2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic" (IEC) | 978-1-6654-1283-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/IEC52205.2021.9476128

Microarray Gene Expression Data for Detection Alzheimer's Disease Using k-means and Deep Learning

Heba M. AL-Bermany
Software Department
College of Information Technology, University of Babylon
Hilla, Iraq
heba.muthanna9@gmail.com

Sura Z. AL-Rashid
Software Department
College of Information Technology, University of Babylon
Hilla, Iraq
sura_os@itnet.uobabylon.edu.iq

***Abstract*— Microarray technology is a novel method to monitor the expression levels of an enormous number of genes simultaneous. These gene expressions are being used to detect various forms of diseases. The problem is not all genes are important; some genes can be redundant or irrelevant. These irrelevant genes add a computational workload to the prediction process. Therefore, this study aims at (1) identifying the most important genes that cause of Alzheimer's Disease (AD) by using feature (gene) selection to reduce the high-dimensional data size. Hence, a process for gene selection is twofold; removing the irrelevant genes and selecting the informative genes, and (2) predicting AD patients based on the selected subset of genes. In this paper, gene selection methods have been implemented, including Analysis of Variance (ANOVA) and Mutual Information (MI). In addition to, the k-means algorithm as a gene selection has been suggested. It is also presumed that the relevant genes have been existed in a same cluster, while the insignificant genes are really not belonging to the any cluster. The proposed system is applied on a high dimensional dataset namely AD dataset that contains 16382 genes. After picking the informative genes, prediction is performed with Convolutional Neural Network (CNN) that is commonly used in multiple prediction tasks. The proposed system performance was evaluated using the accuracy of the prediction. The results show that k-means clustering based gene selection can be performed to produce subset of key genes. The k-means algorithm with CNN model returns 0.929 accuracy based on genes subset from ANOVA method while k-means algorithm and CNN model achieve 0.886 accuracy based on genes subset from MI method. Thus, Genes subset selected is achieved a better accuracy at prediction and a little time of processing.**

Keywords*— *alzheimer's disease, microarray technology, gene expression, gene selection, clustering, predication.

I. INTRODUCTION

AD is a disease that described as degenerative one which leads to decline progressively for the memory and cognition. This causes a damage to the nervous cells inside the brain that associated with language and memory. After 65 years, the symptoms begin to appear and with age, the prevalence is growing sharply [1]. During 2050, many people who have AD is estimated to rise in United States (US) alone from 5.4 to 11 million. Despite these alarming numbers, there is no successful method to detect disease before symptomatic is appeared, which might be the only phase in the disease's progression where we could interfere [2]. Depending on the importance of the AD and failure to own a specific cure for that kind of diseases, a microarray technology has been applied to define the genes that causes the disease. Microarray technologies are significant tools at medical that the biologist used for monitoring gene expression levels in a specific organism [3]. However, the gene expression data produced from the technology of microarray can cause problems with the methods of prediction because the genes number in the microarray data is huge. In data mining, this truth is defined as the dimensionality curse. One of the efficient techniques to address the dimensionality curse is the gene selection, which selects relevant and informative genes. In fact, gene selection is the process by which a subset of informative genes is identified from the original dataset. This genes subset helps researchers to gain significant insight a biological structure for the disease [4]. In this paper, gene selection based on ANOVA and MI method were used for selecting the relevant genes. k-means algorithm also used as a gene selection for producing a minimal dataset consisting only of key genes [5]. Then, CNN model is employed to predict AD based on gene expression data and improved predictive performance. This study combines the gene selection methods with the prediction model to achieve the best accuracy.

The remaining sections of the presented study are arranged as follows: Section II demonstrates the related work. The microarray technology is showed in section III. The gene expression matrix is described in section IV. in our study, the materials and methods are elaborated in section V. The proposed methodology is described extensively in section VI. The result and discussion are described in section VII. The last section is presented the conclusion.

II. RELATED WORK

Chihyun Park, Jihwan Ha, and Sanghyun Park, (2020), used a prediction model known as deep learning algorithm. The proposed system has the ability to predict AD by using data of gene expression and data of DNA methylation. The result showed that applying deep learning can yield the best results in prediction model compared to traditional machine learning algorithms [6]. Karthik Sekaran, and Sudha. M, (2019), used Rhinoceros Search Algorithm (RSA) just as a feature selection. They used Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Multi-layered Perceptron Neural Network (MLP-NN) for prediction. They utilized (GEO: GSE1297) gene expression dataset and the experiments showed that RSA-MLP-NN model was more accurate in defining the distinguish between AD and genes that are normal to achieve the effectiveness [7]. Devi Arockia Vanitha C., Devaraj D., and Venkatesulu M., (2015), apply Mutual Information (MI) to identify the most relevant genes and Support Vector Machine (SVM) as a classifier. The experiments are implemented on two cancer datasets: Colon and Lymphoma. The results presented a proposed system

