INTELLIGENT ALZHEIMER'S DISEASE PREDICTION USING EXPLAINABLE BOOSTING MACHINE

ARCHANA MENON P[1] , R. GUNASUNDARI[2]

[1]Research Scholar, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India, archananirmal1414@gmail.com, ORCID 0000-0003-3103-2200

[2]Professor & Head, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, India, gunasoundar04@gmail.com

ABSTRACT

Alzheimer's Disease (AD), a progressive brain disorder, poses a growing health challenge. Early detection is crucial for providing proper treatment and to prevent its progression. Revolutionary deep learning models used in AD prediction exhibit high performance compared to simpler models while its black box nature makes the model capricious for the clinicians to make decisions. This paper aims to propose a ML model for the accurate detection and prediction of AD in an explainable way. Feature selection techniques are employed to maximize the relevancy of features with the class labels. Among the different glass-box and black-box models inspected to prognosticate AD, Explainable Boosting Machine (EBM) with Chi-square feature selection could generate more accurate and explainable results even in small datasets. The interpretability graph of EBM delivers both global and local explanation for the predicted results and identifies the features responsible for pulling the prediction towards a particular class. EBMs foster trust and transparency in model's decision-making process. The proposed model obliges the medical practitioners to take better and confident decisions.

Keywords: *Alzheimer's Disease, Black-box, Explainability, Explainable Boosting Machine, Glass-box, Feature Selection*

1. INTRODUCTION

Alzheimer's Disease (AD) is a progressive disorder that shrinks brain and damage brain cells destroying memory and mental functions. 60%-70% of AD leads to dementia and AD is one of the leading reason for dementia [1]. AD is a rapidly growing health concern affecting millions of people and the number of AD cases will be rising significantly in the coming years. AD is hard to predict. Early detection of dementia can help patients receive perfect treatment on right time and also to prevent its progression. Traditional diagnostic methods such as cognitive assessments can be expensive, time consuming and subjective and may miss early signs of the disease. The lack of definitive biomarker for AD makes the diagnosis of this disease very challenging. Artificial Intelligence (AI) techniques are used to predict a wide range of diseases including AD. They can analyze vast datasets of medical records, brain scans, genetic data etc. to discover patterns and predict AD risk. Machine Learning (ML) techniques such as Linear Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT) etc. are simple models but are not that accurate. Neural Networks (NN) and other Black Box models perform well but are hard to interpret. These models are computationally expensive too. Ensemble models such as XGBoost (XGB), Gradient Boosting etc. performs better but interpretability is harder in these models.

Most of the AD detection models which produced good accuracy were based on Deep Learning (DL) algorithms. However, DL models are data hungry and they need huge datasets to show accurate results. Acquiring huge datasets in medical field is a really challenging task. Also, It is always hard to balance between accuracy and interpretability. Simple models such as LR and DT are interpretable but gives low accuracy whereas complex models such as ensemble models, NN etc. gives good accuracy by compromising model interpretability. This lack of transparency hinders trust in the model and limits its integration into clinical practice. Clinicians need to understand the factors influencing a model's prediction to feel confident using it for patient care. The dataset may come in varying sizes. With huge number of features, it will be difficult for the ML models to take accurate decisions. For automatically choosing important features from the large set of input features, feature selection techniques can be employed. Doing so, reduces the dimension of the dataset as well as noise by removing unwanted or irrelevant features for the given problem. Feature selection techniques also helps to gain improved accuracy by making the problem simpler and more understandable [2].

The idea behind this paper is to come up with a model which is transparent in its prediction process as well as to produce a good performance accuracy with a smaller number of features. Here, an ML model- Explainable Boosting Machine (EBM)- is proposed to predict AD at an early stage, that prioritizes both accuracy and transparency. EBM is an ML model which combines the power of boosting and interpretability. Unlike complex DL models, the focus here is to develop simpler yet effective methods. By employing feature selection techniques, the model enhances prediction accuracy while maintaining transparency in the decision-making process. The gradient boosting technique optimizes the model by minimizing the loss function. EBM is capable of providing insights on how the model has arrived at such predictions. EBMs can visually explain the predictions. The explainability and flexibility of EBM helps to solve complex ML problems where transparency and interpretability are important. This approach has significant advantages such as follows:

1. Enhanced Trust: Clinicians require transparency to trust, integrate and adopt ML predictions into patient care. EBM builds trust by explanations and reveals the factors most influential in model's prediction.
2. Mitigating Bias: Finding out the rationale behind predictions help to identify and mitigate biases within the data and model.
3. Novel insights: Recognizing the factors influencing model predictions lead to new research avenues.

This paper aims to contribute to the development of accurate, transparent and trustworthy model for better patient care and improved outcomes by leveraging the power of EBM. It also incorporates feature selection techniques such as chi-square and L1 regularization to identify most relevant features contributing to model's prediction of AD risk. This research focuses on utilizing existing data sources and feature selection techniques. This work has been evaluated on a research level and not been validated through clinical trials and also have not delved into the development of new biomarkers for AD. To the best of our knowledge, this is the first work using EBM that uses OASIS brain MRI dataset for Alzheimer's prediction. The paper has the following main contributions:

1. Investigates different categories of ML models and proposes the best model for the accurate prediction of AD.
2. Feature selection techniques selects appropriate features for the models and thereby improves performance accuracy.
3. Investigates various interpretable models for the better understanding of predicted results.
4. Extracts knowledge and generates explanation from the results attained from the explainable model.
5. Overcomes the traditional accuracy-interpretability trade-off problem.

The rest of the paper is organized in the following manner: Section II reviews the related literature, Section III covers the motivation behind this work and other background studies. Section IV details the proposed predictive model and analyses different categories of ML algorithms employed. Section V includes the experiments, results and discussions. The paper concludes with Section VI.

2. LITERATURE REVIEW

Many researches had been carried out for predicting AD, using various ML and Deep Learning (DL) algorithms. Not many researchers have used EBM in their work. A few recent studies which used EBM model in various domains are reviewed here. In 2021, AD prediction was made from MRI Hippocampal subfields using EBM [3]. 200 brain MRIs of patients with Mild Cognitive Impairment (MCI) were divided into equal parts as progressive (pMCI) and stable (sMCI). They got prediction accuracies of 80.5% and 84.2% for EBM without and with pairwise interactions respectively.

In 2020, Harsha Nori & et al experimented multiple classification and regression datasets using EBM and applied Differential Privacy [4]. The major benefits of their DP-EBM model are: (i) it yielded good accuracy (ii) it provided strong differential privacy (iii) it provided exact global and local interpretability and finally (iv) the models can be edited without loss of privacy even after training. A flawed dataset misdirects machine learning models to learn defective knowledge from it which in turn caused incorrect classification or prediction by the model. Flaws in the dataset may not be identified easily. Using visualization of EBM shape function and domain knowledge, Zhi Chen & et al detected some common flaws in data such as data drift, missing values, bias and fairness, confounding and treatment effects, and outliers [5]. This helps the users to detect such problems hidden in their data.

