Indonesian Journal of Electrical Engineering and Computer Science
Vol. 27, No. 3, September 2022, pp. 1689~1697
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v27.i3.pp1689-1697

Alzheimer's disease prediction using three machine learning methods

Shaymaa Taha Ahmed[1], Suhad Malallah Kadhem[2]
[1]Department of Computer Science, Faculty of Basic Education, University of Diyala, Diyala, Iraq
[2]Department of Computer Science, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Jun 1, 2022
Revised Jun 16, 2022
Accepted Jul 4, 2022

Keywords:

Alzheimer's disease
Gene expression
Information gain
Microarraytechnology
Support vector machine

ABSTRACT

Alzheimer's disease (AD) is the most common incurable neurodegenerative illness, a term that encompasses memory loss as well as other cognitive abilities. The purpose of the study is using precise early-stage gene expression data from blood generated from a clinical Alzheimer's dataset, the goal was to construct a classification model that might predict the early stages of Alzheimer's disease. Using information gain (IG), a selection of characteristics was chosen to provide substantial information for distinguishing between normal control (NC) and early-stage AD participants. The data was divided into various sizes; three distinct machine learning (ML) algorithms were used to generate the classification models: support vector machine (SVM), Naïve Bayes (NB), and k-nearest neighbors (K-NN). Using the WEKA software tool and a variety of model performance measures, the capacity of the algorithms to effectively predict cognitive impairment status was compared and tested. The current findings reveal that an SVM-based classification model can accurately differentiate cognitively impaired Alzheimer's patients from normal healthy people with 96.6% accuracy. As discovered and validated a gene expression pattern in the blood that accurately distinguishes Alzheimer's patients and cognitively healthy controls, demonstrating that changes specific to AD can be detected far from the disease's core site.

Corresponding Author:

Shaymaa Taha Ahmed
Department of Computer Science, Faculty of Basic Education, University of Diyala
Diyala, Iraq
Email: mrs.sh.ta.ah@gmail.com

1. INTRODUCTION

Alzheimer's disease (AD) dementia affects 40–50 million individuals worldwide, with the number more than doubling between 1990 and 2016 [1]. Alzheimer's disease (AD) is by far the most common type of dementia and is anticipated to become more widespread as the population ages. The costs are rising in tandem with the growth in its occurrence. Alzheimer's disease is estimated to have cost the globe $604 billion in 2010 [2]. By 2030, Alzheimer's disease is predicted to cost $2 trillion in healthcare worldwide, affecting more than 131 million individuals. As a result, AD is quickly becoming a major global health and economic issue, prompting intensive scientific research to identify underlying genetic risk factors and regulatory markers, as well as to reduce the estimated healthcare burden through early detection, particularly at presymptomatic stages, to lessen the expected cost of healthcare [3], [4]. The late-onset symptoms of AD are the subject of a lot of research neurofibrillary tangles, amyloid plaques, neuronal tangles, and other tangles are examples [5], [6]. Although these discoveries have diagnostic relevance, the overall therapeutic contributions of these late-onset Alzheimer's disease characteristics are unclear [7], [8]. Furthermore, clinical

trials demonstrate that patients with Alzheimer's disease have a wide range of symptoms and respond to different treatments, implying that there are numerous biological origins for the disease. This complicates the investigation of AD even further [9], [10].

Data obtained by high-throughput gene expression describing has provided new paths for a better understanding of complicated disease mechanisms and pathways at the molecular level in recent years [11]. However, identifying embedded patterns in high throughput gene expression data is difficult due to the large dimension, small sample size, and noise. In the context of gene expression summary dataset analysis, the methods for identifying the most explanatory gene subsets through data reduction and feature selection are now divided into two categories [12]: i) method of marginal filtering and ii) method wrapper (embedded) [13], [14]. Univariate and multivariate marginal filtering are the two types of marginal filtering. The paired t-test (TS), information gain (IG), and pearson correlation coefficient (PCC) are examples of univariate filtering procedures [15]-[17]. If there are too many features, there will be over fitting issues, and if there are too few, key features will be missed [18]. As a result, feature selection is a critical component of modeling [19], [20]. Although it may sound ideal for developing a predictive and robust model, the difficulty of selecting highly relevant features present in the gene expression dataset, which has around 16,382 characteristics. In such circumstances, feature selection and dimensionality reduction algorithms aid in identifying core feature(s) that have a significant impact on result prediction. The information gain, chi-squared test, and mean decrease gini test are among the statistical tests utilized in this study to find such gene expressions [21], [22].