which decreases the dimensions of data by selecting the relevant genes and enhancing the classification accuracy [8]. Padideh Danaee, and Reza Ghaeini, (2017), used Principal Component Analysis (PCA) to extract efficient genes from gene expression data. Artificial Neural Network (ANN) has been used for the classification model. The results of the suggested approach showed the relevant genes can be helpful to detect the diseases [9]. Diyar Qader Zeebaree, Habibollah Haron, and Adnan Mohsin Abdulazeez, (2018), proposed Convolutional Neural Network (CNN) for classifying the microarray dataset. The experiments are implemented on ten cancer datasets. The results show that the proposed system has ability for reducing the genes and enhance the cancer classification performance [10],[29-33].

III. MICROARRAY TECHNOLOGY

DNA microarrays are commonly referred to as DNA chips.They are a set of thousands of microscopic spots of DNA fixed to a solid surface . Each spot includes several copies of a same sequence of DNA that is a specific representation of a gene in an organism. The spots are organized into pen groups in a regular manner [11]. The level of expression stores in the form of an image (CEL File) for each gene. Next, from such image and by using specialized software, the data is extracted. Fig.1 shows DNA Microarray Surface.

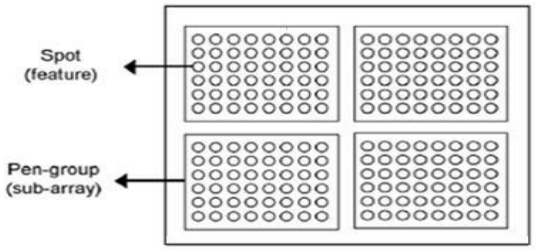

Fig.1. The DNA Microarray Surface [12]

Most microarray manufactures supply a specific software. For example, a package of Limma, can be a collection of methods of analysis to CEL file raw from the data of microarray. Biologists use DNA microarray for monitoring a higher number of levels of gene expression in a given organism under certain conditions simultaneously. By comparing and measuring the levels of gene expression in an unhealthy cell compared with healthy cell, the genes which are responsible for different diseases could be identified [13].

Usually, microarrays store the data from thousands of individual genes' expressions. Atypical manner of representing a dataset generated by experiments of microarray is in the matrix form, named gene expression matrix, whose rows represent the samples with the columns represent the gene expression levels. Thus, there are hundreds of numbers of rows and many or tens of thousands of numbers of columns (genes) [14]. Fig. 2 illustrates a data matrix for the gene expression.

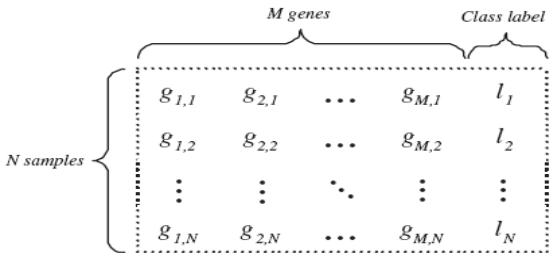

Fig.2. The Structure of Gene Expression Matrix [15].

A. Gene Selection

Generally, the data dimensionality of microarray is very huge but really the genes number associated with the disease is lower. Consequently, in microarray data, there are a huge number of insignificant genes that are not affected much and often cause some problems in the prediction of the diseases. Thus, gene selection becomes the most important method which has been used in data preprocessing to effectively reduce data. It is a method of choosing the most significant genes set from the origin data [16]. In this paper, Analysis of Variance, Mutual Information, and K-means method have been used to select the informative genes.

B. Analysis of Variance (ANOVA)

Analysis of Variance is a statistical test used to compare means between three or more groups (classes) of samples. It aims for reducing the data size by selecting the genes subset from the origin data. It checks how the variations between the groups and within the groups differ. The different equations between groups and within groups are shown in Table I to compute the Sum of Squares (SS) and the Mean Square Error (MSE) that will later be used for the F-test calculation [17]. Where n_i is represented the size sample of a group i, $\bar{x}_i$ represented the mean of group i, $\bar{x}$ represents the grand mean (mean combined for all groups), s represents standard deviation, n represents the samples number, and k is the total groups number. Thus, The F- test is computed by the ratio of mean square error between groups and mean square error within groups as shown in "(1)".

