Enhancing Trust in Alzheimer's Disease Classification using Explainable Artificial Intelligence: Incorporating Local Post Hoc Explanations for a Glass-box Model

Abraham Varghese, PhD* Ben George, PhD* Vinu Sherimon, PhD* Huda Salim Al Shuaily, PhD**

ABSTRACT

Background: Alzheimer's disease (AD) leads to cognitive dysfunction among older people worldwide, making it nearly impossible for them to carry out their daily lives. Due to the inherent characteristics of Alzheimer's disease and its impact on the brain, timely intervention is crucial to delay its onset and mitigate its progression. Currently, the diagnosis of Alzheimer's disease often occurs at a stage where it is too late for effective prevention measures, allowing the disease to cause significant damage to the brain. The use of machine learning and deep learning models is critical for the classification of demented and non-demented cases, but most highly accurate models are non-linear and less transparent, not revealing the logic behind the predictions. Therefore, incorporating interpretability components into the models will make them more transparent and trustworthy. This study is aimed to develop appropriate diagnostic methods capable of assessing Mild Cognitive Impairment (MCI), the early stage of Alzheimer's disease that occurs before the irreversible loss of neurons.

Methods: Explainable artificial intelligence (XAI) refers to AI systems that can provide explanations for their decisions or predictions. In the context of AD classification, explainable AI systems aim to provide insights into the features or characteristics of the model used to make a prediction. This XAI provides a mechanism to understand and interpret the basis of a model's predictions which is more important for improving the trust in the system and its results. As such, a non-linear neural network is employed in this work to distinguish between demented and non-demented cases while local post hoc explanations are incorporated to make it a glass-box model using the XAI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME).

Results: The application of LIME provided valuable insights into the impact of various factors on predictions. Notably, factors such as CDR, Age, and ASF aligned with clinical knowledge and proved instrumental in predicting dementia cases. Conversely, features like nWBV, MMSE, and eTIV adversely affected the predictions, highlighting their significance in identifying non- demented cases. Similarly, exploring SHAP values yielded a comprehensive understanding of the decision-making process employed by the model in detecting Alzheimer's disease.

Conclusion: Through the utilization of explainable artificial intelligence (XAI) methods, this study endeavors to develop a dependable and transparent technique for early detection, monitoring, and personalized interventions in the realm of Alzheimer's disease.

Keywords: Machine learning, Alzheimer’s disease, Interpretable Machine learning, LIME, SHAP, Explainable AI, Neural network

INTRODUCTION

The alarming scale of the global impact caused by Alzheimer's disease (AD) cannot be overlooked. As per the latest Alzheimer's report, it is projected that a new case of dementia will occur approximately every 3.2 seconds worldwide[1]. These statistics highlight the significant and growing burden of Alzheimer's disease on a global scale[2]. According to research, there will likely be a significant increase in the number of people who have dementia, with an estimated increase from 57.4 million people in 2019 to a staggering 152.8 million by the year 2050[2,3]. This significant escalation underscores the urgent need for effective preventive measures, improved treatments, and enhanced support systems to address the growing impact of dementia on a global scale. According to a study published in The Lancet Public Health in 2022, the global cost of dementia in 2019 is estimated to be a staggering US$1 trillion. This figure reflects the significant economic burden associated with dementia on a global scale. The comprehensive impact of this burden has yet to be fully quantified and accounted for, highlighting the complex and multifaceted consequences of Alzheimer's disease on individuals, caregivers, and society.

* College of Computing and Information Sciences
University of Technology and Applied Sciences
Muscat, Sultanate of Oman.
E-mail: abraham.varghese@utas.edu.om

** Deputy Vice-Chancellor Office
University of Technology and Applied Sciences
Muscat, Sultanate of Oman.

Alzheimer's disease is a progressive neurological condition marked by memory loss and a deterioration in cognitive function over time[1,4]. It is the most prevalent form of dementia, a term encompassing various cognitive impairments that affect daily functioning. The buildup of plaques and tangles in the brain, as well as the destruction of neural connections, are linked to the onset of Alzheimer's disease (Alzheimer's Disease Fact Sheet, n.d.). As the disease advances, it leads to the deterioration of memory, thinking skills, and the ability to carry out daily responsibilities. It involves complex changes, including the formation of abnormal protein structures known as amyloid plaques and tangles, which disrupt the normal functioning of nerve cells. In the early stages, Alzheimer's primarily affects specific brain regions like the hippocampus and entorhinal cortex, which are crucial for memory. Over time, it spreads to other areas, causing widespread damage and neuronal loss. Memory problems are typically the initial noticeable symptoms of cognitive decline in Alzheimer's disease, including conditions like Mild Cognitive Impairment (MCI). Individuals with MCI experience greater difficulties with memory compared to others their age, although these challenges may not significantly interfere with their daily activities. MCI can also be associated with other issues such as changes in sense of smell or problems with movement. With proper care and therapeutic interventions, individuals with MCI have the potential to regain their normal cognitive function.

Machine learning models have emerged as valuable tools in improving disease prediction accuracy, including in the context of Alzheimer's disease. These models enable doctors and researchers to better identify individuals who may benefit from preventative care while avoiding unnecessary medication for others[5]. In the field of Alzheimer's disease, machine learning techniques have been applied to various data types, including neuroimaging data, genetic data, and clinical data. These models have shown promise in aiding early diagnosis, predicting disease progression, and identifying individuals at risk of developing Alzheimer's disease. One area where machine learning has made significant contributions in Alzheimer's research is neuroimaging analysis. By utilizing advanced neural network models, machine learning algorithms can analyze brain MRI or PET scans to identify biomarkers and patterns associated with Alzheimer's disease. These models can detect subtle structural or functional changes in the brain that may indicate the presence or progression of the disease[6]. This early detection can help initiate interventions and treatments at a stage when they are most effective.