Henceforth looked at the importance of feature selection in the detection of Alzheimer's disease and discovered a suitable selection approach that can better predict the disease in this study. The information gain (IG) is based on three biomarkers: MRI, PET, and CSF, all of which are recommended by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and related disorders association [23], [24]. To improve the accuracy of the classifications, a method was utilized to choose features, and the top 44 ranked feature subsets were constructed. Then, using the selected features, a support vector machine (SVM) algorithm-based classifier will be built to predict AD from healthy controls (HCs). The predictor's performance is evaluated using a 5-cross validation method. The findings of this study's experiments revealed the efficacy of the proposed feature selection strategy in the diagnosis of Alzheimer's disease.

2. RELATED WORKS

Researchers from a wide range of fields are interested in microarray data analysis. The following are some of the most recent proposals for microarray data analysis in the fields of artificial intelligence, machine learning, and other related fields. Several research projects have recently been carried out because of recent advancements in biomedical and information technology, resulting in several algorithms that are useful for AD prediction. In this section, therefore, look at some of the most recent methods that have been developed employing data mining techniques.

Meng *et al.* [25], used four types of machine learning models which are support vector machine (SVM), Naive Bayes (NB), random forest (RF) and multilayer perceptron neural network (MLP-NN) to analysis gene expression of AD patients and normal people. They used in this experiment a dataset namely gene expression omnibus (GEO: GSE1297) maintained by National Center for Bioinformatics Information (NCBI). The statistical t-test method with the significance of p-value <0.05 is used as a gene selection for selecting the best gene subset. The results indicated the accuracy of the above models as follows: (87.10), (90.32), and (97.66) respectively. Among them, the MLP-NN model performs better than other models, on identifying the distinction between AD and normal genes and proving its efficacy.

Scheubert *et al.* [26], presented a classification system for predicting AD from the dataset GSE5281, which is referred to as the AD dataset. A wrapper of genetic algorithm and support vector machine (GA/SVM) is used as a feature selection method to select a subset of relevant genes that improves the performance of classification. Six different classification methods: Naive Bayes (NB), C4.5 (decision tree), k-nearest neighbor (KNN), random forest (RF), SVM with Gaussian kernel and SVM with linear kernel have been used. The results indicated the accuracy of the above models as follows: (81.4), (78.9), (87.0), (87.0), (85.7) and (91.9) respectively.

Huang *et al.* [27], presented a classification system for predicting AD from the dataset GSE63060 and GSE63061, which is referred to as the AD dataset. Including analysis of variance (ANOVA) and mutual information (MI) is used as a feature selection method to select a subset of relevant genes that improves the performance of classification. Different classification methods: k-means algorithm and convolutional neural network (CNN) have been used. The results indicated the accuracy of the above models as follows: 0.886 and 0.929 respectively. A fundamental challenge in deep learning is determining the network design that provides

the best prediction accuracy. This process involves choosing network hyper parameters, including the number of layers, transformation types, and training parameters.

Eke *et.al.* [28], presented a classification system of AD using gene expression datasets namely: GSE63060 and GSE63061. These two datasets were merged. The least absolute shrinkage and selection operator (LASSO) feature selection method is used to detect the optimal subset. The classification models: support vector machine (SVM), random forest (RF) and logistic ridge regression (RR) have proved predictive in distinguishing between cognitively normal (CN), mild cognitive impairment (MCI), and subjects with AD. SVM, RF, and RR classification models achieved accuracy (0.773), (0.785), and (0.765) respectively.

Niyas and Thiyagarajan [29], reported on the diagnosis of Alzheimer's disease, using machine learning techniques. Bermany and Rashid [30] used a supervised approach of support vector machine (SVM) model to classify image. Jha and Kwon [31], proposed a cluster analysis technique for AD diagnosis from Magnetic Resonance Brain Images in the light of the selection tree. Bi *et al.* [32], proposed the random neural system group to enhance the order of execution. The dataset for their experimentation was selected from the Alzheimer's disease neuroimaging initiative (ADNI). The elman neural network was confirmed to be an ideal base classifier that utilizes the arbitrary neural system cluster that is dependent on the outcome of highlighted selections, with 92.31% of accuracy.

Tanveer *et al.* [32] used on the random support vector machine clustering approach to diagnose AD and to group the disease regions into inferior frontal gyrus, superior frontal gyrus, precentral gyrus, and cingulate cortex. Voyle *et al.* [33] developed multiple models for the classification of AD using various machine learning techniques, namely multilayer perceptron, bagging, decision tree, coactive neuro-fuzzy inference system (CANFIS) and genetic algorithm. The CANIFIS method's classification precision was 99.55%. Plant *et al.* [34] developed a hybrid model using SVM and Bayesian Classifier to detect the brain atrophy patterns as well as to predict AD. A pattern matching index of 92% was obtained for their method. Vega *et al.* [35] developed a new methodology for the classification of MR brain images into normal and AD affected images. Their method is underlined by MRI feature extraction in the wavelet domain followed by dimensionality reduction and SVM classification.