TABLE I. ANOVA TEST

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square Error (MSE)
Between Groups	$\sum_i n_i(\bar{x}_i - \bar{x})^2$	k −1	$\frac{\sum_i n_i(\bar{x}_i - \bar{x})}{k-1}$
Within Groups	$\sum_i (n_i - 1)S_i^2$	n−k	$\frac{\sum_i (n_i - 1)s_i^2}{n-K}$

$$F = \frac{MS_{Between}}{MS_{Within}} \quad (1)$$

Then, F-value is utilized to compute the P-value[18].

C. Mutual Information (MI)

Mutual Information is defined as an index measuring of the correlation between the two random variables. It is an effective filter method for its simplicity and efficiently. MI between the genes and the class label is computed to identify the informative genes. The concept of MI has been derived from entropy of a random variables. The steps of the MI computation are provided as follows [19]:

1) Let X be the random variable that represents the gene and Y represents the class label. Firstly, the entropy calculation of X provides by the entropy H (X), and the entropy calculation of Y provides by the entropy H (Y), both are defined as follows:

$$H(X) = -\sum_{x \in X} P(X)\ log\ p(X) \quad (2)$$

$$H(Y) = -\sum_{y \in Y} P(Y)\ log\ p(Y) \quad (3)$$

2) P(X) and P (Y) are represented the probabilities values of the gene X and the class label Y respectively. Secondly, the

calculation of the joint entropy H (X, Y), that is the quantity of uncertain association between two variables that are random X and Y, has been defined in "(4)" .

$$H(X, Y) = -\Sigma_{x\epsilon X} \ \Sigma_{y\epsilon Y} \ P(X,Y) \log (X,Y) \quad (4)$$

3) P (X, Y) is represented the probability of joint of X and Y values that are occurred with each other. Lastly, The calculation of MI (X, Y) between the two variables X and Y is defined as follows:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \quad (5)$$

4) Then, the result of MI is sorted in descending order and the highest value is the first gene, which will be selected as the most relevant genes [20].

D. K-means Clustering Algorithm

K-means is a commonly used methods of clustering, with its efficiency is affected by the important genes subset [21]. The K-means method has been divided the input dataset into a pre-defined number of K-clusters with determining the cluster's centroid by computing the mean of the data which locate in such cluster. Using randomly cluster centroids, K-means method has been initiated, so that each input dataset has been related to a cluster based on the distance to the centroids of cluster. The input dataset belongs to the cluster in which a centroid was closer for it. The K-means method is including the repeat of two steps, namely the association of the input dataset into the cluster with closest centroid of cluster, while update cluster centroid for the mean of the all data which is located in such cluster. The repetition has been done till a criterion for convergence is achieved [5]. It mostly uses the Euclidean distance for calculating the distance between the data points and the centroid clusters. The calculated distance between two points of data according to Euclidean distance X = $(x_1, x_2, x_3, \ldots, x_n)$ and Y = $(y_1, y_2, y_3, \ldots, y_n)$ is described as follows [22]:

$$Dis(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (6)$$

As the K-Means method requires the predefined clusters number before the technique is implemented. Therefore, elbow method has been used to supply several indicators for estimating the appropriate clusters number, that have been represented fundamental information to analyze the cluster. The Elbow method is a common method to validate consistency inside clusters. The elbow method helps in finding the number of optimized clusters in the dataset. This method is assumed that the k-means algorithm runs on a dataset for a number of k values, and for every k-cluster, the Sum of Square Error (SSE) is computed according to the "(7)" .

$$SSE(K) = \sum_{i=1}^{K} \ \sum_{P}^{Ci} \ (P - M_i)^2 \quad (7)$$

Where p represents a cluster data point, M_i represents the center of C_i and k represents the clusters number [23].

E. Deep Learning Convolutional Neural Network (CNN)

Deep Learning (DL), as an Artificial Intelligence branch, relies over algorithms to simulate the processing of data and thought processes, or for abstraction development. Convolutional Neural Network (CNN) is a widely utilized supervised model of DL. CNN has a distinctive architecture built up of numerous layers which will be sorted by their functionalities. Information was passed into each CNN layer, with the former layer output provides an input for the next layer. A first layer in network is an input layer, whereas an output layer is the last layer. All the layers between the input layer and output layer are referring as the hidden layers. Usually, every layer is consisting of one type of the activation function [24].

