accuracy and stability by mitigating overfitting, as predictions are aggregated from diverse trees. The introduction of feature randomization at each split ensures that only a random subset of features is considered, promoting decorrelation among the trees, and increasing the model's robustness against overfitting. Random Forest's adeptness at handling complex non-linear relationships and intricate feature interactions further contributes to its success.

The presented results include confusion matrix heatmaps, ROC curves, and R2 scores obtained from different black-box models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP). The confusion matrix heatmaps provide insights into the model's classification performance across different classes, showcasing the true positive, true negative, false positive, and false negative predictions. ROC curves illustrate the trade-off between sensitivity and specificity, offering a comprehensive view of each model's discriminatory ability. Additionally, R2 scores quantify the proportion of variance explained by the models.

Figure 4 depicts the confusion matrix of all 5 models. The comparative analysis of the black-box models reveals variations in their performance metrics. The accuracy scores indicate that Random Forest achieved the highest accuracy at 90%, followed by Support Vector Machine

(SVM) at 89%, Gradient Boosting at 88%, Multilayer Perceptron (MLP) at 88%, and K-Nearest Neighbors (KNN) at 79%. In terms of R2 scores, Random Forest again leads with 81%, followed by SVM and MLP both at 84%, Gradient Boosting at 77%, and KNN at 65%.



Confusion Matrix - KNN
Accuracy: 0.79
R2 Score: 0.65

	0	1	2	3	4
0	97	10	0	9	0
1	35	40	7	24	12
2	0	5	93	1	14
3	0	0	0	128	0
4	0	0	5	2	114

Confusion Matrix - SVM
Accuracy: 0.89
R2 Score: 0.84

	0	1	2	3	4
0	105	10	0	1	0
1	16	87	5	6	4
2	0	5	97	0	11
3	0	0	0	128	0
4	0	1	8	0	112

Confusion Matrix - Random Forest
Accuracy: 0.90
R2 Score: 0.81

	0	1	2	3	4
0	106	9	0	1	0
1	8	93	5	4	8
2	0	5	103	0	5
3	0	0	0	128	0
4	0	6	8	0	107

Confusion Matrix - Gradient Boosting
Accuracy: 0.88
R2 Score: 0.77

Confusion Matrix - Multilayer Perceptron
Accuracy: 0.88
R2 Score: 0.84

Confusion Matrix - Gradient Boosting
Accuracy: 0.88
R2 Score: 0.77

	0	1	2	3	4
0	106	6	0	4	0
1	6	95	5	5	7
2	0	4	99	1	9
3	2	2	0	123	1
4	0	1	16	0	104

Confusion Matrix - Multilayer Perceptron
Accuracy: 0.88
R2 Score: 0.84

	0	1	2	3	4
0	107	9	0	0	0
1	13	82	12	6	5
2	0	7	95	0	11
3	0	0	0	128	0
4	0	1	8	1	111

FIGURE 4. Confusion matrix for five different black box models

The insights from the black-box models suggest that Random Forest performs exceptionally well in terms of both accuracy and R2 score, making it a robust choice for this task. SVM and MLP also exhibit strong performance, particularly in terms of R2 score, indicating their capability to capture the variance in the target variable. Gradient Boosting demonstrates good accuracy, but its R2 score is slightly lower compared to Random Forest, SVM, and MLP. KNN, while having a respectable accuracy, shows a comparatively lower R2 score, suggesting limitations in explaining the variance in the target variable.

Figure 5 presents the ROC of the five models. The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are significant metrics in assessing the performance of classification models. The ROC curve provides a graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across various decision thresholds. The analysis through the Receiver Operating Characteristic (ROC) curves highlighted Random Forest's superior discriminatory ability, as reflected in AUC values of 1 for Class 0 and Class 3, 0.96 for Class 1, 0.98 for Class 2, and 0.99 for Class 4. These results signify Random Forest's effectiveness in classifying instances across multiple classes, emphasizing its robustness in handling the complexities of the dataset.