Pathologic complete response (pCR) is an important factor to determine if a patient with rectal cancer (RC) should have surgery after neoadjuvant chemoradiotherapy (nCRT). [6] proposed a model using EBM to predict the pCR of RC patients following nCRT. This model helped to avoid the involvement of a pathologist for analysis and assessment of pCR. The interpretable classifier, EBM is applied by [7] to predict the presence of pests in the field. It considered insect traps, weather predictions and vegetation indices to anticipate the commencement of bollworm harmfulness in cotton fields in Greece. [8] proposed an explainable ML model to forecast the compressive strength of concrete. The experiment was carried out under different mix ratio conditions as the mixing ratio of raw materials impacts on concrete compressive strength. They could also determine the impact of each combination ratio parameter on concrete compression strength.

The ability of non-ductile reinforced shear walls to withstand deformation was predicted by [9] based on experimental data. With the aid of the model, the user could perceive the relationship between wall properties and deformation capacity. They could quantify the individual contribution of wall properties and the correlation among the properties. Similarity between the characteristics of feature importances -such as incomplete, top-weighted and indefinite- are quantified by [10] using the RBO (Rank-biased Overlap) score. They conducted a case study on Parkinson's disease and is classified using EBM with the help of RBO. The experimental results showed that RBO exhibited maximum size of overlapping. They could also demonstrate that when feature importances are similar, for the same accuracy, the model becomes more stable and reliable.

[11] carried out a study on the likelihood of slope failure in 4 separate areas in West Virginia and USA based on digital terrain characteristics. They implemented different ML algorithms among which EBM outperformed others. The complications and risk factors for mothers and babies post birth had been analyzed and gained intelligible interpretations of the features contributing to risk by [12] using EBM. They focused on three types of risks such as shoulder dystocia, severe maternal morbidity and preterm preeclampsia and the results showed that the accuracy of EBM matches with other deep neural networks. This helps the obstetricians to take better and timely decision. Table 1 shows the summary of recent papers reviewed.

Table 1. Review Of EBM Papers From Different Domains

Year	Author	Title	Description
2021	Alessia Sarica & et.al [3]	Explainable Boosting Machine for Predicting AD from MRI Hippocampal Subfields	EBM is used to add intelligibility for AD prediction. It not only produced good accuracy but also helped to identify which hippocampal subfields of brain MRI drove for the particular prediction.
2021	Harsha Nori & et.al [4]	Accuracy, Interpretability and Differential Privacy via Explainable Boosting	Adding differential privacy to EBM for training ML models help to achieve privacy along with good accuracy. Apart from privacy and high accuracy, applying DP to EBM could produce 2 more benefits- the model could bring global as well as local interpretability and the model was able to edit even after training.
2021	Zhi Chen & et.al [5]	Using Explainable Boosting Machines to Detect Common Flaws in Data	EBM discovers dataset flaws such as missing values, data drift, bias, outliers, etc. Using some case studies, it is proved that when data correction is difficult, EBM provides simple tools for correcting problems.
2022	Du Wang & et.al [6]	Interpretable Machine Leraning for predicting pathologic complete response in patients treated with chemoradiation therapy for rectal adenocarcinoma	pCR helps to determine if a parient with rectal cancer should undergo surgery after nCRT. This model predicts pCR of RC patients following nCRT.
2022	Ornela Nanushi & et.al [7]	Pest Presence Prediction Using Interpretable Machine Learning	Considering various factors such as vegetation indices, weather predictions, insect trap, the presence of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

			insects is predicted using interpretable EBM.
2023	Gaoyang Liu & et.al [8]	Concrete compressive strength prediction using an explainable boosting machine model	Forecasted the compressive strength of concrete considering various features. The impact of each mix ratio parameter on the concrete compression strength.
2023	Zeynep Tuna Deger & et.al [9]	Estimate Deformation Capacity of Non-Ductile RC Shear Walls using Explainable Boosting Machine	A transparent model which predicts the estimate of the deformation capacity of reinforced concrete shear walls.
2022	Alessia Sarica & et.al [10]	Introducing the Rank-Biased Overlap as similarity Measure for Feature Importance in Explainable Machine Learning: A Case Study on Parkinson's Disease	Similarity between the characteristics of feature importances is measured using RBO (Rank-biased Overlap). Parkinson's disease is classified using EBM with the help of RBO.
2021	Aaron E.Maxwell & et.al [11]	Explainable Boosting Machines for Slope Failure Spatial Predictive Modeling	Predicts the probability of the occurrence of slope failure 4 different Major Land Resource Areas.
2022	Tomas Bosschieter & et.al [12]	Using Interpretable Machine Learning to Predict Maternal and Fetal Outcomes	Predicts the important risk factors of births in mothers and babies which helps the obstetricians deliver better case.

3. PRELIMINARIES

3.1 Motivation

Deep neural networks and other complex ML algorithms are used by the developers to solve many of the real-time and sophisticated problems. The major advantage of such models is that they produce good performance accuracy. But these models are opaque in nature. i.e., the end users would not understand that why the system has come up with that particular prediction, which are the factors influencing the prediction, what all should be done to change the current prediction etc. Nobody has control or understanding on the inner working of the system. This blackbox approach is becoming a bottleneck to such efficient models. One cannot depend only on classification accuracy for making crucial decisions. Here comes the necessity of explainability of the models.

Certain domains need the model to explain why or how it come up with such a prediction. In healthcare industry, a patient has the right to know why his MRI result is classified as tumor. An applicant should know, based on what criteria is his loan rejected. And a judicial can't simply sentence a person to imprisonment based on the results from AI model. All these scenarios need explanation or transparency to the model which came up with such a prediction. The transparency helps to increase the trust of end users toward the model.

Providing explanations is crucial in ensuring fairness, privacy, reliability and trustworthiness in the models utilized [13]. Balancing interpretability and accuracy consistently entail making compromises. Consequently, crafting a model that is both highly accurate and easy to interpret is a formidable challenge [13].

3.2 Explainability

Explainability or interpretability refers to the extent to which a model can be comprehended by humans [14]. The ML model adopts a black-box methodology and it hardly explains the results which are comprehensible to humans. A model has the responsibility to provide explanations at both the global level, which encompasses the overall behavior and significant factors it considers, and the local level, which delves into the specifics of individual predictions. Global explanations aid in comprehending the model's priorities and can unveil potential decision-making shortcomings, while local explanations reveal the process behind a single prediction [14].