The architecture of CNN is able to deal with the high dimensional data, like gene expression data. CNN contains the input layer, the output layer, and other hidden layers. These layers have generally categorized as Convolution Layer, Max-Pooling Layer, and a Fully-Connected Layer. The overall CNN architecture layer by layer is explained as follows:

- Convolution Layer: It forms the core structure of the CNN model where several of the computation are concerned. Its own parameters contain a set of learnable filters, often called kernels. The major issue of the convolution layer is to identify features found in the input data inside local areas and map their presence for a feature map. The feature map is being obtained for every filter in the layer via repeating the application of the filter through sub-regions of the entire data, i.e., convolved the filter together with the input data, trying to add a bias element, and after that applied an activation function. The input region to which a filter was applied is called a local receptive field. The field receptive size likes the filter size [25]. The convolution operation is shown in Fig.3.

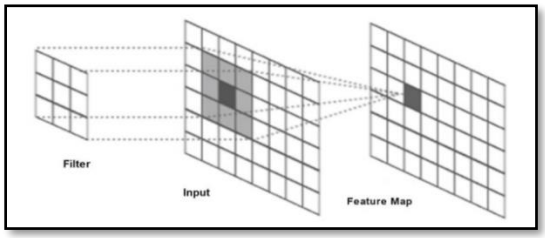

Fig. 3. The convolution operation [25]

The output of the feature maps of the convolution layers are feeding to the activation function named Rectified Linear Unit (ReLU) for converting the linear output to nonlinear. ReLU works for comparing the input value with zero and picks the highest one like a winner. It is mathematically given as:

$$ReLU(z) = \max(0, z) \quad (8)$$

- Max pooling layer: After a convolution layer, there is a pooling layer immediately. It suggests that the convolution layers outputs are pooling layers inputs for the network. Operations of pooling decrease the dimensions of the feature maps across a deception of several functions for summarizing sub-regions, such as taking the highest value. Pooling layers aim to step by step cut the dimensionality of the data, and therefore decrease the parameter's number and also the procedure complexness for the model and thus control the matter of overfitting. Max-pooling is the most popular pooling technique. The operation of pooling layer is discarded minimal important data but is preserved the features that detected in a lower represent. [26]. The operation of max pooling is presented in Fig. 4.

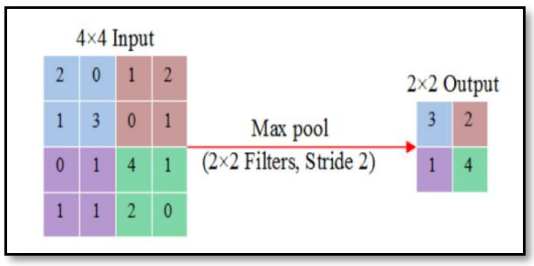

Fig. 4. The operation of max pooling [26]

- Fully-Connected Layer (FC): CNN is consisting of two stages: the stage of feature extraction and prediction stage. The convolution and pooling layer stack operate as the extraction stage of the feature, whereas the prediction stage consists of one or even more fully connected layers follows by a softmax function. The FC also refers as a Dense layer and locates in the last part of the network. FC makes to connect all the neuron from the preceding layer to every neuron of the next layer. The resulted feature map from the last previous layer must be flattened firstly to be feature vector to fully connected with an output layer that composed of neurons equal to number of classes as per to a used softmax or sigmoid activation functions which are specialized for multi-class and binary class prediction respectively utilized in the final CNN layer to predict the trained data [27]. Fig.5 shows the CNN architecture.

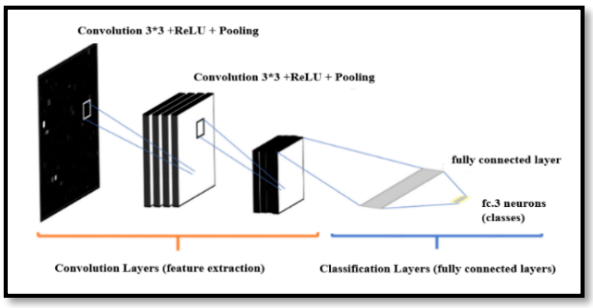

Fig. 5. CNN Architecture [24]

IV. RESEARCH METHODOLOGY

In this section, the suggested approach includes main tasks like loading the raw AD dataset. Then, normalization by using the Min-Max technique, gene selection methods and prediction via Convolutional neural network (CNN) as presented in Fig. 6.

A. Dataset

In this paper, the dataset was obtained from the publicly accessible source of data, called Gene Expression Omnibus (GEO: GSE63060 and GSE63061) retained by the National Center for Bioinformatics Data (NCBI). Then we merged these two datasets to one dataset namely AD dataset. AD dataset contains 16382 genes and 569 samples which composed of 245 patients with AD, 142 Mild Cognitive Impairment (MCI) and 182 Control Subject (CTL). Table II shows basic information on the dataset. The dataset was uploaded from (http:// www.ncbi.nlm.nih.gov/geo/).