The tremendous computational power and storage capacity available today have enabled the development of highly accurate machine learning models. However, these models often lack transparency, making it challenging to comprehend the rationale behind their predictions. The complexity of these models, particularly neural networks with multiple layers and interconnected links, makes it nearly impossible to fully comprehend the decision-making process, even through extensive examination of the model's internal workings[7]. The inability to interpret complex models has posed limitations on the practical application of machine learning methods and has raised concerns about the reliability and trustworthiness of these models. Despite achieving high accuracy, the lack of interpretability undermines the transparency and understanding required to gain insights from the model's predictions.

In essence, the inherent complexity of neural networks and other sophisticated models hinders our ability to fully grasp how they arrive at specific decisions. This challenge has sparked ongoing research and efforts to develop methods for interpreting and explaining the reasoning behind machine learning predictions, aiming to enhance the transparency and trustworthiness of these models. As a result, deep learning is frequently referred to as a "black box." In several application sectors, there is growing concern that these black boxes may be biased. This can have significant implications, particularly in medical contexts. Hence, it is crucial to obtain explanations for the decisions made by AI models. Providing explanations serves two important purposes: building trust and detecting potential biases in the system. It is necessary to have explanations that are contextual and understandable to users. The demand for explainable AI (XAI) arises from ethical concerns surrounding the lack of transparency in AI systems, particularly in the healthcare domain[8]. XAI methods are employed to describe AI models and their predictions, addressing the limitations of black-box machine learning algorithms, for instance, neural networks, which are difficult to interpret[5].

By utilizing explainable AI, the complex interactions between risk factor parameters and their independent implications on outcomes can be better understood[5]. In vital applications like healthcare, it is crucial for AI models to be understandable to humans. Users need to develop confidence in the model and comprehend how it generates its results. This understanding allows medical practitioners to make more informed decisions and assess whether a prediction aligns with relevant features and factors[9]. XAI techniques can be differentiated based on various criteria. One such criterion is post-hoc explainability and ante-hoc explainability, as discussed by[8] and[7]. Post-hoc explainability focuses on explaining the model's predictions in terms that are easily interpretable. It involves training a neural network and then making efforts to describe the behavior of the resulting "black box" network. Contrarily, ante-hoc explainability explicitly incorporates explainability into the design of an AI model. It aims to design neural networks that are inherently explainable. Another criterion is model-agnostic versus model-specific explanations, as described by[7]. Model-agnostic explanations are independent of the type of neural network used and focus solely on the network's input and output. They allow users to understand how changes in the input affect the network's output. In contrast, certain types of models can only be used using model-specific explanation strategies. However, limiting the neural networks that can be used in this approach may overlook networks that could potentially provide more accurate results based on the input data.

There have been several studies that have explored the use of explainable AI techniques for image classification especially in Alzheimer's diagnosis[10]. Explore the application of convolutional neural networks (CNNs) for classifying different pathologies associated with Alzheimer's disease in brain MRI scans. They propose an interpretable CNN model that generates heatmaps, allowing visualization of the specific regions in the brain that are most indicative of each pathology, providing insights into the classification process. The study given in[11], propose three effective methods for generating visual explanations from 3D convolutional neural networks (3D-CNNs) in the context of Alzheimer's disease classification. Their approaches include sensitivity analysis on hierarchical 3D image segmentation and visualization of network activations on a spatial map, demonstrating the ability to identify significant brain regions for accurate diagnosis of Alzheimer's disease. A comprehensive and interpretable model for Alzheimer's disease detection and prediction was proposed in[12]. By leveraging explainable artificial intelligence, the model enables accurate diagnosis and progression detection, empowering physicians with precise decision-making capabilities and providing accompanying explanations for each decision made. The TADPOLE challenge, the most comprehensive to date in terms of participants, subjects, and features, aimed to identify the best predictive data and methods for Alzheimer's disease progression. While tree-based ensemble methods emerged as the most effective, the challenge lacked insight into their contribution to accuracy and interpretability. The study by[13] intensively compares the top three TADPOLE models, investigates their meaningful features and quantifies their contribution to accuracy using SHapley Additive exPlanations (SHAP) values, shedding light on their performance and alignment with clinical knowledge[14]. Presents a comprehensive analysis of a large dataset to explore the early diagnosis of Alzheimer's disease, employing appropriate preprocessing techniques and training an XGBoost model with hyperparameter tuning. The study emphasizes the importance of interpretability using SHAP values, deriving valuable insights and challenging established hypotheses, while achieving a competitive f1-score of 0.84 and providing additional knowledge for physicians in the accurate diagnosis of early-stage Alzheimer's disease. In this paper, our objective is to leverage these interpretability methods to classify cases of dementia and non-dementia. By incorporating XAI techniques into our machine learning models, we aim to not only improve the accuracy of the classification but also provide meaningful explanations for the predictions made. This will enable healthcare professionals and researchers to gain valuable insights into the features and regions of the brain that are indicative of Alzheimer's disease.