We have inherently explainable models such as LR, DT etc. The reason behind the predictions made by these simple models can be understood by the end users easily because they are interpretable models by-design. But they do not perform well always. For complicated algorithms, developers have to use interpretable techniques on the already built models. This approach of achieving interpretability by detaching explanations from the models is referred to as model-agnostic techniques [15]. These techniques are responsible for giving the explanations of the predictions made by the opaque models. Partial Dependency Plot (PDP), Global Surrogate, Local Surrogate (LIME), Shapley values, SHAP etc. are some of the examples of model-agnostic models [15].

3.3 Glassbox v/s Whitebox Models

ML algorithms falls under either glasssbox or blackbox model.

3.3.1 Glassbox models

Glassbox models are crafted to be entirely interpretable, frequently delivering comparable accuracy to cutting-edge methods. They are capable of furnishing explanations at both global and local scales. In the context of glass box models, these explanations are precise, meaning they accurately elucidate the precise process the model followed in making its decision. Such explanations serve as valuable tools for conveying to end users which factors held the greatest sway in a given prediction. LR and DT falls under Glass box model [15].

3.3.2 Blackbox models

Black box models are characterized by their ability to produce results or make decisions without revealing or illustrating the underlying methodology. The internal mechanisms and the specific factors and their weights utilized in these models remain concealed. This opacity leads to a lack of transparency in black-box models. Interpretability methods for black-box models can extract explanations from any machine learning algorithm or framework [15]. These methods are useful in cases of pipelines where not all components are directly interpretable. This encompasses model ensembles, pre-processing steps and intricate models like deep neural networks. Figure 1 [16] shows different ways of explaining AI models.

InterpretML is an open-source package that comprises cutting-edge ML interpretability techniques [16]. It exposes two types of interpretability- glassbox and blackbox. This package also has a visualization platform [16].

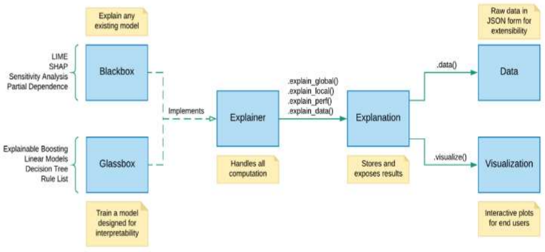

Figure 1. Different Ways Of Explaining AI Models In Interpretml

3.3.3 ML models employed in the study

The ML models implemented and analyzed in this experiment are explained below.

3.3.3.1 Logistic Regression

It is a simple and well analysed model. When the dependent variables are categorical, logistic regression is used [17].

3.3.3.2 Decision Tree

Decision trees are another simple and interpretable models. They are non-parametric supervised algorithms. A decision tree can represent the decisions and decision making visually. They are simple to understand and easy to interpret [18].

3.3.3.3 XGBoost

This algorithm is an implementation of optimized gradient boosted decision trees which focuses on speed and performance. It comes under ensembling techniques which comiles the predictions of multiple weak learners to produce a strongr prediction. It can efficiently handle missing values [19].

3.3.3.4 ExtraTree Classifier

It stands for Extremely Randomized Tree Classifier. This is an ensembling algorithm which aggregates the results of multiple decision trees. Each tree is built using the original sample. Each node will have random k features from the feature-set. Using some mathematical criteria such as Gini Index, Information gain etc. best feature is selected to split to separate the samples of a node into two groups [20].

3.3.3.5 Voting Classifier

In this context, various different ML algorithms are integrated and employed to predict class labels by either taking a majority vote (hard voting) or considering the average predicted probabilities (soft voting) [21].

3.3.3.6 Multi-Layer Perceptron

An MLP is a type of fully connected feedforward Neural Network characterised by its architecture, which includes one input layer, one output layer and the option to have any number of hidden layers in between. MLP solves supervised learning problem [22].

3.3.3.7 Explainable Boosting Machine

EBM developed by Microsoft is an interpretable model which is based on boosting technique. There arises a trade-off between accuracy and interpretability of a model. It is generally considered that simple glass box models are more interpretable while its accuracy remains low and black box models such as complex neural networks yield better predictive performance while it remains opaque. Black-box explainers works on top of black box models and these explainers give shallow explanations. Glass-box models are by-design interpretable. The major highlight of EBM is that it performs well and at the same time deliver informative explanations [23]. Figure 2 [23] shows performance-intelligibility trade-off of EBM and other ML algorithms.

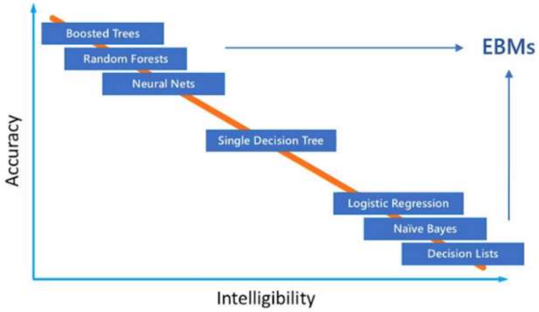

Figure 2. Performance-Intelligibility Trade-Off Of Explainable Boosting Machine

The above graph can be interpreted as follows: Simple models such as LR, DT etc. are not much accurate in general. Complex black box models such as Neural Networks perform satisfactorily but are arduous to interpret. Ensemble models such as Random Forests, boosted trees etc. shows superior performance but are very hard to interpret. EBMs got a space in the top-right corner of the graph which is a consolation for the developers. EBM helps to build models that are both accurate and easy to interpret as well.

EBM belongs to the family of GAMs (Generalized Additive Model). GAM generalizes simple regression models. A GAM considers each feature and it takes the form

$$g(E[y]) = \beta_0 + \sum f_j(\mathrm{x_j}) \qquad (1)$$

In Eq. (1) [24], y is the prediction, $\mathrm{x_j}$ represents input feature, f_j represents feature function, g is a link function which adapts GAM to regression or classification models [24]. GAM ignore the interactions between different features which is considered as its major drawback. Also, GAM's classification accuracy is low.

EBM is built from Generalized Additive Models with Pairwise Interactions (GA^2M). In EBM, each feature function is learned through a combination of bagging and gradient boosting techniques. During boosting, the model is trained by considering one feature at a time in a round-robin fashion, typically with a very low learning rate. Here the feature order does not have any impact on the output [24]. Thus, it eliminates the effects of co-linearity. GAMs are by design explainable. It also detects pairwise interaction between the terms. This model yields both intelligible and accurate results [25] and it takes the form

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}\left(x_i, x_j\right) \qquad (2)$$

(x_i,x_j) in Eq. (2) [23] represents the pairwise interaction of features. EBMs are generalized regressions, parts of which are learned by gradient boosting. In EBM, a lot of trees are trained and each tree explains the error made by the previous one. Initially each tree is built using a single feature which would have a small depth. For each feature from 1 to n, m trees are trained thereby producing a total of n*m different trees. First, we build a tree with $\mathrm{feat_1}$, and update the residuals using gradient boosting and then build the second tree using $\mathrm{feat_2}$ and so on. By aggregating all the trees feature-wise produces a contribution graph for respective features [24]. These graphs display the individual contributions of each feature towards the ultimate prediction. Function f is composed of sums of all small trees. And finally, all the trees are deleted as they are not required for prediction. Now the trained model only consists of contribution graphs. For prediction, the values of contribution graphs would be passed to the function g [23]. EBM uses both bagging and boosting during the training phase. Figure 3 [23] shows how each tree is built, and residuals are carried over, and sums of all tree is taken for building EBM.