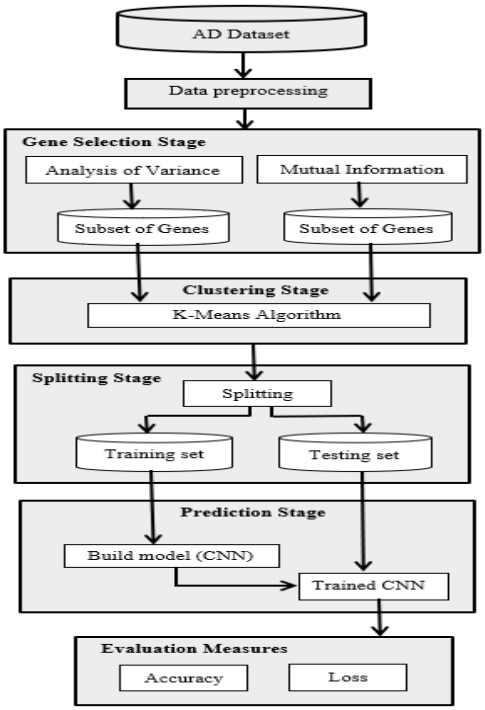

Fig.6. Proposed Approach

TABLE II. DETAILS OF AD DATASET

The Title of the Dataset:	AD dataset
The characteristics of the dataset:	Multivariate
Attribute Properties:	Real
Instances Number:	569
Attributes Number:	16382
Number of Class Labels:	3

B. Preprocessing

Data preprocessing is an essential step because of the noisy nature of the data produced by microarray technology. The dataset needs to be normalized to reduce the expression measurements variation. The normalization Min-Max is used to normalize the data set. The gene values are scaled such that the lowest value by each gene be zero and the highest value will be one.

C. Gene Selection

The main aim of gene selection is for reducing the data dimensions. Gene selection is a technique for selecting a small genes subset from the origin dataset since the genes are typically irrelevant. ANOVA method has been used in the data analysis and draw interest information depending on p-value. ANOVA performs analysis by comparing the given sample dataset and returns a single p-value, which is important. In our study, the p-value was set at 0.05, every value lesser than 0.05 is effective, where as any value higher than that value is unimportant. The p-values have been used for ranking the important genes that have small values. Then, the sorted genes number has been used for more analysis. The other method is to use MI to identify the genes which are informative. MI calculates among genes and class labels in a dataset. Thus, the MI values for all genes arrange in descending order. The initial genes with higher value for MI identify as significant genes and used as input for further analysis.

D. *The Performance of k-Means Clustering for Gene Selection*

In our study, k-means method has been implemented for dividing the genes of the dataset into pre-defined clusters number. The aim from applies k-mean clustering will be to group genes into a cluster wherein data points have the same values. Suppose that there may be D genes in the dataset used and the dataset used size is N. The objective of implementing k-means method on the data is for producing gene sets of size K, where K < D. Such K genes are composed of cluster centroids that result after criteria of convergence have been met.

The procedure of clustering partitions the dataset into the clusters; dividing is then performed in horizontal way. The origin data transforms into a matrix of transpose, which is used as the clustering algorithm input. The outcome of this step is really for reducing the data size where the size became N x k. Such k genes are employed for building the model of prediction in the following step.

E. *Building Prediction Model using Reduced Gene Set*

After completion of the collection of data, preprocessing and gene selection, CNN model configures. In our study, One-Dimensional Convolutional Neural Network model (1D-CNN) is proposed. The architecture of 1D-CNN contains 13 layers. Table III illustrate the1D-CNN model including three convolution layers and three max pooling layers. It also consists of FC layers right after the values have been flattened. In additional to all these basic layers, the 1D-CNN model can include extra layers such as dropout layers to resolve the overfitting problem. Typically, the last FC layer output is feeding to the predictor that outputs the score of class. The softmax function (Exponential Function) has been used for multi-class prediction purposes to return the predicted class probabilities as shown in "(9)" .

$$\text{Softmax}\ (x_j) = \frac{e^{x_j}}{\sum_{k=1}^{c} e^{x_k}} \tag{9}$$

In our study, the categorical cross entropy for calculating loss in training and testing data with Adaptive Moment Estimation (ADAM) optimizer are employed. ADAM optimizer performs by calculating for each parameter the adaptive learning rate. The system was learning at the ratio for the training with testing at 70% and 30% for the dataset respectively.

TABLE III. THE CNN MODEL PROPOSED STRUCTURE.