The subsequent sections of the paper will delve into the materials and methods used in our research (Section 2), outlining the specific techniques and models employed. We will then present the results obtained from our experiments (Section 3), highlighting the efficiency of the XAI techniques in improving the interpretability and accuracy of the classification. The conclusion section (Section 4) will summarize our findings and discuss the implications of our research. Finally, the paper will conclude with a comprehensive list of references, citing the relevant studies that have contributed to the field of explainable AI in Alzheimer's disease classification.

MATERIALS AND METHODS

Dataset: In this research, 150 persons between the ages of 60 and 96 were examined using data from OASIS, more especially the OASIS-2 dataset[15]. Over 373 imaging sessions, each subject underwent at least two examinations. All the participants were right-handed, male, and female; during the investigation, 72 of the subjects were determined to be normal. At the time of their initial visits, 64 of the study participants were identified as having dementia, and 51 as having mild to moderate Alzheimer's disease. In this development cohort, there were 160 men and 213 women, or 57% and 43% respectively. Out of 213 female patients, 84 (46%) were demented and 129 (68% of the total) were not. Similarly, out of a total of 160 men, 61 did not have dementia and 99 did. The characteristics used in the study are Quantum Local Binary Pattern (QLBP), Age, Education, MMSE, CDR, Normalized whole-brain volume (nWBV), Local Binary Pattern (LBP), Atlas Scaling Factor (ASF), and 14 Haralick features.

Variable selection: Feature ranking algorithms such as ANOVA, Chi-square, and the Kruskal-Walli's test are commonly employed to address the curse of dimensionality by identifying the most relevant features. The Kruskal-Wallis test, a non-parametric statistical method introduced in[16], is specifically used to determine statistically significant differences between a categorical independent variable and a continuous dependent variable. It assesses whether one sample stochastically dominates another by pooling and ranking the values of the dependent variable for each group. ANOVA, as described in[17], is a statistical method utilized to compare means across two or more groups and determine if they are significantly different. The Chi-Square test[18], as examines the independence between two events or variables. It measures the discrepancy between expected (E) and observed (O) counts, and higher Chi-Square values indicate greater dependence between features and the response variable, making them valuable for training models. Figure 1 depicts the potential variable using different feature ranking algorithms. It is observed from all the feature ranking algorithm that the features CDR and MMSE are more significant.

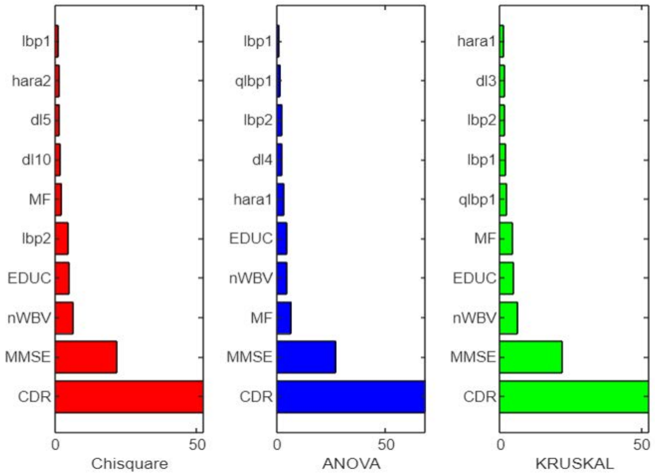


Figure 1: Feature relevance using Chi-Square, ANOVA, and Kruskal-Wallis algorithms

Classification model: Achieving high accuracy is crucial when selecting a classifier, considering the wide range of options available. However, it's important to note that some highly accurate models may lack interpretability. Conversely, linear models may be transparent but not as accurate. It's crucial to comprehend the conflict between accuracy and interpretability to choose an algorithm with confidence. Figure 2 in the study by[19] depicts this trade-off, highlighting the relationship between interpretability and accuracy in different scenarios.

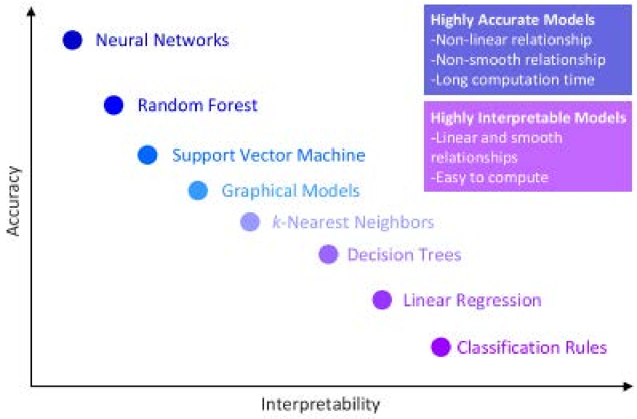


Figure 2: Interpretability vs Accuracy of ML models[19]

From Figure 2, it is evident that neural network, random forest, and support vector machine are very high accurate model with less interpretability and on the other hand decision trees, linear models and classification rules are more interpretable in nature but with less accuracy.

In this study, a neural network model is constructed using the top 10 features from 25 features identified through the Kruskal-Wallis algorithm obtained from 150 observations. The algorithm involves training a neural network with 10 input features derived from Alzheimer's data using the Stochastic Gradient Descent with Momentum (SGDM) optimization algorithm. The network includes one hidden layer, which introduces non-linear transformations to capture complex relationships in the data. The network is initialized with random weights and biases, and the learning rate and momentum are set for the SGDM algorithm. During each iteration, forward propagation is performed to compute the output prediction based on the input features. The loss/cost function is calculated, and then backpropagation is applied to update the weights and biases using the computed gradients and the momentum update rule. This process is repeated until convergence or reaching the maximum number of iterations.