Here, a very small learning rate (residual) is used to prevent bias towards a particular feature. EBM is a fast implementation of GA^2M and is parallelizable. EBMs tend to have a slower training process when compared to modern algorithms. However, they are known for being compact and highly efficient during prediction time [24]. Each feature contribution can be plotted and visualized using f_j which makes EBM interpretable. EBMs are a part of InterpretML [16].

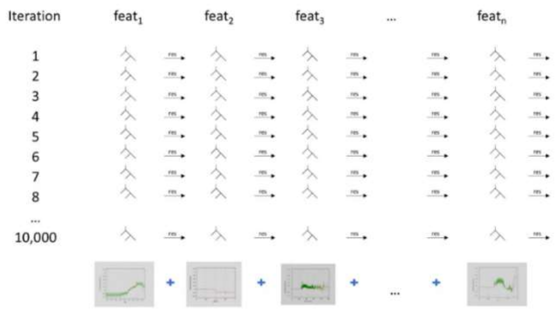

Figure 3. Building EBM

3.4 Feature Selection techniques

Feature selection techniques in ML finds the best set of features to build an optimized model. By incorporating feature selection techniques in the model, the accuracy can be improved as well as the training time and overfitting can be reduced [26]. Most common methods for supervised feature selection are filters, wrappers, embedded and hybrid methods [2]. There are a plenty of techniques available under each method. The feature selection techniques implemented in this study are:

3.4.1 Chi-Square Method

It comes under filter method where the features are selected based on chi-square score regardless of the ML model employed. It performs a simple statistical test for the categorical features in the dataset [2]. From a given set of data, Chi-square evaluates the independence of each feature to the target variable and evaluates the significance of each feature.

LASSO Regularization (L1)

Regularization adds a penalty term to the differet parameters of the ML model to prevent overfitting and improve generalizations. L1 even shrinks some coefficients to zero, thereby removing redundant or irrelevant features and prevents a model becoming too complex [26]. This technique follows an embedded approach and possess the benefits of both the filter and wrapper methods.

4. METHODOLOGY

4.1 Overview

This research employs a ML model with explainability design o detect the Ad in an accurate and interpretable way. The data collected form OASIS dataset contains demographic information of human beings, cognitive test scores, brain scan measurements etc. As a next step, clean and pre-process the data by handling missing values, outliers and scaling numerical features. And then employ feature selection techniques such as chi-square and L1 regularization for identifying relevant information contributing to the AD risk prediction. Choose EBM and train the model and obtain the feature importance scores. Evaluate the model performance using standard metrics, analyze the feature importances and interpret the results globally and locally. This study proposes different categories of ML algorithms for the prediction of AD. Figure 4 shows the ML algorithms employed for model training. They are as follows:

- Simple and by design interpretable models, which comes under the category of Glass-box models, such as LR, DT, and EBM.
- As ensemble techniques and neural networks produces greater accuracy in general, these models are also employed in this experiment. But these Black-box models are not interpretable. XGBoost, Extra Tree Classifier and Voting Classifier are used as Ensemble techniques.
- MLP is used to build a basic Neural Network.

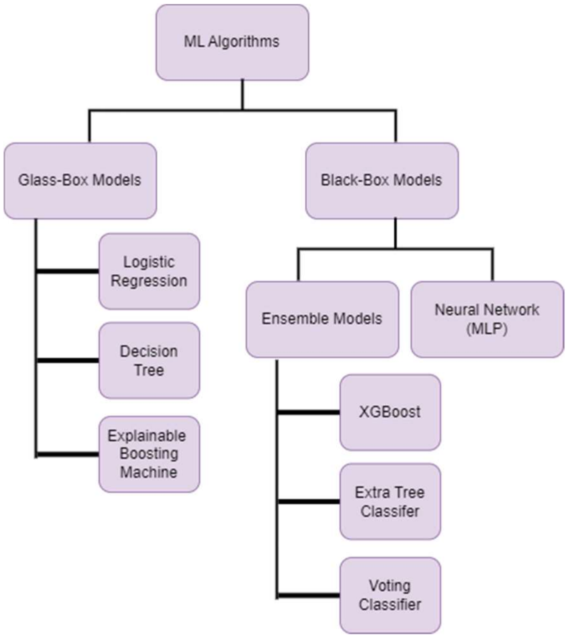

Figure 4. ML Algorithms Employed For Training The Model

4.2 Dataset Description

The Alzheimer’s dataset consists of brain MRI longitudinal data from Open Access Series of Imaging Studies (OASIS) [27]. OASIS refers to a collection of neuroimaging datasets that are accessible to the public for use in research purposes. A total of 150 subjects are examined and 14 important attributes such as age, SES, Years of Education, Socioeconomic status, Gender, and other important features from MRI values such as MMSE, CDR, eTIV, nWBV, and ASF [27], which are useful for the classification task, are considered for building the dataset. Each person is classified into either demented or non-demented category.

4.3 The Proposed Predictive Model

The proposed framework is summarized in Figure 5. The longitudinal MRI dataset from OASIS [27] is taken for the prediction of AD. Data is preprocessed by handling the missing values and encoding the categorical data. Then the feature selection techniques such as Chi-square and L1 regularization are applied on this categorical data. Afterwards, the dataset is split for training and testing purposes. The model is trained with various categories of ML algorithms such as Glass-box Models – LR, DT and EBM, and Black-box models- XGBoost, RF, ExtraTree Classifier, Voting Classifier, and MLP. Both the Glassbox and Blackbox algorithms are implemented and EBM outperformed the rest of the algorithms in terms of performance accuracy and interpretability. EBM is a tree-based model that employs cyclic gradient boosting and Generalized Additive Model principles, and it automatically detects interactions between variables [24]. They are notable for achieving a level of accuracy that is comparable to state-of-the-art black-box models, all while maintaining complete interpretability. EBM shows an accuracy comparable to gradienr boosting algorithms (XGBoost, LightGBM) and is interpretable as well. Hence, EBM is considered for building the framework which facilitates better accuracy and interpretability in AD prediction.

