

Alzheimer's Disease Diagnosis by Using Dimensionality Reduction Based on KNN Classifier

M. BALAMURUGAN, A. NANCY and S. VIJAYKUMAR

Department of Computer Science, Bharathidasan University, Tamilnadu, India.

*Corresponding author E-mail: nancyabraham88@gmail.com

<http://dx.doi.org/10.13005/bpj/1299>

(Received: November 22, 2017; accepted: December 04, 2017)

ABSTRACT

Data mining is fast developing technology in extensive sort of applications. One of the essential data mining areas is medical data mining. Healthcare industry is a kind of industry, where huge amount of information's and are more sensitive. That information is required to be handle very carefully without any mischievousness. There is a wealth of data presented in healthcare but there is no effective analysis tool to discover hidden relationships in data. There are numerous data mining methods that have been utilized as a part of healthcare industry but now the investigation has to be going on the performance of several classification techniques. In this paper, they proposed the Novel dimensionality reduction based KNN Classification Algorithm for analyzing and classifying the Alzheimer disease and Mild Cognitive Impairment are present in the datasets. National Alzheimer's Coordinating Centre (NACC) having the Researcher's Data Dictionary - Uniform Data Set (RDD-UDS) is gives dataset for the researchers to analyzing clinical and statistic information's. From this research work, that gives more accuracy percentage, sensitivity percentage and specificity percentage to provide a better result.


Keywords: AD- Alzheimer's disease, Mild Cognitive Impairment (MCI), National Alzheimer's Coordinating Centre (NACC), Researcher's Data Dictionary - Uniform Data Set (RDD-UDS),K nearest neighbor(KNN).

INTRODUCTION

Data mining is a most dominant technology which having high potential to support discovering of hidden predictive data from huge datasets. These extracted data's are load in a data warehouse which are stored and managed into multidimensional databases. In current days, the applications of data mining in healthcare system play a vital role because the health region contains rich information and it became an essential technology. This technology is

used to extract the information from the database at any time, which needs for processing. The understanding of this techniques leads to improve the efficiency and enhance the feature of disease analysis in the certain datasets^{1,2}. Dementia is a type of syndrome; it contains various set of symptoms includes functional degradation in sense of space, memory loss, decision taken, manipulating ability, abstract thinking, and attention. The patients may have complex actions because of internal stimulus, identity changes, misunderstandings



This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted Non Commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

or hallucinations. The seriousness of such manifestations may impact patients' relationships and capacities to work³.

Alzheimer's Disease (AD) is the greatest eminent type of disease in dementia; it has common symptoms of memory loss, thinking capacity, changes in behaviors, moods and other cognitive abilities have severe enough to affect the day by day life. Memory harms are usually one of the starting signs of Alzheimer's disease; but various persons can have the different symptoms initially. It has different sorts of stages:

- Preclinical
- Early(mild)
- Middle(moderate)
- Late(severe)

In the preclinical stage it does not emit any symptoms but the brain gets some toxic. The early stage gets the symptoms of slight changes which are not noticeable. In the middle stage, there is loss of memory, confusions, and people may get the difficulty in identifies the family and friends. The

final stage having more severe issues with the loss of communication ability, sleeping problem and loss more weight. Mild cognitive impairment (MCI) causes light changes in cognitive abilities but that are noticed by the ways of expressing them or other people. These changes are not affecting the daily life of that individual.

LC-kNN classifier is executed to find the Alzheimer's disease (AD). This classifier distinguished the ailment by using arranges of high compactness (low dimension) and discriminability. The INs representations beside with the versatile idea of the utilized metric separation are assumed to be the key elements of the LC-kNN classifier. The examination of a component that declines the data vectors dimension by considering the classes⁴. The objective of this work to discover and classifying the Alzheimer's disease and Mild Cognitive Impairment by using the dimensionality reduction based KNN classification algorithm.

The sensitive techniques of cerebrospinal fluid (CSF) biomarkers and brain imaging were utilized to discover the initial step of disease to improve the progress. Here lack of improvement in the biomarkers present in blood which has been replicated in huge amount of studies and that useful for clinically discovers the difficulties facing by individuals during the innovation of AD. A few serum markers have been described which may emerge from inflammatory actions in the central nervous system in the early course of AD.