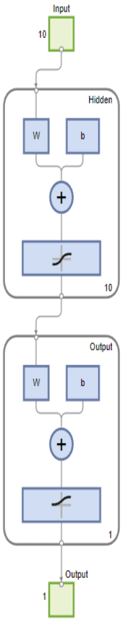
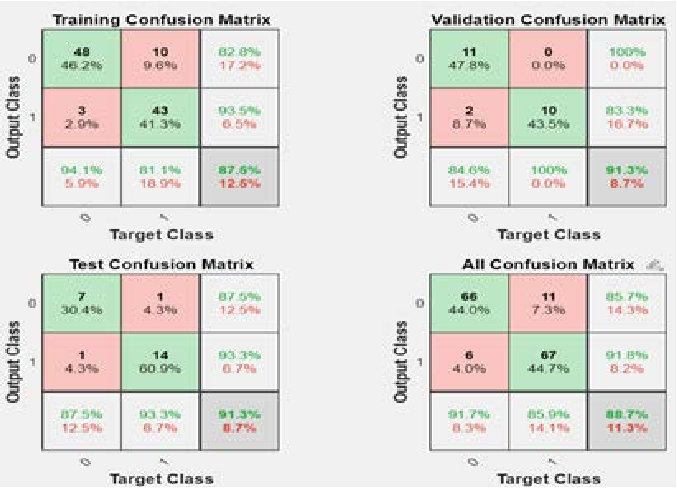


Figure 3: The neural network architecture

To evaluate the model's performance, 15% of the data is reserved for testing, while training and validation are conducted on the remaining portion. Cross-validation with a k-fold value of 5 is employed to mitigate overfitting risks. Figure 3 visualizes the neural network architecture, highlighting how the 10 selected input features are connected to the single output. On the other hand, Figure 4 presents the confusion matrix, which displays the performance of the model in terms of true positives, false positives, true negatives, and false negatives.



Training Confusion Matrix

	0	1	
0	48 46.2%	10 9.6%	82.8% 17.2%
1	3 2.9%	43 41.3%	93.5% 6.5%
	94.1% 5.9%	81.1% 18.9%	87.5% 12.5%

Validation Confusion Matrix

	0	1	
0	11 47.8%	0 0.0%	100% 0.0%
1	2 8.7%	10 43.5%	83.3% 16.7%
	84.6% 15.4%	100% 0.0%	91.3% 8.7%

Test Confusion Matrix

	0	1	
0	7 30.4%	1 4.3%	87.5% 12.5%
1	1 4.3%	14 60.9%	93.3% 6.7%
	87.5% 12.5%	93.3% 6.7%	91.3% 8.7%

All Confusion Matrix

	0	1	
0	66 44.0%	11 7.3%	85.7% 14.3%
1	6 4.0%	67 44.7%	91.8% 8.2%
	91.7% 8.3%	85.9% 14.1%	88.7% 11.3%

Output Class
Target Class

Figure 4: The confusion matrix in classifying dementia from non-dementia

Figure 5 showcases the Receiver Operating Characteristic (ROC) curve, providing insights into the model's trade-off between true positive rate and false positive rate at different classification thresholds. Additionally, Figure 5 also highlights the best validation epoch, indicating the point during training where the model achieved optimal performance. As per the confusion matrix, the training accuracy is 87.5, validation accuracy is 91.3, testing accuracy is 91.3 and the overall accuracy is 88.7. Figure 5 a show the cross-entropy values for the various accuracies mentioned above. From the figure, it can be observed that the validation accuracy is high as much as 0.17691 from epoch 12. Figure 5 b shows the receiver operating characteristic curve for the selected neural network. The ROC value is 0.933 which is comparatively high for the classification of demented and non-demented cases.

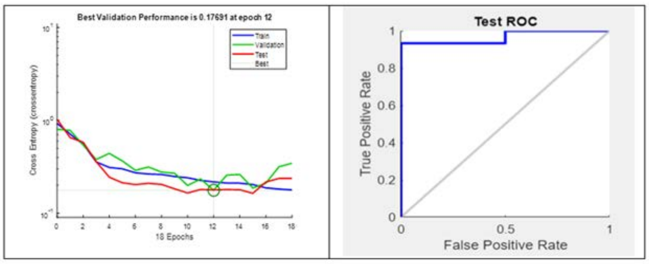


Figure 5: a) Best validation performance b) ROC

Interpretable models: Based on the input features along with the hyper parameters, the classification model gives training, validation, and testing accuracy. Confusion matrix gives both demented and non-demented true positive / true negative rates. A decision can be made based on all these factors whether a model is required. Even though the model provides 91.3% accuracy for the testing data, it is not addressing many details of the internal structure of the model. The current model lacks the ability to address important issues such as determining the most influential feature for achieving high accuracy, understanding the underlying reasons for the model's outcomes, identifying success and failure criteria, and establishing the level of trust in the model. These concerns arise due to the model's limited interpretability, which hinders our ability to gain meaningful insights. To overcome these limitations, it is crucial to develop a model that can provide explanations for its decisions, offer transparency on feature importance, clarify the reasons behind outcomes, define success and failure thresholds, and inspire confidence in its reliability.