Type of Layer	Output Shape	No. of parameters
Convolution_ 1 (Conv1D)	(None, 16380,16)	64
MaxPooling_1 (MaxPooling1D)	(None, 8190,16)	0
Convolution_ 2 (Conv1D)	(None, 8188,32)	1568
MaxPooling_2 (MaxPooling1D)	(None, 4094,32)	0
Convolution_ 3 (Conv1D)	(None,4092,64)	6208
MaxPooling_3 (MaxPooling1D)	(None, 2046, 64)	0
Flatten	(None, 130944)	0
Dense_1	(None, 128)	67043840
Droput_1	(None, 128)	0
Dense_2	(None, 256)	525312
Droput_2	(None, 256)	0
Dense_3	(None, 512)	262400
Dense_4	(None, 3)	771
Total number of trainable parameters	67, 840, 163	

F. *Evaluation Measures*

A common performance metrics have been used, like accuracy of prediction and loss. The accuracy has been used to assess a model's overall predictive capacity which considering four parameters called True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), (see (10)). That performs to look at the sample's numbers with correct prediction according to a ratio for the total samples test number.

$$\text{accuracy} = \frac{\text{TP+FN}}{\text{TP+TN+FP+FN}} \tag{10}$$

According to the proposed approach the error score is calculated using a loss function. Categorical cross-entropy identifies the loss during the outcomes of categorical are a non-binary, which is more than two. The outcome can be: (YES / NO / MAYBE) as seen "(11)" .

$$\text{Loss} = \sum_i y_i' \ log\ (y_i) \tag{11}$$

In which y_i' is the target label, and y_i is the output of the predictor.

V. RESULT AND DISCUSSION

The key objective of this paper is for reducing the data dimensionality to a lowest possible while maintaining high accuracy, so the process of neglecting the weak relevant genes and the selection of strong relevant genes have been applied. The first reduction of genes with ANOVA where the number of genes remaining after this process is (7044) out of (16382) genes. The other reduction of genes with MI where the number of genes remaining after this process is (9829) out of (16382) genes. Table IV provides the description of the reduced dataset.

TABLE IV. THE DESCRIPTION OF THE REDUCED DATASET

Dataset	Methods	The Genes Number		Samples
		Original	*Reduced*	
AD	ANOVA	16382	7044	569
	MI		9829	

In this paper, another algorithm was applied on the reduced dataset namely k-means algorithm to divide the genes which are obtained from the previous phases. It is known that the k-means algorithm needs a predefinition for the clusters number before the algorithm is run. Therefore, elbow method has been used to decide the optimal number of clusters. A range of 1 to 15 of k was used for the elbow method in our study. It illustrates that the optimal K of data takes place when k = 4. The graph of the Elbow method for k-means algorithm is seen in Fig. 7 and Table V shows the values of the elbow plot.

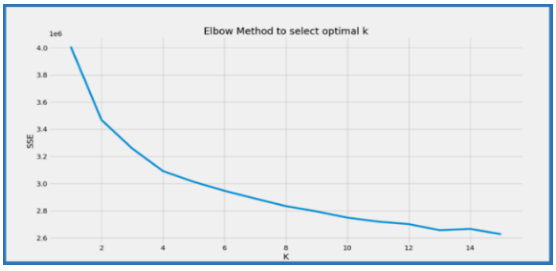

Fig. 7. Elbow Method of K-Means Algorithm

TABLE V. SOME VALUES OF ELBOW METHOD

Clusters Number	Sum of Square Error	Clusters Number	Sum of Square Error
k=1	4008036.33	k=9	2796416.72
k=2	3468161.58	k=10	2764150.02
k=3	3258169.68	k=11	2733994.48
k=4	3091961.66	k=12	2687419.39
k=5	3014809.77	k=13	2675903.77
k=6	2941299.81	K=14	2643798.00
K=7	2889355.87	K=15	2634975.36
K=8	2833166.83		

The k-means algorithm has been applied onto two subsets of genes from ANOVA and MI methods for gene selection. This method partitions each of the subsets into 4 clusters based on the elbow method scores. By using k-means algorithm, the numbers of genes are reduced from 7044 to minimum of 2500 genes based on the subset from ANOVA. The same process was done in which the genes number are reduced from 9829 to minimum of 4600 genes depending on the subset from MI. After this process, the selected genes which are the most informative genes are passed as input to the prediction model.

The proposed CNN model takes the informative genes as an input. These selected genes pass through the number of convolution layers and pooling layers, fully connected layers as well as the dropout layers. The last fully connected layer uses a softmax activation function which has 3 neurons to predict 3 classes on the output layer. In addition, the size of 10 for the epoch is used. Generally, the number of epochs is the number of times the model will cycle through the data. The more epochs we run, the more the model will improve, up to a certain point. After that point, the model will stop improving during each epoch.

