Deep Learning approach for Microarray Alzheimer's Data Classification

HebaMuthanna Mohammad Ali [1] ,Sura Zaki.AL-Rashid [2]

[1] M.Sc. student, software department, college of information technology, University of Babylon, Hilla, Iraq. E-mail: heba.muthanna9@gmail.com.

[2] Ph.D. lecturer, in software department, college of information technology, University of Babylon, Hilla, Iraq. E-mail: sura_os@itnet.uobabylon.edu.iq.

Article Info
Volume 83
Page Number: 2016 - 2029
Publication Issue:
July-August 2020

Abstract:

Alzheimer's disease (AD) has a quite complex genetic architecture and is an important progressive neurodegenerative disease. A major objective to the research of biomedical is to discover the genes that are risk then to explain the task of these genes in development of the disease. For such reason it is necessary to expand the set of genes which are associated to AD. Genes take central part in all biological processes. Microarray technology has provided genes with a huge number to measure many expression levels simultaneously. Microarray datasets characteristically have genes which are a huge number and samples with small size. This truth is described as a dimensionality curse that has a complex task. A promised method which named gene selection is solving such issue and has a major turn for creating an effective diagnosis of Alzheimer's. In such study, methods of gene selection have been implemented, including Principle Component Analysis (PCA) and Singular Value Decomposition (SVD). Such methods have the ability to reduce the number of insignificant and genes that are redundant in the original datasets. After that, deep learning (DL) via Convolutional Neural Network (CNN) serves as a classifier to predict AD. CNN which consists of six-layer having various parameters for the dataset has been used. The empirical results are showed with AD dataset that PCA-CNN model achieves 97.24% accuracy and loss 0.4614 while SVD-CNN model achieves 98.99% accuracy and loss 0.2588. Thus, the proposed system is suitable for decreasing the genes dimensions by means of selecting subset of informative gene and enhance the classification accuracy.

Article History
Article Received:*06 June 2020*
Revised: *29 June 2020*
Accepted: *14 July 2020*
Publication: *25 July 2020*

Keywords:*Alzheimer's Disease,MicroarrayTechnology, Gene Expression Data, Gene Selection, Classification, Deep Learning.*

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. This is a common shape for dementia [1]. AD accounts 60-80 per cent from all dementia states. During 2050, many numbers of people who have AD is estimated to rise in the US alone from 5.4 million to between 11 and 16 million, and dementia is estimated to cost $2 trillion globally by 2030. Despite these shocking numbers, there is no successful method to detect disease before symptomatic is appeared, which may be the only phase in the disease's progression where we could interfere [2]. Based on the importance of AD with insufficiency in a specific curing to such disease, a new technology by using microarray has been applied to define the genes that causes the disease. Microarray technologies are an important medical tool that the biologist uses to monitor gene expression levels in a given organism. Microarray data analysis focuses in helping to identify a best handling to several diseases and precise medicinal diagnosing for many genes through various empirical cases [3]. However, the data of gene expression generated of microarray technologies are concerning to the high curse of dimensionality. Genes can be either redundant or irrelevant, and thus be removed without deducting much information loss. The major problem with analysis of microarray data is that genes are in huge numbers and samples are in small ones. This may lead to a decrease in prediction accuracy and an increase in overfitting problems [4]. The technique to solve this problem is the use of gene selection method, which extracts only the optimal set of features (genes) for building classification model. Actually, gene selection is a method to select a small subset of genes from the original set which only contains the informative genes. This subset of genes allows researchers to gain substantial vision about the genetic nature of disease. This method has the ability to reduce the computation of costs and increase the efficiency of classification for AD [5]. Gene selection can be implemented using various algorithms such as PCA and SVD. These algorithms are typical unsupervised methods to analyze the microarray data of gene expression, which provide details on the total framework to the dataset which is analyzed. Recently, they are implemented on very large datasets to generate a low dimensional gene expression data before classification happens [6]. Classification of microarray data is not a trivial task. There are multiple approaches being used by the bioinformatics community to diagnose and classify the microarray data with the aid of machine learning systems [7].

In this study, a deep learning model is employed to predict AD by using data of gene expression. DL is a sub-set from machine learning. DL such as CNN is taking a big amount of data for learning the behavior of genes via training set, and predicting the unseen class label. Moreover, applying CNN model might improve predictive performance. Also, we emphasize not only on the gene selection methods but also the accuracy of the classification after applying gene selection.

The remaining sections of the presented study are organized in the following manner: The related work is showed in Section 2. The background of microarray technology is described in Section 3. Section 4 elaborates the materials and methods used in our study. In this section, dataset is explained and also gene selection, classification. The proposed methodology is described extensively in section 5. Our simulation and results are described in section 6. The last section is focused on the conclusion and remarks for future research.