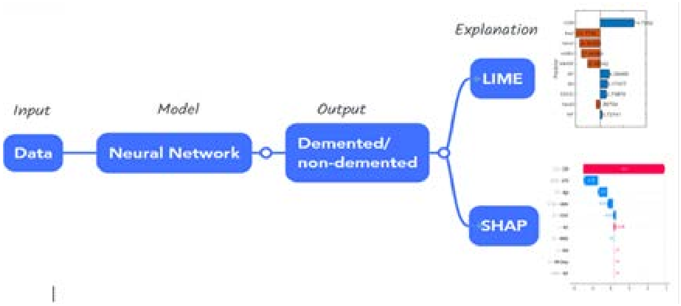


Figure 6: Overview of the method

As part of our research, we propose an approach to provide interpretable explanations for the complex neural network model. Our aim is to offer insights into the decision-making process of the model while remaining locally true to the classifier. By "interpretable explanation," we mean providing understandable and meaningful insights into why the model makes certain predictions or classifications. We want to address the need for transparency and trust in complex models by offering explanations that can be easily understood and validated. These explanations will help users and stakeholders gain confidence in the model's predictions and understand the factors that contribute to its decision-making. The overview of the proposed method is given in Figure 6.

Local Model-Agnostic Explanations

Local Interpretable Model Agnostic Explanations (LIME): LIME (Local Interpretable Model-Agnostic Explanations) is a technique that aims to interpret specific predictions made by a model by locally estimating its behavior around those predictions[20-22]. It provides an interpretable representation of the model's decision-making process, which differs from the original features and is designed to be understandable to humans. The explanation produced by LIME is obtained by the following formula:

$$\xi(x) = \underset{q \in G}{argmin} L(f, q, \pi x) + \Omega(q)$$

In LIME, an explanation is defined as a model q that belongs to a class of potentially interpretable models, such as q linear models or decision trees. The complexity of the explanation is measured by $\Omega(q)$. The classification function $f(x)$ represents the probability that a given instance x belongs to a specific class. The proximity measure $\pi_x z$ captures the closeness between an instance z and x, defining the locality around x. To ensure interpretability and local fidelity, LIME minimizes the measure $L(f, q, \pi\_x)$, which quantifies the degree to which q approximates f within the defined locality. By minimizing $L(f, q, \pi\_x)$ and keeping $\Omega(q)$ sufficiently low for human interpretability, LIME produces an explanation for a given prediction using the formula.

SHapley Additive exPlanations (SHAP): The SHAP (SHapley Additive exPlanations) approach is a powerful method for interpreting machine learning model outputs, providing a clear understanding of feature importance. The Shapley values theory, which has its roots in cooperative game theory, serves as its foundation. To explain how SHAP works, let's consider a classification problem. SHAP determines the contribution of each feature to a prediction provided by the model given the prediction. It does this by evaluating the impact of including each feature in all possible subsets of features and measuring the change in prediction accuracy. This process captures the interaction effects between features, providing a holistic view of their importance. Mathematically, the Shapley value for a specific feature is defined as the average marginal contribution of that feature across all possible feature combinations. It can be expressed as: $\Phi i(f) = \frac{\sum (f(S \cup \{i\}) - f(S))}{n * (n-1)}$

where: $\Phi i(f)$ represents the Shapley value for feature i,
$f(S \cup \{i\})$ denotes the model's output when feature i is included in the subset S,
$f(S)$ is the model's output when feature i is excluded from the subset S,
n , is the total number of features?
By calculating the Shapley values for each feature, we obtain a comprehensive understanding of their individual contributions to the model's predictions. Using SHAP plots or summary plots, this data can be represented graphically, offering clear insights into the model's decision-making process, and assisting in the detection of biases or faults.

RESULTS AND DISCUSSION

LIME

The interpretation of LIME is done with Linear model based on 2 input instances, one from demented cases and other from non-demented case. In this case, Figure 7 displays the predictions obtained using LIME with a linear model for an instance from both demented and non-demented cases. The important aspect is that both the black box model (the original complex model being explained) and the linear model (a simpler and more interpretable model used for explanation) give the same prediction for the given input instances. This consistency provides confidence in the accuracy of the original neural network model. The plot in Figure 7 shows 10 predictions for the query point, providing insight into the variability of predictions. The horizontal bar graph presents the sorted predictor importance values, indicating which features are most influential in determining the prediction. In this case, LIME identifies CDR (Clinical Dementia Rating), MMSE (Mini-Mental State Examination), and nWBV (normalized Whole Brain Volume) as three important predictors for the query point.

