Journal of IMS Group
Volume 19, No.01, January-June, 2022, pp. 1- 6

ISSN No. 0973-824X

Early Detection of Dementia using Machine Learning Techniques

* Rahul Singhal

Abstract

Today, as many as 7% of adults aged sixty and above suffer from dementia. Over four million individuals in India suffer from dementia in some form. Dementia affects at least 44 million people worldwide, making it a global health concern that must be tackled. Dementia is a condition of the mind rather than a disease. It is defined as a significant deterioration in mental function from a prior higher level that interferes with a person's everyday activities. Abnormal brain alterations create the disorders included under the umbrella term "dementia." This condition causes a deterioration in thinking abilities, also known as cognitive capacities, that is severe enough to interfere with everyday living and independence. It also has an impact on one's conduct, emotions, and relationships. Early identification and diagnosis of dementia can help in halting disease progression and minimize stress and morbidity in patients and caregivers. The objective of this work is to implement different machine learning-based models to identify dementia using gathered data and brain MRI scans in general practice. The approach might be beneficial for detecting persons who may have dementia but have not been formally diagnosed or who have a tendency for it. To improve the results, appropriate feature engineering and data preparation were used. Finally, using suitable performance assessment parameters, the outcomes from all of the models were compared.

Keywords: Dementia, Feature Engineering, Machine Learning

1. Introduction

Dementia is a phrase that refers to a set of symptoms that impact your memory, and reasoning tothe point that they interfere with your regular activities. It is a condition, not a disease, with symptoms that are shared by a variety of brain disorders. With time, it will deteriorate.

Medications, on the other hand, may assist to slow down the deterioration and alleviate symptoms such as behavioural changes. Memory loss is a common symptom of dementia, although it can be caused by a variety of factors. Memory loss isn't always a marker of dementia; however, it is generally one of the first symptoms. Alzheimer's disease is the most prevalent cause of progressive dementia in older people, although dementia can also be caused by a varietyof other conditions. One

***Rahul Singhal**
Department of Computer Science, Jaypee Institute of Information Technology, Noida, India
rahulsinghal1904@gmail.com

prevalent misconception concerning memory loss is that it invariably indicates dementia in you or a loved one. Memory loss can be caused by a variety of factors. A diagnosis of dementia isn't always made on the basis of memory loss. It's also true that some memory alterations are typical as people become older since some neurons in the brain die naturally as we get older. This form of memory loss, on the other hand, isn't functionally detrimental; it doesn't interfere with everyday living.

Table 1 : Types of Dementia

Primary:	**Secondary:**	**Reversible dementia:**
Dementia is the primary ailment in these diseases and illnesses.	Dementia caused by a different illness or condition	Other illnesses or reasons can generate similar symptoms

The advancement of medical diagnosis utilizing magnetic resonance imaging (MRI) is extensively utilized for the treatment of neurological problems; it allows for the acquisition of ever more functional and anatomical information from the brains of patients, with remarkable accuracy in time and space. The typical time between symptom start and physician contact is four years, which is mainly due to the patient's insecurity over having a memory impairment. Physicians are rarely able to arrest the course of dementia and decrease detrimental behavioral changes at this stage of the disease. A simple way of diagnosing dementia early in its development might encourage people to seek examination and treatment sooner rather than later.In this paper, we implement different machine learning-based models to detect dementia using brain MRI scans from the OASIS-Brains.org website which contains Longitudinal MRI data of patients suffering from dementia.

