JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 586-592
https://publishoa.com
ISSN: 1309-3452

Predictive Diagnostic Analysis for Early Detection of Alzheimer's disease Using Machine Learning

[1]K.C.Veena, [2] Mrs. R. Kavi Priya, [3] Dr. D. Sumathi

Department of Computer Science and Engineering, Kuppam Engineering College, Kuppam, India

Abstract - The accurate diagnosis of Alzheimer's disease (AD) plays an important role in patient treatment, especially at the disease's early stages, because risk awareness allows the patients to undergo preventive measures even before the occurrence of irreversible brain damage. Although many recent studies have used computers to diagnose AD, most machine detection methods are limited by congenital observations. Alzheimer's, an irreparable brain disease, impairs thinking and memory while the aggregate mind size shrinks which at last prompts demise. Early diagnosis of AD is essential for the progress of more prevailing treatments. Machine learning (ML), a branch of artificial intelligence, employs a variety of probabilistic and optimization techniques that permits PCs to gain from vast and complex datasets. As a result, researchers focus on using machine learning frequently for diagnosis of early stages of AD. This paper presents a analysis and critical evaluation of the recent work done for the early detection of AD using ML techniques. Several methods achieved promising prediction accuracies, however they were evaluated on different pathologically unproven data sets from different imaging modalities making it difficult to make a fair comparison among them. Moreover, many other factors such as pre-processing, the number of important attributes for feature selection, class imbalance distinctively affect the assessment of the prediction accuracy. To overcome these limitations, a model is proposed which comprise of initial pre-processing step followed by imperative attributes selection and classification is achieved using association rule mining. Furthermore, this proposed model based approach gives the right direction for research in early diagnosis of AD and has the potential to distinguish AD from healthy controls.

Index Terms - : Alzheimer's disease prediction, Heterogeneous, Random forest, Data mining, Early diagnosis etc.

I. Introduction

AD is a grievous neurodegenerative disease which is one of the chronic diseases that needs early detection, so that the treatment can be effective. Usually it starts slowly but with time worsens. This disease primarily affects the older people [1] and can become the cause for dementia [2]. The task of detecting this disease at the early stage is very difficult but equally important, so there is a necessity of intelligent system for supporting the clinicians in the early diagnosis of this disease. To address the mentioned problem, this paper elaborates the machine learning concept. Due to the handiness of improved technology, data exponentially increases and the world becomes a data rich society. This enormous amount of data takes the learning algorithms of machine at paramount height. Analyzing the immense data to get the fruitful result is the area of budding research. The target of all learning technologies is to retrieve the unseen patterns that can be further helpful in taking decision. Learning techniques are abundantly found in different sectors like media, health, agriculture, etc. Discovering and tackling of data is the herculean effort without intelligent learning models. Intelligent methods of learning are very useful for getting unseen patterns, trends, and relation between the features. The objective is to elaborate learning paradigms to model as well as to analyze the MRI data [1]. The vital role of learning methods in analyzing data motivates this research for comprehensive literature. As success of task directly depends on decision and correctness of decision depends on data as well as the proper analyses of data. This motivates the researchers either involved in academic or industry to work on techniques of learning models. Thus, this article revisits the present research and performed a systematic study. To know the potential of various well-known algorithms of machine learning such like LR, SVM, DT, RF, and boosting adaboost, the experiment conducted on all the mentioned learning approaches with the help of longitudinal MRI data. This dataset provides a useful aid for the diagnosis with the tracking of AD. The efficiency of the above mentioned models is measured in terms of accuracy, recall, and AUC (Area Under Curve). Additionally, time taken by the respective learning methodologies is also calculated.

The organizational framework of this study divides the research work in the different sections. The literature review is presented in section 2. Further, in section 3, Problem statement discussed. Moreover, in next section IV, briefly explain about Data and data set and data processing and in section V ,the description of machine learning and different learning mechanisms are mentioned. In section VI, experimental work is shown with diverse learning models.Conclusion and future work are presented by last sections VI.