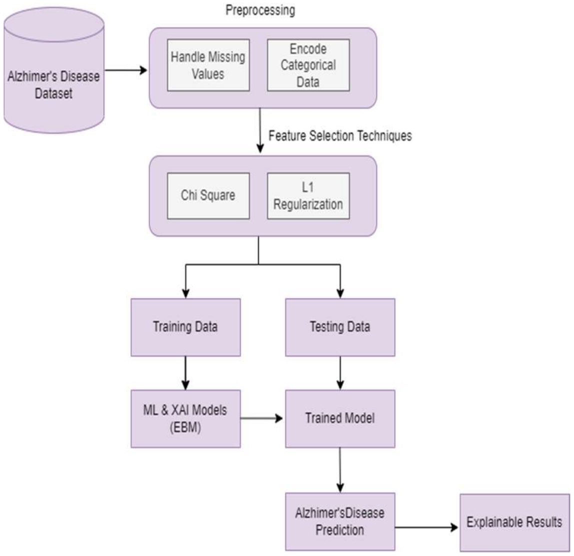

Figure 5. Proposed Predictive Model

Algorithm 1: Interpretable Alzheimer’s prediction using EBM

```
Input: Patient Details $f(x_1, x_2,...x_{14})$
Output: The class, $C_f$ of the patient; Interpretability graphs $Gg$ for global and $G_l$ for local explanation to the prediction
Step 1: Perform pre-processing on $f(x_1, x_2,...x_{14})$ to obtain $f_p(x_1, x_2,...x_{10})$
    Step 1.1: Remove 4 fields which has no corelation to the label
    Step 1.2: Encode Categorical data
        Step 1.2.1: Replace the values of the attribute M/F, with 0 & 1 respectively,
        Step 1.2.2: Replace the values of the label Nondemented and Demented with 1 & 0 respectively
Step 2: Input $f_p(x_1, x_2,...x_{10})$ for feature selection. The new reduced dataset after applying feature selection techniques will take the form $f_{fs}(x_1, x_2,...x_n)$.
Step 3: Input $f_{fs}(x_1, x_2,...x_n)$ into the trained EBM model and obtain the class $C_f$
Step 4: Obtain the global interpretability graph $Gg$ from EBM
Step 5: Obtain the local interpretability graph $G_l$, from EBM
Return $C_f$, $Gg$ and $G_l$
```

Algorithm 1 details the step involved in categorising a patient into demented or non-demented class and giving the explanation for the reason behind the paticular prediction. General patient details collected such as age, gender, social status, etc. along with the details such as MMSE, CDR etc. collected from MRI images are considered as input attributes. A total of 14 attributes are there in the dataset and hence the i ut function looks like $f(x_1, x_2, \dots x_{14})$. Among these 14 attributes, only 10 are considered for training the model. Rest 4 attributes – MRI ID (Subject ID uniquely identifies each patients and MRI ID is a duplicate information), Visit (No. of visits made by the patient), MR Delay (the delay made after taking 1st MRI), Hand (Dominant hand-left/right)- doesn't have an impact on the prediction. Since they have zero corelation with the label, these 4 attributes are removed as a part of the pre-processing step. Categorical values such as M (Male) and F (Female), Nondemented and Demented are encoded with 0 and 1 respectively. After pre-processing, the dataset is left with 10 attributes and is given to the feature selection algorithms Chi-square and L1 regularization which gives a new reduced featureset as its output $f_{fs}(x_1, x_2, \dots x_n)$, where *n* represents the new reduced number of features. Then this new featureset $f_{fs}(x_1, x_2, \dots x_n)$ is given as an input to the EBM model for training. When a new unseen input comes, it is given to the trained EBM model for predicting the AD. The model then classifies the person either into demented or non-demented class C_f. EBM also generates 2 type of interpretability graphs Gg and G_l for both global and local explanation respectively. Gg graph shows the positive and negative contribution of each feature to the model which in turn is responsible for the overall prediction performance. G_l graph explains the reason behind the prediction for a given single input. It shows individual feature contributions which are responsible for the particular prediction. EBM not only considers individual feature contribution but also the interaction bteween features. The details of the implementation of Algorithm 1 is given in the Section 5.

5. EXPERIMENTS, RESULTS & DISCUSSIONS

To assess and compare the effectiveness of EBM, some of the glassbox and blackbox algorithms are implemented here in OASIS MRI Brain image dataset for Alzheimer's prediction.

5.1 Experiments

Algorithms from different categories are employed to compute the efficiency of the proposed model.

5.1.1 Logistic Regression

Since this model is simple, the results are also interpretable. It had shown the least accuracy of 60% without any feature selection and 65% in Chi-square tecnique.

5.1.2 Decision Tree

This is another by design explainable model. This algorithm generated a performance accuracy of 71% and 75% without feature selection and with Chi-square feature selection respectively.

5.1.3 XGBoost

A boosting technique which is extensively used by the researchers in different domains produced an accuracy of 86% and 89% respectively without feature selection and with Chi-square feature selection respectively.

5.1.4 ExtraTree Classifier

This ensemble classifier produced an accuracy of 89% and 90% respectively without and with feature selection techniques.

5.1.5 Voting Classifier

This model generated an accuracy of 90% without feature selection and 92% with Chi-square feature selection, which is the second highest performance accuracy among all other models.

5.1.6 Multi Layer Perceptron

A Simple Neural Network is also employed in this experiment. It could produce only a low accuracy of 61% without feature selection but 84% with Chi-square feature selection which is a greater improvement.

5.1.7 Explainable Boosting Machine

This model is by design interpretable and produced the highest performance accuracy 93% without any feature selection techniques and 95% and 94% respectively with Chi-square and L1 feature selection techniques respectively. The major highlight of this model is that the results of this model are explainable. It could generate both global and local explanations.

5.2 Results & Discussions

Table 2 shows the results of classification performance accuracy of the ML models employed in OASIS Brain MRI dataset for AD prediction. The feature selection methods applied on the dataset after preprocessing are Chi-square and L1 regularization. Chi-square calculates the chi-scores and ranks the features. Lesser the chi-square value, the more the feature is independent of the class. By analyzing the Table 2, we can see the positive effect of feature selection techniques applied for AD prediction. L1 regularization reduces the variance of the model and yields good results. Both Chi-square and L1 regularization techniques could improve the prediction accuracy of each models. The idea behind incorporating feature selection techniques is that not all features are useful in predicting the AD. The dataset may contain redundant and irrelevant features which might lead to the overfitting of the models. Applying feature selection techniques have reduced the problem of over-fitting and also established a prime search area for the classification. It can also be observed that all the models could produce greater accuracy with Chi-square feature selection compared to L1 regularization.