The remaining of this paper structured as follows: The comprehensive descriptions of the related works on the detection of Alzheimer disease under data mining domains are discussed in section II. The execution process of Novel dimensionality reduction based KNN Classification algorithm is defined in section III. The performance analysis of KNN with existing approaches delivered in section IV. At last, the conclusions about the diagnoses of AD using KNN algorithm described in section V.

Related Works

This section describes the detection of Alzheimer's disease and Mild cognitive impairment (MCI) from the datasets. Lu, *et al*⁵ involved in the huge amount of datasets that were classified based

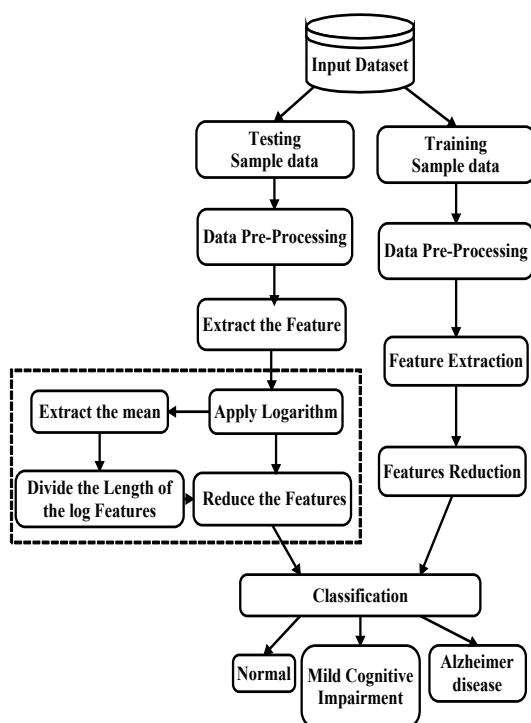


Fig. 1: Workflow of proposed system

on the symbolic learning methods in decision tree. The given number of data's were discovered the high quality rules. While substantial work on using neural networks for classification has been reported, none of them can generate rules with the quality comparable to those generated by Neuro Rule. Here, issues was to reduce the training time of neural networks. Gironi, *et al.* [6] established the technique of Artificial neural networks (ANNs) that used to manipulate the simulated models by central nervous system. The computational models inspired by the network and had the ability for learning mechanisms and detection of patterns. This work has to reveal the correlation between immunological and oxidative stress markers in AD and MCI by the application of ANNs.

Al-nuaimi, *et al*⁷ presented an approach of Tsallis entropy has theoretical systems information for measuring deviations in the EEG. It provides beneficial vision of brain functions and it played an important role in initial stage of detection and diagnosis of dementia by using decision-support tool. It has a maximum temporal resolution, non-invasive, and the minimum of cost. The dementia causes the damage which affects the brain cells and this leads to modify the features of EEG. Information of theoretical methods developed as a possible way to compute the variations in EEG as biomarkers of dementia. Ortiz, *et al*⁸ shown the process of neurodegenerative, that produce the changes in connection of brain network and the physical changes in pattern of brain activation. Sparse Inverse Covariance Matrix (SICE) method used to work out an experimental observation of brain network with the help of Gaussian graphical. This allowed to build a graphical view of the brain connectivity, by exhibiting the inter-regional brain network. The conditional independence was not able to evaluate by the covariance-based methods based on the variables to factor out the impact of other regions.

Savio and Graña⁹ discovered the variations of significant morphological in brain anatomy that were relevant to the brain atrophy of AD. Then it suggested the method of Computer Aided Diagnosis (CAD) which chosen from the fields of deformation, and moderated GM, that gained from non-linear registration process. The discriminant localization site was reliable with the literature concerning images

in biomarkers of AD. The various classification results among deformation measures were not significant. Bhat, *et al*¹⁰ stated that the analysis of AD was still a challenging task. Hence a study of recent research automated electroencephalography based on the diagnosis of AD was presented in this work. Neuroprotective and symptomatic methods were as anti-oxidants and neurotransmitters have verified to be effective in the treatment of AD symptoms and found the delay in development process. Here, the lack of accuracy in EEG based Diagnosis of AD. Kapur, *et al*¹⁰ described about the collection of data's which was produced in the daily life, in that data's were extracted at the time of needed for processing otherwise no use of it. Data that we have to gauge students' potential based on various indicators like previous performances and in other cases their background to gave a comparative account on what method was the best in achieving that end. They compared six algorithms were Random Forest, Naive Bayes, Naive Bayes Multinomial, K-star, IBk. In that study, all the algorithms exposed that the amount of accuracy was low for that to be implemented on a large measure.