Although the random forest model demonstrates its ability to predict different classes effectively, it lacks the capability to provide explanations for its predictions. To address this limitation, we employed explainable AI techniques to elucidate the reasoning behind the predictions.

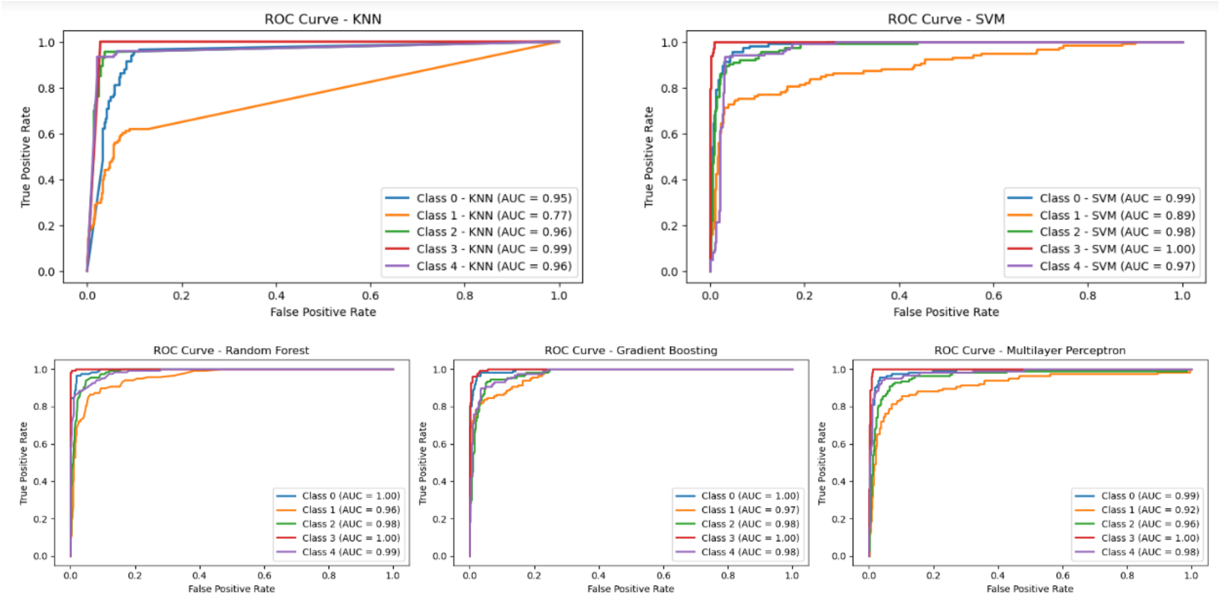


FIGURE 5. ROC and AUC for different models

Figure 6 illustrates the explanation for a specific instance where the predicted class matches the actual class (NL to MCI). This explanation clearly outlines the contribution of each feature in predicting the transition from Mild Cognitive Impairment (MCI) to Dementia, as both the actual and predicted classes belong to Class 3. The LIME explanation offers valuable insights into the model's decision, highlighting influential features and their respective impacts. Positive weights associated with features like 'CDRSB_diff,' 'RAVLT_learning_diff,' and 'ADAS11_diff' indicate their positive influence on predicting the transition. Conversely, negative weights for features such as 'RAVLT_forgetting_diff' suggest a counteractive effect.