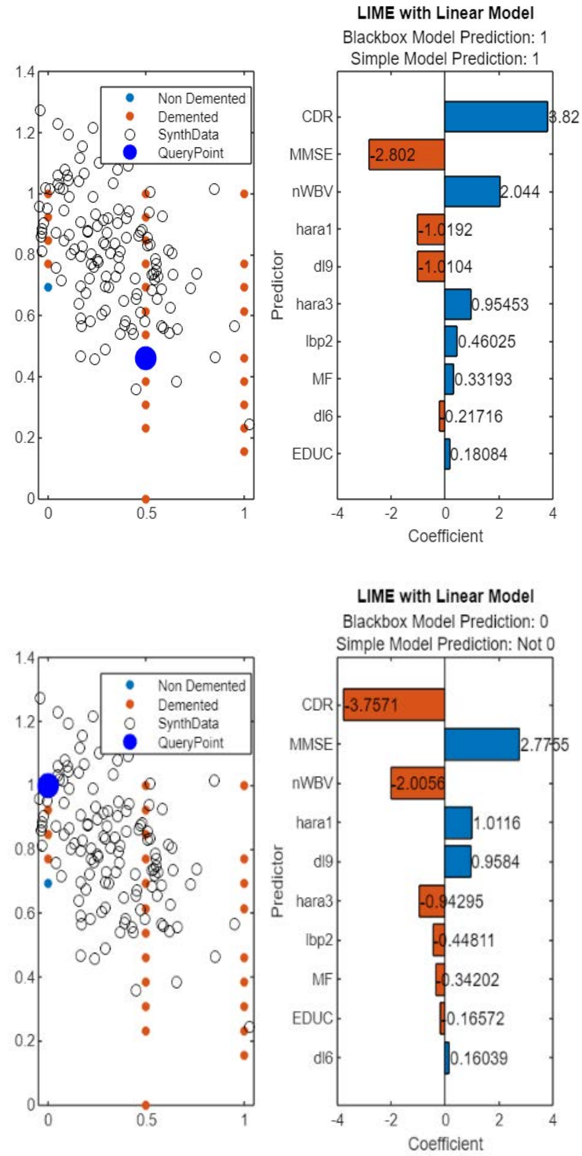


Figure 7: a) LIME prediction with linear model for non-demented case b) LIME prediction with linear model for demented case

The interpretation reveals that when the CDR value increases, indicating a higher severity of dementia, the model tends to predict demented cases. Conversely, when the CDR value decreases, indicating a lower severity of dementia, the model tends to predict non-demented cases. Similarly, as the MMSE score increases, reflecting better cognitive function, the model leans towards predicting non-demented cases. On the other hand, as the MMSE score decreases, indicating poorer cognitive function, the model leans towards predicting demented cases. These findings reinforce the accuracy of the neural network model and highlight the meaningful relationship between the predictors (CDR, MMSE, nWBV) and the prediction of dementia. By understanding and analyzing these predictor relationships, researchers and practitioners can gain insights into the decision-making process of the neural network model and build trust in its reliability.

It's important to note that the provided information focuses on a specific example, and the interpretability of complex machine learning models can vary depending on the dataset, model architecture, and specific domain. Therefore, it's always crucial to carefully interpret and validate the results in the context of the problem being addressed.

SHAP

The contribution of each feature to a particular prediction generated by a machine learning model is quantified by SHAP values. They offer a method to comprehend how the model generates its predictions and are specific to each prediction. The average marginal contribution of each feature over all potential coalitions that incorporate that characteristic is represented by a unique SHAP value for each occurrence. A positive SHAP value for a feature indicates that the expected probability of that instance rises as the value of the characteristic rises. If a feature has a negative SHAP value, it signifies that increasing its value will reduce the likelihood that the instance will occur. The following plots were created based on the decision tree model for all the test records and one demented record as a special case for more explainability.

Summary Bar Plot

The feature importance is plotted using a bar plot. The features are ranked according to how much influence they have on the model's prediction. The average absolute SHAP value for each feature is represented by the x-axis.

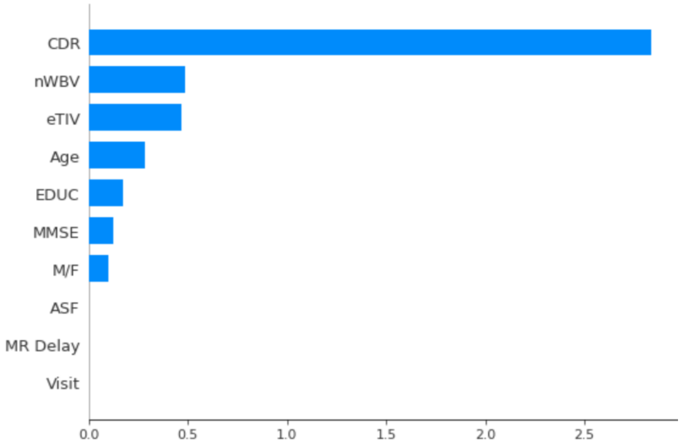


Figure 8: Mean SHAP value for all features

Figure 8 shows the average SHAP value of all selected features. From the figure, it is obvious that the CDR is the only most significant feature compared to all the features with mean SHAP values of more than 2.5. This is followed by nWBV and eTIV with an average SHAP value of 0.5 and followed by Age, EDUC, MMSE and Gender. The remaining features do not contribute to the prediction of the disease.

Summary Dot Plot

The directionality impact of the attributes is visualized using dot plot charts. The x-axis of this graph exhibits the SHAP value, and the y-axis displays all the features. In the graph, each point corresponds to a single SHAP value for a prediction and a feature. A higher value for a feature is shown by red. Features with a lower value are indicated in blue. We can generalize the directionality influence of the characteristics according to the distribution of the red and blue dots.