Walker, *et al*¹¹ talked about the difficulty of handling the microarray data that come from two known classes were as Alzheimer and normal. They suggested three various techniques used to identify the genes associated with the Alzheimer disease (AD). It has the major process which utilized gene expression data for disease classification and diagnosis. Here, only considered the data's for the process. Seixas, *et al*¹² described about the early analysis of this kind of disease permits to taken the treatment in advance and recover quickly. For this they developed the Bayesian network decision model for helpful in analysis of dementia, AD and Mild Cognitive Impairment (MCI). Bayesian networks were suitable for demonstrating the uncertainty and causality, which were existed in medical domains. Candás, *et al*¹³ detected the disorders and diseases using abnormal human activities under free-living conditions was a reliable process. They proposed Automatic data mining method based on physical activity measurements used to evaluate the activities of human beings. But they don't consider real time analysis of operation to get the effective performance. Bang, *et al*¹⁴ several research work were established to improve a dementia identification

method in the area of computer-aided Diagnosis (CAD) technique. This work implemented a quad phased data mining modeling containing of 4 segments. In Proposed Module, substantial for analytical measures were selected the effective way of discovery. Trambaiolli, *et al.*¹⁵ intended to estimate the significance of FS methodologies implemented to Alzheimer's Disease (AD) EEG-based analysis and compare the selected features with previous medical findings. This was estimated by seeing the leave-one-subject-out exactness of Support Vector Machine classifiers made from the datasets defined by the certain features. Zhang, *et al.*¹⁶ developed a hybrid classification system for distinctive NC, MCI, and AD based on physical MRI images. They utilized MRI data, demographics, clinical analysis, and resultant anatomic capacity as the training data. The CDR was used as the target data. They used PCA to decrease the dimensionality of the feature vectors of the MRI data and the resultant principal mechanisms reserved significant data. Bull, *et al.*¹⁷ proposed the open-source SAMS structure which used for data collection, text collection, and analyze the methods. It faced a number of challenges, but the primary challenges was derived to deploy it on a real users' home computers and to gather data as from their used computers to do everyday things. But it did not allow the resources to establish a SAMS product configuration lines for every type, brand and version of desktop computer, operating system, web browser and desktop application. Adeniyi, *et al.*¹⁸ applied a K-NN classification method with Euclidean distance technique has accomplished to generate suitable and a quite better classifications and recommendations of the customer at any time. The system performs classification of users on the simulated active sessions extracted from testing sessions by collecting active users' click stream and matches this with similar class in the data mart. This utilized to produce a set of recommendations to the customer in a Real-Time basis. But it has less efficiency.

Zhang, *et al.*¹⁹ Normal elder control used for the early analysis or identification of Alzheimer's disease (AD). In this Computer-aided diagnosis (CAD) technique for MR brain pictures based on Eigen brains and mechanism learning with the accurate discovery of AD and

AD- related brain diseases. Here, there was a lack of classification accuracy during the process. Khan, *et al.*²⁰ Considered the spatial data streams classifications and the training datasets were frequently changed. Training data's were arrived and that data's added to the training set. Here, proposed a k-nearest neighbor(KNN) classification, it finds the k nearest neighbors based on certain distance metric by finding the distance of the target data point from the training dataset. The problem associated with KNN classifiers were it significantly increases the classification time. Ferreira, *et al.*²¹ selected the most related neuropsychological and sorts of demographic for the prediction of prognostic with the help of genetic algorithm. This selection model has the ability of conversion predicted for dementia with the 88 percent of sensitivity. It helps for improve the treatment process in effective manner. Suk, *et al.*²² introduced a multitask learning method for feature selection in computer-aided Alzheimer's disease (AD) or Mild Cognitive Impairment (MCI) diagnosis. During the distribution process the neuroimaging data has numerous peak or modes. To detect the multi peak distributional characteristics and define the subclass based on the outcomes of clustering method.

Proposed work

This section evaluates the implementation details of proposed Novel dimensionality reduction based KNN Classification algorithm for achieving the classification result of Alzheimer disease. The simultaneous achievement of large size of data samples and better classification depends on following models in proposed work as shown in Fig. 2 and are explained as follows.

- Data Pre-Processing
- Dimensionality reduction based classification

The aim of this research work is to detect the Alzheimer Disease (AD) from the datasets by using Novel dimensionality reduction based KNN Classification algorithm. Input Datasets having more number of data's which contains all type of information's including the Normal, Mild cognitive Impairment (MCI), and the Alzheimer's disease (AD). Collect data's from the datasets that are used for testing and training Feature Extraction method used

for the selection purpose which extracts the required data's for further process. Then apply logarithm for the data's to compute the mean value. purpose. The data's performing a data preprocessing method for filtering the unwanted data's from the datasets.

The extracted mean values are divided by the size of log features and executed the feature reduction operation. This is also repeated for the training process. At last, the outcome of Training and testing process are classified into Normal, (MCI), and (AD). These are based on analysis of information in the datasets.

Data Preprocessing

Preprocessing guarantees effective operation of disease analysis. The main purpose of data preprocessing is to reduce the amount of features used for classification method. Data preprocessing techniques involving the following steps:

- Data collections
- processing the information's

Table 1: Proposed measures of different classes

Parameters	Class 1	Class 2	Class 3
TP	2000	2000	1136
TN	3138	3141	4000
FP	0	0	7
FN	5	2	0

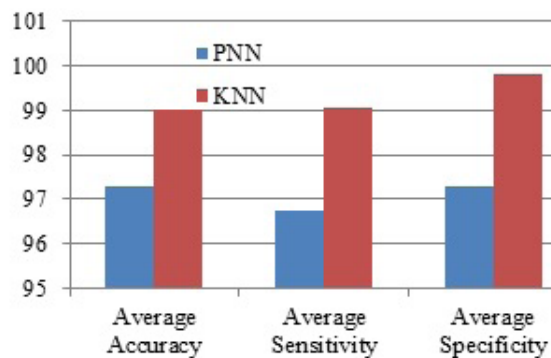


Fig. 2: Average accuracy, sensitivity and specificity

- Extract the unwanted Basic information's such as, DOB, Sex, Education etc.,
- Select health oriented information's.

From the datasets, processing all the data's to separate the relevant information's. Then the unwanted information's are eliminated from the dataset; the remaining features which retain sufficient information for analyzing the person's health condition are as Normal, Mild cognitive Impairment (MCI), and the Alzheimer's disease (AD). Feature extraction is a common method that contains all type of data's are as both relevant and irrelevant for the disease analyzing purpose. In this some irrelevant features are extracted from the overall features and it provides a new set of features for the further process.

Dimensionality reduction based Classification

Dimensionality reduction techniques used to manage the interesting tasks in data mining. Due to the trouble of the data and ordering in real-world applications, it appears not a simple task to construct a common dimensionality reduction technique. Decreases in dimensionality can initiate with dual features: decreasing the amount of samples or decreasing the amount of features. To get reliable methods for dimensionality reduction falls into two different sorts of models: feature extraction and feature subset selection. Feature extraction denotes to raising new features with a linear or nonlinear conversion from the original input space to a feature space, while feature subset selection is to discover certain informative features from the original set and

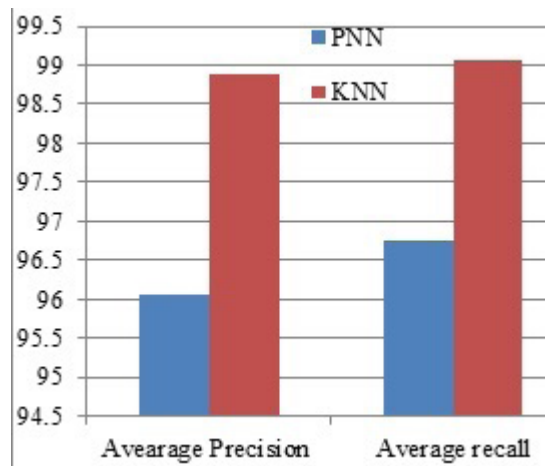


Fig. 3: Precision and Recall

remove another's. Data reduction is to decrease the dimensionality and difficulty of the information and data conversion to change the data into appropriate form for mining etc. The classification techniques based on the process of Feature Reduction. It has a significant part in data mining for classification of data's. The feature reduction process is used for data selection and defining the dimension of feature set properly. This process enhances the efficiency and the performance in data mining. In this work, three classes are separated for the analysis

- Class 1- Alzheimers Disease (AD)
- Class 2- Normal
- Class 3- Mild Cognitive Impairment (MCI)

The dimensionality reduction based KNN Classification algorithm used for identifying the Alzheimer's disease. From the input of dataset, the data's are taken for the testing and training process. Initially, take logarithm for the data to get a log feature value. Then Extract the mean value of log features with the help of dividing its length. Likewise, divide the mean value using the length of log features. Take the mean value from the calculation which is greater than the log features. The above process will be repeated for Training Features and store the result as $Train_{Feat}$. After, Evaluate the Euclidean Distances between the points.

The equation (6) used for calculating the distance between the Training and Testing feature. Based on $Dist_{feat}$, the training features are sorted. From the sorted list the k points are selected and those points are the k closest training samples to $Test_{feat}$. The nearest neighbors are represented as k. Based on the maximum number of vote to find the class for $Train_{feat}$.

Performance Analysis

This section assesses the execution of Alzheimer's disease analysis in data mining systems utilizing proposed dimensionality reduction based KNN Classification algorithm. The performance of this proposed method is analyzed by using the Datasets. Alzheimer Disease is the most familiar kind of Dementia which represents an probable of 60 to 80 percent cases. The indications of Alzheimer disease contain difficulty in the memory of recent conversations, laziness, depression etc. Alzheimer was characterized as the moderate dynamic mind

sickness which starts a long time before side effects develop The Researcher's Data Dictionary - Uniform Data Set (RDD-UDS) is gives dataset for researchers for analyzing clinical and statistic information from the National Alzheimer Coordinating Center (NACC). The RDD-UDS unites data from the first information accumulation mechanisms for all past and present UDS frame variants. The primary target of Data mining utilizing clinical dataset is to bring out the hidden patterns from the clinical information.

Performance Metrics

Accuracy

The closeness of a measured value to the standard or known value is termed as accuracy. It is otherwise stated as a weighted arithmetic means of precision and the recall.

$$Accuracy = \frac{TP + TN}{P + N} \text{ OR } \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity

The true positive rate is also called as sensitivity is defined as the fraction of positives which are appropriately recognized.

$$Sensitivity = \frac{TP}{TP + FN}$$

Where TP denotes the number of true positives that correctly classified the DR region and FN denotes the amount of false negatives that incorrectly classified the normal DR region.

Specificity

A true negative is called as specificity which evaluates the fraction of negatives that are appropriately recognized.

$$Specificity = \frac{TN}{TN + FP}$$

Where, TN represents the amount of true negatives that are properly classified in a normal portion and FP represents the amount of false positives that incorrectly classifies the DR regions.

Precision

The precision is the portion of the retrieved document that is related to the query and it called

for the positive predictive rate (PPR).

$$Pr = \left\{ \frac{\{\text{relavant document} \cap \text{retrieved document}\}}{\text{retrieved document}} \right\} \text{ or}$$

$$\frac{TP}{TP + FP}$$

Recall

The recall is termed as sensitivity, which is the proportion of the portion of recovered and relevant instances.

$$\text{Recall} = \left\{ \frac{\{\text{relavant document} \cap \text{retrieved document}\}}{\text{relavant document}} \right\} \text{ or}$$

$$\frac{TP}{TP + FN}$$

TP, TN, FP, FN

The four parameters deliberate such as TP, TN, FP, and FN is utilized to classify the normal, MCI and AD present in the Datasets.

Table I furnishes the Parameters of TP, TN, FP, and FN that having the outcome measure of disease in the datasets. From the results, it is obvious that Novel dimensionality reduction based KNN Classification algorithm outperforms the existing PNN scheme.

Average Accuracy, Sensitivity and Specificity

The parameter metrics such as Accuracy, sensitivity, and specificity are analyzed. Fig. 3 shows the comparative analysis of the proposed Novel dimensionality reduction based KNN Classification algorithm and the existing PNN classifier algorithm.

Hence, the comparative analysis of existing PNN classifier with the proposed Novel dimensionality reduction based KNN Classification algorithm improves the Accuracy, sensitivity and specificity performance values.