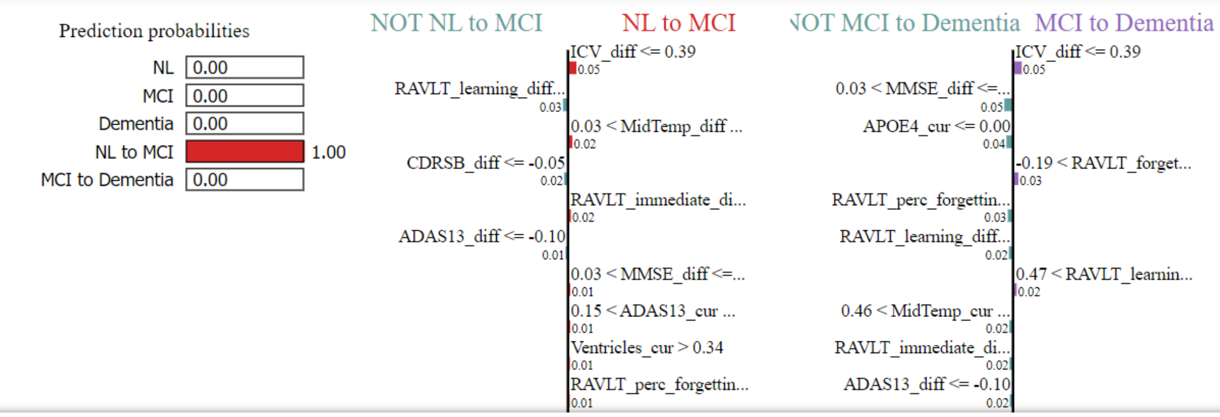


FIGURE 6. LIME explanation of an individual instance

The identified features, including 'CDRSB_diff,' 'RAVLT_learning_diff,' 'ADAS11_diff,' and 'RAVLT_forgetting_diff,' bear clinical significance in predicting the transition from Mild Cognitive Impairment (MCI) to Dementia. An increase in 'CDRSB_diff' signifies a deterioration in cognitive and functional abilities, aligning with expected disease progression. Elevated 'RAVLT_learning_diff' suggests improved learning, potentially indicative of effective interventions or compensatory mechanisms. A decrease in 'ADAS11_diff' reflects enhanced cognitive function, hinting at positive responses to treatments targeting Alzheimer's-related cognitive decline. The negative change in 'RAVLT_forgetting_diff' implies reduced forgetting, a critical aspect of memory decline. Collectively, these insights offer clinicians valuable information for monitoring and understanding the progression from MCI to Dementia, aiding in tailored interventions and treatment assessments. These insights underscore the relevance of cognitive and clinical measurements in the model's prediction, emphasizing the specific factors contributing to its sensitivity in identifying the progression from MCI to Dementia.

The SHAP explanation of an instance with predicted class and actual class MCI to dementia is given in Figure 7. The water plot curve uses color differentiation to represent the individual and cumulative Shapley values for each feature in the instance. The blue color corresponds to the Shapley values for the individual instance, showcasing the impact of each feature in isolation on the model's prediction. Positive (above the baseline) and negative (below the baseline) segments indicate the direction and magnitude of influence, respectively. On the other hand, the red color represents the cumulative Shapley values, illustrating the overall contribution of each feature to the model's prediction. Again, positive (above the baseline) and negative (below the baseline) segments indicate the direction and magnitude of the combined impact. By combining both representations in a single water plot curve, we can visually assess how each feature contributes individually (blue) and collectively (red) to the model's decision for the given instance. This dual-color scheme provides a comprehensive understanding of the intricate dynamics of feature contributions in the interpretability of the model's predictions.