Li *et al.* [36] proposed a current strategy based on the principal component analysis (PCA). It utilizes models of continuous selection and dropout as well as the model of restricted boltzmann machine (RBM), which is a profound teaching method. Ma *et al.* [37] created a technique to analyze AD from medical images using ML-based multimodal data fusion.

3. MATERIALS AND METHODS

The dataset is described in this section, techniques for preparing data, as well as the categorization model's machine learning algorithm. The section too includes a description of weka's statistical model performance evaluator, which can be used to analyze and compare the robustness and dependability of created models. It is shown in Figure 1.

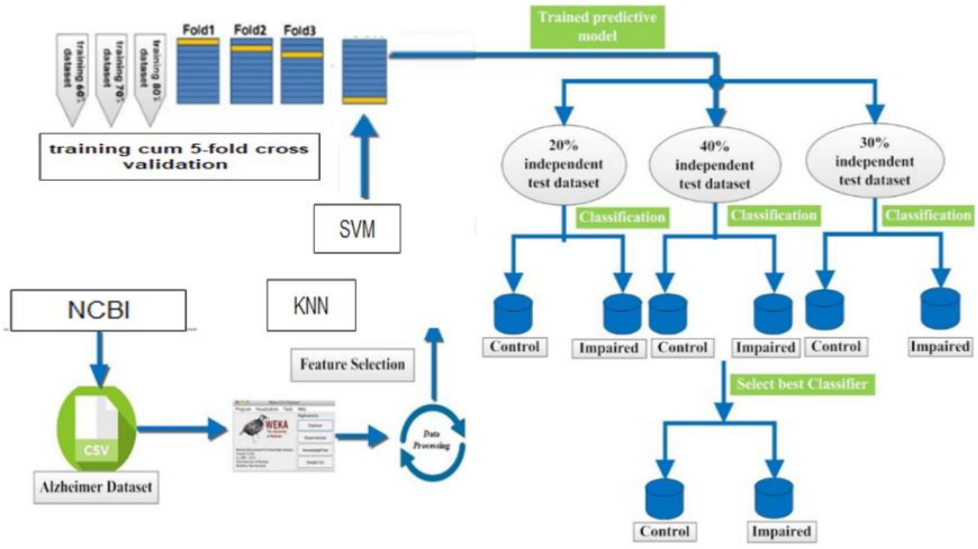


Figure 1. An instance of the steps involved in creating a classification model for predicting AD in its early stages

3.1. Dataset

There are numerous biological datasets available. The dataset for this paper is collected from the gene expression omnibus (GEO), which is a publicly available data source. The National Center for Bioinformatics Information (NCBI) released it on August 5, 2015. The dataset's accession numbers are (GSE63060 and GSE63061) as provided by the AddNeuroMed Cohort. To expand the amount of the samples, these two datasets were merged into one, the AD dataset. The amount of gene expression in the AD dataset was monitored using microarray technology. It has appropriate columns for (16382 genes) and rows for (569 samples). It is made up of (245 patients with AD, 142 MCIs, and 182 CTLs).

3.2. Feature cleaning/selection of features for reducing redundant data

Selection of features/dimensionality reduction is a common technique for improving model correctness or boosting performance on very large datasets. For the following reasons, the feature selection method must be used in this scenario:

- For some machine learning algorithms, inputting above 16,382 features may take too long to train.
- Models are simplified to make them easier to interpret for researchers and users.
- Reduced over fitting improves generalization (formally, reduction of variance).
- Many features in the data could be redundant (highly correlated, linearly dependent) or irrelevant. These features can be disabled without causing significant data loss.

The entropy (information gain) of a random collection of samples determines its impurity. The projected decrease in entropy owing to dividing the samples before splitting the feature node is known as information gain. It is a method of determining the connection between inputs (samples on X axis) and outputs (entropy on Y axis). As a result, the greater the knowledge gained the better as shown in Figure 2 [38].

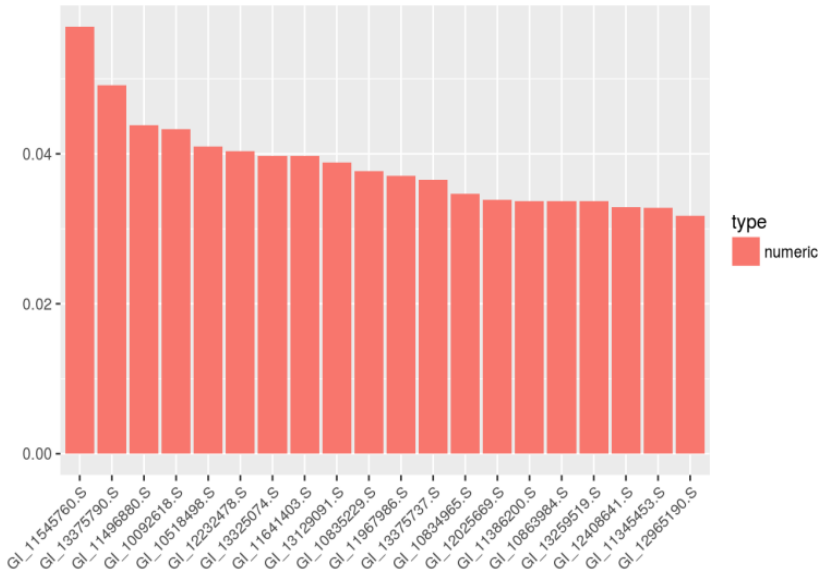


Figure 2. Information gain bar chart

3.3. Select features and combine

The number of significant features after using various feature cleaning/selection procedures is by using IG can be obtained from 44 features. This represents a significant reduction in dimensionality from over 16,382 features. These 44 characteristics are utilized to create a prediction model.

3.4. Machine learning algorithms for model building

Subjects are classified using a machine learning algorithm based on related properties. In this research, three well-known ML algorithms, specifically Naive Bayes (NB), support vector machine (SVM), and K-NN, were used to categorize people into impaired and healthy control groups based on selected attributes. The prediction ability of each model was assessed and compared using several statistical methods. The following is a quick rundown of the machine learning approaches used to create a classification model that can tell the difference between MCI sufferers and healthy people in the current study:

3.4.1. Naive Bayes (NB)

The Naïve Bayes (NB) approach is established on the premise of each of the training dataset's predictive features (X1, X2... Xn) is conditionally independent. The Bayes theorem is used by the NB method to classify attributes in the test dataset. It calculates the probability of an attribute being categorized in any of the provided classes in the past. Prior experience determines an attribute's prior probability, according to the Bayes rule; as a result, the participants in the test case are divided into classes based on several qualities' conditional probabilities. Second, the percentage of subjects in any of the classes with similar traits determines the possibility of a topic being classified in one of the classes. In terms of NB analysis, in a dataset, a subject's final classification is established by multiplying prior and probability information about an attribute to generate a posterior possibility. A subject is assigned to a group if he or she has a higher possibility of having traits in that class [36]. The following is a description of the NB algorithm: Suppose that the likelihood of a person "X" with certain characteristics is X = < x1...,xn> belonging to the impaired class, which is denoted by the letter "*h*" and is represented as follows:

$$P(h1/xi) = \frac{(P(xi/h1)\,P(h1)}{(P(xi/h1)\,P(h1) + (P(xi/h2)\,P(h2)} \quad (1)$$

P (h1) is the previously related probability with class *h1*, and *P (h1|xi)* is the posterior probability. As a result, have "n" different hypotheses as shown in (2).

$$P(h1/xi) = \frac{(P(xi/h1)\,P(h1)}{(P(xi))} \quad (2)$$

As a result:

$$P(xi) = \sum_{j=1}^{n} P(xi/hj)\,P(hj) \quad (3)$$

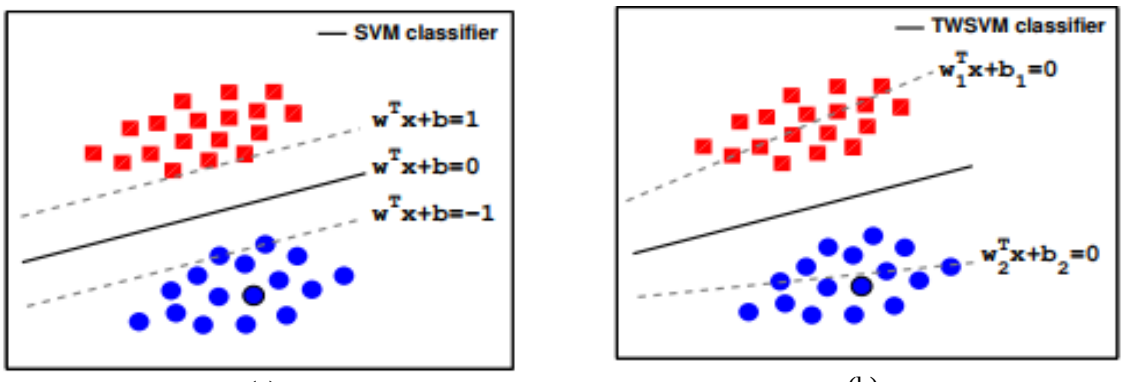
3.4.2. Support vector machine (SVM)

SVM is a supervised machine learning technique that can tackle issues in classification and regression. Support vector (frontier) is simply the coordinates of individual observation in the hyper-plane which best segregates or differentiates the two classes [37]. The decision function of SVM is stated as, after solving a convex optimization problem.

$$f\,(x) = sin\,(wT\,x + b) \quad (4)$$

Where w is the weight vector and b are biased.

Pros. with high dimensional spatial data, one of the most efficient and effective supervised machine learning algorithms. Clearer margin of separation between the support vectors results in better prediction. It's especially useful when the number of dimensions exceeds the number of samples [38] which is the perfect machine learning algorithm for the gene expression dataset (16,382 dimensions, 569 samples). In addition, it is a very memory efficient algorithm [39]. Cons. computation time is usually longer than a normal machine learning algorithm. It performs with noise data such as data with lots of highly correlated features [40]. Show in Figure 3.



(a) (b)

Figure 3. SVM and TWSVM classifiers are shown on a graph [39]

3.4.3. K-nearest neighbors (K-NN)

The K-NN approach is a straightforward data mining tool that may be used to solve both classification and regression problems. Based on the majority classes of its K neighbors, the K-NN classification algorithm gives an object to a certain class. The number of neighbors to be considered for polling is defined by the value of K, which is a positive integer. The value of K in this analysis is 11, which was chosen using the trial and error method [40]. KNN classifier shown in Figure 4.

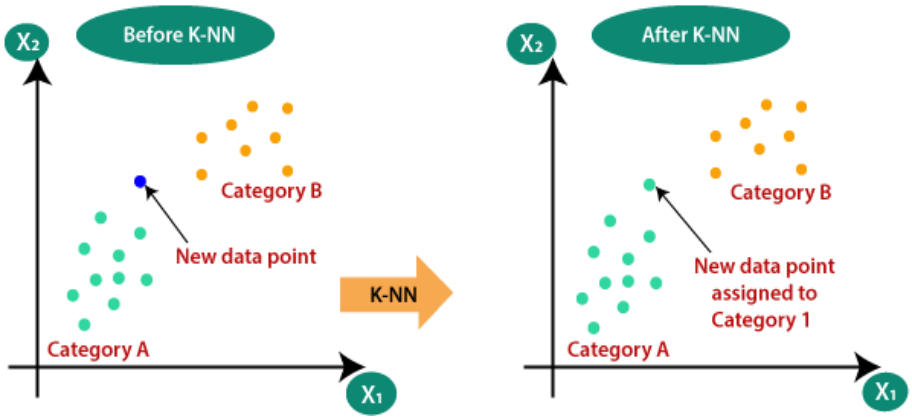


Figure 4. KNN classifier [41]

4. RESULT ANALYSIS /DATA PREPROCESSING

To properly feed data to each machine learning algorithm, data preprocessing is essential. Furthermore, the quality of the data has a significant impact on the performance of the machine learning algorithm [41]. As the gene expression dataset's format isn't the greatest for fitting into the algorithms, data transformation has been done. This gene expression dataset is now formatted as row (gene expression) x column (individual), but it should be formatted as row (individual) x column (individual) (gene expression). The classification performance of the classifiers during the validation and testing period is shown in Table 1.

Table 1. Performances of three proposed models

Classifier	Accuracy
K-NN	89%
Naive Bayes	85%
SVM	96%

Evaluation of performance: Over-fitting occurs when a model's parameters are discovered and tested on the same dataset, resulting in flawless accuracy when training with seen data but considerably erroneous results when training with unseen data. Cross-validation (CV) is used in this work to avoid over-fitting, while accuracy and area under the curve are used to quantify performance. Accuracy: each prediction has two values: one is the chance of having a sickness, which is '1', and the other is the probability of not having a disease, which is '0.' The likelihood of not having a sickness, which is '0,' is another value. The accuracy is based on a 0.1 threshold, which means that if the value has 0.1, it is a "1," else it is a "0." The cost of a false positive and false negative is assumed to be equal; hence the threshold is set at 0.1.

Area under the curve (AUC): AUC is a receiver operating characteristics (ROC) curve-based measuring of accuracy that shows the tradeoff between sensitivity and specificity. Sensitivity and specificity have an inverse relationship (increasing sensitivity results in decreasing specificity). The optimum and most accurate AUC curve spans the entire ROC space from the bottom right to the top left. The ROC space curve with a 45-degree diagonal has 50% predictive power, which is a randomly determined classification.

Cross validation (CV): To test the model, k-fold cross-validation (CV) divides the training set into k smaller sets called validation sets and utilizes the rest of the data set to train the model. It switches to the next smaller subsets to test the model, and so on, with each iteration. The average of each score is the result.

- Each model in this study is trained with k=5.
- To test the performance measure, which in this case is accuracy, a smaller set is employed.

5. DISCUSSION

All the classification models in this study were on the training dataset, trained with 5 folds cross-validation. To avoid over-fitting, the models' cross validation is used using the training dataset. The classifiers' performance is assessed using an unknown test dataset. Using the IG algorithm also chose the most discriminative traits that would most help in the diagnosis of early-stage AD. To identify the most significant features, use the SVM classifier on the data as shown in Figure 5.

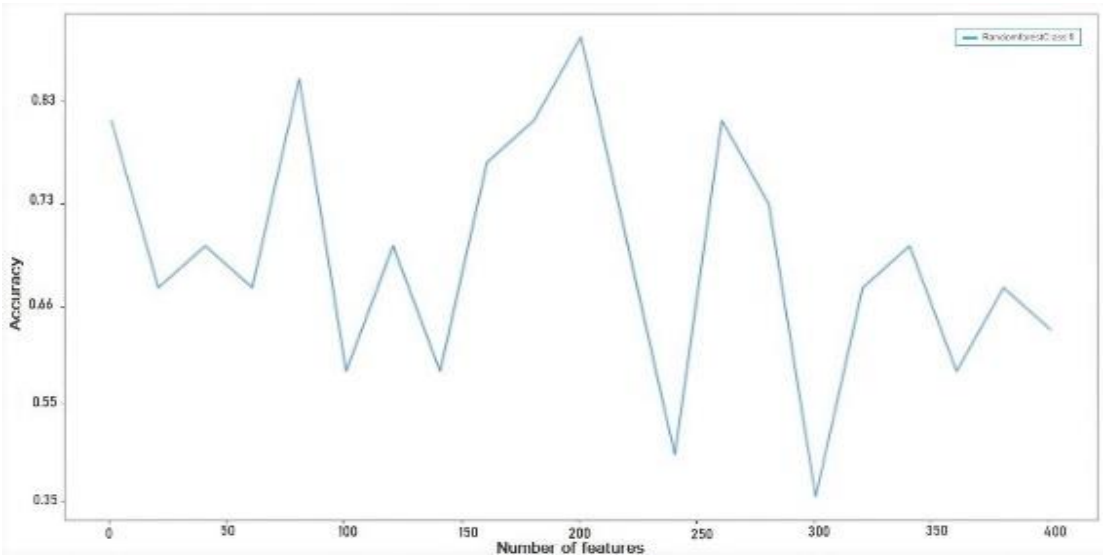


Figure 5. Applying SVM classifier to select features

The accuracy of the results varies depending on the number of features employed. With 44 features, it had the best accuracy of 96.6. By lowering the threshold term (i.e., set to less than 125 mean) in this approach of feature selection, the mean cross-validation accuracy improves as the number of features increases.

6. CONCLUSION

We suggested an approach to improve AD diagnosis prediction by incorporating a selection of features for the predictor. We introduce three methods of machine learning classifiers (SVM, K-NN and Naive Bayes) with method of feature selection and show the results which are the correctly classified and incorrectly classified and can identify the most appropriate features for SVM model training, 5-fold cross-validation was used in this study, and low variance of prediction error was attained, demonstrating the robustness of our method. In medicine and healthcare studies, machine learning and data mining techniques are particularly useful for the early identification and diagnosis of a variety of disorders. The biggest advantage of our method over previous feature selection models is that the training system has automatically, for a better prediction, replenished the required features. The best classification methods as we saw above were in SVM with feature selection information gain (96.6) accuracy. However, past research has linked several of the qualities chosen in our technique to Alzheimer's disease or other psychological disorders, demonstrating the model's efficacy. Furthermore, the findings demonstrated that machine learning and data mining approaches can be utilized to accurately detect, predict, and diagnose a variety of diseases. Increase the number of examples for AD and NC classes so that the model may be trained with enough and balanced data for all classes to increase the accuracy of the AD stages categorization.

REFERENCE

[1] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Syst. Appl.*, vol. 140, p. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.
[2] K. Nishiwaki, K. Kanamori, and H. Ohwada, "Gene selection from microarray data for Alzheimer's disease using random forest," *Int. J. Softw. Sci. Comput. Intell.*, vol. 9, no. 2, pp. 14–30, 2017, doi: 10.4018/ijssci.2017040102.
[3] N. Jameel and H. S. Abdullah, "Intelligent feature selection methods: a survey," *Eng. Technol. J.*, vol. 39, no. 1B, pp. 175–183, Mar. 2021, doi: 10.30684/etj.v39i1b.1623.

[4] R. A. Saputra, C. Agustina, D. Puspitasari, R. Ramanda, D. Pribadi, and K. Indriani, "Detecting Alzheimer's disease by the decision tree methods based on particle swarm optimization," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012025.