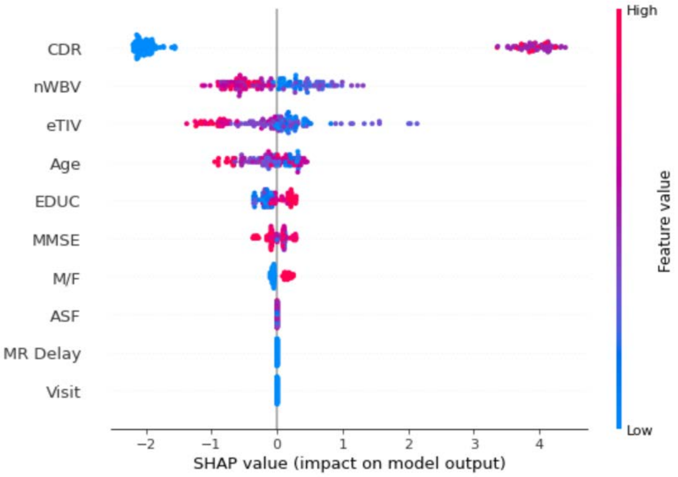


Figure 9: SHAP values and feature importance

From the above Figure 8, the CDR feature is the most significant among all the features. The inference from the Figure is that a higher value of CDR leads to higher chance of dementia and lower value leads to lower chance of dementia. The next significant features are nWBV and eTIV, where these are opposite of the CDR feature. The higher value of nWBV and eTIV leads to lower chance of dementia and lower values lead to higher chance of dementia.

Waterfall Plot

The SHAP waterfall charts are a technique to see how various features contribute to the predictions of a classification model. A bar is used to symbolize each feature, and its length indicates how much it contributed to the prediction. The features are ordered in descending order of contribution, from top to bottom, with the one with the biggest contribution at the top. Whether the feature has a favorable or unfavorable impact on the forecast is indicated by the color of the bar.

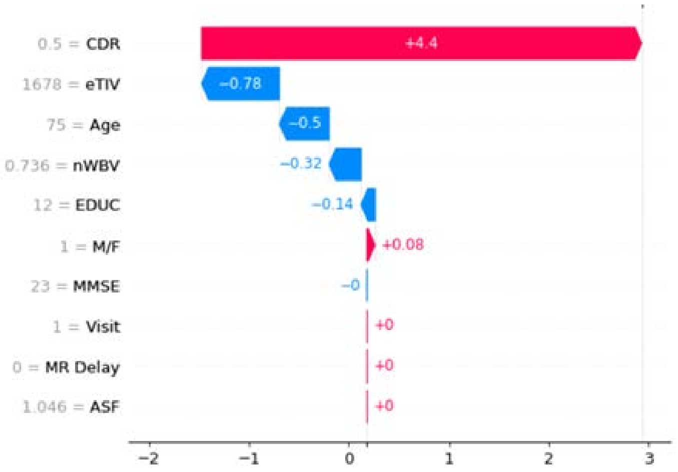


Figure 10: Waterfall plot for test record no 1

The waterfall plot shows in Figure 10 the SHAP values on X axis and all features with their corresponding value for a selected test data is available in Y axis. The red color represents the demented cases and blue color represent the non-demented outcome. From the plot, the feature CDR has the highest impact in producing the predicted output as demented and which is same as the actual outcome. The other feature contributing to the correct outcome is gender. The features eTIV, age, nWBV and EDUC are contributing for the non-demented outcome which are incorrect with the actual prediction.

Force Plot

The standard technique for displaying specific model predictions is a SHAP force plot. The probability that a person has dementia or not is predicted. Red arrows indicate feature effects (SHAP values) that drive the prediction value higher, and blue arrows indicate feature effects that drive it lower. Each arrow's size indicates the strength of the characteristic's effect. According to Figure 11, the base value, which is equal to 2.93, represents the model's average prediction over the training set. From the figure it is again explicitly shown that the CDR feature is the main contributor for the correct prediction for the selected feature set.

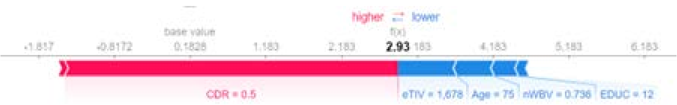


Figure 11: Force plot for test record no 1

Decision Plot

Decision plots are straightforward to interpret since they are direct representations of SHAP values. The model's base value is indicated by the straight vertical line in the decision plot. The prediction appears as a colored line. For reference, feature values are presented next to the prediction line. The prediction line depicts how the SHAP values add up from the plot to the ultimate score of the model at the top. The figure shows that the major contributors for the prediction of the correct outcome demented are the CDR followed by eTIV, Age and nWBV.

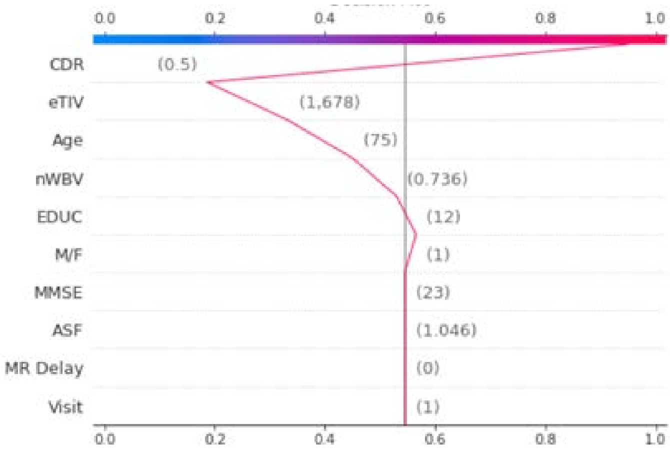


Figure 12: Force plot for test record no 1

CONCLUSION AND FUTURE

In this study, the focus was on developing a machine learning model that not only achieves high accuracy in detecting Alzheimer's disease but also ensures interpretability of its predictions. Alzheimer's disease poses significant challenges for the elderly population, leading to increased disorientation, memory loss, and other cognitive impairments. Therefore, it is crucial to have a reliable and transparent diagnostic tool that can aid in early detection and appropriate intervention.

One of the main hurdles in achieving interpretability lies in the inherent non-linearity of highly accurate machine learning models. These models, while effective in making accurate predictions, often lack transparency, making it difficult for non-technical stakeholders, such as clinicians, caregivers, and patients, to comprehend the underlying reasoning behind the model's predictions. To address this, this research employed two popular explainable AI techniques, namely LIME and SHAP, which offer insights into the decision-making process of complex models. By utilizing LIME, the researchers were able to interpret the model's predictions based on two input instances: one from demented cases and the other from non-demented cases. The linear model derived from LIME provided a clear understanding of the impact of various features on predictions. It was observed that features like CDR, Age, and ASF played a positive role in predicting dementia cases, aligning with clinical knowledge. Conversely, features such as nWBV, MMSE, and eTIV had a negative impact on the predictions, indicating their relevance in differentiating non-demented cases. Similarly, the analysis of SHAP values provided a comprehensive understanding of the model's decision-making process in Alzheimer's disease detection. The findings reaffirmed the significance of features like CDR, nWBV, and eTIV in predicting the disease, while highlighting the limited impact of certain other features. By leveraging the interpretability offered by SHAP values, the developed model gains transparency and reliability, aiding in the diagnosis and management of Alzheimer's disease.

Moreover, the researchers recognized the importance of expanding the scope of the work to enhance the model's accuracy and applicability. Suggestions for future research included validating the model with larger datasets to assess its generalizability across diverse populations. Additionally, incorporating multimodal data, such as brain imaging or genetic markers, could further improve the accuracy and interpretability of the model. The involvement of healthcare professionals and experts in the validation process would help ensure the model's effectiveness in real-world clinical settings. Overall, the combination of accuracy and interpretability in the detection of Alzheimer's disease holds great promise. By leveraging XAI techniques, incorporating diverse datasets, and addressing practical considerations, this research aims to provide a reliable and transparent tool for early detection, monitoring, and personalized interventions in the context of Alzheimer's disease.

Authorship Contribution: A.V., B.G. and V.S. contributed towards conception and design, acquisition, analysis, interpretation of data, drafting the article, revising it critically for important intellectual content and approving the manuscript to be published. H.S.S contributed towards overall review of the paper.

Potential Conflicts of Interest: None

Competing Interest: None.

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.

Acceptance Date: 05 June 2023

REFERENCES

1. What is Alzheimer’s Disease? Symptoms & Causes | alz.org. https://www.alz.org/alzheimers-dementia/what-is-alzheimers (accessed Jun. 15, 2023).
2. ADI - Dementia statistics. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ (accessed Jun. 15, 2023).

3. The Lancet Public Health: Global dementia cases set to triple by 2050 unless countries address risk factors. Institute for Health Metrics and Evaluation 2022. https://www.healthdata.org/news-release/lancet-public-health-global-dementia-cases-set-triple-2050-unless-countries-address (accessed Jun. 15, 2023).
4. https://www.facebook.com/NIHAging. Alzheimer's Disease Fact Sheet. National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet (accessed Jun. 15, 2023).
5. Khedkar S, Subramanian V, Shinde G, et al. Explainable AI in Healthcare. Rochester NY 2019.
6. Lee G, Nho K, Kang B, et al. Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 2019;9(1):1952.
7. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79(1):102470.
8. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017;2.
9. Shevskaya NV. Explainable Artificial Intelligence Approaches: Challenges and Perspectives, in 2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS) 2021;540-3.
10. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline | Nature Communications. https://www.nature.com/articles/s41467-019-10212-1 (accessed Jun. 15, 2023).
11. Yang C, Rangarajan A, Ranka S. Visual Explanations from Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. AMIA. Annu Symp Proc 2018;2018:1571-80.
12. El-Sappagh S, Alonso JM, Islam SMR, et al. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep 2021;11(1):2660.
13. Hernandez M, Ramon-Julvez U, Ferraz F, et al. Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis. PLOS ONE 2022;17(5):e0264695.
14. In-depth insights into Alzheimer's disease by using explainable machine learning approach | Sci Rep https://www.nature.com/articles/s41598-022-10202-2 (accessed Jun. 16, 2023).
15. Han SK. OASIS 2: online application for survival analysis 2 with features for the analysis of maximal lifespan and healthspan in aging research. Oncotarget 2016;7(35):56147-52.
16. McKight PE, Najab J. Kruskal-Wallis Test, in The Corsini Encyclopedia of Psychology, John Wiley & Sons, Ltd 2010;1-1.
17. St»hle L, Wold S. Analysis of variance (ANOVA). Chemom Intell Lab Syst 1989;6(4):259-72.
18. Ugoni, Walker BF. The Chi Square Test. Comsig Rev 1995;4(3):61-4.
19. Morocho-Cayamcela ME, Lee H, Lim W. Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. IEEE Access 2019;7(1):137184-206.
20. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Jan. 22, 2023. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
21. Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier. arXiv, Aug. 09, 2016. Accessed: Jan. 22, 2023.
22. Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. Proc AAAI Conf Artif Intell 2018;32(1):11491.