Comparative Analysis of Machine Learning Approaches for Early Detection of Alzheimer's Disease

Akileshwaran S[1], Sathish kumar R[2], Malathi A[3], Maheswari M[4], Roselin Mary S[5]

1. Student, Computer science and engineering, Anand Institute of Higher Technology, Chennai, India.
2. Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.
3. Assistant Professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.
4. Assistant Professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.
5. Head of Department, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.

ABSTRACT

The purpose of this survey is to compare the accuracy of Machine Learning algorithms to predict Alzheimer's disease early. Machine learning algorithms is comparatively applied on variety of biomarkers correlated with disease in order to examine the efficiency of those Machine Learning techniques. Based on this analysis the foremost algorithm can be employed to perceive the Alzheimer's disease in advance. In this paper we are going to find better ways to predict the Alzheimer's disease when other chronic conditions are present.

Keywords: *Random Forest (RF), Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), K-Nearest Neighbor (KNN), XG-Boost, Machine Learning (ML), Features, Classifiers*

1.INTRODUCTION

A disease's likelihood of being cured greatly increases if it is diagnosed as soon as possible. In recent years, the use of Machine Learning (ML) is surging through all fields of science, and the field of neurology is definitely undergoing a revolution thanks to it. Medical science has benefitted from the application of Machine Learning to improve the prediction and detection of Alzheimer's disease.

Through this effort, we aim at finding the most accurate technique for detecting different brain diseases which can be employed for future betterment. The symptom of that woman was memory loss, language problem and unpredictable behavior. Then her brain was tested and noticed that anomalous clumps and collection of fibers are created in the brain, then the research is started on that and given the name of the doctor to the disease as Alzheimer. A neurofibrillary fiber is a fibrous clump of abnormal protein, known as an amyloid plaque. The disease states from the hippocampus and spread over the brain, as the result of this the death of neuron occurs and the tissues in the brain shrunk so that fully memory loss will cause.

To find better ways to manage dementia when other chronic conditions are present. By developing prediction model for early detection of Alzheimer's disease, we can help doctors to have a better chance of helping asymptomatic patients by preventing further complications. This means that the diagnosis criteria and treatment plan for Alzheimer's disease needs to be revised. Determining whether inflammatory reactions are persistent is critical for diagnosing and treating Alzheimer's disease.

2. RELATED WORKS

The sorts of data used and the efficiency of machine learning approaches in predicting early stages of Alzheimer's disease have been highlighted as recent trends in machine learning [1]. The "MRI and Alzheimer's" dataset, which was provided by the Open Access Series of Imaging Studies (OASIS) project, was used to predict Alzheimer's disease or dementia in adult patients. SVM is the best model among the other models in the system. It has better accuracy, recall, area under the curve, and F1 score [2]. As rapid progress in neuroimaging techniques has created large-scale multimodal neuroimaging data, the application of deep learning to early diagnosis and automated categorization of Alzheimer's disease (AD) has recently gotten a lot of interest.[3]. A deep convolutional neural network-based pipeline for Alzheimer's disease diagnosis utilizing magnetic resonance scans, as well as a four-way classifier to predict AD [4]. A method for promoting end-to-end learning of a volumetric convolutional neural network (CNN) model for four binary classifications [5]. Non-amyloid biomarker panels based on blood for early detection of Alzheimer's disease Apart from that, this notion focuses on identifying the performance of novels like A2M, ApoE, BNP, Eot3, RAGE, and SGOT in order to determine mild cognitive impairment (MCI)[6]. Multiple Deep learning and Machine learning techniques were reviewed. A novel multiclass classification strategy that utilizes one-versus-one error correction output codes classification and pairwise t-test feature selection to handle the outlier identification problem [8]. The pathological hallmarks of AD brains are early stage -amyloid oligomers (AOs) and late-stage A plaques. The intention of this initiative is to detect A abnormalities in the early and late phases of Alzheimer's disease [9] Classical applications such as graph partitioning, graph visualization, and graph coarsening have recently been utilized in Graph Convolutional Neural Network (GCNN) architecture to perform graph pooling. This modified GCNN architecture is then used as a graph signal classifier to detect early-stage Alzheimer's disease [10]. Review of contemporary machine learning and deep learning approaches for detecting four brain diseases, including Alzheimer's disease (AD), brain tumors, epilepsy, and Parkinson's disease, in order to determine the most accurate technique for detecting different brain diseases that can be used in the upcoming years [11]. A metabolite-corrected artery input function (AIF) is required for quantitative analysis of PET brain imaging data in order to estimate distribution volume and related outcome measures. PET studies that collect arterial blood samples add risk, cost, measurement inaccuracy, and patient discomfort.[12]. Machine learning algorithms use psychological MMSE parameters including age, number of visits, and education to predict Alzheimer's disease. Support vector machine and decision tree techniques used [13]. As the combined high-order network (CHON) constructs FCN by combining static, dynamic, and high-level information, whereas the GCN is utilized to integrate non-image information to improve the classifier's performance [14]. ResNet18 and DenseNet201 were utilized to perform the AD multiclass classification challenge. [15]. A survey, analysis, and critical critique of recent work on the early diagnosis of Alzheimer's disease using machine learning techniques [16]. Thus, by referencing the work done in related articles, we were finally able to conceive a survey with the help of our dataset [17] for developing our front-end model for the early detection of Alzheimer’s disease.