2. Literature review

Cuingnet et al. (2011) examined 10 approaches using MRI scans from the ADNI dataset and foundthat whole-brain methods have the highest accuracy, with a sensitivity of 81 percent and a specificity of 95 percent. For SVM and regularized logistic regression, Mathotaarachchiet et al. (2017) utilized the ADNI-GO/2 research using PET images. SVM, Bayesian-network classifier withinverse tree structure (BNCIT), Naïve Bayes, decision tree, multiple layer perceptron, logistic regression, and discriminant analysis were implemented on MRI images obtained from the Washington University by Chen et al (2010). Kloppel et al. (2009) employed SVM to identify Alzheimer's disease using real-world MRI data. Bhagyashree et al. (2017) used Jrip, Random Forest, Nave Bayes, and J48, in a pilot exploratory research in Mysore Studies of Natal Effects on Health and Aging. Maroco et al. (2011) compared seven non-parametric classifiers with Quadratic Discriminant Analysis, Logistic Regression, and Linear Discriminant Analysis. Williams et al. (2013) used SVM, Decision Tree, NN, and Naive Bayes to predict CDR scores and clinical diagnoses where missing data were substituted with average values. It was found that Naive Bayes gave the best accuracy. On data with 24 attributes from a database gathered from several neuropsychologists, four classification approaches were tested, and it was discovered that Nave Bayes provided the best accuracy by Shree et al. (2014). Aruna et. al. (2016) employed MRI images from the OASIS dataset for SVM classification with the Gabor Filter, Gray Level Co-occurrence Matrix, Independent Component Analysis, and feature fusion with the RBF classifier, all of which showed high accuracy, precision, and recall. Tohka et al. (2016) examined several feature selection strategies for dementia using SVM and Logistic Regression for anatomical brain MRI and it was found that removing age increases the average accuracy of all the classifiers. The OASIS dataset's MRI scans are subjected to Voxel-Based Morphometry in a paper by Chyzhyk et al. (2010). Rodman et al. (1996) used C4.5 rules, C4.5, Nave Bayes, and IB1 to assess data collected from two simple cognitive and functional skills tests, FAQ and BOMC, in order to improve dementia detection.

3. Dataset Description

In this work, longitudinal MRI data from OASIS Brains.org was used. The dataset comprises of longitudinal MRI data from 373 participants ranging in age from 60 to 96 years old.

Demographics Information has the following attributes:

- Gender
- Handedness
- Age
- Years of education (EDUC)
- Socioeconomic status (SES) as assessed by the Hollings head Index of Social Position and classified into categories from 1(highest status) to 5 (lowest status)

Clinical Information has the following attributes:

- Mini-Mental State Examination score(MMSE) It ranges from 0 for worst to 30 for best

- Clinical Dementia Rating (CDR): It is 0 for no dementia, 0.5 for very mild AD, 1 for mild AD, 2 for moderate AD

Derived Anatomic Volumes have the following attributes:

- Estimated total intracranial volume (eTIV)

- Normalized whole-brain volume (nWBV): The automatic tissue segmentation technique labels a percentage of all voxels in the atlas-masked picture as grey or white matter.

- Atlas scaling factor(ASF): A scale factor that converts native-space brain and skull to the atlas target.

4. Proposed Methodology

This section discusses the methodology proposed in this study.

4.1 Exploratory data analysis

Exploratory data analysis has been carried out to better understand the dataset. From the Group variable, we find that 39% of them are Demented,51% are Nondemented and 10% are converted. Then, we analyzed the most important categorical feature i.e., Clinical Dementia Rating (CDR). From Fig. 1 we see that a CDR value of 0 has the highest frequency (200 patients).

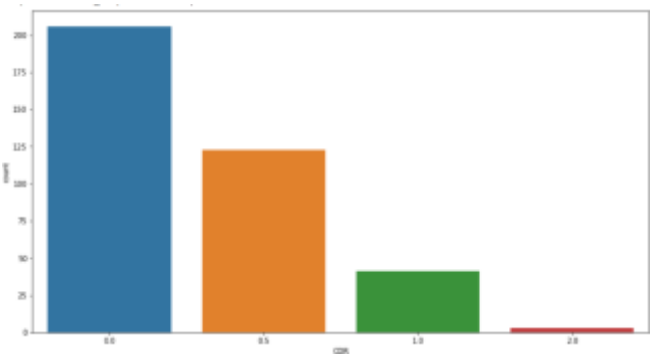

Figure 1. Frequency distribution for different CDR values

In Fig. 2, MMSE values of 30 have the highest count and normal MMSE is the highest at 86%. From Fig. 3 we can see that Visit and MR Delay are showing a close correlation of 0.92. These correlated columns convey similar information to the learning algorithm and therefore, should be removed.