II. Literature survey

Health sector is the critical area with the huge volume of data that needs special attention, so role of machine learning gains paramount importance to analyze the available records and on that basis predict the desired result. For the analytics of big data, the athmaja et. al write the survey related to the different algorithms of machine learning [3]. Various learning approaches are studied and presented to analyze big data [3].

In this literature, bkassiny et. al. write the survey in which learning problems are categorized for cognitive radios and artificial intelligence importance is stated to achieve the real cognitive systems for communication [4]. Various algorithms are studied and classified in main two categories, first is decision-making, and other is feature classification [4].

Liu et. al. explains and present a detailed survey in a systematic manner concern with the security threats along with the defensive techniques for machine learning [5]. This systematic survey is performed with two aspects, first one is training phase, and another is testing.

Ji et. al. performs the survey that gives overview comprehensively regarding techniques of tensor along with the applications related to machine learning [6]. This survey presents tensor basic knowledge which includes its operations, decomposition, several algorithms based on tensor, and its applications in machine and deep learning [6].

Das et. al. has performed a survey on different types of techniques of machine learning to prevent the systems from the intrusion, survey is performed by the [7]. With the evolution of the computational technologies, the data generated and transmitted by them increased exponentially. This network traffic requires effective surveillance, analysis by using intrusion detection with prevention system [7].

With the purpose of diagnosis, various learning algorithms are widely used in the biomedical area. To classify the skin disease with the help of texture as well as color features, this literature authors compare the different algorithms of machine learning [8]. This article addresses the problem of skin disease with the help of learning approaches [8]. This paper experimental result shows the higher accuracy of LDA (linear discriminant analysis) as well as for SVM.

Qian et. al. accomplishes the survey for the applications in the field of deep learning comprehensively [9]. This task performs on network layers. This study aids the readers in knowing the protocols related to the DL-enhanced in wireless network. Deep learning is the powerful tool of machine learning that is capable to handle the complex data for pattern recognition. The intelligence can be added to the wireless network using this learning approach [9]. The paramount importance of biomedical field attracts the attention to the analysis of the available medical data for the diagnosis and prediction of disease.

This research motive is seriously analyzing the longitudinal MRI data by using various learning paradigms for predicting early AD. Learning models are trained for identifying the patients on the basis of fourteen different features available in the MRI dataset. This dataset categorized into three labels: Nondemented, Demented, and Converted.

III. Problem statement

ALZHEIMER'S DISEASE

- Alzheimer's disease (AD) is a neurodegenerative disorder of uncertain cause and pathogenesis that primarily affects older adults and is the most common cause of dementia.
- The earliest clinical manifestation of AD is selective memory impairment and while treatments are available to ameliorate some symptoms, there is no cure currently available.
- Brain Imaging via magnetic resonance imaging (MRI), is used for evaluation of patients with suspected AD.
- MRI findings include both, local and generalized shrinkage of brain tissue. Below is a pictorial representation of tissue shrinkage.

- Some studies have suggested that MRI features may predict rate of decline of AD and may guide therapy in the future.
- However in order to reach that stage clinicians and researchers will have to make use of machine learning techniques that can accurately predict progress of a patient from mild cognitive impairment to dementia.
- We propose to develop a sound model that can help clinicians do that and predict early alzheimer's.

IV. DATA

The team has found MRI related data that was generated by the Open Access Series of Imaging Studies (OASIS) project that is available both, on their website that can be utilized for the purpose of training various machine learning models to identify patients with mild to moderate dementia.

A.DATA SET

Alzheimer is an important neurological disease, which needs s special attention. To detect AD, this study uses longitudinal dataset. This dataset consists of fifteen columns and 373 records. Out of fifteen columns, fourteen are features and one column represents label. Labels are classified into three categories, named are Nondemented, Demented, and Converted. First of all, various machine learning models are trained using longitudinal MRI dataset. After that, testing is performed and as a result Random Forest shows the best performance.

B.DATASET DESCRIPTION

- We will be using the longitudinal MRI data.
- The dataset consists of a longitudinal MRI data of 150 subjects aged 60 to 96.
- Each subject was scanned at least once.
- Everyone is right-handed.
- 72 of the subjects were grouped as 'Nondemented' throughout the study.
- 64 of the subjects were grouped as 'Demented' at the time of their initial visits and remained so throughout the study.
- 14 subjects were grouped as Non-demented at the time of their initial visit and were subsequently characterized as 'Demented' at a later visit. These fall under the 'Converted' category.

C. EXPLORATORY DATA ANALYSIS (EDA)

In this section, we have focused on exploring the relationship between each feature of MRI tests and dementia of the patient. The reason we conducted this Exploratory Data Analysis process is to state the relationship of data explicitly through a graph so that we could assume the correlations before data extraction or data analysis. It might help us to understand the nature of the data and to select the appropriate analysis method for the model later.

Table 1: The minimum, maximum, and average values of each feature for graph implementation are as follows.

	Min	**Max**	**Mean**
Educ	6	23	14.6
SES	1	5	2.34
MMSE	17	30	27.2
CDR	0	1	0.29
eTIV	1123	1989	1490
nWBV	0.66	0.837	0.73
ASF	0.883	1.563	1.2

D. DATA PREPROCESSING

We identified 8 rows with missing values in SES column. We deal with this issue with 2 approaches. One is just to drop the rows with missing values. The other is to replace the missing values with the corresponing values, also known as 'Imputation'. Since we have only 150 data, I assume imputation would help the performance of our model.

a. ***Removing rows with missing values***

Dropped the 8 rows with missing values in the column, SES

df_dropna = df.dropna(axis=0, how='any')

pd.isnull(df_dropna).sum()

b. Imputation

Scikit-learn provides package for imputation [6], but we do it manually. Since the *SES* is a discrete variable, we use median for the imputation.

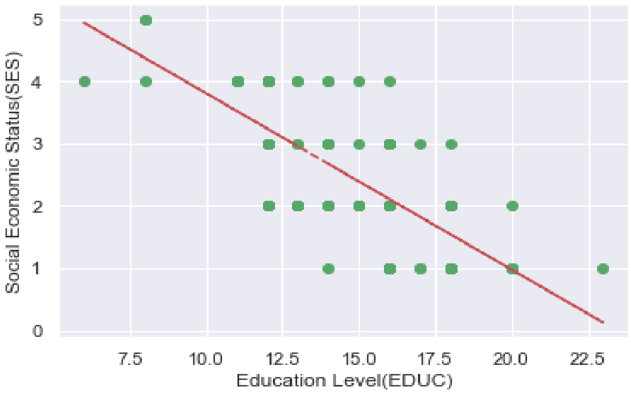


Figure 1: Imputation

c. Splitting Train/Validation/Test Sets

From sklearn. model_ selection import train_ test_ split

From sklearn import preprocessing.

From sklearn. Preprocessing import Min Max Scaler

From sklearn. Model_ selection import cross_ val_ score

d. Cross-validation

We conduct 5-fold cross-validation to figure out the best parameters for each model, Logistic Regression, SVM, Decision Tree, Random Forests, and AdaBoost. Since our performance metric is accuracy, we find the best tuning parameters by *accuracy*. In the end, we compare the accuracy, recall and AUC for each model.

V. Model

A. PERFORMANCE MEASURES

We use area under the receiver operating characteristic curve (AUC) as our main performance measure. We believe that in case of medical diagnostics for non-life threatening terminal diseases like most neurodegenerative diseases it is important to have a high true positive rate so that all patients with alzheimer's are identified as early as possible

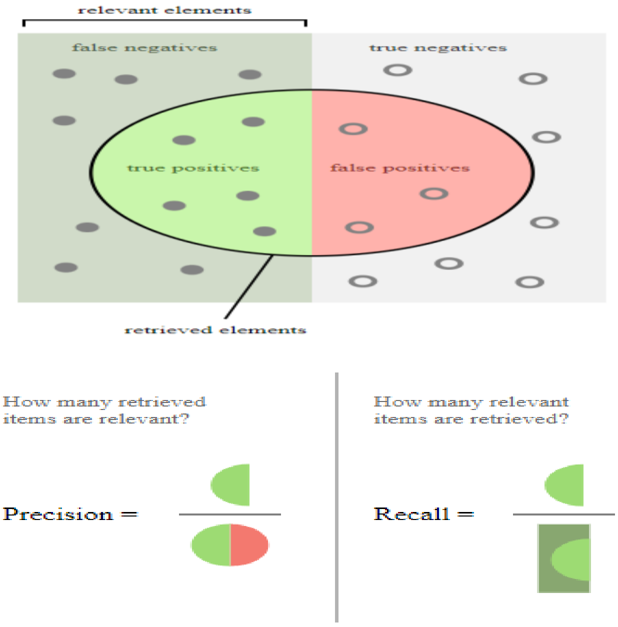


Figure 2: Sown precision and Recall

But we also want to make sure that the false positive rate is as low as possible since we do not want to misdiagnose a healthy adult as demented and begin medical therapy. Hence AUC seemed like a ideal choice for a performance measure. We will also be looking at accuracy and recall for each model.

In the figure 2, you can think relevant elements as actually demented subjects. Precision and Recall .

B. LOGISTIC REGRESSION

The parameter C, inverse of regularization strength.

Tuning range: [0.001, 0.1, 1, 10, 100]

C. SVM

C: Penalty parameter C of the error term. [0.001, 0.01, 0.1, 1, 10, 100, 1000]

gamma: kernel coefficient. [0.001, 0.01, 0.1, 1, 10, 100, 1000]

kernel: kernel type. ['rbf', 'linear', 'poly', 'sigmoid']

D. Decision Tree

Maximum depth. [1, 2, ..., 8]

8 is the number of features

E.RANDOM FOREST CLASSIFIER

n_estimators(M): the number of trees in the forest

max_features(d): the number of features to consider when looking for the best split

max_depth(m): the maximum depth of the tree.

F.ADABOOST

1. Retrains the algorithm iteratively by choosing the training set based on accuracy of previous training.
2. The weight-age of each trained classifier at any iteration depends on the accuracy achieved.
3. How do we select the training set?
4. How to assign weight to each classifier?

Let’s explore these questions, mathematical equation and parameters in behind them.

Ada-boost, like Random Forest Classifier is another ensemble classifier. (Ensemble classifier is made up of multiple classifier algorithms and whose output is combined result of output of those classifier algorithms).

In this chapter, we shall discuss about details of Ada-boost classifier, mathematics and logic behind it.

Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier.

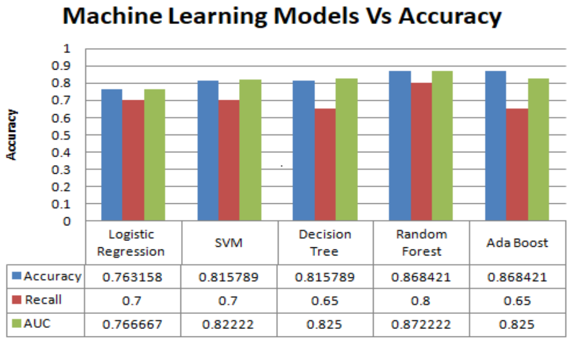
VI. EXPERIMENTAL RESULTS

This section shows the analyses of different learning models through the experimental work. The paper conducts the experiments on four well known different learning models using longitudinal MRI dataset taken. This dataset basically contains features to classify the Alzheimer’s disease. LR, SVM, DT, RF, and boosting adaboost are the famous learning models. The potential of these models is compared with respect to their accuracy, recall, and AUC. These learning approaches outcome is compared and shown in Table 2. The time required by the respective classifiers is also analyzed by using Table 1.

Table II: Machine Learning Models Comparison

S.No	Parameter	Accuracy	Recall	AUC	Time Taken in Sec
1	Logistic Regression	0.763158	0.70	0.766667	0.085
2	SVM	0.815789	0.70	0.82222	6.784
3	Decision Tree	0.815789	0.65	0.82500	0.068
4	Random Forest	0.868421	0.80	0.87222	483.8
5	ADA BOOST	0.868421	0.65	0.82500	2.615

The Table 2 presents the comparison of various learning models in reference of their respective accuracy. As compared with other classifiers, Random Forest and Adaboost achieve higher accuracy. Additionally, Random Forest gains higher recall and AUC. Pictorial analysis based on accuracy, recall, and AUC is represented by using Figure 4.



	Logistic Regression	SVM	Decision Tree	Random Forest	Ada Boost
🟦 Accuracy	0.763158	0.815789	0.815789	0.868421	0.868421
🟥 Recall	0.7	0.7	0.65	0.8	0.65
🟩 AUC	0.766667	0.82222	0.825	0.872222	0.825

Figure 3. Learning Models v/s Accuracy, Recall, and AU

Figure 3 graphically depicts the comparison of five well known learning classifiers with reference to their respective accuracy. All the models are showing good accuracy, minimum is 76.31% gained by logistic regression. However, random forest and ada boost achieves better outcomes in reference of accuracy, recall, along with AUC strength.

Figure 4 Analyze graphically the time required by the different classifiers. This analysis is helpful to judge the time complexity.

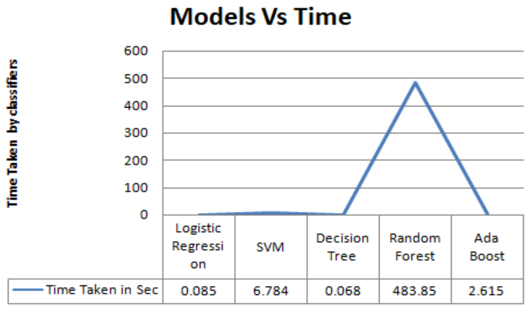


Figure 5. Machine Learning Models v/s Time Requirement

VII. CONCLUSION

The paper presents the comprehensive overview of AD and how various learning approaches can analyze them. With the advancement in computational technology, data is enhancing day by day. It becomes difficult to handle this bulk of data. Machines as well as deep learning models explained in this study are the tools to tackle and analyze this bulky data. These learning models are able to analyze the data and can further classify or predict the result. Different learning model's analysis through experimental work is also shown in this article. The potential in terms of accuracy, recall, AUC, and time requirement to execute is analyzed efficiently in tabular as well as graphical form. With the comparative study of five ML models on longitudinal MRI datasets, this paper concludes:

- Random Forest and AdaBoost achieves high accuracy as compared with others. This is because AdaBoost has the capability to turn weak classifier to strong one. Random Forest strength to overcome the overfitting problem, makes it better than the others.
- Random Forest gets high recall or true positive rate due to reduction of overfitting problem.
- AUC is a performance measure parameter, which is high with Random Forest.
- Along with good performance, Random Forest classifier takes more time to execute. This classifier training speed is low as compared to others.

Future work

Along with the overview of machine and deep learning models, this research gives data pre-processing details. This is performed by gathering information by surveying various papers. In the coming time, this research can be very fruitful and utilized in applying learning models in different areas like health, agriculture, banking, etc. For analyzing the data, these learning models are great achievements for the scientists involved in both academic and industry. This article is useful for the researchers who are working in this direction and further can come up with more yielding outcomes.

REFERENCES

1. J. Escudero, E. I feachor, J.P. Zajicek, C. Green, J. Shearer, S. Pearson, “Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease”, IEEE transactions on biomedical engineering, Vol. 60, no. 1, (2013), pp. 164-168.
2. Q. Zhou, M. Goryawala, Cabrerizo, J. Wang, W. Barker, D.A. Loewenstein, R. Duara, M. Adjouadi, “An optimal decisional space for the classification of Alzheimer's disease and mild cognitive impairment”, IEEE Transactions on Biomedical Engineering, Vol. 61, no. 8, (2014), pp. 2245-2253.
3. S. Athmaja, M. Hanumanthappa, V. Kavitha, “A survey of machine learning algorithms for big data analytics”, Proceedings of the IEEE International Conference on Innovations in Information, Embedded and Communication Systems, Coimbatore, India, (2017) March 17-18.
4. M. Bkassiny, Y. Li, S.K. Jayaweera, “A survey on machine-learning techniques in cognitive radios”, IEEE Communications Surveys & Tutorials, Vol. 15, no. 3, (2012), pp. 1136-1159.
5. Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, V. C. M. Leung, "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven

View," IEEE Access, vol. 6, (2018), pp. 12103-12117.

6. Y. Ji, Q. Wang, X. Li, J. Liu, "A Survey on Tensor Techniques and Applications in Machine Learning" IEEE Access, Vol. 7, (2019), pp. 162950-162990.
7. S. Das, M.J. Nene, "A survey on types of machine learning techniques in intrusion prevention systems", Proceedings of IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, (2017) March 22-24.
8. P. R. Hegde, M. M. Shenoy, B. H. Shekar, "Comparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features", Proceedings of IEEE International Conference on Advances in Computing, Communications and Informatics, Bangalore, India, (2018) Sep 19-22.
9. Q. Mao, F. Hu, Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey", IEEE Communications Surveys & Tutorials, Vol. 20, no. 4, (2018), pp. 2595-2621.
10. P. Sherkhane, D. Vora, "Survey of deep learning software tools", Proceedings of IEEE International Conference on Data Management, Analytics and Innovation (ICDMAI), Pune, India, (2017) Feb 24-26.
11. Q. Mao, F. Hu, Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey", IEEE Communications Surveys & Tutorials, Vol. 20, no. 4, (2018), pp. 2595-2621.
12. L. Jiao, J. Zhao, "A Survey on the New Generation of Deep Learning in Image Processing", IEEE Access, Vol. 7, (2019), pp. 172231-172263.
13. T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, A. Liu, "Big Data Cleaning Based on Mobile Edge Computing in Industrial Sensor-Cloud", IEEE Transactions on Industrial Informatics, Vol. 16, no. 2, (2020), pp. 1321-1329.
14. A.C. Kakouri, C.C. Christodoulou, M. Zachariou, A. Oulas, G. Minadakis, C.A. Demetriou, C. Votsi, E. Zamba-Papanicolaou, K. Christodoulou, G.M. Spyrou, "Revealing clusters of connected pathways through multisource data integration in huntington's disease and spastic ataxia", IEEE journal of biomedical and health informatics, Vol. 23, no. 1, (2018), pp. 26-37.
15. Z. Wen, M. Zhou, "Evaluating the use of data transformation for information visualization", IEEE transactions on visualization and computer graphics, Vol. 14, no. 6, (2008), pp. 1309-1316.
16. R. M. Gahar, O. Arfaoui, M.S. Hidri, N.B. Hadj-Alouane, "A Distributed Approach for HighDimensionality Heterogeneous Data Reduction", IEEE Access, Vol. 7, (2019), pp. 151006-151022.
17. S. Garcia, J. Luengo, J.A. Sáez, V. Lopez, F. Herrera, "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no. 4, (2012), pp. 734-750.
18. G. Krummenacher, C.S. Ong, S. Koller, S. Kobayashi, J.M. Buhmann, "Wheel defect detection with machine learning", IEEE Transactions on Intelligent Transportation Systems, Vol. 19, no. 4, (2017), pp. 1176-1187