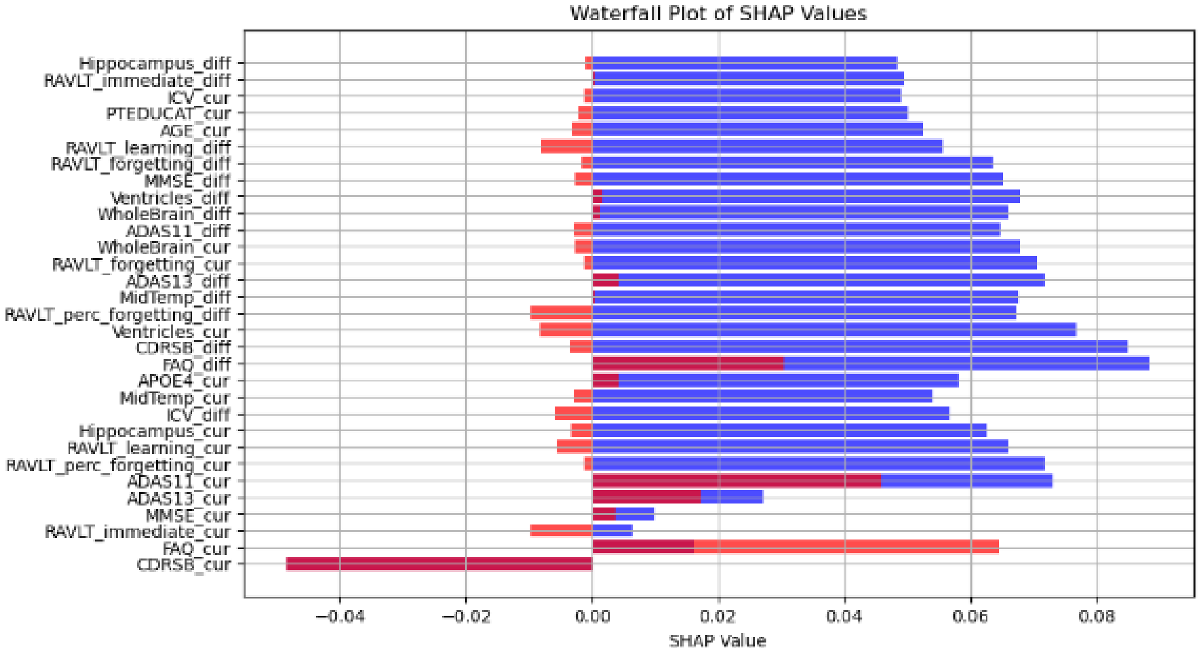


FIGURE 7. SHAP value of an instance

The global feature importance analysis is performed on a RandomForestClassifier using SHAP. The TreeExplainer from SHAP is utilized to compute Shapley values, representing the impact of each feature on the model's output. The mean absolute Shapley values across all instances are then calculated to determine the global importance of each feature. The results are sorted, and the top features are visualized along with their corresponding importance. The summary_plot function generates a bar plot illustrating the ranked feature importance, providing insights into the relative contribution of each feature to the model's overall predictions. Figure 8 gives the global feature importance by considering all features.

Among the features, 'CDRSB_cur' (Clinical Dementia Rating Scale - Sum of Boxes at the current visit) emerges as the pivotal indicator, encapsulating the comprehensive clinical status of dementia. Its prominence lies in assessing the overall severity of cognitive decline, with higher scores indicating a more advanced stage of the disease. Following closely is 'FAQ_cur' (Functional Activities Questionnaire at the current visit), which delves into the individual's ability to perform daily activities independently, offering valuable insights into functional independence. The subsequent features, including 'ADAS13_cur' and 'MMSE_cur', contribute to cognitive assessment, pinpointing specific cognitive domains affected by the disease. 'MMSE_diff', measuring changes over time, and 'ADAS11_cur', emphasizing memory and language domains, further enrich our understanding of cognitive progression. The dynamics of 'CDRSB_diff' highlight shifts in overall clinical symptoms, while memory-related features like 'RAVLT_immediate_cur' and 'RAVLT_perc_ forgetting_diff' elucidate immediate recall and memory retention alterations. Collectively, these features serve as crucial markers in unraveling the multifaceted landscape of dementia progression and cognitive decline.

