Alzheimer's Disease Gene Prediction based on Hyprid SVM-based Classifcation Methods

Hala AlShamlan
Information Technology Department
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

Rehab AlJurayyad
Information Technology Department
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

Samar F. Omar
Information Technology Department
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

Abstract
Alzheimer's disease (AD) is a progressive neurodegenerative brain disorder with complex genetic architecture. This disease is the focus point of many bioinformatics kinds of research, where the key goal of these researches is to classify the genes involved in the processes of Alzheimer's and to explore the function of these risk genes in the progress of the disease. For this purpose, we here seek out the best model to detect the genes related to AD with the usage of several feature selection methods. In this study, we compared the efficiency of the feature selection methods with SVM classifier including mRMR, CFS, Chi-Square Test, F-score, and GA was compared. The accuracy of SVM classifier has been calculated with validation methods such as 10-fold cross-validation. We applied these methods to the public AD gene expression dataset consist of 696 samples and 200 gene. The results show that mRMR and F-score obtain high accuracy around 84% with number of genes between 40 to 80.

Keywords
Data Mining, Genetic Disease Prediction, Alzheimer Disease. Gene Expression

1. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia that accounting for up to 80 percent of cases of dementia. It is a progressive and irreversible loss of neurons and brain functions which often affects elderly people also has currently no cure for it [1] [2]. In the early stages of the disease, the memory loss is mild, but with the late-stage of Alzheimer's, the patient loses the ability to even carry on a conversation. The global impact of dementia According to a recent report [3], In 2018 there are around 50 million people worldwide who have been diagnosed with dementia. As well as, the annual number of new cases of AD and other kinds of dementia is projected to get triple by 2050 that will reach 152 million cases, where there will be one new case of dementia every 3 seconds[3].

While the researchers believe that there is not a single cause of Alzheimer's disease. There are multiple factors that may contribute to the development of Alzheimer's disease, such as genetics, age, lifestyle, and environment[4]. In fact, up to 5% of Alzheimer's disease cases are due to genetic inheritance, where genes being responsible of the appearing of an early onset AD [5]. In these cases, mutations in different genes such as APP, APOE, PSEN1 and PSEN2 have been connected to the disease between 1987 - 1993 [4][5].

Early and accurate diagnosis of AD plays an important role in prevention, treatment, and patient care, especially in the early stage. Therefore, the diagnosis of AD has been mostly performed by analyzing brain images such as: MRI, and PET. Nevertheless, it is important to know which genes are usually caused by AD and altered in its patients, especially in its early stage, in order to control of AD progression. A recent study in Bioinformatics and Biomedicine fields focused on applying different Statistical, machine learning, data mining techniques to find useful data patterns and discover to detected several disease[6]. This study aimed at developing an intelligent data mining model for the prediction of AD based on gene expression.

The rest of the paper is organized as follows. Section II presents a brief background of AD, machine learning techniques and Features selection methods. Section III introduces the work that has been done in the field of Genetic Disease Prediction. Section IV we describe the dataset we have used in this study and the proposed model architecture. Section V presents our experiment stages and the results we obtained. In section VI. We present future work we will going to do and conclude the output of this study.

2. MATERIALS AND METHODS:

Background

A. Alzheimer's Disease

"People don't see it as a disease. They think it is simply part of getting old. Awareness is low and stigma is high" - Paola Barbarino - CEO of Alzheimer's Disease International (ADI)

Manuscript received November 5, 2025
Manuscript revised November 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.11.10

Alzheimer's Disease (AD) is an irreversible, progressive brain disorder. It was first described in 1906 by Dr. Alois Alzheimer. Dr. Alzheimer diagnosed the symptoms that his patient experienced that includes memory loss, paranoia, and psychological changes. Dr. Alzheimer noticed in the autopsy that there was shrinkage of the patient's brain [7]. The disease is the most common cause of dementia. Dementia slowly destroys memory and cognitive functioning - thinking, remembering, and reasoning -, and, ultimately, the ability to carry out regular daily life [8].

Nowadays, AD is ranked the sixth leading cause of death around the world [8]. In most people with Alzheimer's, symptoms first appear in their mid-60s. Estimates vary, but experts suggest that 3.23% of the population in Saudi Arabia, most of them age 65 or older, may have dementia caused by Alzheimer's [9].