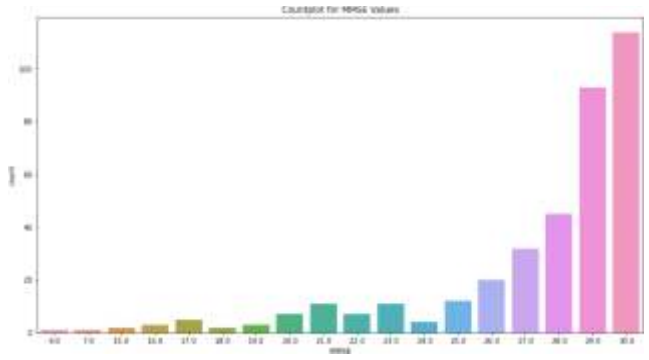

Figure 2. Count plot of MMSE values

Figure 3. Heatmap showing correlation among different features

4.2 Data Pre-processing

Data pre-processing is the process of transforming raw data into an understandable format. It involved removing the irrelevant columns such as 'Subject ID', 'MRI ID', 'Hand',' Visit', and 'MR Delay'. The columns SES and MMSE contained null values, and they had to be replaced with mean, median, or mode instead. It was found that SES has only integer values so we have to replace it with median or mode. Mean, median, and mode were close to 2 so we replaced null values of SES with 2. One-hot encoding was performed on the 'Group' and 'M/F' columns and then the numeric and categorical variables were concatenated to form the input data to be sent to the machine learning model.

4.3 Feature Engineering and Model implementation

Feature Engineering is carried out to extract features from raw data and keep the necessary features for efficient implementation of machine learning models. The dataset was split into 70 % training data and 30% testing data. Several machine models with appropriate feature selection and hyperparameter tuning were implemented for detecting dementia. I began by implementing Logistic Regression wherein the performance of the Logistic regression model was improved by selecting only the significant features (we find the variable having p values<0.05 to get statistically significant variables). Hence 'M/F', 'SES', 'eTIV', 'ASF', and 'CDR' were chosen as the input attributes for the model. RFE is another feature selection approach that fits a model and eliminates the weakest features until the desired number of features is attained. RFE approach was used for selecting the six best features for the Logistic Regression model. We obtained 'M/F', 'EDUC', 'SES', 'ASF', 'MMSE', 'CDR' as the most significant features through this approach. It was found that Logistic Regression with RFE gave the best accuracy out of the three LR approaches. Similarly, Naive Bayes classifier, KNN, Decision Tree and Random Forest, XGBoost, and AdaBoost were also implemented.

Hyperparameter tuning and feature selection was done to improve the results. Then, I proceeded with building an ensemble model where the predictions of multiple classification models such as Random Forest, KNN, and Naive Bayes were used as additional features to train a meta- classifier.

5. Experimental Results

Different machine learning models are evaluated and the results are tabulated below:

Table 2. Accuracy scores of different machine learning models in detecting dementia

S.No.	Model	Accuracy
1	Naive Bayes	96.42%
2	Logistic Regression	92.85%
3	Logistic Regression (With Feature Selection)	92.85%
4	Logistic Regression (With Recursive Feature Elimination Approach)	96.42%
5	KNN (with Hyperparameter Tuning)	95.53%
6	Decision Tree (with Hyperparameter Tuning)	95.53%
7	Random Forest (with Hyperparameter Tuning)	96.42%
8	AdaBoost	96.42%
9	XGBoost (with Hyperparameter Tuning)	95.53%
10	Ensemble Model (Random Forest, KNN, Naive Bayes)	95.53%

Table 3. Precision, Recall, and F1 Score values for different machine learning models in detecting dementia

S.No.	Model	Precision	Recall	F1-Score
1	Naive Bayes	1.00	0.9310	0.9642
2	Logistic Regression	0.98	0.8750	0.9245
3	Logistic Regression (With Feature Selection)	0.9827	0.8906	0.9344
4	Logistic Regression (With Recursive Feature Elimination Approach)	1.00	0.9310	0.9642
5	KNN (with Hyperparameter Tuning)	0.9814	0.9298	0.9549
6	Decision Tree (with Hyperparameter Tuning)	1.00	0.9122	0.9541
7	Random Forest (with Hyperparameter Tuning)	1.00	0.9285	0.9629
8	AdaBoost	0.9807	0.9444	0.9622
9	XGBoost (with Hyperparameter Tuning)	1.00	0.9122	0.9541
10	Ensemble Model (Random Forest, KNN, Naive Bayes)	1.00	0.9122	0.9541

From Table 2, It can be inferred that Naive Bayes, Logistic Regression with Recursive Feature Elimination, Random Forest, and Adaboost produce the best accuracy (96.42%) while simple Logistic Regression gave the worst accuracy for detection of people suffering from dementia. It can be seen that implementation of feature selection and RFE improves the Accuracy, Precision, Recall, and F1- Score considerably in the case of Logistic Regression. Because the penalty of making a false negative forecast on dementia detection is significantly higher than making a false positive, the emphasis on recall is very important. AdaBoost gives the best Recall score (0.9444) while simple Logistic Regression gives the least score (0.8750) as given in Table 3.

6. Conclusion

Dementia is the seventh biggest cause of mortality worldwide, as well as one of the top causes of impairment and reliance among the elderly. Dementia has medical, cognitive, social, and financial repercussions for individuals suffering from dementia, their caregivers, families, and society as a whole. As a result, having a system that can take big quantities of data and automatically identify people who may have dementia in order to support focused screening may be highly valuable and assist enhance diagnosis rates. Machine learning algorithms help in solving this problem. This paper has focused on the comparison of results derived from implementing different machine learning models for the identification of dementia. Naive Bayes, Logistic Regression with RFE, Random Forest, and Adaboost gave the best results (Accuracy of 96.42%). It was also observed that feature selection and hyperparameter tuning helped in optimizing the performance of models. Dementia detection at an early stage can considerably reduce the number of patients with dementia who are undetected. The scope of this study can be extended further by combining with other datasets and also trying other modalities. Deep learning method can be also be incorporated to learn better from the given information.

Reference

- Aruna, S. K., & Chitra, S. (2016). Machine Learning Approach for Identifying Dementia from MRI Images. International Journal of Computer and Information Engineering, 9(3),881–888. https://publications.waset.org/10004510/machine-learning-approach-for- identifying-dementia- from-mri-images

- Bhagyashree, S. I. R., Nagaraj, K., Prince, M., Fall, C. H. D., & Krishna, M. (2017). Diagnosis of Dementia by Machine learning methods in Epidemiological studies: a pilotexploratory study from south India. Social Psychiatry and Psychiatric Epidemiology, 53(1), 77–86. https://doi.org/10.1007/s00127-017-1410-0

- Chen, R., & Herskovits, E. H. (2010). Machine-learning techniques for building a diagnostic model for very mild dementia. NeuroImage, 52(1), 234–244. https://doi.org/10.1016/j.neuroimage.2010.03.084

- Chyzhyk, D., & Savio, A. (2010). Feature extraction from structural MRI images based on VBM: data from OASIS database. http://www.ehu.eus/ccwintco/uploads/3/38/GIC-UPV-EHU-RR- 2010-10-14

- Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., & Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage, 56(2), 766–781. https://doi.org/10.1016/j.neuroimage.2010.06.013

- Kloppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., & Frackowiak, R. S. J. (2008). Automatic classification ofMR scans in Alzheimer's disease. Brain, 131(3), 681–689. https://doi.org/10.1093/brain/awm319

- Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011).Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMCResearch Notes, 4(1). https://doi.org/10.1186/1756-0500-4-299

- Mathotaarachchi, S., Pascoal, T. A., Shin, M., Benedet, A. L., Kang, M. S., Beaudry, T., Fonov, V. S., Gauthier, S., & Rosa-Neto, P. (2017). Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiology of Aging, 59,80–90. https://doi.org/10.1016/j.neurobiolaging.2017.06.027

- Rodman, W., Datta, P., Dillencourt, M., & Pazzani, M. (1996). Improving Dementia Screening Tests with Machine Learning Methods. Retrieved May 19, 2022, from http://www.ics.uci.edu/~pazzani/Publications/alz-res.pdf

- Shree, S. R. B., & Sheshadri, H. S. (2014). An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques. 2014 IEEE International Conference on Computational Intelligence and Computing Research. https://doi.org/10.1109/ICCIC.2014.7238300

- Tohka, J., Moradi, E., & Huttunen, H. (2016). Comparison of Feature Selection Techniquesin Machine Learning for Anatomical Brain MRI in Dementia. Neuroinformatics, 14(3), 279–296. https://doi.org/10.1007/s12021-015-9292-3

- Williams, J., & Weakley, A. (2013). Machine learning techniques for diagnostic differentiation of mild cogni- tive impairment and dementia. https://eecs.wsu.edu/~cook/pubs/aaai13.2.pdf