3.EXISTING SYSTEM

Previously, users had to manually enter MRI images into the system, after which the value was calculated using image restoration, linear filtering, pixelation, grey scaling, and template matching. Later values from the image were extracted, and those values were trained and evaluated against a dataset to precisely establish the ranges for each decisive value. Thus, utilizing generated values from feed photographs as input values, integer and float data types, we were able to detect the presence of the disease. Finally, the output indicates whether or not the patient has been diagnosed with dementia.

4.PROPOSED SYSTEM

By presenting our survey, we aim to establish a front end. In our dataset, we want to train the supervised ML classifiers. We created a survey to collect data on the accuracies of different supervised ml classifiers. We were able to select the classifier with the highest accuracy by eliminating a few features in various combinations that were present in our dataset. We also acquire values from demographics, neuro-physicians' clinical tests, and MRI scan reports, and submit them to the backend for processing. Finally, we can forecast whether or not a person has acquired the risk of developing dementia.

5. IMPLEMENTATION

The system is divided into four sections. Dataset Analysis is the first module, and it is the process of evaluating, cleansing, transforming, and modelling data with the goal of identifying relevant information through informing conclusions and helping decision-making. The second module is Dataset Preprocessing to cleaning the data, which increases the accuracy and efficiency of a machine learning model using Synthetic Minority Oversampling Technique (SMOTE). The third is Model testing and Training, in this module we use supervised classification algorithms for training and testing our dataset using machine learning algorithms. The fourth module is Model Deployment, in this Module we developed User interface.

5.1 MACHINE LEARNING

Machine Learning is a process of training a computer to apply its past experience to solve a problem given to it. The machine can process, analyze and make abstracts based on a large amount of data. Its unique and intelligent behavior allows it to discover correlations and insights that are not readily apparent to the human eye, making abstractions from experience.

5.2 CLASSIFICATION

Data samples to be assessed that is transforming raw datasets into machine readable data this method is known as preprocessing or data cleaning. Besides the removal of a characteristics of data to be performed called feature extraction. After the extraction of features, the data can be labeled. The method by which the machine takes decisions of labeling data is called a classifier or classification. Certain Machine Learning techniques are used in this paper such as RF, GNB, SVM, XGB, KNN, LR.
It uses several layers of nonlinear processing units. The output of a unit is imparted as input to the next unit. Throughout the ordered structure of data movement, each level transforms the data it receives into more condense data to be imparted to the next level. ML classifiers to detect brain diseases can be classified as shown in Figure 1.

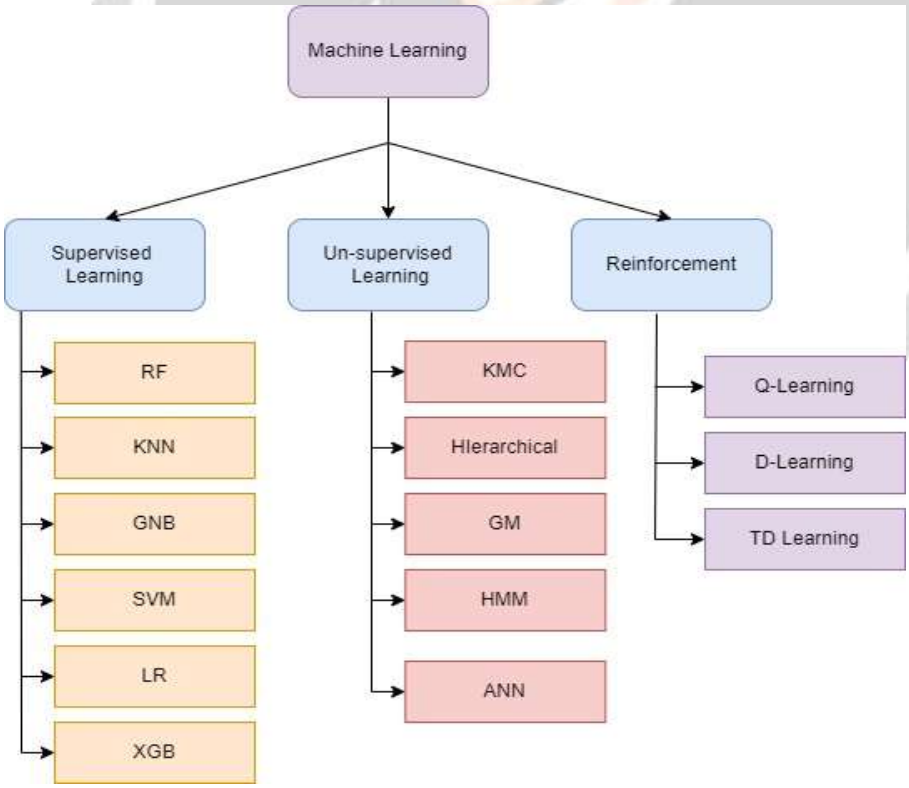

Figure A: Classification of ML

5.3 DATASET

Dataset being collected from the OASIS longitudinal in Kaggle repository. Kaggle(www.kaggle.com) is a repository containing over 50,000 publicly available datasets. Labels indicating the presence and absence of tumors are marked by “yes” and “no”, respectively. The dataset is an open-source .csv data set that can be used by anyone. Initially, it consisted of 374 patients' data in rows and features in 15 columns, all of them being right-handed and aged 18 to 96 years. Both male and female patients were present. One hundred of them aged above 60 were diagnosed with very mild to moderate AD. MRIs should be done with three to four T1-weighted scans, with high contrast to noise ratio. Here, the total volume of the brain and the estimation of the intracranial volume used for analyzing normal aging and Alzheimer's disease.

5.4 ML TECHNIQUES

Machine learning algorithms employ computer methods to "learn" information directly from data rather than depending on a model based on a preconceived equation. As the number of samples available for learning grows, the algorithms adapt their performance.

5.4.1 Gaussian Naive Bayes

It calculates affiliation probabilities for each class, such as the likelihood that a certain record or data point belongs to that class. The most likely class is defined as the one having the highest probability. When working with continuous data, one common assumption is that the continuous values associated with each class follow a normal (or Gaussian) distribution.

5.4.2 Support Vector Machine

This analyzes data for classification and regression analysis. It creates a hyperplane that separate to classes, it can create a hyperplane or set of hyperplanes in high dimension space. We want to optimize the margin between the data points and the hyperplane.

5.4.3 K-nearest neighborhood (KNN)

A distance metric is at the heart of this classifier. The more accurately that metric captures label similarity, the better. It’s not an invariable technique used to find matching ratings and average ratings of top of KNN. It’s a dominant technique to understand and to execute. A peculiarity of the KNN algorithm is that it’s sensitivity to local structure of the data.

5.4.4 XG-Boost

Gradient boosting is an AI method utilized in classification and regression assignments, among others. A loss function should be improved, which implies bringing down the loss function better than the result. Decision trees are utilized in this limit the loss function. After training, if we want to predict for a new data point then we will use constructed trees or models to get all values to solve the equation.