AD is complex, and up to now there is no medication or other intervention can successfully treat it. Present approaches concentration on helping people maintain mental function, manage behavioral symptoms, and slow down certain problems, such as memory loss. Researchers are looking to develop therapies targeting specific genetic, molecular, and cellular mechanisms so that the actual underlying cause of the disease can be stopped or prevented [8].

B. Supervised Machine Learning

Supervised Machine Learning (SML) is a machine programmed to find particular patterns in massive data. SML has different ways to adjust this data by adjusting the algorithm to make predictions and many other tasks [10]. This term directly related to the fields of programming, IT, and math. It applied in all kinds of sectors of government, marketing, medicine, and any business which collects data and wants to make a decision based on this data. Subsequently, it employed in consumer choices, weather forecasting, and website calculations. In this paper, we concentrated on some types of SML [10]. Although there are many, many categories and aspects to SML, we would only generally describe the following: Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, K-nearest neighbor, Decision Tree and Naive Bayes.

1) Support Vector Machine (SVM)

SVM is a discriminative classifier formally defined by a separating hyperplane. The algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space, this hyperplane is a line dividing a plane into two parts. Each class is separated on each side of the place [10]. Hyperplane is a line that linearly separates and classifies a set of data. Generally, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. Hence, when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it [10].

Given the solutions $\hat{\beta}_0$ and $\hat{\beta}$, the decision function can be written as

$$\hat{G}(x) = \text{sign}\left[\hat{f}(x)\right] = \text{sign}\left[x^T\beta + \hat{\beta}_0\right] \quad (1)$$

One aspect of SVM is their accuracy. SVM works well on smaller cleaner datasets. It can be more efficient because it uses a subset of training points. The cons are that it isn't suited to larger datasets as the training time with SVM can be high [10].

C. Feature Selection

Features selection is a process of removing irrelevant features in the dataset, where the chosen algorithm automatically selects those features that contribute most to the prediction variable or output in which you are interested. Having features selection before fitting the data into the classifier can enhance the accuracy by reducing training time and overfitting.

1) mRMR

It stands for minimum Redundancy Maximum Relevance. mRMR aims to selects the genes that shown low correlation among the other genes (Minimum Redundancy) but still having a high correlation to the classification variable (Maximum Relevance). For classes c=(ci,....ck) the maximum relevance condition is to maximize the total relevance of all features in

$$max_{S\subset\Omega}\frac{1}{|S|}\sum_{i\in S} I(c, f_i)$$

The Minimum redundancy condition is

$$min_{S\subset\Omega}\frac{1}{|S|^2}\sum_{i,j\in S} I(f_i, f_j)$$

2) CFS

CFS stands for Correlation-based Feature Selection algorithm. CFS selects attributes by using a heuristic which measures the usefulness of individual genes for predicting the class label along with the level of inter-correlation among them. The highly correlated and irrelevant features are avoided. The method calculates the merit of a subset of k features as:

$$Merit_{S_k} = \frac{kr_{cf}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Here, $\overline{r_{cf}}$ is the avarage value of all feature-classification correlations, and $\overline{r_{ff}}$ is the avarage value of all feature-feature correlation. The CFS criterion is defined as follows:

$$CFS = max_{S_k}\left[\frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2\left(r_{f1f2} + \cdots + r_{fifj} + \cdots + r_{fkf1}\right)}}\right]$$

3) Chi-Square Test

Chi-Square Test is a statistical algorithm, is used by classification methods to check the correlation between two variables. In the following equation, High scores on $\chi 2$ indicate that the null hypothesis (H0) of independence should be eliminated and thus that the occurrence of the term and class are dependent:

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

4) F-score

F-score is a simple statistical algorithm for feature selection. F-score can be used to measure the discrimination of two sets of real-number.

5) GA

Genetic Algorithm (GA) is one of the common wrapper gene selection methods. It is usually applied to discrete optimization problems. The main goal of GA is discovering the best and perfect solution within a group of potential solutions. This method reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation. Each set of solutions is named as population. Populations consist of vectors, i.e., chromosomes or individuals. Every item in the vector is referred to as the gene[11]

3. RELATED WORKS

This section will investigate different approaches used to detect AD. As well as highight the derived benefits from previous work that used machine learning algorithms with genetic. Fron their experiment 18 AD datasets have been invastigated, In Table 1 we have listed all the datasets. All the works have been summaries in Table 2.