[5] O. D. Madeeh and H. S. Abdullah, "An efficient prediction model based on machine learning techniques for prediction of the stock market," in *Journal of Physics: Conference Series*, Mar. 2021, vol. 1804, no. 1, doi: 10.1088/1742-6596/1804/1/012008.

[6] E. Bonilla Huerta, B. Duval, and J. K. Hao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," *Neurocomputing*, vol. 73, no. 13–15, pp. 2375–2383, 2010, doi: 10.1016/j.neucom.2010.03.024.

[7] W. M. S. Abedi, I. Nadher, and A. T. Sadiq, "Modified deep learning method for body postures recognition," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 2, pp. 3830–3841, 2020.

[8] S. T. Ahmed, Q. K. Kadhim, H. S. Mahdi, and W. S. A. Almahdy, "Applying the MCMSI for online educational systems using the two-factor authentication," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 13, pp. 162–171, 2021, doi: 10.3991/ijim.v15i13.23227.

[9] A. R. Abbas and A. O. Farooq, "Skin detection using improved ID3 algorithm," *Iraqi J. Sci.*, vol. 60, no. 2, pp. 402–410, Feb. 2019, doi: 10.24996/ijs.2019.60.2.20.

[10] A. T. Sadiq and K. S. Mohsen, "Modify random forest algorithm using hybrid feature selection method," *International Journal on Perceptive and Cognitive Computing*, vol. 4, no. 2, pp. 1-6, 2018.

[11] I. Abed, "Lung cancer detection from X-ray images by combined backpropagation neural network and PCA," *Eng. Technol. J.*, vol. 37, no. 5A, pp. 166–171, May 2019, doi: 10.30684/etj.37.5a.3.

[12] A. T. Sadiq and S. A. Chawishly, "Intelligent methods to solve null values problem in databases," *J. Adv. Comput. Sci. Technol. Res.*, no. December, 2016.

[13] A. T. Sadiq and N. H. Shukr, "Classification of cardiac arrhythmia using ID3 classifier based on wavelet transform," *Iraqi J. Sci.*, vol. 54, no. 4, pp. 1167–1175, 2013.

[14] S. Gauthier *et al.*, "Management of behavioral problems in Alzheimer's disease," *Int. Psychogeriatrics*, vol. 22, no. 3, pp. 346–372, 2010, doi: 10.1017/S1041610209991505.

[15] Ei, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, 2018, doi: 10.3390/molecules23123140.

[16] M. S. Croock, S. D. Khudhur, and A. K. Taqi, "Edge Detection and Features Extraction for Dental X-Ray," *Eng. Tech.Journal*, vol. 34, no. 13, pp. 2420–2432.

[17] J. H. Assi and A. T. Sadiq, "NSL-KDD dataset classification using five classification methods and three feature selection strategies," *J. Adv. Comput. Sci. Technol. Res.*, vol. 7, no. 1, pp. 15–28, 2017.

[18] C. Jian *et al.*, "Microglia Mediate the Occurrence and Development of Alzheimer's Disease Through Ligand-Receptor Axis Communication," *Front. Aging Neurosci.*, vol. 13, no. September, pp. 1–11, 2021, doi: 10.3389/fnagi.2021.731180.

[19] C. Park, Y. Yoon, O. Min, S. J. Yu, and J. Ahn, "Systematic identification of differential gene network to elucidate Alzheimer's disease," *Expert Syst. Appl.*, vol. 85, pp. 249–260, 2017, doi: 10.1016/j.eswa.2017.05.042.

[20] S. T. Ahmed and S. M. Kadhem, "Early Alzheimer ' s Disease Detection Using Different Techniques Based on Microarray Data : A Review," *iJOE – Vol. 18, No. 04, 2022*, no. Mci.

[21] A. R. T. Silva *et al.*, "Transcriptional Alterations Related to Neuropathology and Clinical Manifestation of Alzheimer's Disease," *PLoS One*, vol. 7, no. 11, 2012, doi: 10.1371/journal.pone.0048751.

[22] S. T. Ahmed and S. M. Kadhem, "Using machine learning via deep learning algorithms to diagnose the lung disease based on chest imaging: a survey," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 16, p. 95, 2021, doi: 10.3991/ijim.v15i16.24191.

[23] L. Mesrob *et al.*, "Identification of atrophy patterns in Alzheimer's disease based on SVM feature selection and anatomical parcellation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5128 LNCS, pp. 124–132, 2008, doi: 10.1007/978-3-540-79982-5_14.

[24] S. Z. A.-R. H. M. Ali, "Deep learning approach for microarray alzheimer's data classification," *Test Eng. Manag.*, vol. 83, no. 0193-4120 Page No. 2016–2029, pp. 2016–2029, 2020.

[25] G. Meng, X. Zhong, and H. Mei, "A systematic investigation into Aging Related Genes in Brain and Their Relationship with Alzheimer's Disease," *PLoS One*, vol. 11, no. 3, pp. 1–17, 2016, doi: 10.1371/journal.pone.0150624.

[26] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen, "Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC Bioinformatics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2105-13-266.

[27] X. Huang *et al.*, "Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning," *BMC Neurol.*, vol. 18, no. 1, pp. 1–8, 2018, doi: 10.1186/s12883-017-1010-3.

[28] C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Early Detection of Alzheimer's Disease with Blood Plasma Proteins Using Support Vector Machines," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 1, pp. 218–226, 2021, doi: 10.1109/JBHI.2020.2984355.

[29] K. P. M. Niyas and P. Thiyagarajan, "Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2020.12.009.

[30] H. M. Al-Bermany and S. Z. Al-Rashid, "Microarray gene expression data for detection alzheimer's disease using k-means and deep learning," in *2021 7th International Engineering Conference "Research and Innovation amid Global Pandemic" (IEC)*, Feb. 2021, pp. 13–19, doi: 10.1109/IEC52205.2021.9476128.

[31] D. Jha and G.-R. Kwon, "Alzheimer disease detection in MRI using curvelet transform with K-NN," *J. Korean Inst. Inf. Technol.*, vol. 14, no. 8, p. 121, 2016, doi: 10.14801/jkiit.2016.14.8.121.

[32] M. Tanveer *et al.*, "Machine learning techniques for the diagnosis of alzheimer's disease: A review," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 16, no. 1s, 2020, doi: 10.1145/3344998.

[33] N. Voyle *et al.*, "A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis," *J. Alzheimer's Dis.*, vol. 49, no. 3, pp. 659–669, 2015, doi: 10.3233/JAD-150440.

[34] C. Plant *et al.*, "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease," *Neuroimage*, vol. 50, no. 1, pp. 162–174, 2010, doi: 10.1016/j.neuroimage.2009.11.046.

[35] R. Ocampo-Vega, G. Sanchez-Ante, M. A. De Luna, R. Vega, L. E. Falcón-Morales, and H. Sossa, "Improving pattern classification of DNA microarray data by using PCA and logistic regression," *Intell. Data Anal.*, vol. 20, no. s1, pp. S53–S67, 2016, doi: 10.3233/IDA-160845.

[36] H. Li *et al.*, "Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017, doi: 10.1038/s41598-017-13700-w.

[37] G. Ma *et al.*, "Differential expression of mRNAs in the brain tissues of patients with Alzheimer's disease based on GEO expression profile and its clinical significance," *Biomed Res. Int.*, vol. 2019, pp. 1–10, 2019, doi: 10.1155/2019/8179145.

[38] M. Barati and M. Ebrahimi, "Identification of Genes Involved in the Early Stages of Alzheimer Disease Using a Neural Network Algorithm," *Gene, Cell Tissue*, vol. 3, no. 3, 2016, doi: 10.17795/gct-38415.
[39] S. Mahajan, G. Bangar, and N. Kulkarni, "Machine Learning Algorithms for Classification of Various Stages of Alzheimer ' s Disease: A review," *Int. Res. J. Eng. Technol.*, vol. 07, no. 08, pp. 817–824, 2020.
[40] N. A. Al-Thanoon, O. S. Qasim, and Z. Y. Algamal, "Improving nature-inspired algorithms for feature selection," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2021, doi: 10.1007/s12652-021-03136-6.
[41] A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," *J. Basic Appl. Sci.*, vol. 13, pp. 459–465, 2017, doi: 10.6000/1927-5129.2017.13.76.

BIOGRAPHIES OF AUTHORS

Shaymaa Taha Ahmed M.Sc. (2015) in (India), currently a postgraduate student studying PhD in Department of Computer Sciences, University of Technology, Baghdad, Iraq. Affiliation: University of Diayla Dept.: computer science/College: basic of education specialization: Computer science/information system. Research Interests: Cloud Computing-Deep Learning-Machine learning-AI–Data mining. She can be contacted at email: Shaymaa.taha.ahmed@basicedu.uodiyala.edu.iq or mrs.sh.ta.ah@gmail.com.

Suhad Malallah Kadhem PhD in Computer Sciences/2003/Computer Science Department/ University of Technology. Scientific Specialization: Natural Language Processing. Scientific Title: Assistant Professor/2012. Scientific Research Interest: Natural Language Processing, Machine translation, Artificial intelligence, Information Security, Information Hiding, machine learning, deep learning and. She can be contacted at email: 110102@uotechnology.edu.iq.