5.4.5 Logistic Regression

It is an algorithm for predictive analysis that relies on the concept of probability. This method is used to for binary classification problems. This algorithm is based on predictive analysis which is used to describe data. It also describes the relationship between one or more nominal or ratio-level independent variables and one or more department binary variables.

5.4.6 Random Forest

Random Forest is an adjustable, effortless to Machine Learning algorithm. It’s one of the incredible and most effective Machine Learning algorithms uses both classification and regression. Ensemble learning methods such as classification and regression produce mode of prediction mean by creating a multitude during training.

6. RESULT AND DISCUSSION

The results obtained by dropping the couples of features from the OASIS dataset are as discussed here

Algorithm/Dropped Features	eTIV & ASF	eTIV & nWBV	ASF & CDR	MMSE & SES	CDR & SES	nWBV & Age	Group & ASF	Group & Age
Random Forest	99.1	98.9	98.2	97.2	96.05	95.8	90.09	89.19
Support Vector Machine	50.45	68.47	50.45	50.45	50.45	50.45	50.45	50.45
Gaussian Naïve Bayes	97.3	97.4	97.6	96.8	95.5	95.4	90.09	89.19
K-Nearest Neighbor	44.14	83.79	55.86	53.19	48.65	48.75	48.65	60.36
Logistic Regression	91.89	98.1	89.19	92.79	69.37	89.19	70.27	89.19
XG-Boost	98.1	96.1	97.9	95.8	95.89	95.1	90.09	89.19

Table-1: Accuracy by dropping couple of features in different combinations.

Algorithm/Dropped Features	EDUC & MRI ID	Sub ID & Visit	MR Delay & Visit	EDUC & SES	MMSE & MR Delay	MRI ID & Sub ID	M/F&Visit
Random Forest	88.78	88.13	87.57	87.12	86.87	85.97	85.25
Support Vector Machine	50.45	50.45	50.45	50.45	50.45	50.45	50.45
Gaussian Naïve Bayes	87.83	88.1	86.87	83.76	86.07	85.04	85.14
K-Nearest Neighbor	50.45	47.75	55.86	54.05	53.15	50.45	48.65
Logistic Regression	76.58	72.07	86.4	82.7	82.99	79.28	71.17
XG-Boost	85.71	88.1	85.6	86.62	85.45	84.76	84.95

Table-2: Accuracy results generated by dropping couple of features in different combinations

6.1 GRAPHS FOR RESULTS

The graphs were drawn for combination of dropped features based on the tabulation above.