Paylakhi et al., [12] proposed an ADG identification approach a novel hybrid method to identify relevant genes of AD. The proposed approach consists of 3 stages. In the first stage, they applied a Fisher criterion method as a filtering method to reduce the noise and eliminate redundant genes from the dataset. Then SAM technique is applied to identify genes with notable changes in their expression in order to increase the number of relevant genes. In the final stage, they used Genetic Algorithm GA for optimization of the feature selection and Support Vector Machine SVM for classification. The proposed approach was tested on the AD GSE1297 dataset consists of 31 samples provided by the Gene Expression Omnibus (GEO) database. The result shows that the proposed method obtains 94.55% Accuracy with 44 genes identified. Analysis of the 44 identified genes by GO and KEGG led to the identification of AD-related terms and pathways.

Voyle et al., [13] investigated the robustness of Pathway Level Analysis of Gene Expression (PLAGE) across gene-expression samples from different population. The study used a combined dataset which contains 748 subjects of peripheral-blood gene expression and obtained from Add Neuro Med (ANM) and Dementia Case Registry (DCR) cohorts. A Random Forest (FR) models were used with recursive feature elimination in particular to differentiation of patients with AD or mild cognitive impairment (MCI) from healthy elderly controls (CTL). The authors state that, no observable difference in the performance metrics of gene expression and PLAGE model.

Yu Miao et al., [14] proposed an ADG identification algorithm that combines feature selection, cascading classifier, and majority voting. As well as multiple classifier integration such as support vector machine SVM, random forest RF and, extreme learning machine ELM. In addition, a feature selection algorithm called ReliefF is applied to select the most relevant attributes and improve accuracy. The proposed method was tested on the AD dataset of 22,283 genes of 31 patients provided by The National Center for Biotechnology Information (NCBI). The result shows that the proposed method obtains 78.77% Sensitivity, 83.1 % Specificity and, 74.67% Accuracy. Also, based on the proposed method they list the top 13 genes predicted ADGs with a probability higher than 85%.

Another related work is found in the study elaborated by Martínez-Ballesteros et al., [15]. Their work was focused on analyzing the gene expression profiles related to AD using three integrated machine learning techniques (Decision Trees, Quantitative Rules, and Hierarchical Cluster). The purpose was to identify genes highly related to AD, through changes in their expression levels between control and AD samples. To fulfill this purpose, sixth phases were applied in a dataset consists of 33 samples and 1663 genes provided by Dunckley et al [16]. In the first three phases, C4.5 algorithm with minimum threshold the accuracy and GarNet algorithm were applied to establish the best configurations settings for the model. While in the fourth and fifth phases gene groups were validated using hierarchical cluster analysis, including to statistical tests and biological knowledge integration process based on the information fusion obtained from prior phase. Finally, they validated the results using additional data and a permutation test in the sixth phase. Their proposed method successfully characterized more than 90 genes whose expression is modified during AD progression.

On the other hand, Park et al., [17] concentrated on construct the AD disease-specific genes networks and disease mechanism using SML techniques. The study presents a novel approach for gene-gene interaction (GGI) identification, where a Random Forest algorithm was applied over heterogeneous gene expression profiles to determine the significant GGIs. They obtained a dataset consist of 257 normal gene and 439 AD gene by selecting two associated gene-expression profile (GSE33000 and GSE44770). To define their features from an expression profile, they extracted 22 features some of them are statistical

measurements of gene expression and two correlation-based similarity measures, PCC and MI. And to assign the label for the gene pairs various interactome datasets and gene sets were utilized. To evaluate their proposed method, they compared the classification performance among various ML algorithms which are Naïve Bayes, SVM, ANN and PART, and found that their proposed model outperformed all the algorithms.

In addition to that, Huang et al., [18] published one of Genome-Wide Association Studies (GWAS) on AD to identify disease-risk genes in the progress of disease and the complex networks of GGI. In their study a prediction algorithm for AD candidate genes was implemented based on SVM classifier. A dataset consists of 22,646 brain-specific gene expression network collected then distributed among four datasets (AD-associated set, Control dataset, AD-predicted dataset and Non-mental-health dataset). And to find the significant differences among all four datasets a set of features were extracted. Thereafter, uses the coefficients specific to predict the level of potential AD association for every gene in the complex networks. As a result, the proposed algorithm classified AD candidate genes with an accuracy of 94% and the area under the receiver operating characteristic (ROC) of 84.56%.

In 2019 Park et al., [19] published another study that proposed AD prediction model based on a deep neural network using a multi-omics dataset. The authors integrating two heterogeneous omics datasets large-scale gene expression and DNA methylation datasets. They obtained a large-scale gene expression dataset by integrated two gene profiles (GSE33000 and GSE44770) [20] [21] to increase the sample size. The integrated dataset was consisting of 257 normal control and 439 AD samples. In addition, DNA methylation dataset consists of 8 normal and 74 AD samples provided by Smith et al.,[22].In this study, the authors applied a feature selection method that consists of two steps to reduce the number of features. Firstly, they identify differentially expressed genes (DEG) and differentially methylated positions (DMP) and used the "Limma package" for analysis of both DEG and DMP. Second, they integrate DEGs and DMPs by intersecting. Also, they develop a prediction model based on a deep neural network (DNN) model and applied the Bayesian optimization on the proposed model to investigate the optimized hyperparameters of it. The result of the proposed prediction model shows that outperformed conventional machine learning algorithms such as Random Forest, SVM, and Naïve Bayesian. Where the highest accuracy of the conventional machine learning algorithm achieved was 0.7. for Random Forest while the average accuracy of the proposed prediction model was 0.823.

Table 1 Summary of related works on machine learning applications in Alzheimer's disease

Ref.	Year	Application	Methodology		Dataset	Perfomance
			Feature Selection	Classification		
[12]	2016	Identification of genes related to Alzheimer's disease with classification	– Fisher criterion – SAM – GA	SVM	D1	Accuracy = 94.55%
[13]	2016	Identification of genes related to Alzheimer's disease with classification	Recursive feature elimination	Random forest algorithm	D2,D3	Sensitivity = 61.0% Specificity = 70.3 % Accuracy = 65.7%
[14]	2017	Identification of genes related to Alzheimer's disease with classification	ReliefF	– SVM – RF – ELM	D1	Sensitivity = 78.77% Specificity = 83.1 % Accuracy = 74.67%
[15]	2017	Identification of genes related to Alzheimer's disease	– GarNet – C4.5 – QAR	– Decision Trees – Quantitative Rules – Hierarchical Cluster	D5	Best Accuracy = 89.03%
[18]	2018	Identification of gene-gene interactions of Alzheimer's disease	Sequence-based features of all genes	SVM	D7,D8, D9	Accuracy = 94% Receiver Operating Characteristic (ROC) = 84.56%
[17]	2018	Identification of gene-gene interactions of Alzheimer's disease	22 Features (statistical measurements of gene expression, correlation-based similarity measures)	Random forest algorithm	D5,D6, D7,D10	Best results. Accuracy = 91.6% Precision = 91.6% Recall = 91.6% F-Measure = 91.6% ROC area = 96.5%

[19]	2019	Alzheimer's disease AD prediction model	– differentially expressed genes (DEG) differentially methylated positions (DMP) – "Limma package" for analysis – integrate DEGs and DMPs by intersecting	Deep neural network DNN	D5,D6	Average Accuracy = 82%.

4. Materials and Methods

In this section, we presented various methods that we used to construct our proposed model which developed in Python. We started by elaborating on the gene expression dataset, and then we demonstrated our methods for gene selection and classification. Various type of filter gene selection method was built into the classifier to search for an optimal subset of genes. Subsequently, the classifier identifies the genes that highly related to AD. The below subsections explained more our wrapper model, and Figure 1 shows the overview of the model.