Average Precision and Recall

The parameter metrics such as precision and recall are analyzed. Fig.4 shows the comparative analysis of the proposed Novel dimensionality reduction based KNN Classification algorithm and the existing PNN classifier algorithm.

CONCLUSION

This paper proposed the Novel dimensionality reduction based KNN Classification Algorithm analyzed and classified the Alzheimer's disease present in the datasets. By this algorithm they separated it in 3 classes are as Class 1 having the Alzheimer's Disease (AD), class 2 having Normal result, Class 3 having Mild Cognitive Impairment (MCI). The data's are taken from the National Alzheimer's Coordinating Centre (NACC) has the Researcher's Data Dictionary - Uniform Data Set (RDD-UDS). The proposed technique taken the data from dataset a testing and training operations are performed to evaluate the outcome of this technique. The comparative analyses between the existing PNN classification techniques with the proposed KNN classification showed that high amount of average accuracy, Sensitivity, Specificity Precision, Recall, Jaccard and Dice coefficients and also reduce the data dimensionality and computational complexity. The future work, the feature extraction and classification algorithm will use to improve the classification accuracy rate.

REFERENCES

1. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology*, **5**: pp. 241-266 (2013).
2. I. ĀRANU, "Dta mining in healthcare: decision making and precision," *Database Systems Journal BOARD*, p. 33: (2016).
3. A. So, D. Hooshyar, K. W. Park, and H. S. Lim, "Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques," *Applied Sciences*, **7**: p. 651(2017).
4. S. Bhat, U. R. Acharya, N. Dadmehr, and H. Adeli, "Clinical neurophysiological and automated EEG-based diagnosis of the Alzheimer's disease," *European neurology*, **74**: pp. 202-210 (2015).
5. H. Lu, R. Setiono, and H. Liu, "Neurorule: A connectionist approach to data mining," *arXiv preprint arXiv:1701.01358*, (2017).
6. M. Gironi, B. Borgiani, E. Farina, E. Mariani, C.

- Cursano, M. Alberoni, *et al.*, "A global immune deficit in Alzheimer's disease and mild cognitive impairment disclosed by a novel data mining process," *Journal of Alzheimer's Disease*, **43**: pp. 1199-1213(2015).
7. A. H. Al-nuaimi, E. Jammeh, L. Sun, and E. Ifeachor, "Tsallis entropy as a biomarker for detection of Alzheimer's disease," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 4166-4169 (2015).
 8. A. Ortiz, J. Munilla, I. Álvarez-Illán, J. M. Góriz, J. Ramírez, and A. s. D. N. Initiative, "Exploratory graphical models of functional and structural connectivity patterns for Alzheimer's Disease diagnosis," *Frontiers in computational neuroscience*, **9** (2015).
 9. A. Savio and M. Graña, "Deformation based feature selection for computer aided diagnosis of Alzheimer's disease," *Expert Systems with Applications*, **40**: pp. 1619-1628 (2013).
 10. B. Kapur, N. Ahluwalia, and R. Sathyaraj, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms," *International Journal of Advanced Research in Computer Science*, **8** (2017).
 11. P. R. Walker, B. Smith, Q. Y. Liu, A. F. Famili, J. J. Valdés, Z. Liu, *et al.*, "Data mining of gene expression changes in Alzheimer brain," *Artificial intelligence in medicine*, **31**: pp. 137-154 (2004).
 12. F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade, "A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment," *Computers in biology and medicine*, **51**: pp. 140-158 (2014).
 13. J. L. C. Candás, V. Peláez, G. López, M. Á. Fernández, E. Álvarez, and G. Díaz, "An automatic data mining method to detect abnormal human behaviour using physical activity measurements," *Pervasive and Mobile Computing*, **15**: pp. 228-241 (2014).
 14. S. Bang, S. Son, H. Roh, J. Lee, S. Bae, K. Lee, *et al.*, "Quad-phased data mining modeling for dementia diagnosis," *BMC medical informatics and decision making*, **17**: p. 60 (2017).
 15. L. Trambaiolli, N. Spolaôr, A. Lorena, R. Anghinah, and J. Sato, "Feature selection before EEG classification supports the diagnosis of Alzheimer's disease," *Clinical Neurophysiology*, **128**: pp. 2058-2067 (2017).