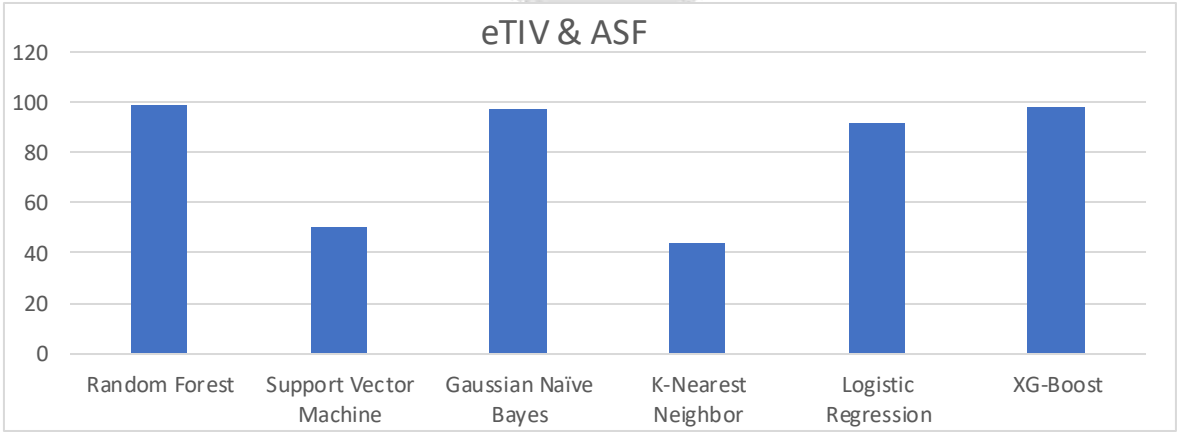

Figure 1: Accuracies generated by dropping eTIV & ASF

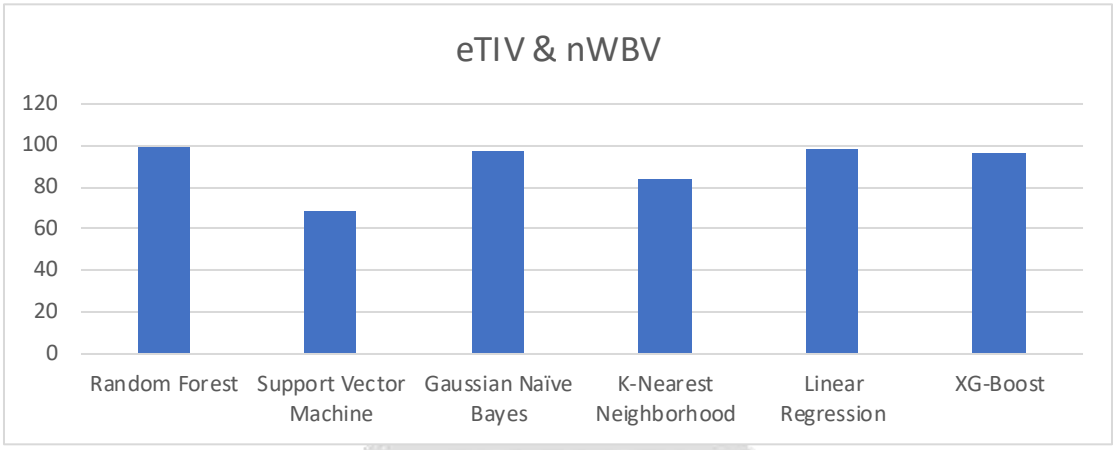

Figure 2: Accuracies generated by dropping eTIV & nWBV

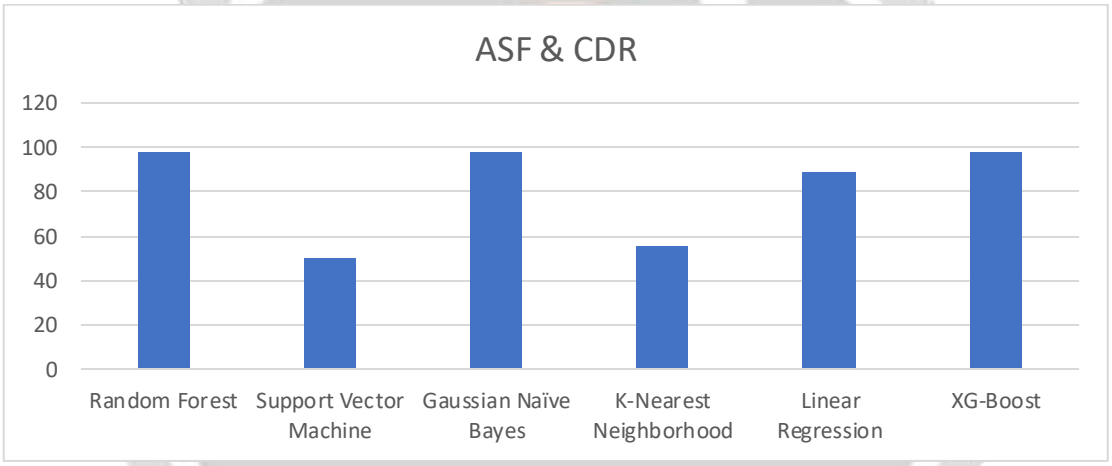

Figure 3: Accuracies generated by dropping ASF & CDR

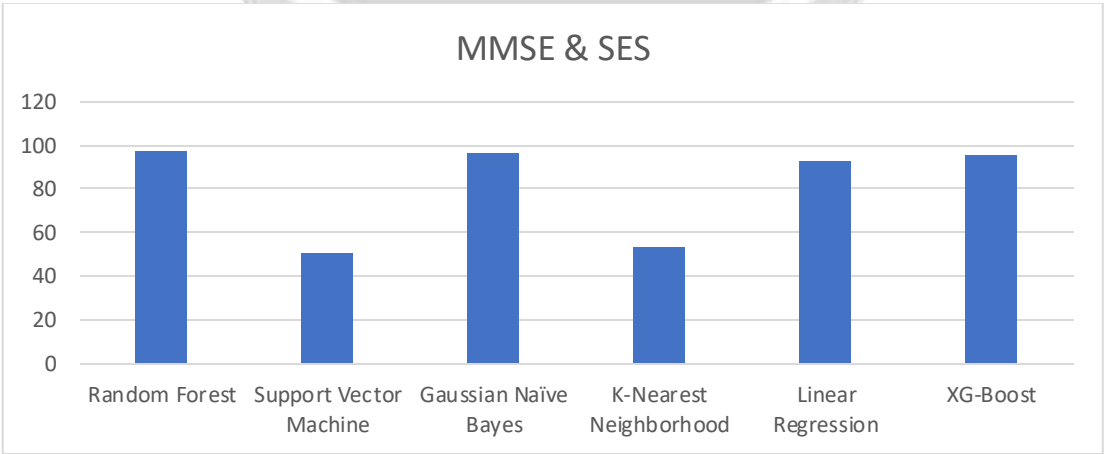

Figure 4: Accuracies generated by dropping MMSE & SES

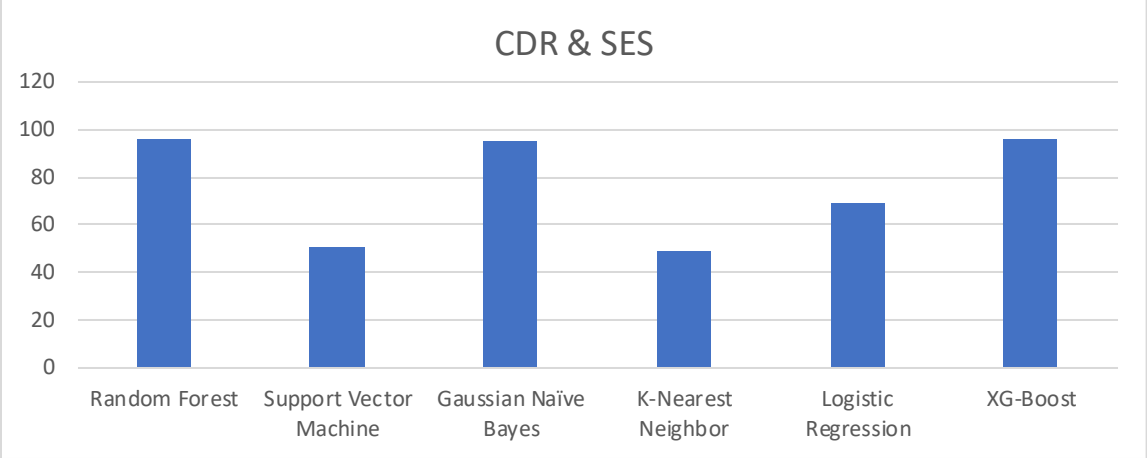

Figure 5: Accuracies generated by dropping CDR & SES

6.2 RESULTS

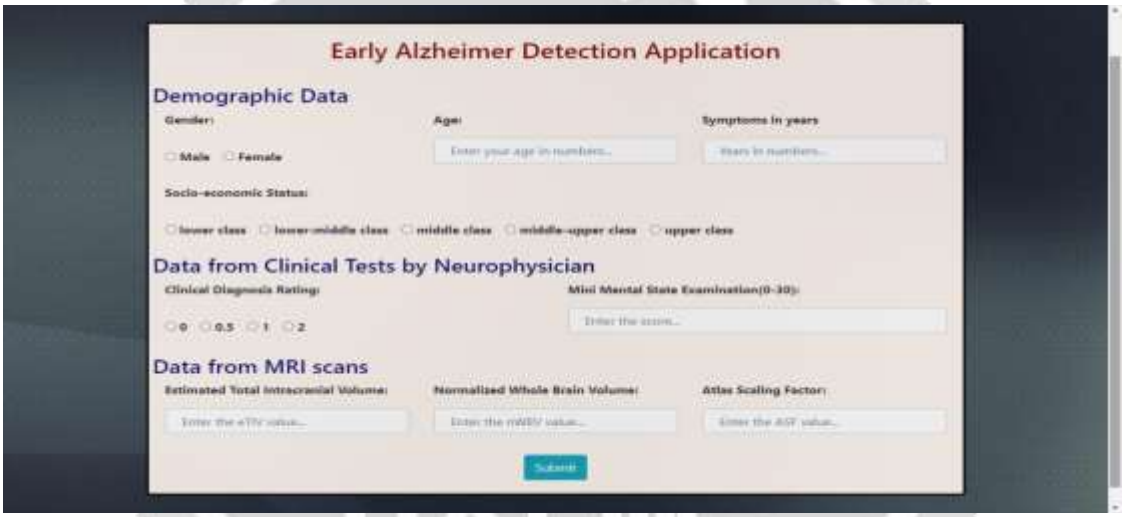

Figure 6: User Interface for data entry.

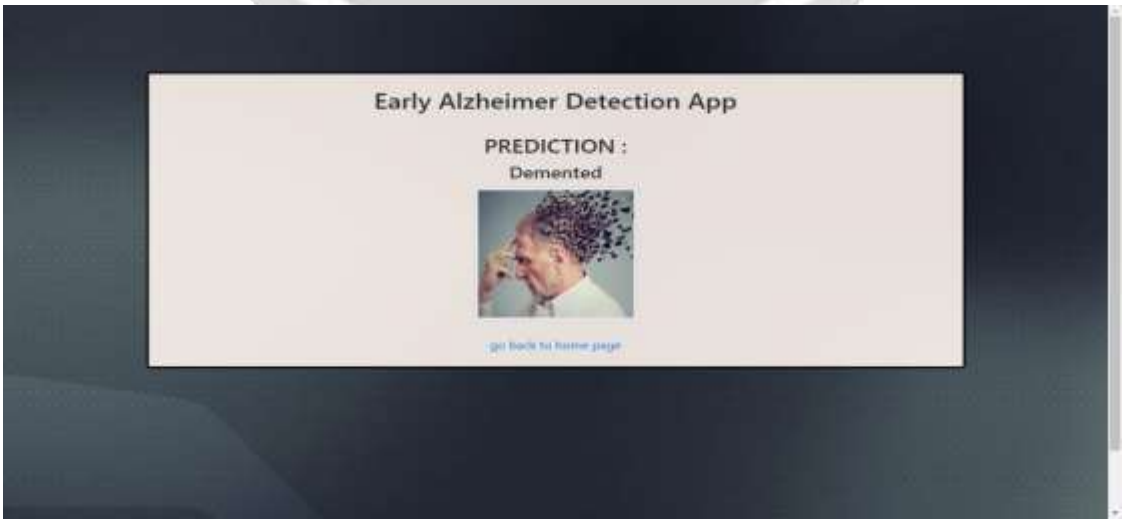

Figure 7: User Interface for detecting Demented Individual

Figure 8: User Interface for detecting Converted Individual

Figure 9: User Interface for detecting Non-Demented Individual

Using the tabulations mentioned above, we were able to plot the graphs based on those. In figure 1, we plotted the graph by dropping a couple of features, namely eTIV & ASF. In figure 2, we plotted the graph by dropping a couple of features, namely eTIV & nWBV. In figure 3, we plotted the graph by dropping a couple of features, namely ASF & CDR. In figure 4, we plotted the graph by dropping a couple of features, namely MMSE & SES. In figure 5, we plotted the graph by dropping a couple of features, namely CDR & SES.

Thus, by looking at the above bar graphs and tables we come to a conclusion that, out of a number of classifiers which were being simultaneously trained and tested against the dataset. Out of these trained and tested classifiers, it has been surveyed that, out of all the classifiers that we have tested, by discarding a combination eTIV and ASF of and with a 7:3 splitting ratio, the Random Forest classifier produced the highest accuracy of 99.1%, amongst all other classifiers.

Thus, based on our overall survey, we were able to develop our front-end model. For our User Interface, in figure 6, it depicts our home page, containing fields for obtaining input values, from 3 categories of data, namely Demographic Data, Data from clinical tests conducted by a Neuro-physician

and Data from MRI scans. In figure 7, after submitting all the obtained values, we were able to witness the particular patient was demented. In figure 8, after submitting all the obtained values, we were able to witness the particular patient was non-demented. In figure 9, after submitting all the obtained values, we were able to witness the particular patient was converted.

7. CONCLUSION

In this paper, we have presented a survey of comparative analysis of ML techniques to detect Alzheimer's disease early. In the process of doing so, we tested all of our classifiers by dropping a couple of features in combinations and we witnessed Random Forest generating the highest accuracy of 99.1% amongst all others. As a result, the Random Forest classifier emerges as the best option for detecting Alzheimer's disease in its earlier stages.

8. REFERENCES

[1]. Khan, Aunsia, and Muhammad Usman. "Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper." 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K).

[2]. Bari Antor, Morshedul, et al. "A comparative analysis of machine learning algorithms to predict alzheimer's disease." Journal of Healthcare Engineering 2021 (2021).

[3]. Jo, Taeho, Kwangsik Nho, and Andrew J. Saykin. "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data." Frontiers in aging neuroscience 11 (2019): 220.

[4]. Farooq, Ammarah, et al. "A deep CNN based multi-class classification of Alzheimer's disease using MRI." 2017 IEEE International Conference on Imaging systems and techniques (IST). IEEE, 2017.

[5]. Oh, Kanghan, et al. "Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning." Scientific Reports 9.1 (2019): 1-16.

[6]. 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.

[7]. Al-Shoukry, S., Rassem, T. H., & Makbol, N. M. (2020). Alzheimer's diseases detection by using deep learning algorithms: a mini-review. IEEE Access, 8, 77131-77141.

[8]. Jimenez-Mesa, Carmen, et al. "Optimized One vs One approach in multiclass classification for early Alzheimer's disease and mild cognitive impairment diagnosis." IEEE Access 8 (2020): 96981-96993.

[9]. Dong, Celia M., et al. "Early Detection of Amyloid β Pathology in Alzheimer's Disease by Molecular MRI." 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020.

[10]. Padole, Himanshu, Shiv Dutt Joshi, and Tapan K. Gandhi. "Graph wavelet-based multilevel graph coarsening and its application in graph-CNN for alzheimer's disease detection." IEEE Access 8 (2020): 60906-60917.

[11]. Khan, Protima, et al. "Machine learning and deep learning approaches for brain disease diagnosis: Principles and recent advances." IEEE Access 9 (2021): 37622-37655.

[12]. Mikhno, Arthur, et al. "Toward noninvasive quantification of brain radioligand binding by combining electronic health records and dynamic PET imaging data." IEEE journal of biomedical and health informatics 19.4 (2015): 1271-1282.

[13]. Neelaveni, J., MS Geetha Devasena, and G. Gopu. "A comparative study on the application of machine learning algorithms for neurodegenerative disease prediction." Handbook of Decision Support Systems for Neurological Disorders. Academic Press, 2021. 283-302.

[14]. Jain, Varun, et al. "A Novel AI-Based System for Detection and Severity Prediction of Dementia Using MRI." IEEE Access 9 (2021): 154324-154346.

[15]. Song, Xuegang, Ahmed Elazab, and Yuexin Zhang. "Classification of mild cognitive impairment based on a combined high-order network and graph convolutional network." Ieee Access 8 (2020): 42816-42827.

[16]. Odusami, Modupe, Rytis Maskeliūnas, and Robertas Damaševičius. "An Intelligent System for Early Recognition of Alzheimer's Disease Using Neuroimaging." Sensors 22.3 (2022): 740.

[17]. Dataset source - EDA for predicting Dementia - https://www.kaggle.com/code/sid321axn/eda-for-predicting-dementia/data?select=oasis_longitudinal.csv.