ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF RESEARCH IN PHARMACEUTICAL SCIENCES

Published by JK Welfare & Pharmascope Foundation Journal Home Page: www.ijrps.com

A detailed investigation in determining Alzheimer's disease and its risk factor using different classification techniques

Mahendran Radha*, Anitha M, Jeyabaskar Suganya

Department of Bioinformatics, School of Life Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

Article History:

Received on: 15 Nov 2020
Revised on: 12 Dec 2020
Accepted on: 15 Dec 2020

Keywords:

Alzheimer's Disease,
Dementia,
Classification Techniques,
Linear SVM,
WEKA

ABSTRACT

The prevalence of genetic disorders has recently crept surprisingly high. Neurodegenerative complications, specifically, pose physical and mental stress to parents and caretakers. These complications may be witnessed in the case of dementia. The general dementia type that accounted for between 60 to 80 per cent of psychiatric illnesses was Alzheimer's disease. At an earlier stage, illness detection serves as a critical task that helps the diseased person to enjoy a decent quality of life. It has become a much necessitated strategy towards relying on automated techniques like data mining approach for early diagnosis and assessment of risk factors concerned with Alzheimer's. There has been an unprecedented growth of interest concerned with devising novelized approaches proposed in recent times for classifying the disease. However, there is still a grave need for developing an efficacious approach for better prognosis and classification. Data mining is carried out using different machine-learning approaches to assess the risk factors for Alzheimer's disease. Through the present research, and we compared numerous classification methods such as Decision Tree, Linear SVM, KNN, Logistic Regression, Radial SVM, and Random Forest, and finally reported the most outstanding approach in terms of its accuracy.

*Corresponding Author

Name: Mahendran Radha
Phone: +91 9003237145
Email: mahenradha@gmail.com

ISSN: 0975-7538

DOI: https://doi.org/10.26452/ijrps.v12i1.4149

Production and Hosted by

IJRPS | www.ijrps.com

INTRODUCTION

Approximately 44 million people have dementia (shree *et al.*, 2014). There are 38 million people with Alzheimer's disease who are struggling. One of the forms of dementia is Alzheimer's disease (Viswanathan *et al.*, 2009; Sosa *et al.*, 2009). Alois Alzheimer's, a German neurologist and physician, discovered Alzheimer's disease in 1906 (Sandeep *et al.*, 2015). Multiple risk factors that lead to the progression of the disease are distinct (shree *et al.*, 2014). Height, Down syndrome, consumption of alcohol and smoke, food style, cholesterol, etc. The signs of this disorder are interpersonal coordination, decision making, total lack of memory and failure of gestures, bad judgment, and irregular moods. The three different steps of ADD care are visiting the general surgeon, doing neuropsychological assessments, and taking MRI scans (Saling *et al.*, 2007). By 2001, more than 11 million people were uniformly afflicted by Alzheimer's disease. There are approximately about 36 (35.6) million people suffered with AD or other manifestations of dementia, rising to about 66 (65.7) million by 2030 and rising to around 115 (115.4) million by 2050 (Alzheimer's Association, 2010). The number of people with demen-

tia is predicted to double by 2030 and to triple by 2050. Although as medical specialists such as physicians, there is a significant difference between them; medical practitioner never reveals to the outside world their system of prediction of a specific illness. Therefore this crisis could be overcome by a prediction approach with expert experience and lead to reliable disease prediction outcomes. We use different kinds of machine learning algorithms for this research.

Literature survey

In over 60 to 80% of dementia cases, Alzheimer's disease accounted. Such disorders remain undiagnosed at an early stage (Sandeep *et al.*, 2017a,b). There are 3 main diagnostic stages via a general practitioner. Step one is consultation. The 2nd stage includes multiple neuropsychological assessments after MRI scans are taken in the 3rd stage (Thies and Bleiler, 2013). AD requires a screening test can be used, regardless of culture, gender, education and religion, for the subjects. The Dementia Research Group 10/66 formed a network in 1998 and dedicated itself to studies of the highest standards in those areas. This phase is also dependent on the psychologist's mood. In addition it is not easy to prevent human error. This crisis could be resolved by machine based research. So, researchers discovered information using a data mining approach. Using methods such as analytics, artificial intelligence and machine learning, data mining can be performed. As different scholars have used data mining was explored for the study of various diseases (shree *et al.*, 2014). Use decision tree and Bayesian classification when evaluating the data sets of patients with heart disease (Soni *et al.*, 2011). Classification algorithms have been used to classify Parkinson's disease (Tarigoppula *et al.*, 2013). The machine learning approach used in the classification of Alzheimer's disease, vascular dementia and Parkinson's disease (Joshi *et al.*, 2010). The whole work illustrates the efficacy of assuming that the risk factor for proper classification of AD, VD and PD is very significant. It was determined from assessing 180 related investigations. The study showed a precision of 99.33 per cent obtained using perceptron multilayer and random forest (Tarigoppula *et al.*, 2013). The machine learning investigation governing Alzheimer's disease was discussed in (Escudero *et al.*, 2013).

MATERIALS AND METHODS

Architectural framework

The workflow of the present study is represented in Figure 1.

Dataset Array

The data obtained here is the most significant. The 750 medical reports obtained by various neuropsychologists comprised of datasets. The four age ranges are 65-69, 70-75, 76-79, and over 80 years of age.

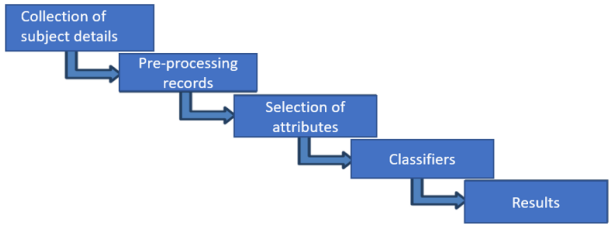

Figure 1: Block diagram indicating the working flow.

Preprocessing

It is a stage the missing and incorrect values can be verified. Data preprocessing is not carried out here as there is no risk of missing data.

Options for Attributes

A selection of attributes is the main stage, and certain attributes create a great difference in decision making. The data set comprises 8 attributes that represent the main risk factors related to AD namely, Family history, Age, Environmental toxins, Gender, Head injury, Factors including High BP and cholesterol level, Low education Level, and Lifestyle.

	Family history	Gender	Age	Env. toxins	Head injury	Factors	Low edu. level	Lifestyle	Outcome
Family history	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Gender	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
Age	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
Env. toxins	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Head injury	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
Factors	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
Low edu level	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Life style	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1

Figure 2: Correlation matrix representing the datasets.

Classification Techniques (classifiers)

WEKA Tool

The next stage is the grouping. This is done to understand just how the material is being categorized.

For research, the WEKA tool is used. The classification algorithm runs several times to maximize precision. WEKA has two successful assessors for learning. The first one is a classifier and cross-validation is the second one.

Table 1: Accuracies before standardization.

Model	Accuracy
Radial SVM	0.6510416666666666
KNN	0.7291666666666666
Decision Tree	0.7552083333333334
Linear SVM	0.7708333333333334
Logistic Regression	0.7760416666666666
Random Forest	0.8072916666666666

Table 2: Accuracy after standardization and selecting correlated features.

Model	New Accuracy	Accuracy	Increase
Linear Svm	0.78125	0.7708333333333334	0.01041666666666663
Radial Svm	0.7708333333333334	0.6510416666666666	0.11979166666666674
Logistic Regression	0.7760416666666666	0.7760416666666666	0.0
KNN	0.7291666666666666	0.7291666666666666	0.0
Decision Tree	0.7291666666666666	0.7552083333333334	-0.02604166666666674
Random Forest	0.7708333333333334	0.8072916666666666	-0.03645833333333326

Table 3: Cross Validation Scores.

Model	CV Mean
Linear SVM	0.78125
Radial SVM	0.7708333333333334
Logistic Regression	0.7760416666666666
KNN	0.7291666666666666
Decision Tree	0.7291666666666666
Random Forest	0.7708333333333334

Linear SVM (Linear support vector machine)

Linear SVM is the recently discovered classification technique for large dataset data mining. Compare to other techniques, Linear SVM is the best performer.

Radial SVM (Radial support vector machine)

Radial SVM is a common kernel feature used in different learning algorithms that are kernelized. It is commonly used in the Methodology of Support Vector Machines (Chang *et al.*, 2010). Do the optimization of an SVM model that can forecast bankruptcy. While the RBF kernel is commonly used in the fitting of data for its stability, other common kernels, such as polynomial or sigmoid, are (Joshi *et al.*, 2010).

Logistic Regression

For predicting binary classes it is statistical method. The target variable is dichotomous. Dichotomous means there are only two possible classes. It calculates the probability of an event occurrence.