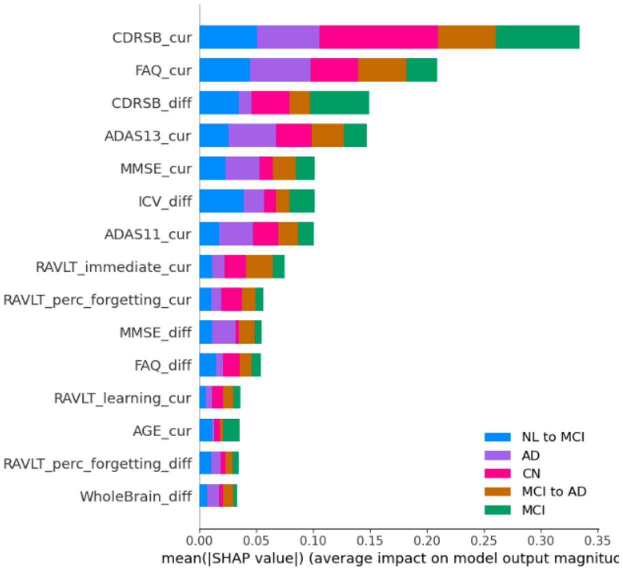


FIGURE 8. Global feature importance from SHAP

In the further analysis of the top features for each class, 'CDRSB_cur' emerges as a significant predictor, showcasing class-specific trends. The decreasing trend for Class 0 (CN) suggests that higher cognitive function is associated with a higher likelihood of being classified as CN. Conversely, for Class 1 (MCI) and Class 2 (AD), the increasing trends indicate that higher values of 'CDRSB_cur' are linked to an elevated probability of being classified as MCI or AD, respectively. Additionally, exploring 'FAQ_cur,' the Functional Activities Questionnaire, reveals insights into functional abilities. Lower 'FAQ_cur' values are associated with a higher probability of being classified as CN, while increasing values align with a higher likelihood of MCI or AD. Figure 9 illustrates these trends for the 2 most dominating features. Extending this interpretation to the other features for each class allows for a comprehensive understanding of the model's perspective on disease progression. Clinically, these insights are valuable for timely interventions and close monitoring, providing a nuanced approach to patient care tailored to different stages of cognitive decline. It's essential to note that while these trends offer valuable associations, they do not imply causation, and clinical judgment remains paramount in decision-making.

Table 3 gives the observations on the partial dependence plot of the feature CDRSB_cur and FAQ_cur.

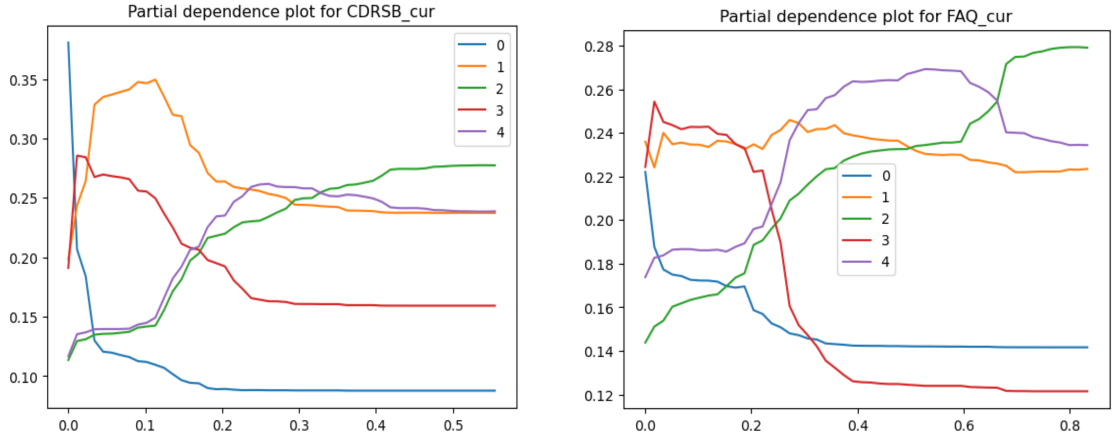


FIGURE 9: Partial dependence plot for CDRSB_cur and FAQ_cur

TABLE 3. Observations on the partial dependence plot of the feature CDRSB_cur and FAQ_cur

Class	Trend with ‘CDRSB_cur’	Interpretation
0 (CN)	Decreasing	Higher values of ‘CDRSB_cur’ are associated with a decreased predicted probability of belonging to class CN.
1 (MCI)	Increasing	Higher values of ‘CDRSB_cur’ are associated with an increased predicted probability of belonging to class MCI.
2 (AD)	Increasing	Higher values of ‘CDRSB_cur’ are associated with a higher predicted probability of belonging to class AD.
3 (NL to MCI)	Fluctuating	The relationship between ‘CDRSB_cur’ and the predicted probability of NL to MCI transition is more complex and fluctuates. Further analysis may be needed.
4 (MCI to AD)	Increasing	Higher values of ‘CDRSB_cur’ are associated with an increased predicted probability of transitioning from MCI to AD.
Class	Trend with ‘FAQ_cur’	Interpretation
0 (CN)	Decreasing	Higher values of ‘FAQ_cur’ are associated, on average, with a decreased predicted probability of belonging to class CN.
1 (MCI)	Fluctuating	The relationship between ‘FAQ_cur’ and the predicted probability of belonging to class MCI is more complex, with some peaks and valleys. Further analysis may be needed.
2 (AD)	Increasing	Higher values of ‘FAQ_cur’ are associated, on average, with a higher predicted probability of belonging to class AD.
3 (NL to MCI)	Fluctuating	The relationship between ‘FAQ_cur’ and the predicted probability of NL to MCI transition is complex, with some peaks and valleys, similar to Class 1. Further analysis may be needed.
4 (MCI to AD)	Decreasing	Higher values of ‘FAQ_cur’ are associated, on average, with a decreased predicted probability of transitioning from MCI to AD.

DISCUSSION

The integration of time-aware modeling and Explainable AI techniques represents a synergistic approach to enhance Alzheimer’s disease progression prediction. Time-aware modeling involves not only considering the current values of key predictors but also incorporating the temporal aspect by evaluating the rate of change in these predictors over time. With the use of this methodology, a deeper understanding of the structure and progression of the disease can be obtained, facilitating the identification of minute changes that might point to the beginning or advancement of Alzheimer’s disease.

It is apparent from the analysis of several black-box models, including Multilayer Perceptron (MLP), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Random Forest, that the models behave much better when time-aware features are included. The confusion matrix heatmaps, ROC curves, and R2 scores highlight Random Forest as the top performer, emphasizing its ability to capture the complexities of Alzheimer’s progression. Despite the robust predictive capability of Random Forest, its black-box nature necessitates the incorporation of Explainable AI techniques. LIME and Shapley values serve this purpose by offering interpretable insights into the decision-making process of the model. ‘CDRSB_cur’ stands out as a critical characteristic, suggesting that it plays an important part in predicting the course of the disease. The examination of global feature importance highlights the relevance of ‘CDRSB_cur’ and presents additional significant contributions, including ‘FAQ_cur’, ‘ADAS13_cur’, and ‘MMSE_cur’.

The rate of change of ‘CDRSB’ (‘CDRSB_diff’) is significant as it indicates how time-aware characteristics affect the interpretability of the model. Identifying the time-related component of clinical measurements improves the model’s ability to identify small changes over time, leading to a more thorough knowledge of the course of Alzheimer’s disease. This increases prediction accuracy and makes it easier to create an understandable model. The patterns in partial dependence plots for 'FAQ_cur' and 'CDRSB_cur' that have been found across classes further highlight how important these traits are in predicting cognitive decline. The interpretability of the model is enhanced by including both the current values and the rate of change in important variables over time. This gives physicians practical insights. By bridging the gap between clinical application and prediction accuracy, this method opens the door to more effective interventions and individualized care in the context of Alzheimer's disease improvement.