The proposed approach using ANOVA with CNN model is achieved accuracy at (0.758) while k-means algorithm with CNN model can enhance the accuracy at (0.929) based on a subset of relevant genes obtained from ANOVA method. The average accuracy for the MI method is achieved (0.668) using CNN model while k-means with CNN model can enhance the accuracy performance at (0.886) based on a subset of informative genes obtained from MI method. Thus, it can be seen that our proposed gene selection algorithm using k-means selects the genes which are sufficient to describe the target class and deletes the genes that do not have any positive effect on the performance of the prediction model. Moreover, the number of the genes theirs result for k-means method show that the accuracy of the model will increase and the complexity of the model will be decreased accordingly. Table VI illustrates the comparative results of accuracy and loss to CNN model for all methods.

TABLE VI. THE COMPARATIVE RESULTS FOR THE PREDICTION ACCURACY AND LOSS

Method	Number of Genes	CNN	
		Accuracy	Loss
Raw data	16382	0.612	0.833
ANOVA	7044	0.758	0.596
K-means based on ANOVA	2500	0.929	0.199
MI	9829	0.668	0.743
K-means based on MI	4600	0.886	0.384

To build our model, the proposed approach is implemented in Python 3.6 with PyCharm 2018 IDE and in Keras Deep Learning Library. This conducted the prediction training on the proposed CNN on a PC where the processor is Intel Core i7, and 2.40 GHz speed with the RAM of 8 GB.

VI. CONCLUSION

In our study, Convolutional Neural Network was proposing for predicting the multi-class AD dataset.This dataset is normalized using min-max normalization. It contains a huge number of genes, making it difficult to analysis. The dimensionality curse is an important challenge we faced in analysis of this dataset. Therefore, two filtering methods of gene selection which are analysis of variance and mutual information are used for obtaining the smallest number of genes to enhance the prediction performance. Furthermore, k-means algorithm is used in our study as a gene selection to reduce the irrelevant genes, since it tries to group related genes into one cluster together. Categorical cross entropy has been used for computing the error magnitude through the training and testing procedure, as it is a standard loss function and is suggested for non-binary prediction issues. ADAM is applied to the purpose of optimization. To validate the performance of the proposed approach, the measures of evaluation namely accuracy and loss have been implemented. The results indicate that the proposed approach can reduce the irrelevant genes, and it achieves higher accuracy of prediction with short processing time.

ACKNOWLEDGMENT

This study was supported through Babylon University, College of Information Technology, department of Software. Many thanks for their support and cooperation.

REFERENCES

[1] K. Tejeswinee, S. G. Jacob, and R. Athilakshmi, "Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's and Parkinson's Disease," Procedia Computer Science, vol. 115, pp. 188-194, 2017.

[2] X. Li, H. Wang, J. Long, G. Pan, T. He, O. Anichtchik, R. Belshaw, D. Albani, P. Edison, E. K. Green, and J. Scott, "Systematic Analysis and Biomarker Study for Alzheimer's Disease," Scientific Reports, vol. 8, no. 1, pp. 1-14, 2018.

[3] K. Guckiran, I. Canturk, and L. Ozyilmaz, "DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO," Journal of Natural and Applied Sciences, vol. 23, pp. 126-132, 2019.

[4] R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "A novel gene selection algorithm for cancer classification using microarray datasets," BMC Medical Genomics, vol. 12, no. 1, pp. 1-12, 2019.

[5] D. P. Ismi, S. Panchoo, and M. kusno, "K-means clustering based filter feature selection on high dimensional data," International Journal of Advances in Intelligent Informatics, vol. 2, no. 1, pp. 38-45, 2016.

[6] 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. 10, 2020.

[7] K. Sekaran, and M. Sudha, "Diagnostic gene biomarker selection for alzheimer's classification using machine learning," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 12, pp. 2348-2352, 2019.

[8] D. A. Vanitha, D. Devaraj, and M. Venkatesulu, "Gene Expression Data Classification using Support Vector and Mutual Information-based Gene Selection," Procedia - Procedia Computer Science, vol. 47, pp. 13-21, 2015.

[9] P. Danaee, and R. Ghaeini, "A deep learning approach for cancer detection and relevant gene identification," pp. 219-229, 2017.

[10] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene Selection and Classification of Microarray Data Using Convolutional Neural Network," International Conference on Advanced Science and Engineering, no. December 2018, pp. 145-150, 2018.