KNN (K Nearest Neighbor)

The K Nearest Neighbor algorithm has been used in various data analysis because of its simplicity and high accuracy (Xiong *et al.*, 2007). It has been accepted as one of top 10 algorithms in data mining. Estimating k value by 10 fold cross validation, 97.4% of accuracy has been obtained (Wu *et al.*, 2008).

Decision Tree

It is a tree structure that is flowchart like. When the internal node represents a function (or attribute), the branch represents a law of choice, and the outcome is expressed by each leaf node. In decision making, this flowchart-like form supports. Like a flowchart map, it's a hallucination that imitates thinking at the human level. But it is easy to grasp and interpret only decision trees.

Random Forest

It is a managed algorithm for learning. For classification and regression, Random Forest is used. The algorithm is the algorithm that is most versatile and simple. Woodland is composed of trees. It lies at the base of the Boruta algorithm, which is a dataset that selects essential attributes.

RESULTS AND DISCUSSION

From the observed classification methods that are used for the datasets (i.e.) Risk factors which are major contributors for AD were taken into account. The accuracy from the classification of algorithms before standardization was represented in Table 1. The percentage of test set tuples that are appropriately identified by the classifier is the accuracy of the classifier on a given test set. Post-standardization precision and collection of correlation features are calculated in Table 2 and its cross-validation scores represented in Table 3. From the result, the process was used for determining the model showcasing the best accuracy. From the determined Datasets of AD based on major risk factors is represented in the Correlation matrix Figure 2.

CONCLUSIONS

Different data mining classification methods have been compared and graded. The specificity of the execution of each procedure is observed. Linear Support Vector Machine, Radial Support Vector Machine, Logistic Regression, K Nearest Neighbor, Decision Tree, and Random Forest are the following classifiers used in the prediction of Alzheimer's disease risk factors. Among them, compared with other classifiers, Linear SVM demonstrated better accuracy. This study shows the Linear SVM classification method serves as the best protocol for the prediction of various genetic disorders. We conclude with this analysis linear SVM classification technique can use another genetic disease risk prediction process.

Conflict of Interest

The authors declare that they have no conflict of interest for this study.

Funding Support

The authors declare that they have no funding support for this study.

REFERENCES

Alzheimer's Association 2010. Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 6(2):158–194.

Chang, Y. W., *et al.* 2010. Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research*, 11(4):1471–1490.

Escudero, J., *et al.* 2013. Machine learning based method for personalized and cost effective detection of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, 60(1):164–168.

Joshi, S., *et al.* 2010. Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. *International Journal of Engineering and Technology*, 2(4):350–355.

Saling, M., *et al.* 2007. *Early Diagnosis of Dementia*. Alzheimer's Australia, Pfizer, Australia, Pfizer, Australia. (Accessed On March 2007).

Sandeep, C. S., *et al.* 2015. A review on the early diagnosis of Alzheimer's disease (AD) through different tests, techniques and databases. *AMSE Journals Modelling C*, 76(1):1–22.

Sandeep, C. S., *et al.* 2017a. The early diagnosis of Alzheimer Disease using CAMD, TREAD and NAAC databases. *International Journal for Science and Advance Research in Technology*, 3(3):366–371.

Sandeep, C. S., *et al.* 2017b. The online datasets used to classify the different stages for the early diagnosis of Alzheimer's Disease. *International Journal of Engineering and Advanced Technology*, 6(4):38–45.

shree, S. B., *et al.* 2014. An Approach in the Diagnosis of Alzheimer Disease - A Survey. *International Journal of Engineering Trends and Technology*, 7(1):41–43.

Soni, J., *et al.* 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 17(8):43–48.

Sosa, A. L., *et al.* 2009. Population normative data for the 10/66 Dementia Research Group cognitive test battery from Latin America, India and China: a cross-sectional survey. *BMC Neurology*, 9(1):48–57.

Tarigoppula, V., *et al.* 2013. Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms. *International Journal of Engineering and Innovative Technology*, 3(3):212–227.

Thies, W., Bleiler, L. 2013. 2013 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 9(2):208–245.

Viswanathan, A., *et al.* 2009. Vascular risk factors and dementia: How to move forward? *Neurology*, 72(4):368–374.

Wu, S. H., *et al.* 2008. Toward the optimal itinerary-based KNN query processing in mobile sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 20(12):1655–1668.

Xiong, L., *et al.* 2007. Mining multiple private databases using a KNN classifier. *Proceedings of the 2007 ACM symposium on Applied computing*, pages 435–440.