CONCLUSION

This study concludes by addressing the urgent need for reliable and understandable models for projecting the course of Alzheimer's disease. A complex and therapeutically applicable predictive model is developed by the merging of Explainable AI approaches with time-aware modelling. Random Forest emerges as the best performer in terms of accuracy, ROC curves, and R2 scores when several black-box models, such as K-Nearest Neighbors, Support Vector Machine, Random Forest, Gradient Boosting, and Multilayer Perceptron, are analyzed. In the context of Alzheimer's disease, this research assists in narrowing the gap between interpretability and predictive accuracy. Time-aware modelling, Explainable AI methods, and an emphasis on important predictors provide a comprehensive knowledge of the disease's dynamic nature. The results may open the door to more efficient, prompt, and tailored interventions, thereby raising the standard of living for Alzheimer's patients and their families. This research offers a basis for the creation of transparent and clinically applicable models for the prediction of Alzheimer's disease progression as artificial intelligence in healthcare continues to evolve.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the Research Council (TRC) of the Sultanate of Oman under the Block Funding Program BFP/RGP/ICT/21/148.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Adadi, A., & Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

ADNI | ACCESS DATA. n.d. https://adni.loni.usc.edu/data-samples/access-data/

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based alzheimer's disease classification. *Frontiers in Aging Neuroscience 11*. https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194

Chun, M. Y., Park, C. J., Kim, J., Jeong, J. H., Jang, H., Kim, K., & Seo, S. W. 2022. Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. *Frontiers in Aging Neuroscience 14*. https://www.frontiersin.org/articles/10.3389/fnagi.2022.898940

Dubois, B., Picard, G., & Sarazin, M. 2009. Early detection of Alzheimer's disease: New diagnostic criteria. *Dialogues in Clinical Neuroscience 11*(2): 135–139. https://doi.org/10.31887/DCNS.2009.11.2/bdubois

Fabrizio, C., Termine, A., Caltagirone, C., & Sancesario, G. 2021. Artificial intelligence for alzheimer's disease: Promise or challenge? *Diagnostics 11*(8): 1473. https://doi.org/10.3390/diagnostics11081473

Gaur, L., Masud, M., & Jhanjhi, N. 2022. *Explanation-driven HCI Model to Examine the Mini-Mental State for Alzheimer's Disease*.

Jerome H., F. 2001. Greedy function approximation: a gradient boosting machine on JSTOR. *The Annals of Statistics 29*(5): 1189–1232.

Kamal, Md. S., Northcote, A., Chowdhury, L., Dey, N., Crespo, R. G., & Herrera-Viedma, E. 2021. Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. *IEEE Transactions on Instrumentation and Measurement 70*: 1–7. https://doi.org/10.1109/TIM.2021.3107056

Lundberg, S. M., & Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. 2022. Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review 55*(5): 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Muslikh, A. R., Andono, P. N., Marjuni, A., & Santoso, H. A. 2023. Systematic literature review of data distribution in preprocessing stage with focus on outliers. *2023 International Seminar on Application for Technology of Information and Communication* *(iSemantic)* 328–333. https://doi.org/10.1109/iSemantic59612.2023.10295291

N. V., C., K. W., B., & L. O., H. 2002. SMOTE: Synthetic minority over-sampling technique: *Journal of Artificial Intelligence Research 16*. https://doi.org/10.1613/jair.953

Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. "Why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

S., E.-S., J. M., A., S. R., I., A.M., S., & K. S., K. 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports* 11(1): 2660. https://doi.org/10.1038/s41598-021-82098-3

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. 2017. *Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization*. 618–626. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

Shad, H. A., Rahman, Q. A., Asad, N. B., Bakshi, A. Z., Mursalin, S. M. F., Reza, Md. T., & Parvez, M. Z. 2021. Exploring alzheimer's disease prediction with XAI in various neural network models. *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)* 720–725. https://doi.org/10.1109/TENCON54134.2021.9707468