The performance of glass box models, black box models and neural network are analysed in this experiment. The experimental results show that EBM outperformed all the other models. Different categories of ML algorithms such as glass box (LR, DT, EBM), ensemble techniques (XGB, ExtraTree Classifer, VC) and Neural Network (MLP) are investigated in this experiment to find out which algorithm gives more accurate and interpretable results. From Table 2, we can infer that among all the algorithms implemented, EBM trounced all the other models with a performance accuracy of 95% in AD classification and Voting Classifier acquired the second highest accuracy of 92%. EBM algorithm is preferred to other algorithms because of the fact that EBM models are self-explainable whereas the prediction made by the voting classifier and Extra Tree Classifier are difficult to understand. We could generate both global and local explanations from EBM. Moreover, a practical reasoning on why a person is correctly or incorrectly classified to a particular class is also provided by EBM.

Table 2. Results Of Classification Performance Accuracy Of ML Models In Alzheimer's Disease Dataset

Sl. No.	Model Name	Alzheimer's Disease Performance Accuracy		
		Without feature selection	**Chi-square**	**L1 Regularization**
I	**Glass Box Models**			
1.	Logistic Regression	0.601905	0.651292	0.620689
2.	Decision Tree Classifier	0.714762	0.754540	0.726207
3.	Explainable Boosting Machine	**0.932381**	**0.954335**	**0.943552**
II	**Black Box Models**			
4.	XGB Classifier	0.864571	0.892483	0.870000
5.	Extra Tree Classifier	0.890762	0.904550	0.902552
6.	Voting Classifier	0.902381	0.924335	0.913103
7.	MLP Classifier	0.61410	0.845851	0.682758

Even though we have black-box explainers which work on top of complex black box models, they deliver very shallow understanding of the model, whereas glass-box models are by design interpretable. The complex shape function from gradient boosting functions in EBM ensures the accuracy of the model. The aggregate value of individual shape functions shows the sole contribution of each feature for final prediction and there lies interpretability of the model. EBM not only gives us result, but provides the end user with extra information about the predicted result. Here, it shows which are the features primarily responsible (both positively and negatively) for a particular prediction. The intelligent results help clinicians to take better decisions. EBM uses light memory and predicts the results faster.

5.3 Comparison of the proposed model with existing models

Table 2 shows the results of some of the earlier works carried out for the diagnosis and classification of AD using ML, DL and other visualization techniques. It is clear that the proposed EBM model has shown an improved performance accuracy compared to the other models. It can be observed from the Table 2 that the proposed EBM model can classify AD more convincingly than the other existing models discussed.

Table 2. Comparison Of The Proposed Model With Existing Models

Reference	**Year**	**Model**	**Performance Accuracy**
Proposed Study	2024	**EBM**	**94.35%**
[28]	2021	CapNet	92.29%
[29]	2022	ResNet-50	90%
[30]	2023	XGBoost-SHAP	87.57%
[31]	2020	RF Classifier	86.84%
[32]	2021	VGG16-LIME	86.82%
[33]	2021	BrainNet3D	80%
[34]	2017	Deep CNN	73.75%

The application of EBM in various domain to get accurate and interpretable results is proven in the literatures [3-12]. It is proved that the application of EBM in AD risk prediction is also very effective [28-34].

5.4 Interpretability Graphs

OASIS Brain MRI dataset is considered for the illustration of Interpretability graphs.

5.4.1 Global Explanations

This explains the overall contribution of features to the model. Contribution of induvidual features and its relation to the model would be clearly shown here. These explanations shows us what a model finds important. It also helps us to identify the potential flaws of a model in decision making [35]. The global importance of each feature is calculated by taking the mean absolute contribution of a feature or its interaction to the predictions averaged across the training dataset [35]. Figure 6 shows the global feature importances of the proposed model. The graph in Figure 6 exhibits 15 most important features taken by the proposed model. These features are helpful to find out whether a person is healthy or demented. Thus, EBM helps to understand data.

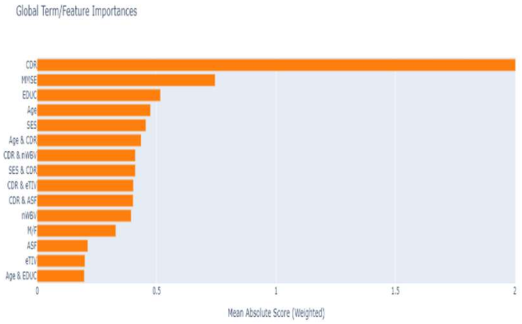

Figure 6. Global Feature Importances Of The Proposed Model

The contribution (score) of the term CDR to predictions made by the model is depicted in the graph below Figure 7. Each graph is centered vertically such that average prediction on the train is 0 [25]. From the graph, it is clear that CDR influences a lot after the value 0.2.

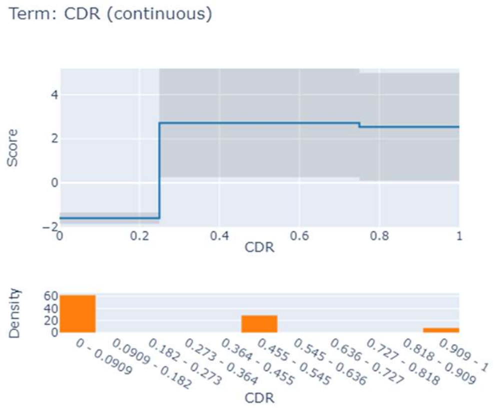

Figure 7. Contribution Of The Term CDR To The Predictions Of The Model

The contribution (score) of the term Age to predictions made by the model is depicted in the graph below in Figure 8. Here, it is evident that there is no influence of Age until 68. After that the Age has its highest positive influence between 80 and 85 and then it dips. Since this is a classification problem, the scores are on a log scale.

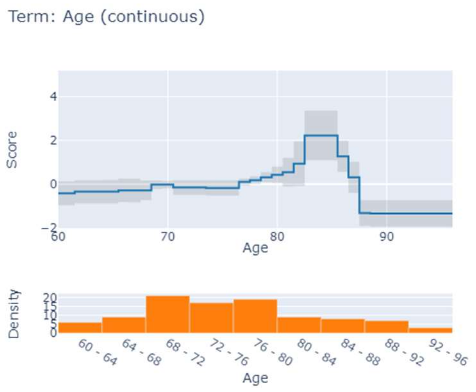

Figure 8. Contribution Of The Term Age To The Predictions Of The Model

The contribution of the features Age & CDR to predictions made by the model is depicted in the graph below Figure 9. Its understood from the graph that just after the value 0.2 in CDR, its risky and hence it classifies the patient as demented. If both Age and CDR are high, then it indicates a bad sign. Here, we can notice the interaction between terms.

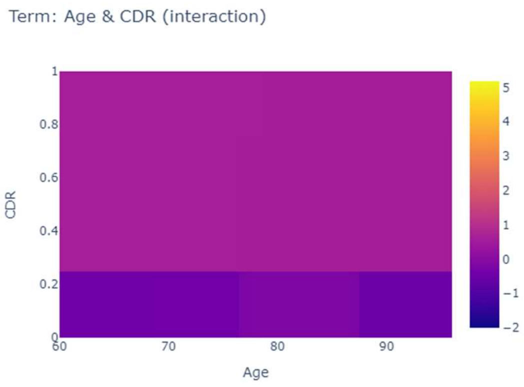

Figure 9. Contribution Of The Term Age& CDR To The Predictions Of The Model

Likewise, the contribution (score) of each terms and interactions can be viewed in the graph.

5.4.2 Local Explanations

This explains individual predictions (at local level). To make individual predictions, the model makes use of each term graph as a reference. It identifies the contribution per term, and sums them together with the learned intercept for making the prediction [28]. These graph explans why a person is classified as demented or non-demented. It shows which all features are responsible for that particular prediction. The graphs given below gives local explanations for different instances.

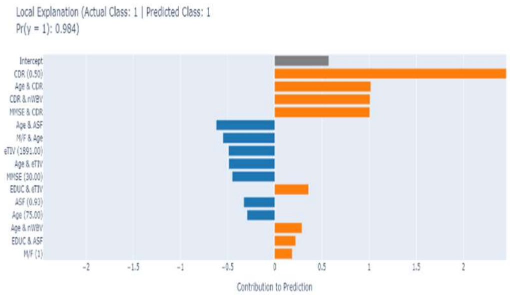

Figure 10. Local Explanation For The Observation Where Actual Class=1 And Predicted Class=1

For the first observation shown in the Figure 10, the actual and predicted class is 1 (i.e., demented) with a predicted probability of 0.984. Some terms such as Age, ASF and other interactions pull the prediction towards non-demented class. But the final prediction is indeed correct i.e., the patient is demented.

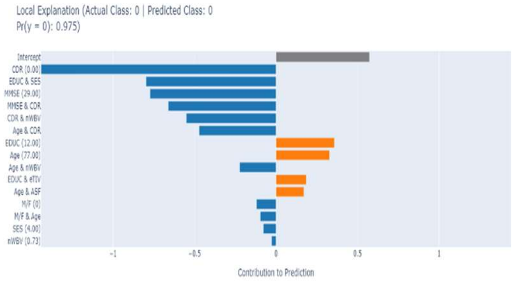

Figure 11. Local Explanation For The Observation Where Actual Class=0 And Predicted Class=0

For another observation in Figure 11, the actual and predicted class is 0 (i.e., non- demented) with a predicted probability of 0.975. We can notice that the features of the patient such as Education, Age and other interactions pull the prediction towards demented class. But the final prediction is correct again which is non-demented.

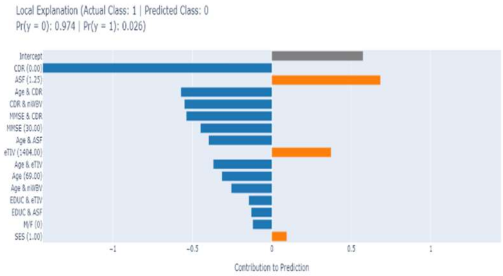

Figure 12. Local Explanation For The Observation Where Actual Class=1 And Predicted Class=0

Next observation is a case of misclassification where the predicted class is 0 (i.e., non- demented) but the actual class is 1. In Figure 12, we can see that the terms other than ASF, eTIV and SES pulled the prediction towards non-demented class which is responsible for misclassification. Thus, EBM helps to understand the reason behind misclassification and enables us to debug the data. In the same manner, we can produce global and local explanations for stroke and Parkinson's Disease prediction.

6 CONCLUSION

Predicting a disease in an accurate and interpretable manner is a challenging task. This research proposes a novel approach for AD detection using EBM. The model demonstrates promising results in accurately predicting AD risk while simultaneously providing interpretable insights into the factors driving those predictions. This focus on explainability fosters trust and facilitates integration with clinical workflows. Glass box models such as Logistic Regression, DT and EBM are employed as well as black-box models such as XGBoost, Extra Tree Classifier, Voting Classifier and Multilayer Perceptron are implemented for the experiment. By incorporating a good feature selection algorithm, Chi-square technique, we could enhance the performance accuracy of the model. The results suggest that EBM with Chi-square feature selection delivered superior prediction performance and ameliorates the transparency by giving more insights and intuitions to the predictions. Hence this model can be used to improve the diagnosis of AD. As the results are self-explainable to the clinicians, it helps them to take decisions confidently. Intelligibility from the graphs of EBM help to create new knowledge. It also aids in understanding and debugging the data. This paper addressed performance-interpretability trade-off in ML models. The contribution of this work lies in advancing the field of XAI for AD detection. By offering a transparent and interpretable model, this research has the potential to revolutionize early AD diagnosis, enabling timely intervention and improved patient outcomes.

As an extension of this work, EBM can also be employed for predicting other neurodegenerative diseases relative to traditional methods as it exhibits very good performance accuracy and this in turn improves the quality of treatments. In the future, the model's performance can be explored on even larger datasets and potentially integrate it with existing clinical decision support systems. We should explore more XAI techniques for AD prediction and evaluate model's generalizability in real world clinical settings. Subsequently, EBMs can be employed and tested in other domains where explainability is mandatory and performance is a key.

REFERENCES

[1] World Health Organization. (n.d.). *Mental health: Neurological disorders*. World Health Organization. https://www.who.int/news-room/questions-and-answers/item/mental-health-neurological-disorders (Accessed Jan 16, 2024)

[2] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

[3] Quattrone, A., Quattrone, A. (2021). Explainable Boosting Machine for Predicting Alzheimer's Disease from MRI Hippocampal Subfields. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds) Brain Informatics. BI 2021. Lecture Notes in Computer Science(), vol 12960. Springer, Cham. https://doi.org/10.1007/978-3-030-86993-9_31.

[4] Nori, H., Caruana, R., Bu, Z., Shen, J.H. &amp; Kulkarni, J.. (2021). Accuracy, Interpretability, and Differential Privacy via Explainable Boosting. *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 139:8227-8237 Available from ttps://proceedings.mlr.press/v139/nori21a.html.

[5] Chen, Zhi & Tan, Sarah & Nori, Harsha & Inkpen, Kori & Lou, Yin & Caruana, Rich. (2021). Using Explainable Boosting Machines (EBMs) to Detect Common Flaws in Data. 10.1007/978-3-030-93736-2_40.

[6] Wang Du, Lee Sang Ho, Geng Huaizhi, Zhong Haoyu & et al, *Interpretable machine learning for predicting pathologic complete response in patients treated with chemoradiation therapy for rectal adenocarcinoma,* Frontiers in Artificial Intelligence, Vol. 5, 2022, https://www.frontiersin.org/articles/10.3389/frai.2022.1059033, doi. 10.3389/frai.2022.1059033, ISSN=2624-8212

[7] O. Nanushi, V. Sitokonstantinou, I. Tsoumas and C. Kontoes, "Pest Presence Prediction Using Interpretable Machine Learning," *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, Nafplio,

Greece, 2022, pp. 1-5, doi: 10.1109/IVMSP54334.2022.9816284.

[8] Gaoyang Liu, Bochao Sun, Concrete compressive strength prediction using an explainable boosting machine model, Case Studies in Construction Materials, Volume 18, 2023, e01845, ISSN 2214-5095, https://doi.org/10.1016/j.cscm.2023.e01845.

[9] Değer, Zeynep & Taskin, Gulsen & Wallace, John. (2023). Estimate Deformation Capacity of Non-Ductile RC Shear Walls using Explainable Boosting Machine. 10.48550/arXiv.2301.04652.

[10] Sarica, A., Quattrone, A., Quattrone, A. (2022). Introducing the Rank-Biased Overlap as Similarity Measure for Feature Importance in Explainable Machine Learning: A Case Study on Parkinson's Disease. In: Mahmud, M., He, J., Vassanelli, S., van Zundert, A., Zhong, N. (eds) Brain Informatics. BI 2022. Lecture Notes in Computer Science(), vol 13406. Springer, Cham. https://doi.org/10.1007/978-3-031-15037-1_11

[11] Maxwell, A. E., Sharma, M., & Donaldson, K. A. (2021). Explainable boosting machines for slope failure spatial predictive modeling. *Remote Sensing*, *13*(24), 4991. https://doi.org/10.3390/rs13244991

[12] Bosschieter, Tomas & Xu, Zifei & Lan, Hui & Lengerich, Benjamin & Nori, Harsha & Sitcov, Kristin & Souter, Vivienne & Caruana, Rich. (2022). *Using Interpretable Machine Learning to Predict Maternal and Fetal Outcomes*. 10.48550/arXiv.2207.05322.

[13] Archana P. Menon, Dr. R. Gunasundari, Study of Interpretability in ML Algorithms for Disease Prognosis, *Revista Geintec-Gestao, Inovacao e Technologias*, Vol. 11 No. 4 Aug 2021, Pages 4735-4749, ISSN NO: 2237-0722. Available: https://www.revistageintec.net/index.php/revista/article/view/2500

[14] O. Conor Sullivan, *Interpretability in Machine Learning*, Towards Data science, Oct.2020. Accessed on: Jun. 12, 2021. [Online]. Available: https://towardsdatascience.com/interpretability-in-machine-learning-ab0cf2e66e1

[15] Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2019. https://christophm.github.io/interpretable-ml-book/.

[16] Nori, Harsha and Jenkins, Samuel and Koch, Paul and Caruana, Rich, "InterpretML: A Unified Framework for Machine Learning Interpretability." *ArXiv* abs/1909.09223 (2019)

[17] Julien I.E. Hoffman, Chapter 33 - Logistic Regression, Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition), Academic Press, 2019, Pages 581-589, ISBN 9780128170847, https://doi.org/10.1016/B978-0-12-817084-7.00033-4

[18] 1.10. Decision Trees. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/tree.html (Accessed Jan 16, 2024)

[19] GeeksforGeeks. (2023, February 6). XGBoost. GeeksforGeeks, https://www.geeksforgeeks.org/xgboost/ (Accessed Jan 16, 2024)

[20] Sklearn.tree.extratreeclassifier. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html (Accessed Jan 16, 2024)

[21] Sklearn.ensemble.VotingClassifier. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html (Accessed Jan 16, 2024)

[22] 1.17. neural network models (supervised). scikit. (n.d.). https://scikit-learn.org/stable/modules/neural_networks_supervised.html (Accessed Jan 16, 2024)

[23] YouTube. (2020, May 16). The science behind InterpretML: Explainable boosting machine. YouTube. https://www.youtube.com/watch?v=MREiHgHgl0k (Accessed April 16, 2023)

[24] Explainable boosting machine — interpretml documentation. https://interpret.ml/docs/ebm.html, (Accessed Jan 16, 2024)

[25] Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13: ACM Press.

[26] S. Picek, A. Heuser, A. Jovic and L. Batina, "A Systematic Evaluation of Profiling Through Focused Feature Selection," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 12, pp. 2802-2815, Dec. 2019, doi: 10.1109/TVLSI.2019.2937365.

[27] 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.

[28] S. Basheer, S. Bhatia and S. B. Sakri, "Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset," in IEEE Access, vol. 9, pp. 42449-42462, 2021, doi: 10.1109/ACCESS.2021.3066213.

[29] Bardwell, J., Hassan, G.M., Salami, F., Akhtar, N. (2022). Cognitive Impairment Prediction by Normal Cognitive Brain MRI Scans Using Deep Learning. In: Aziz, H., Corrêa, D., French, T. (eds) AI 2022: Advances in Artificial Intelligence. AI 2022. Lecture Notes in Computer Science(), vol 13728. Springer, Cham. https://doi.org/10.1007/978-3-031-22695-3_40

[30] Yi, Fuliang & Yang, Hui & Chen, Durong & Qin, Yao & Han, Hongjuan & Cui, Jing & Bai, Wenlin & Ma, Yifei & Zhang, Rong & Yu, Hongmei. (2023). XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. BMC Medical Informatics and Decision Making. 23. 10.1186/s12911-023-02238-9.

[31] Salehi, Waleed & Baglat, Preety & Gupta, Gaurav. (2020). Multiple Machine Learning Models for Detection of Alzheimer's Disease Using OASIS Dataset. IFIP Advances in Information and Communication Technology. 617. 614-622. 10.1007/978-3-030-64849-7_54.

[32] H. A. Shad et al., "Exploring Alzheimer's Disease Prediction with XAI in various Neural Network Models," TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON), Auckland, New Zealand, 2021, pp. 720-725, doi: 10.1109/TENCON54134.2021.9707468

[33] Saratxaga CL, Moya I, Picón A, Acosta M, Moreno-Fernandez-de-Leceta A, Garrote E, Bereciartua-Perez A. MRI Deep Learning-Based Solution for Alzheimer's Disease Prediction. Journal of Personalized Medicine. 2021; 11(9):902. https://doi.org/10.3390/jpm11090902

[34] Islam, J., Zhang, Y. (2017). A Novel Deep Learning Based Multi-class Classification Method for Alzheimer's Disease Detection Using Brain MRI Data. In: Zeng, Y., et al. Brain Informatics. BI 2017. Lecture Notes in Computer Science(), vol 10654. Springer, Cham. https://doi.org/10.1007/978-3-319-70772-3_20

[35] Oleszak, M. (2022, January 27). Explainable boosting machines. Medium. https://medium.com/towards-artificial-intelligence/explainable-boosting-machines-c71b207231b5 (Accessed April 16, 2023)