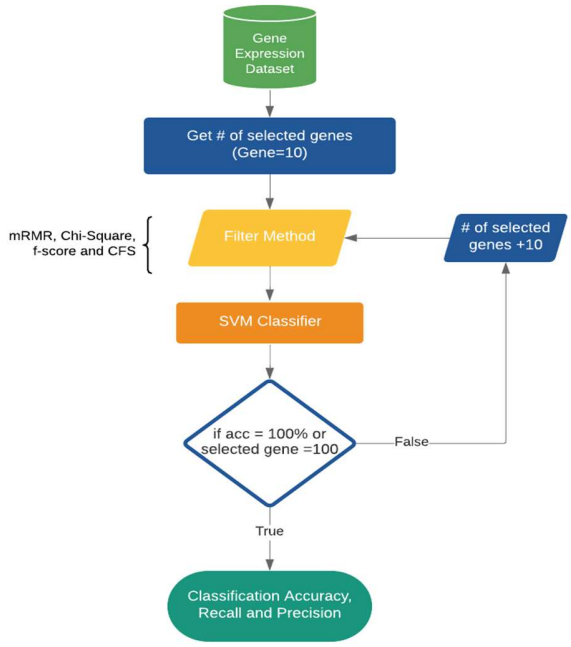

Figure 1 The proposed AD Classification Model

D. Datasets

In this study, we used a dataset published by Park et al., in [19]. The authors obtained their dataset from integrated two large-scale gene expression profiles GSE33000 and GSE44770 [20],[21]. The integrated dataset was composed of 19,488 genes, 257 Normal, and 439 AD samples. This dataset was normalized by the z-score [1]. Z-score is a transformation method that standardizes microarray datasets across a wide range of experiments. It is a useful method that facilitates data comparison and analysis. The final dataset after the normalization has consisted 696 samples with 200 genes, 257 were Normal samples while 439 were AD. The output format of the final dataset was tsv format where fields are separated by tab. We have changed the file format to .csv (comma separated value) since its more efficient for classification, machine learning and data analysis in Python.

E. The Proposed Model

We have built our model based on the wrapper generic selection methods as it evaluated all possible subset of the genes and select the subset that produces the best result for SVM classifier. We have started with calculating the number of samples and genes, and to measure the best performance of SVM we have selected n number of genes (n: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100). To fit the samples into SVM we have used cross validation performance metrics. Cross validation technique support our model to detect the over-fitting. Overfitting occurs when the model or the algorithm fits the data too well. Using cross validation, we have split our dataset into 10 folds. Then, we trained the classifier with these 10 folds of samples for n genes. Eventually, we have tested the predicted labels from our model by calculating the evaluation metrics and measure the performance using the following measurements.

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1907322/

Table 2 SVM Simulation in Alzheimer's disease

# of genes	SVM									
	mRMR			CFS	Chi-Square			F-score		
	Accuracy	Precision	Recall	Accuracy	Accuracy	Precision	Recall	Accuracy	Precision	Recall
10	0.79	0.80	0.88	0.78	**0.83**	0.83	**0.93**	0.82	0.84	0.89
20	0.80	0.82	0.88	0.78	0.82	0.82	0.92	**0.84**	0.85	**0.90**
30	0.83	0.84	0.89	0.78	0.82	0.83	0.91	0.83	0.85	**0.90**
40	**0.84**	**0.86**	**0.90**	0.78	0.82	0.84	0.90	**0.84**	0.85	**0.90**
50	**0.84**	0.85	**0.90**	0.78	0.82	0.84	0.89	**0.84**	**0.86**	0.89
60	**0.84**	0.85	**0.90**	0.78	0.81	0.83	0.89	**0.84**	**0.86**	**0.90**
70	**0.84**	**0.86**	0.89	0.78	0.82	**0.85**	0.88	**0.84**	**0.86**	0.89
80	**0.84**	0.85	**0.90**	0.78	0.81	0.84	0.87	**0.84**	**0.86**	0.89
90	0.82	0.85	0.88	0.78	0.81	0.84	0.86	0.83	**0.86**	0.87
100	0.83	**0.86**	0.88	0.78	0.81	**0.85**	0.85	0.83	**0.86**	0.87
All	0.81	0.85	0.84	0.78	0.78	0.84	0.81	0.81	0.85	0.85

Classification Accuracy

$$Accuracy = \frac{Correct\ Predictions}{Total\ Number\ of\ Examples}$$

Precision

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Recall

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive}$$

These steps have been repeated 11 times based on n to find the optimal subset of genes classification and at the 11th time it was applied on the entire number of genes to calculate the overall accuracy of our dataset.