[11] M. Barati, and M. Ebrahim, "A Gene Expression Profile of Alzheimer's Disease Using Microarray Technology," 2016.

[12] M. M. Babu, "An Introduction to Microarray Data Analysis," p. 225–249, 2004.

[13] K. Raza, "Analysis of microarray data using artificial intelligence-based techniques," Handbook of Research on Computational Intelligence Applications in Bioinformatics, no. August 2015, pp. 216-239, 2016.

[14] M. Muszynski, and S. Osowski, "Data mining methods for gene selection on the basis of gene expression arrays," International Journal of Applied Mathematics and Computer Science, vol. 24, no. 3, pp. 657-668, 2014.

[15] L. Dey, and A. Mukhopadhyay, "Microarray Gene Expression Data Clustering using PSO based K-means Algorithm," International Journal of Computer Science and its Applications, no. May 2014, pp. 232-236, 2005.

[16] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," Procedia Computer Science, vol. 23, pp. 5-14, 2013.

[17] D. Chen, and D. Hua, " Gene Selection for Multi-Class Prediction of Microarray Data," Proceedings of the IEEE Computer Society Bioinformatics Conference, pp. 492-495.

[18] K. M. Poornima, and S. T. Jayakumari, " Neural Network based Technique for Parkinson's Disease Classification using ANOVA as Feature Selection Model," International Journal of Engineering Research & Technology, vol. 3, no. 27, pp.1-5, 2015.

[19] S. B. Guo, M. R. Lyu, and T. M. Lok, "Gene Selection Based on Mutual Information for the Classification of Multi-class Cancer," Springer, pp. 454-463, 2014.

[20] W. Zhongxin, S. Gang, Z. Jing, and Z. Jia, " Feature Selection Algorithm Based on Mutual Information and Lasso for Microarray Data," The Open Biotechnology Journal, vol. 10, no. 1, pp. 278-286, 2016.

[21] M. Shah, and S. Nair, "A Survey of Data Mining Clustering Algorithms, "International Journal of Computer Applications, vol. 128, no. 1, pp.1-5, 2015.

[22] T. Chandrasekhar, K. Thangavel, and E.Elayaraja, "Effective Clustering Algorithms for Gene Expression Data," International Journal of Computer Applications, vol. 32, no. 4, pp.25-29, 2011.

[23] P. Bholowalia, and A. Kumar," EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," International Journal of Computer Applications, vol. 105, no. 9, pp. 17-24, 2014.

[24] N. M. Khalifa, M. N. Taha, D. E. Ali, A. Slowik, and A. E. Hassanien," Artificial intelligence technique for gene expression by tumor RNA-Seq Data: A novel optimized deep learning approach," IEEE Access, vol. 8, pp. 22874-22883, 2020.

[25] M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, "Advances in Deep Learning," Srinagar, India, Springer, 2020.

[26] S. Sakib, N. Ahmed, A. J. Kabir, and H. Ahmed, "An Overview of Convolutional Neural Network: Its Architecture and Applications," no. November, 2018.

[27] H. Lee, and J. Song," Introduction to convolutional neural network using Keras; an understanding from a statistician," Communications for Statistical Applications and Methods, vol. 26, no.6, pp. 591-610, 2019.

[28] M. Younis, S. Y Ameen, S. B. Sadkhan, " evaluation of using genetic algorithm in cryptoanalysis", Journal of University of Babylon, 2006, 12 (5), 982-989.

[29] S. B. Sadkhan, "A Proposed Genetic Based Method to Solve Frequency Assignment Problem for HF Band", 2019 4th Scientific International Conference Najaf (SICN), 48-53/

[30] S.B.Sadkhan, A.N. Abbas, "Watermarked and Noisy Images identification Based on Statistical Evaluation Parameters", Journal of Zankoy Sulaimani- Part A (JZS-A), 2013, 15 (3).

[31] S.B. Sadkhan, N.A. Abbas, "Performance Evaluation of Speech Scrambling Methods based on Statistical Approach", Atti della Fondazione Giorgio Ronchi 2011 66 (5), 601-614.

[32] S.B. Sadkhan, A.J. Alnaji, N.A. Muhsin, "Performance Evaluation of Blind Source Separation Algorithms Based on Neural Networks", 2006, 5th International Symposium on Communication Systems, Networks And Digital Signal Processing, Greece.

[33] S.B. Sadkhan, A.Q. Hameed, H.A. Hamed, "A proposed identification method for multi-user chirp spread spectrum signals based on adaptive Neural-Fuzzy Inference System (ANFIS)", 2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA