Analysis of Alzheimer's Disease Detection using Supervised ML Algorithms

1st MD. Asaduzzaman Sarker Anik
Department of Computer Science and Engineering
BRAC University
md.asaduzzaman.sarker.anik@g.bracu.ac.bd

2nd SK. Mamunur Rashid
Department of Computer Science and Engineering
BRAC University
sk.mamunur.rashid@g.bracu.ac.bd

3rd MD. Tanvir Zahid
Department of Computer Science and Engineering
BRAC University
md.tanvir.zahid@g.bracu.ac.bd

4th Mehnaz Ara Fazal
Department of Computer Science and Engineering
BRAC University
mehnaz.ara.fazal@g.bracu.ac.bd

5th Sania Azhmee Bhuiyan
Department of Computer Science and Engineering
BRAC University
sania.azhmee.bhuiyan@g.bracu.ac.bd

6th Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
annajiat@gmail.com

***Abstract*—Alzheimer's disease, a common disease in the older generations which turns out cognition impairment and loss of memory. We need to make sure that we diagnose this disease with proper medical criteria that are responsible for this disease detection. For this project, we are using ten of the famous influential machine learning algorithms for the diagnosis of this disease. We made sure about the risks that comes with it while detecting the disease. Random Forest algorithm performs excellent in all criteria of AD detection. Furthermore, we investigated a bit more with Explainable AI (XAI) to identify which factors are more responsible for this disease to occur or not. This research is good for early AD detection which can timely intervene for the patient's outcomes that plays an important role on the patients.**

***Index Terms*—Explainable AI, Alzheimer's Disease**

I. Introduction

AD is a severe disorder which affects gradually in the brain and cannot be undone. This is a deadly disease which affects our memory and social interaction ability which means it affects the cognitive function of our brain. This hampers our daily life in a negative way. The symptoms of this disease is slight loss of memory and gradually destroys the cognitive function of our brain. The downside of this disease is that there is no cure yet to be found. However, some medications and habit changes can improve the quality of life under medical supervision. World Health Organization reported that there about around 55 million patients who are diagnosed with AD [1]. Our goal is to minimize its effects at an early stage of the disease. This prevention process at the early stage can prevent to risks lives and also has economic importance [2]. Moreover, this early prevention can lessen the problems of people associated with the patients which has some financial importance. People over the age of 65 are at the maximum risks of this disease. Some factors which triggers to cause this disaease are obesity, high blood pressure, smoking [3].

II. Related Works

Shankle et al., [9] implemented six machine learning algorithms to the dataset of five hundred and seventy eight patients. Using Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition the patients were categorized into three groups. Furthermore, the paper relied on some user characteristics like age, education and feedback. However, taking feedback from subjects who are suffering from dementia may not always produce accurate predictions.

Alvarez et al.(2009) [11] utilized a self acting tool that is supported by Principal Component Analysis in order to decrease the vector size of the features. They also used Support Vector Machine to enhance the Alzhaimer's disease prediction capability. This paper uses Single-photon emission computed tomography images for this process. This provides a significant amount of features, five lacs to be exact. On the contrary, only one hundred features are used for the processing. For this reason, lower data problem arises and it is complicated to use this less numbers data on Support Vector Machine.

In the next paper Escudero et al.(2012) [10] proposes a cost effective personalized machine learning approach. In order to create a classifier which is weighted for every individual patient, the paper quantifies the bio signature arrangements.

Hyunseokc(2018) [8] utilized data associated with Magnetic resonance imaging(MRI) which was developed by the Organization for the Advancement of Structured Information Standards(OASIS). The primary emphasis of this paper was to observe the fluctuation between cognitive impairment(dementia) and MRI results. Furthermore, visualization was also a part of

their process, by which they examined the data and manifested the dependency of result and features. The resulting data was evaluated by employing Support Vector Machine, Decision Trees Classifier, AdaBoose and Logistic Regression. However, the use of only MRI data is the pitfall of this paper since there are many other data points that should be factored in while diagnosing AD.

Previous research (Al-Shoukry et al., 2020) [5] has given a brief rundown of various methods in order to identify AD. These methods include various deep learning, consisting of RNN(Recurrent Neural Network), CNN(Convolutional Neural Network) and a couple of alternatives. In order to discover characteristics and patterns they used MRI output with the help of neural network algorithms. These algorithms according to the research have superior precision and quickness of detection. They also surpass the accuracy derived from conventional ML algorithms for Alzheimer's disease detection. Unfortunately, these algorithms necessitates selection bias and huge data.

Another paper by Eke et al., 2020 [4] tries early detection of AD by deploying Support Vector machine. It exerts biomarkers of non amyloid kind. This method produces excellent results, achieving accuracy of 96%. This model is very effective in positively detecting AD patients while significantly mitigating false positives. Likewise, their finding concludes that five biomolecules named Apolipoprotein E, Beta-2 microglobulin, Complement component 3, Intercellular Adhesion Molecule 1 and Vascular cell adhesion protein 1 are heavily associated with AD.

III. Dataset

We collected our data from Open Access Series of Imaging Studies (OASIS) data set 1 for this project [6]. This data set contains some collections of MRI results from the people with variations in age and brain conditions. This data set has two results for cross-sectional and longitudinal MRI scans particularly for elderly people. The cross-sectional data has about 418 subjects including mild to normal AD. This data also varies across different age groups. For each individual, they scanned about 4 T1-weighted MRIs in a single pass. Furthermore, longitudinal data has about 152 subjects containing elderly peoples' scans. This data is collected across a year for the people of age from 60 to 96. Moreover, this data set contains information about the non demented, those who were non demented before but throughout the time they turned to demented patients. This data set contains 3 classes: demented, non-demented and converted. For simplicity, we combined these data sets and assumed the converted class as a dementia patient.

We flushed some unwanted features from our data set which are irrelevant to our project. This includes- Subject ID, MRI ID, Visit and MR Delay. Table 1 contains the feature names used in this project for a better understanding.

In the first figure, we analyzed the distribution of each features or classes within our data set such that there prevails a normalized numbers of non-demented, converted and

TABLE I
Description of the Dataset

Feature	Description
Gender	Gender
Educ	Total Education years
SES	Status in terms of social position
MMSE	A Score given for Mini-Mental State Examination
CDR	A rating for Clinic based Dementia
eTIV	Approximate intracranial volume
nWVB	Volume of the whole brain in a normalized form
ASF	Scaling factor for Atlas bone

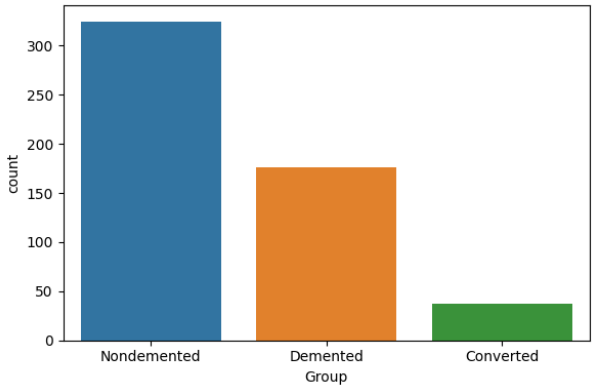

Fig. 1. Group Distribution in the data set

demented people. This figure shows that there is an imbalance distribution in the classes. Moreover, the second figure shows that women have less dementia rate than men in terms of gender-based classification.

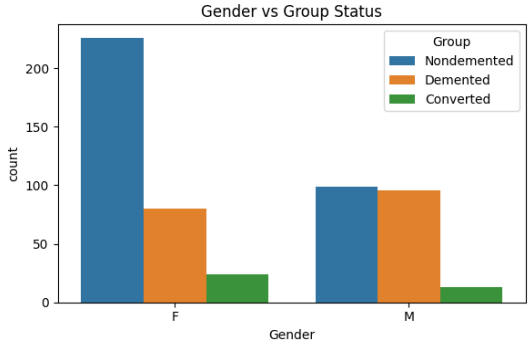

Fig. 2. Genderwise Group Distribution in the dataset

Furthermore, we generated a heatmap to get a glimpse of correlation between the features and target classes. We plotted the map using Pearson Coefficient for Correlation. The third figure shows this correlation. We omitted the feature 'subject's hand' since this is not correlated to the target classes.

IV. Proposed Methodology

Our target is to predict correctly is an individual is affected with AD or not. For that, we used machine learning based classifiers. We tried to predict each person from any one of the

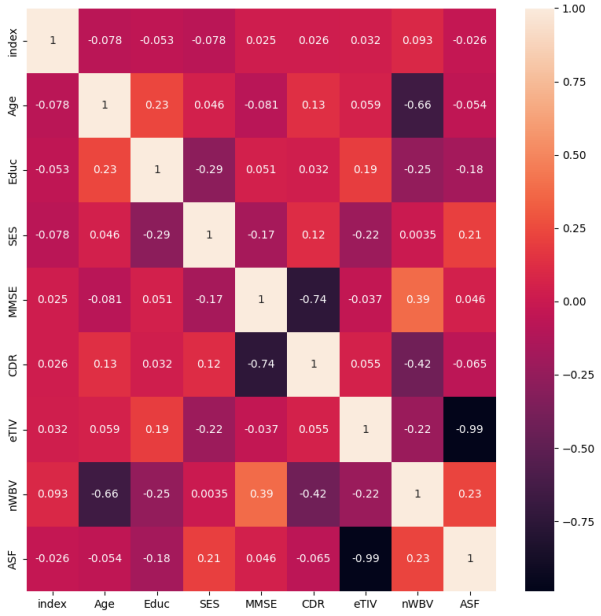

Fig. 3. Correlation heatmap

following two classes: demented and non-demented. contains the flow diagram that we proposed for this project.

A. Data Collection

We used the popular OASIS website to collect our data [6]. This is a common and reliable source for testing AD projects.This data set contains some collections of MRI results from the people with variations in age and brain conditions. This data set has two results for cross-sectional and longitudinal MRI scans particularly for elderly people. The cross-sectional data has about 418 subjects including mild to normal AD. This data also varies across different age groups. For each individual, they scanned about 4 T1-weighted MRIs in a single pass. Furthermore, longitudinal data has about 152 subjects containing elderly peoples' scans. This data is collected across a year for the people of age from 60 to 96. Moreover, this data set contains information about the non demented, those who were non demented before but throughout the time they turned to demented patients.

B. Data Preprocssing

Firstly, we re-termed the converted feature as demented as they turned to demented from non-demented person. Secondly, we omitted some unrelated features for our project which are ID, Subject ID, Delay, MRI ID, Visit, and MR Delay. Thirdly, we combined our data sets for simplicity and less run-time. Fourthly, we redefined the Gender and Group feature into numerical forms for faster calculations. Finally, from the heatmap we found out that Hand feature is negligibly correlated to our target classes hence, we dropped that feature too. These steps helps us to prepare our data set for extensive implementations for the upcoming parts of our project.

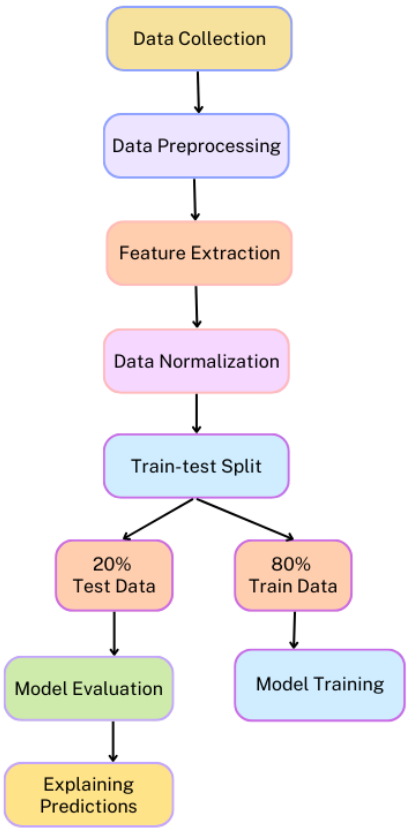

Fig. 4. Flow Diagram of Our Proposed Method

C. Feature Extraction

We implied a general decision tree to find out the top seven most important features among the columns. Here, we used the Gini Index or Mean Decreasing Impurity. The selected features are listed as follows: CDR, Age, Education, MMSE, ATS, Socioeconomic Status and nWBV.

D. Data Normalization

We normalized our data using the Standard Normalization technique to transform the data from values for 0 to 1. This helps the machine to evaluate and learn quickly instead of scanning words. However, it also ensures that no single features dominate the learning period and hampers the train-test process.

E. Train-test split

We divided our data set in 80:20 ration for the train:test split respectively. To elaborate, 80% (430 instances) of this data is assigned to training period whereas the rest 20% (108 instances) is for the testing period.

F. Model Training

In this step we trained our model by applying different machine learning algorithm on training data. The algorithms that we applied include , logistic regression, support vector machines, , gradient boosting, random forests, Adaboost,

voting classifier and k nearest neighbor. We tried to find the patterns between input features and the output classes through these algorithms. We also optimized the model by tuning hyper-parameter. Finally we selected those features which has the most predictability score.

1) Logistic Regression: A common statistical machine learning model. This model is used for classification and predictive analysis. Logistic regression is applied when the dependent variable is categorical. As it estimates the probability of an event, the boundary of the result ranges from 0 to 1. Here 1 is the highest probability. The formula of this classifier is given below:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

2) Random Forest: Random Forest is generally used for both classification and regression tasks. It uses multiple decision trees and it can add randomness to the model by growing the trees. The random forest takes the average of the predictions while improving the accuracy.

3) Support Vector Machine (SVM): This algorithm can also be used for both classification and regression analysis. SVM is based on the concept of hyperplane. The hyperplane is derived from the support vectors. Cosine Similarity, Manhattan distance etc. are used to derive the hyperplane with a maximum margin if possible which means the maximum distance between data points of both classes.

4) AdaBoost: Here weights are set to the classifiers and learning is progressed in each iteration. Weights are used to classify by this classifier. Several weak classifiers are merged together to build one or more strong classifiers.

5) XGBoost: Extreme Gradient Boosting is a distributed gradient boosting library. It can train and test models on large amounts of scalable training data. which is used for scalable training data. It combines multiple weak learners to build a strong one. It can handle the data with missing values and can improve its performance on sparse data.

6) KNN: K-nearest neighbor is a machine learning algorithm used in both classification and regression. The model is not explicitly trained during the learning phase. This is a method by which we can define testing scenarios by electing. In this method, we define the test conditions interval from the dataset values which are in the closest proximity and elect based on those proximity.

7) Gaussian Naive Bayes: This is a method for categorizing or classification of systems which is Gausian spread and centers around possibility-based techniques. This type of Naive Bayes is particularly unique in a sense that the outputs of the forecasting are uninterrupted and they converse on a Gausian spread or distribution.

8) Gradient Boosting: The next algorithm is one with boosting capability that can anticipate or forecast the optimal model based on the previous models. It works in such a way that the outcome of the next prediction becomes less erroneous. While forecasting if a certain projection elevates error rate, it gradually trimmed those false predictions as the updated models approached the intended goal result.

9) Voting Classifier (hard): The currently discussed model is an ordinary classifier which is an ensemble. The estimated result is derived from the counted available votes with superior quantity which disregards the efficiency.

10) Voting Classifier (soft): This is another voting classifier. However it has soft constraint which distincts it from the hard classifier. In this case estimation is done by combining all the output from each models and finally choosing the final output that has the highest forecasting or probability.

G. Model Evaluation

When the model passes the training stage we move on to the next, evaluating. In this state we look through the prediction capability of our models and assess them. Accuracy and recall were used as evaluating matrices for the models. The first one, accuracy calculated the accuracy as a whole for the questioned models. It's a great indicator for the general execution capability of the models. The next one is called recall. It excels in indicating a model's capability to verify or pinpoint an actual manifestation, for example: whether a patient has Alzhemer's disease. When a model has an elevated recall, we can conclude that the model is adequate in recognising patients. The mathematical expression for computing these matrix values are provided below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

TP: true positives, TN: true negatives, FP: false positives and FN: false negatives.

H. Explaining Predictions

For this project, we built a disease detection model. As a result, we also need to explain the predictions made by the model. Here, we used LIME to get a insight about which factors are most important in decision making persistently. We highlighted the dominant features that influenced the model using the LIME and provided explanations for each of the instances.

V. Result Analysis

Summary of our result is displayed in Table 2. We used the Recall metric for our performance measure since this is a disease prediction problem. However, we also have the accuracy metric if in case anyone needs to identify the overall performance of the model. In our research, Random Forest Classifier has achieved the highest percentage of recall (100%) for both demented and non-demented individuals. It also achieved maximum accuracy which fits this algorithm to predict overall correctness of this model. However, we

TABLE II
PREDICTION SCORES FROM THE ALGORITHMS

Model	Recall (ND)	Recall (D)	Accuracy
Random Forest	**100%**	**100%**	**100%**
Logistic Regression	98%	90%	93%
KNN	96%	93%	90%
Gaussian NB	90%	93%	96%
SVM	100%	89%	91%
XGBoost	100%	96%	96%
AdaBoost	97%	99%	90%
Gradient Boosting	99%	95%	91%
Voting Classifier(Hard)	89%	99%	94%
Voting Classifier(Soft)	94%	92%	93%

noted that all models predicted better in non-demented patients compared to demented patients.

For further analysis of our project, we used the LIME [7] to get a better insight. LIME is a technique for understanding the local explanations altogether. Here, we tried to understand the decisions made by each model based on what features they preferred for prediction. For example, the Random Forest, being the highest score, we took some of its predictions to find out what it prioritized more. The figure 5 and 6 shows that the Random Forest preferred the CDR and MMSE features compared to other features. This also applies in real life because these two features are crucial for AD detection. From the figures, we get an idea that a person with higher CDR has the probability to be affected by this disease (dementia) whereas the lower CDR indicates lesser chance (non-dementia).

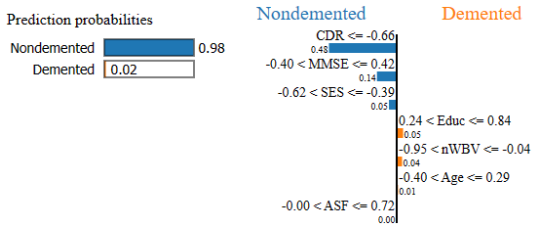

Fig. 5. Nondemented patient correctly detected

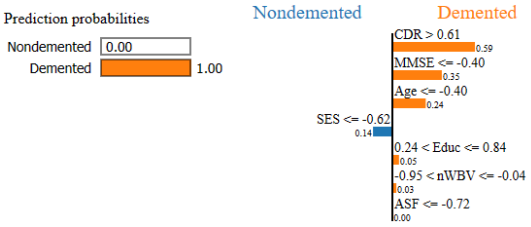

Fig. 6. Demented patient correctly detected

VI. CONCLUSION

In order to offer a cure, we must specify the core of an issue. When considering a disease, this is especially true since, first we have to focus on preclusion. This is where our proposed model comes in handy. Instead of going through the painful process of diagnosis and treatment we need to focus on prevention and early detection. Not only that, we need to create an explainable model for the domain experts to understand. Our goal was to use ten machine learning algorithms to classify Alzheimer's disease. Then with the help of lime we made those models explainable for the domain experts to understand. In the future we would like to make those models even more explainable. Furthermore, expanding the features and applying CNN on those features to achieve even greater results is something we are willing to explore in the future.

REFERENCES

[1] "Dementia." World Health Organization, www.who.int/news-room/fact-sheets/detail/dementia. Accessed 16 May 2023.

[2] Castro, Diego M., et al. "The economic cost of Alzheimer's disease: Family or public-health burden?." Dementia & Neuropsychologia 4 (2010): 262-267.

[3] Website, Nhs. "Causes." nhs.uk, 24 Apr. 2023, www.nhs.uk/conditions/alzheimers-disease/causes.

[4] Eke, Chima S., et al. "Early Detection of Alzheimer's Disease with Blood Plasma Proteins Using Support Vector Machines." IEEE journal of biomedical and health informatics 25.1 (2020): 218-226.

[5] Al-Shoukry, Suhad, Taha H. Rassem, and Nasrin M. Makbol. "Alzheimer's diseases detection by using deep learning algorithms: a mini-review." IEEE Access 8 (2020): 77131-77141.

[6] Marcus, Daniel S., et al. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults." Journal of cognitive neuroscience 19.9 (2007): 1498-1507.

[7] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

[8] Hyunseokc. (2018). DETECTING EARLY ALZHEIMER'S. Kaggle. https://www.kaggle.com/code/hyunseokc/detecting-early-alzheimer-s.

[9] Shankle, W. R., Mani, S., Pazzani, M. J., & Smyth, P. (1997). Dementia screening with machine learning methods. In Intelligent Data Analysis in Medicine and Pharmacology (pp. 149-165). Boston, MA: Springer US.

[10] Escudero, J., Ifeachor, E., Zajicek, J. P., Green, C., Shearer, J., & Pearson, S. (2012). Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease. IEEE transactions on biomedical engineering, 60(1), 164-168.

[11] Alvarez, I., Górriz, J. M., Ramírez, J., Salas-Gonzalez, D., López, M., Segovia, F., ... & Prieto, B. (2009). Alzheimer's diagnosis using eigenbrains and support vector machines. In Bio-Inspired Systems: Computational and Ambient Intelligence: 10th International Work-Conference on Artificial Neural Networks, IWANN 2009, Salamanca, Spain, June 10-12, 2009. Proceedings, Part I 10 (pp. 973-980). Springer Berlin Heidelberg.