5. Experiment and Results

In this section we will present our experiment. First, we validate the dataset and spilt into training and testing sets. Then, we evaluated SVM classifier using 10-fold cross validation. SVM obtain 0.79 % accuracy. After that we evaluated several feature selections models with SVM classifier. we presented our genetic classification result of the classifier with every filter used in our study. To analysis the measurements we obtained from several feature selection models, we have rounded all the values and then arranged it in Table 2. We can notice from the table that the performance of mRMR and F-score outperform all other filters, especially when the number of genes between 20 and 80. Even though both have the same results, the F-score gives better results than the mRMR. As the lowest accuracy, we obtained from mRMR was 0.70 while in F-score, it was 0.81. On the other hand, the accuracy of CFS gives the same results for all selected number of genes, and because of the long-running time, we couldn't get the precision and recall. In addition to that, we have trained the SVM directly as well with the entire dataset and found 0.79 accuracy, 0.86 precision, and 0.81 recall. In addition, we apply we apply GA as feature selection method with SVM and we get 81 %. These results demonstrated the significant of using the gene selection filter method as it increased the performance of the model and enhance the accuracy. The best number of selection genes was between 40 to 80 which give the highest accuracy 84 % with mRMR and F-score. Figure 2 shows a comparison between the accuracy across our model. The pilot lines in the figure emphasized the similarity of the performance between mRMR and F-score. Also, it observed that all the classifiers produced a high accuracy in the beginning while the line of the efficiency goes down when the number of selected genes increased. In addition, The result obtained by mRMR and F-score outperform deep learning model Deep neural network DNN proposed by Park et al., [19] which gained 82% accuracy on the same gene expression dataset.

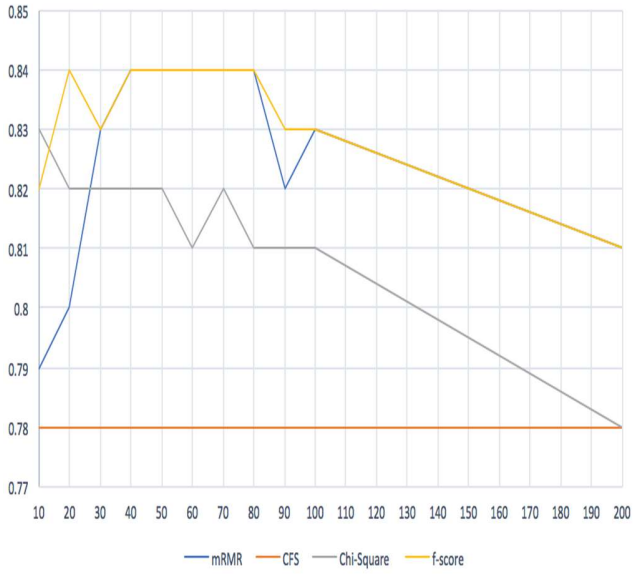

Figure 2 Four Feature Selection method comparison

6. Conclusion and Future Work

In this project, we are using different classifiers and feature selection methods for prediction Alzheimer's disease. We used the gene expression dataset modified by Park et al. [19]. The Results show a superiority of the mRMR and F-source where it provided very good results 84% accuracy with genes number 40-80. Where the results of the feature selection methods mRMR and F-score outperformed the GA, Chi-Square Test , CFS. In the future, we are going to further different datasets and compare the effect of the dataset on different models. As well as, we are going to use a hybrid approach instead of Filters or wrapper models.

References

[1] "What Is Dementia?," *Alzheimer's Disease and Dementia*. [Online]. Available: https://alz.org/alzheimers-dementia/what-is-dementia. [Accessed: 01-Sep-2019].

[2] "Dementia." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed: 05-Sep-2019].

[3] "World Alzheimer Report 2018 - The state of the art of dementia research: New frontiers," *NEW FRONTIERS*, p. 48.

[4] "Genetics," *Alzheimer's Disease and Dementia*. [Online]. Available: https://alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-risk-factors/genetics. [Accessed: 02-Oct-2019].

[5] C. R. Jack and D. M. Holtzman, "Biomarker Modeling of Alzheimer's Disease," *Neuron*, vol. 80, no. 6, pp. 1347–1358, Dec. 2013.

[6] "A review of microarray datasets and applied feature selection methods - ScienceDirect." [Online]. Available: https://www-sciencedirect-com.sdl.idm.oclc.org/science/article/pii/S0020025514006021. [Accessed: 02-Oct-2019].

[7] "History of Alzheimer's: Major Milestones." [Online]. Available: https://www.alzheimers.net/history-of-alzheimers/. [Accessed: 03-Oct-2019].

[8] "Seniors' Health - Overview of Alzheimer's." [Online]. Available: https://www.moh.gov.sa/en/HealthAwareness/EducationalContent/Health-of-Older-Persons/Pages/Overview-of-Alzheimer.aspx. [Accessed: 03-Oct-2019].

[9] "HomePage - الجمعية السعودية الخيرية لمرض الزهايمر." [Online]. Available: http://alz.org.sa/. [Accessed: 03-Oct-2019].

[10] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. New York, NY: Springer, 2009.

[11] \.Ismail Babaoglu, O. Findik, and E. ílker, "A Comparison of Feature Selection Models Utilizing Binary Particle Swarm Optimization and Genetic Algorithm in Determining Coronary Artery Disease Using Support Vector Machine," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3177–3183, Apr. 2010.

[12] S. Paylakhi, S. Z. Paylakhi, and S. Ozgoli, "Identification of Alzheimer disease-relevant genes using a novel hybrid method," *Progress in Biological Sciences*, vol. 6, no. 1, pp. 37–46, Jun. 2016.

[13] on behalf of the AddNeuroMed consortium *et al.*, "A Pathway Based Classification Method for Analyzing Gene Expression for Alzheimer's Disease Diagnosis," *JAD*, vol. 49, no. 3, pp. 659–669, Oct. 2015.

[14] Y. Miao, H. Jiang, H. Liu, and Y. Yao, "An Alzheimers disease related genes identification method based on multiple classifier integration," *Computer Methods and Programs in Biomedicine*, vol. 150, pp. 107–115, Oct. 2017.

[15] M. Martínez-Ballesteros, J. M. García-Heredia, I. A. Nepomuceno-Chamorro, and J. C. Riquelme-Santos, "Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources," *Information Fusion*, vol. 36, pp. 114–129, Jul. 2017.

[16] T. Dunckley *et al.*, "Gene expression correlates of neurofibrillary tangles in Alzheimer's disease," *Neurobiology of Aging*, vol. 27, no. 10, pp. 1359–1371, Oct. 2006.

[17] C. Park, J. Kim, J. Kim, and S. Park, "Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles," *PLoS ONE*, vol. 13, no. 7, p. e0201056, Jul. 2018.

[18] X. Huang *et al.*, "Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning," *BMC Neurol*, vol. 18, no. 1, p. 5, Dec. 2018.

[19] 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 Systems with Applications*, vol. 140, p. 112873, Feb. 2020.

[20] M. Narayanan *et al.*, "Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases," *Molecular Systems Biology*, vol. 10, no. 7, p. 743, Jul. 2014.

[21] B. Zhang *et al.*, "Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease," *Cell*, vol. 153, no. 3, pp. 707–720, Apr. 2013.

[22] R. G. Smith *et al.*, "Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology," *Alzheimer's & Dementia*, vol. 14, no. 12, pp. 1580–1588, Dec. 2018.

[23] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *PNAS*, vol. 101, no. 7, pp. 2173–2178, Feb. 2004.

[24] "AlzGene." [Online]. Available: http://www.alzgene.org/. [Accessed: 03-Oct-2019].

[25] "OMIM - Online Mendelian Inheritance in Man." [Online]. Available: https://www.omim.org/. [Accessed: 03-Oct-2019].

[26] "GIANT: Genome-scale Integrated Analysis of gene Networks in Tissues." [Online]. Available: http://giant.princeton.edu/. [Accessed: 04-Oct-2019].

[27] "IntAct molecular interaction database in 2012 | Nucleic Acids Research | Oxford Academic." [Online]. Available: https://academic.oup.com/nar/article/40/D1/D841/2903045. [Accessed: 19-Oct-2019].