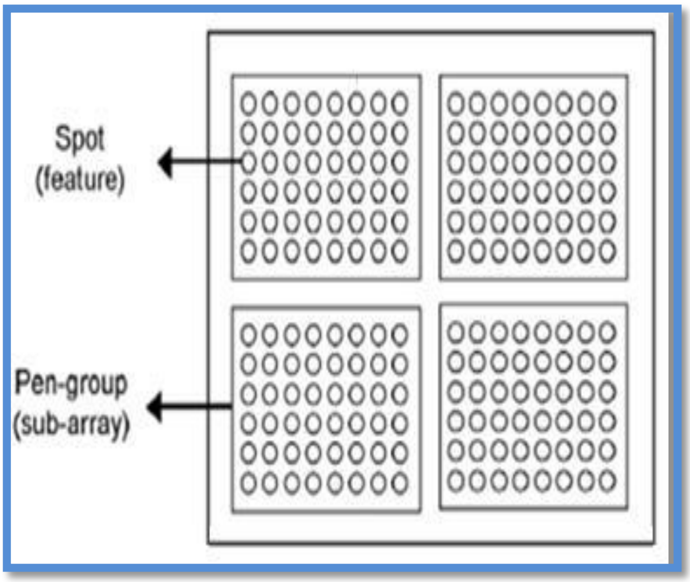
2. Related Work

Chihyun et al. (2020) used a classification model using an artificial intelligent predication system known as deep learning algorithm. The proposed system has the ability to predict Alzheimer's disease by using data of gene expression which its scale is large with 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 [8]. Karthik et al. (2019) used Rhinoceros Search Algorithm (RSA) in the role of a feature selection. They used four supervised machine learning methods like Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Multilayered perceptron Neural Network (MLP-NN) for classification. They used (GEO: GSE1297) gene expression datasets in their experiments. 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 [9]. Devi et al. (2015) apply Mutual Information (MI) to identify the most relevant genes and Support Vector Machine (SVM) as an efficient gene classification process. The experiments are implemented on two cancer microarray datasets: Colon and Lymphoma. The experiments result presented a proposed system which decreases the input dimension of genes to select a subset of relevant genes and improving the accuracy of classification [10]. Lena et al. (2012) used three feature selection approaches: Information Gain (IG), Random Forest (RF), and a wrapper of Genetic Algorithm and Support Vector Machine (GA/SVM). Also, six different classification methods: C4.5 (decision tree), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), SVM with Linear Kernel, and SVM with Gaussian Kernel have been used. The proposed approach was implemented on AD dataset (GEO: GSE5281). The results of the proposed model enable us to select sets that has small genes which are suitable jointly to train the classifier [11]. Xiaoyan et al. (2018) proposed a machine learning approach (SVM). The suggested method has been examined based on integrate the data of gene expression with a specific network data of gene in the human brain, to detect the AD genes which are in full spectrum across an entire genome. The method results give an analysis of genes that are associated with AD and evaluated the reliability of these genes [12]. Padideh et al. (2017) used Principle Component Analysis (PCA) to extract efficient genes from gene expression data which has high dimensionality. After that, the extracted representation of the performance was evaluated via artificial neural network (ANN). ANN is a supervised classification model used to classify genes which are essential to the disease diagnosis. The results of the proposed approach showed the interacting genes might be helpful to detect the diseases [13].

3.Microarray Technology

DNA microarrays are commonly referred to as DNA chips or biochips. They are a collection of thousands of microscopic spots of DNA fixed to a solid surface [14]. Each spot includes several copies of the same sequence of DNA which is a specific representation of a gene in an organism. The spots are organized into pen groups in a regular manner[15]. 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[16]. Figure 1 shows DNA Microarray Surface.

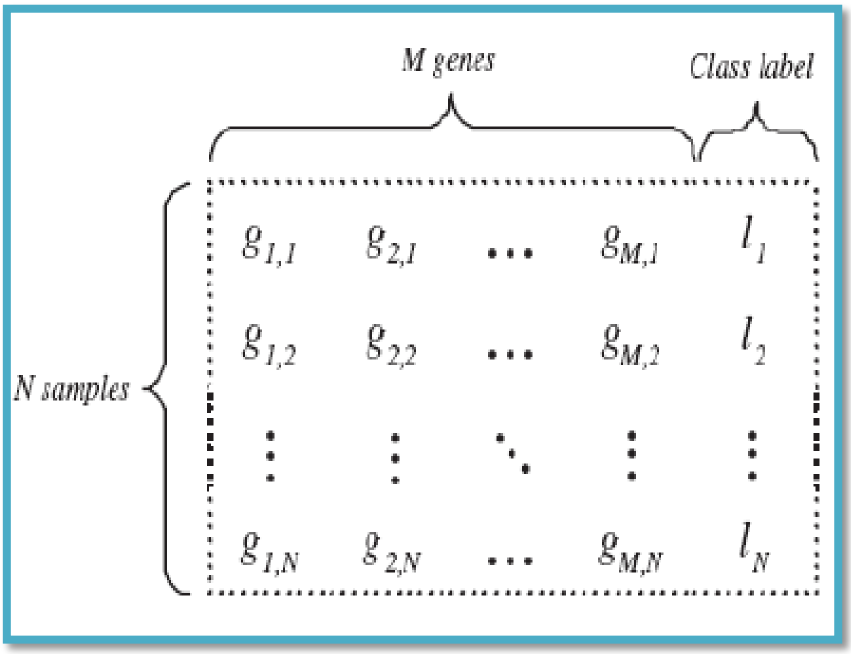
Figure 1: The DNA Microarray Surface [16]



Several microarray manufactures supply their specific software. For example, the package namely Limma is the most commonly used, a module of analysis tools for microarray data that is raw CEL file [15]. Biologists utilize DNA microarray for monitoring a large number of gene expression levels in a given organism under certain conditions or simultaneously. By comparing and measuring the level of gene expression in an unhealthy cell compared with healthy cell, genes that are responsible for different diseases could be identified[17].

DNA microarray usually store data from thousands of different gene expressions. Atypical way of representing the dataset generated by DNA microarray experiments is to create a matrix, also named matrix of gene expression, whose row represents to the experimental condition (sample, time point, etc.) and the column represents the expression levels of genes. Thus, there are hundreds of numbers of rows and many or tens of thousands of numbers of columns (gene sequence or genes). The basis for any analysis is represented by the collection of this data [18].

Figure 2 shows the data matrix for gene expression.



$$\begin{array}{c|ccccc} & \multicolumn{4}{c}{M\ genes} & Class\ label \\ \hline N\ samples & g_{1,1} & g_{2,1} & \cdots & g_{M,1} & l_1 \\ & g_{1,2} & g_{2,2} & \cdots & g_{M,2} & l_2 \\ & \vdots & \vdots & \ddots & \vdots & \vdots \\ & g_{1,N} & g_{2,N} & \cdots & g_{M,N} & l_N \end{array}$$

Figure 2: The data matrix for gene expression forms m gene columns, and n sample rows. The class label is the last column, i.e. which sample goes to which classifier [19].

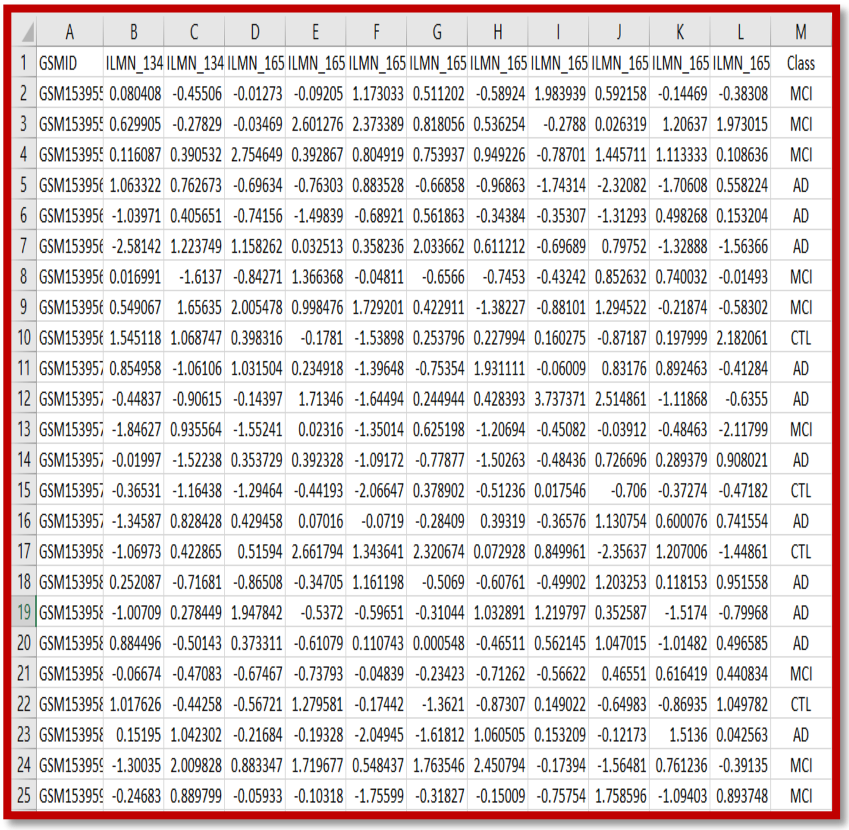
4. Materials and Methods

4.1 Dataset

In this work, 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 1 shows basic information on the dataset. The dataset was uploaded from (http:// www.ncbi.nlm.nih.gov/geo/).

Table 1: Details of AD dataset

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



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GSMID	ILMN_134	ILMN_134	ILMN_165	ILMN_165	ILMN_165	ILMN_165	ILMN_165	ILMN_165	ILMN_165	ILMN_165	ILMN_165	Class
2	GSM153955	0.080408	-0.45506	-0.01273	-0.09205	1.173033	0.511202	-0.58924	1.983939	0.592158	-0.14469	-0.38308	MCI
3	GSM153955	0.629905	-0.27829	-0.03469	2.601276	2.373389	0.818056	0.536254	-0.2788	0.026319	1.20637	1.973015	MCI
4	GSM153955	0.116087	0.390532	2.754649	0.392867	0.804919	0.753937	0.949226	-0.78701	1.445711	1.113333	0.108636	MCI
5	GSM153956	1.063322	0.762673	-0.69634	-0.76303	0.883528	-0.66858	-0.96863	-1.74314	-2.32082	-1.70608	0.558224	AD
6	GSM153956	-1.03971	0.405651	-0.74156	-1.49839	-0.68921	0.561863	-0.34384	-0.35307	-1.31293	0.498268	0.153204	AD
7	GSM153956	-2.58142	1.223749	1.158262	0.032513	0.358236	2.033662	0.611212	-0.69689	0.79752	-1.32888	-1.56366	AD
8	GSM153956	0.016991	-1.6137	-0.84271	1.366368	-0.04811	-0.6566	-0.7453	-0.43242	0.852632	0.740032	-0.01493	MCI
9	GSM153956	0.549067	1.65635	2.005478	0.998476	1.729201	0.422911	-1.38227	-0.88101	1.294522	-0.21874	-0.58302	MCI
10	GSM153956	1.545118	1.068747	0.398316	-0.1781	-1.53898	0.253796	0.227994	0.160275	-0.87187	0.197999	2.182061	CTL
11	GSM153957	0.854958	-1.06106	1.031504	0.234918	-1.39648	-0.75354	1.931111	-0.06009	0.83176	0.892463	-0.41284	AD
12	GSM153957	-0.44837	-0.90615	-0.14397	1.71346	-1.64494	0.244944	0.428393	3.737371	2.514861	-1.11868	-0.6355	AD
13	GSM153957	-1.84627	0.935564	-1.55241	0.02316	-1.35014	0.625198	-1.20694	-0.45082	-0.03912	-0.48463	-2.11799	MCI
14	GSM153957	-0.01997	-1.52238	0.353729	0.392328	-1.09172	-0.77877	-1.50263	-0.48436	0.726696	0.289379	0.908021	AD
15	GSM153957	-0.36531	-1.16438	-1.29464	-0.44193	-2.06647	0.378902	-0.51236	0.017546	-0.706	-0.37274	-0.47182	CTL
16	GSM153957	-1.34587	0.828428	0.429458	0.07016	-0.0719	-0.28409	0.39319	-0.36576	1.130754	0.600076	0.741554	AD
17	GSM153958	-1.06973	0.422865	0.51594	2.661794	1.343641	2.320674	0.072928	0.849961	-2.35637	1.207006	-1.44861	CTL
18	GSM153958	0.252087	-0.71681	-0.86508	-0.34705	1.161198	-0.5069	-0.60761	-0.49902	1.203253	0.118153	0.951558	AD
19	GSM153958	-1.00709	0.278449	1.947842	-0.5372	-0.59651	-0.31044	1.032891	1.219797	0.352587	-1.5174	-0.79968	AD
20	GSM153958	0.884496	-0.50143	0.373311	-0.61079	0.110743	0.000548	-0.46511	0.562145	1.047015	-1.01482	0.496585	AD
21	GSM153958	-0.06674	-0.47083	-0.67467	-0.73793	-0.04839	-0.23423	-0.71262	-0.56622	0.46551	0.616419	0.440834	MCI
22	GSM153958	1.017626	-0.44258	-0.56721	1.279581	-0.17442	-1.3621	-0.87307	0.149022	-0.64983	-0.86935	1.049782	CTL
23	GSM153958	0.15195	1.042302	-0.21684	-0.19328	-2.04945	-1.61812	1.060505	0.153209	-0.12173	1.5136	0.042563	AD
24	GSM153959	-1.30035	2.009828	0.883347	1.719677	0.548437	1.763546	2.450794	-0.17394	-1.56481	0.761236	-0.39135	MCI
25	GSM153959	-0.24683	0.889799	-0.05933	-0.10318	-1.75599	-0.31827	-0.15009	-0.75754	1.758596	-1.09403	0.893748	MCI

Figure 3: Screenshot of the AD dataset.

4.2 Gene Selection Methods

A sequence of microarray experience yields observance of the expression of differential across multiple conditions into many of genes. Thus, microarray data has high dimensionality problem because most of the genes are irrelevant for classification process. Therefore, methods of gene selection are effective at the removal of redundant and irrelevant genes and can minimize data dimensionality [20]. This method can also reduce computational costs and enhance the efficiency of AD classification, i.e., the aim of the method of gene selection is at finding a small subset of genes which achieves high result [5]. In such study, several gene selection methods such as PCA and SVD have been used to identfy informative genes which are associated with the diagnosis of disease directly.

4.2.1 Principal Component Analysis (PCA)

Principle Component Analysis (PCA) is a typical unsupervised approach to analyze the data of gene expression, and provide details about the entire form for the analyzed data. PCA is one of the most effective methods of gene-selection [21]. The objective of using PCA is to decrease data of high dimensionality to a new subset of smaller dimensions than the original. PCA test is used to Select/Extract relvant gene information in large dataset that is helful for further analysis [22]. It is a well-known fact that gene selection by using PCA helps to overcome overfitting, improves accuracy, and maintains model simplicity, while enhancing classification accuracy. The required steps for performing PCA are shown in Algorithm-1.

Algorithm1: Principal Component Analysis [23]

```
Algorithm: PCA
Input: Data Matrix
Output: Reduced set of genes
            ←

          ←

          ←
```

4.2.2 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a mathemtical method for data driven which could be utilized for the data of gene expression to minimize dimension. SVD is a common approach for multivariate data analysis, such as PCA. One single microarray datset may contain thousands of measurements of gene. With SVD, the numbers of redundant information can easily be reduced in the data for gene expression and identify informative genes to enhance classification accuracy [24]. Therefore, SVD as a gene selection technique has been conducted for reducing the dimensions of the dataset.A major purpose at implementing SVD is for defining and extracting the consititution of structural inside the data and as well for determining strong association concerning gene expressions [25].SVD applications contain determining the rank (r), range, approximating a matrix, and null space of a matrix.

The SVD of matrix A$\in C^{M\times N}$with rank (A) = r is defined with:

$$A = U\Sigma\sigma V^{T} \qquad (1)$$

Where, U $\in C^{M\times M}$ = [$u_1,\ldots,u_M$]indicates a unit matrix include Left Singular A Vectors, V $\in C^{N\times N}$ = [$v_1,\ldots,v_N$] indicates a unit matrix include Right Singular A Vectors, and $\Sigma \in \mathbb{R}_{+}^{M\times N}$ indicates singular values matrix of A along its diagonal with diagonal entries $\sigma_1 \geq \ldots\sigma_r > \sigma_{r+1} = \ldots = \sigma_{min\,(M,N)} =$ 0and zeros otherwise.

Rank-K estimation of A is defined using SVD with:

$$A \approx A_k = U_k \Sigma_k V_k^T \qquad (2)$$

Where $K < r$, $U_k$and $V_k$involve the initial K columns from U and V respectively, Then, Σ_k denotes a K × K core submatrix from Σ. Eq.1 is sometimes referred to as the A truncated SVD [26].

4.3 Deep learning-based microarray AD data classification

Deep learning (DL), as an Artificial Intelligence branch, relies over algorithms to simulate the processing of data and thought processes, or for abstraction development. DL uses algorithm layers for processing, analyzing and finding hidden patterns in data.Information is passing during every layer in the deep network, and the output of the previous layer provide as the next layer input. The first netwok's layer is the input layer while the last network's layer is the output layer.The other layers placed between the input and output layers have been called the hidden layers for the network.Usually, every layer is easy, regular and including one type of activation function. It is now seen as an useful method for developing automated diagnostic systems to achieve higher results, expand the scope of disease and execute applicable real-time medical diagnosis for classification systems for diseases [27]. The popular architecture used to create deep learning model and discussed in this study is Convolutional Neural Network (CNN). CNN is the commonest supervised DL modelsthat is used for classification AD based on gene expression data.

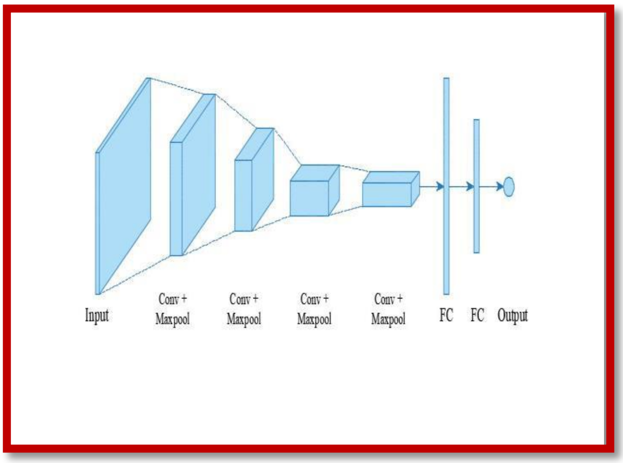
4.3.1 Convolutional Neural Networks (CNN)

Convoltutional Neural Network (CNN) is an example of DL technique that mimics function of brain for processing the information. In this paper, multilayerd CNN is proposed to classify microarray gene expression data. CNN is proposed because of its ability to deal with huge amount of data and improve classification accuracy.Furthermore, CNN is also effective in integrating closely linked datasets which enhance performance in classification process.This is because of its ability to detect latent aspects of AD from comparable kinds [28]. The broad range of application fiels of deep CNN can can attribute for the advantages below:

- CNN integrate the selection mechanism with classification processes to a single unit of learning.They learn how the features are optimized directly from a raw input during the training process.
- CNN has the ability to process inputs that are huge and has a high efficiency of computational because CNN neurons are connecting to bound weights.
- CNN is resistant at tiny input data transformation involving scaling, encoding, distortion and skewing.
- CNN can respond to varying input sizes.

1D CNN have been recently suggested and directly accomplished modern levels of performance for many applications, like early diagnosis, structural monitoring of health and classification of data for personalized biomedical.A further major benefit is that the application of time that is real and hardware which its cost is low can be possible for the easy and compress design of 1D CNN which only achieves 1D convolutions [29]. CNN is comprised of an input and output layer, and many other hidden layers. These layers are generally divided into three types: convolutional, pool, or dense, and short for fully connected (FC). Figure 4 shows a CNN architecture.

Figure 4: A basic architecture of a convolutional neural network [30]

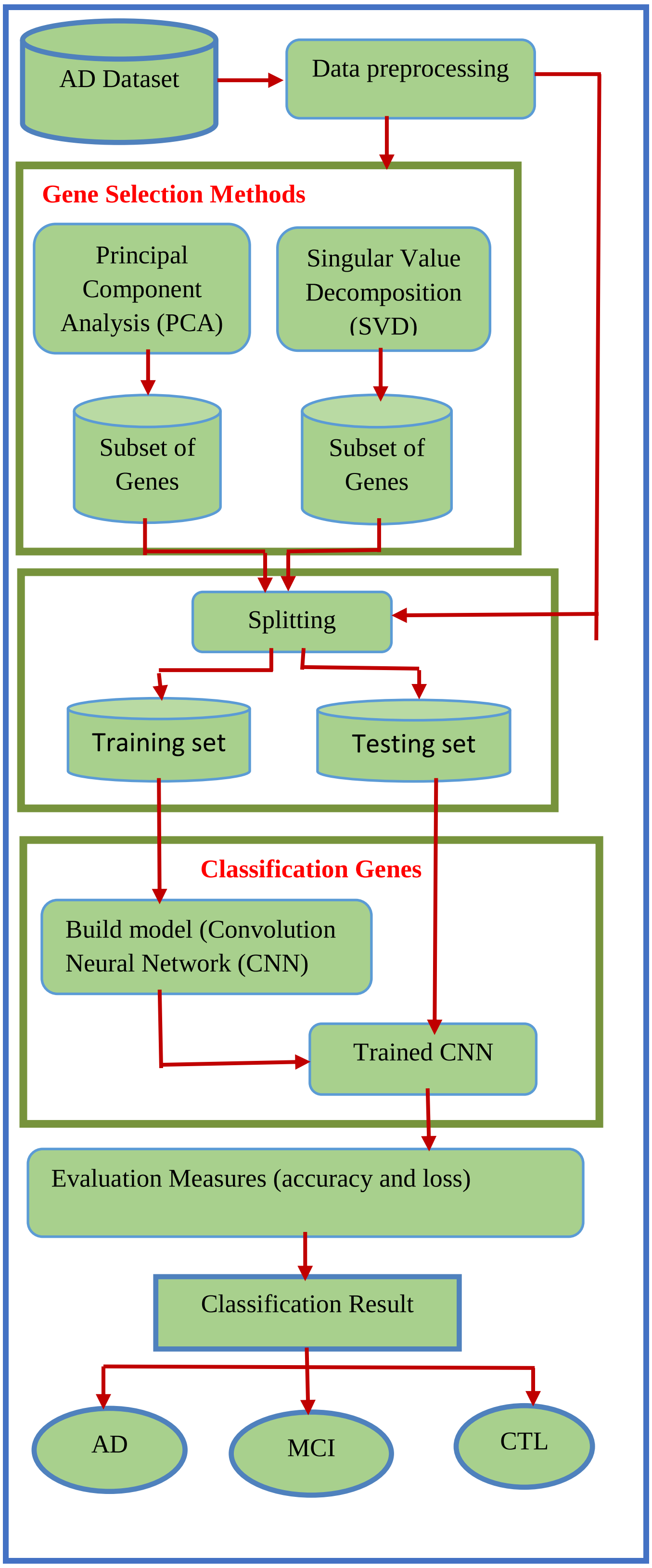


In our study, As a vector, the gene expression is taken by the CNN model with applying 1D kernal at the vector of input.This model is composed of two convolution layers, two dense layers and a single flatten layer. The output layer for the last dense layer is the prediction layer. For simplicity's sake we denominate this model 1D CNN [31].

5. Methodology

In this section, the proposed methodology includes main tasks like loading the raw microarray AD data set. Then, normalization by using the Min – Max technique, gene selection methods and classification via Convolutional neural network (CNN) as presented in Figure 5.

Figure 5: Proposed Approach



5.1 Pre-processing Stage

Data preprocessing is an essential step before starting the experiments 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 can be used to normalize the data set. The values of gene expression are calculated such that the lowest value by each gene be zero and the highest value will be one [32].

5.2 Gene Selection Stage

The main goal for the methods of gene selection is for decreasing the dimensionality at computational space of dataset. It is a method of selecting a small subset of genes from the original dataset because genes are usually irrelevant. These methods are always used before the machine learning algorithms and selecting genes based on particular performance measure regardless of the machine learning techniques. In this work, gene selection methods PCA and SVD have been applied to identify a subset of genes which is directly used in classification.

5.3 Classification Stage

At this stage, deep convolutional neural network model is applied for classifying the gene expression data. Upon accomplishment of data collection, preprocessing and gene selection, CNN model has been configured. A convolutional CNN is chosen consisting of many layers. The architecture of the convolutional layer was used because of its ability to handle high and multi-dimensional data, such as gene expression data [28]. A new method with a 1-Dimentional convolutional has been proposed in this paper. The filter size is 64 kernels with size 3 and the non-linearity activation which is called Rectified Linear Units (ReLU) was used with convolution layer. ReLU could be demonstrated as in eqn. (3)

$$g\ (y) = \begin{cases} 0 \text{ for } y < 0 \\ y \text{ for } y \geq 0 \end{cases} \qquad (3)$$

We used a six-layer CNN model for the dataset, and it is shown in Table 2. The fact is that, the shape of the input layer in the next layer has the same neurons number for the input layer. The size of 100 for the epoch is used. In our study the total number of trainable parameters adds up to 1,576,035. Additionally, to improve network performance, SOFTMAX activation function is employed at the end of the final layer. Now, a SoftMax function could be demonstrated as in eqn. (4).

$$\text{SoftMax}(C)_j = \frac{e^{c_j}}{\sum_{k=1}^{k} e^{c_k}} \qquad (4)$$

The predefined objective function namely categorical cross-entropy to calculate the loss in training and testing data with adaptive moment estimation (ADAM) optimizer is employed [33]. ADAM optimizer performs at calculating for each parameter the learning rate. The system was learned for the ratio of the training and testing is 70% and 30% of the data respectively.

Table 2:Summary of 1D CNN model structure

Type of layer	Output Shape	No. parameters
Conv1d-1 (Conv1D)	(None, 16380,16)	64
Conv1d-2 (Conv1D)	(None, 16378,32)	1568
Dense-1 (Dense)	(None, 16378,32)	1056
Dense-2 (Dense)	(None, 16378,32)	1056
Flatten-1 (Flatten)	(None, 524096)	0
Dense-3 (Dense)	(None, 3)	1572291
total number of trainable parameters		1,576,035

5.4Evaluation Measures

A common performance metrics have been used, like accuracy of classification and loss. The accuracy is 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 (5)). This performs to look at the sample’s numbers with correct classification according to a ratio for a total number of the samples test.

$$\text{Accuracy} = \frac{\text{TP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \qquad (5)$$

According to the proposed approach the error score is calculated by using a loss function. As seen in (6), where m is the genes number,y_i is the class label which is actual, and $y_i'$is a predicted one. Categorical cross-entropy identifies the loss during the outcomes of categorical are a non-binary, which is >2. The outcome can be: (YES / NO / MAYBE) or (class1/ class2/…../class n).

$$\text{Loss} = -\sum_{i}^{m} y_i' \, log_2 \, y_i \qquad (6)$$

6. Result and Discussion

The suggested methodology was conducted for proving the efficacy of selecting the informative genes using PCA and SVD gene selection methods. Assessing the behavior of the classification algorithm to classify data of gene expression and identify optimal gene numbers. Firstly step, the gene expression data are read. Then, dataset is normalized using Min-max normalization. The gene selection methods have been applied to get a smaller genes number which should be close to the samples number.Table 3 presents description for the AD dataset according to the original genes number and the genes which are selected by using PCA and SVD. Because of the most genes in the original dataset are irrelevant at the predicting of the class label. Therefore, the proposed gene selection methods give lower informative genes which maximizing the classification performance through neglect the genes which are not relevant.

Table 3: The summary of the selected data

Methods	Samples	Original Genes	Selected Genes
PCA	569	16382	565
SVD			500

The proposed approach using PCA with CNN model can enhance classification accuracy on AD dataset (97.24%) when compared to raw dataset with no gene selection methods and other gene selection algorithms. Table 4 illustrates comparative results of average classification accuracy and loss. The other gene selection method using SVD performs well with CNN model (98.99%) when compared against other gene selection algorithms.

Table 4: Comparative results of average Accuracy and Loss on the AD dataset

Methods	CNN	
	Accuracy	Loss
Raw data	83.92%	0.6952
PCA	97.24%	0.4614
SVD	98.99%	0.2588

To build our model, the proposed framework was implemented in Anaconda Python 3.6 and in Keras Deep Learning Library. This conducted the classification 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.

7. Conclusion

In this work, a convolutional neural network model was proposed for classifying multi class microarray dataset. The dataset is normalized using Min-max normalization. The PCA and SVD are used as gene selection methods for overcoming the dimensionality curse and different issues that related to the data nature. To validate the proposed approach performance, the measures of evaluation namely accuracy and loss have been implemented. Categorical cross-entropy is used because it is a common loss function and also is suggested for problems with non-binary classification. To the purpose of optimization, ADAM is applied. The dataset results reveal that the proposed approach performance could lower the dimensional data problem by obtaining a subset including informative for increasing the accuracy of classification. In fact, the proposed system is not only providing a small subset for AD classification but also achieving higher accuracy of classification with short processing time. With regard to the future work, our planning is to develop the proposed system and applying it on datasets which show lower accuracy than the raw dataset with no applying gene selection. Even though the proposed method can reduce data dimensions and hence expected to remedy the over-fitting problem, it still needs further improvement to perform well on every dataset.

8.Acknowledgment

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

9.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 et al., "Systematic Analysis and Biomarker Study for Alzheimer's Disease," Scientific Reports, vol. 8, no. 1, pp. 1-14, 2018.
3. M. Panda, "Elephant Search with Deep Learning for Microarray Data Analysis," no. 2006, 2017.
4. M. N. F. Fajila, and R. D. Nawarathna, "New Feature Selection Method for High Dimensional Gene Data," 2018.
5. 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.
6. M. Lenz, F. J. Muller, M. Zenke, and A.Schuppert, "Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data," Scientific Reports, vol. 6, no. November 2015, pp. 1-11, 2016.
7. K. Güçkıran, I. Canturk, and L.Ozyilmaz, "DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature," vol. 23, no. 1, 126-132, p. 7, 2019.
8. 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," Elsevier, vol. 140, p. 10, 2020.
9. 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.
10. C. Devi ArockiaVanitha, 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.
11. L. Scheubert, M. Lustrek, 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, 2012.
12. X. Huang et al., "Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning," BMC Neurology, vol. 1, no. 18, pp. 1-8, 2018.
13. P. Danaee, and R.Ghaeini, "A deep learning approach for cancer detection and relevant gene identification," pp. 219-229, 2017.
14. M. Barati, and M. Ebrahim, "A Gene Expression Profile of Alzheimer's Disease Using Microarray Technology," 2016.
15. M. M. Babu, "An Introduction to Microarray Data Analysis," p. 225–249, 2004.AL-Mshanji and S. Z. AL-Rashid, "Improving Clustering Algorithm for Gene Expression Data Using Hybrid Algorithm," An international journal of advanced computer technology, vol. 8, 2019.
16. 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.

17. 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.
18. 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.
19. R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "Deep gene selection method to select genes from microarray datasets for cancer classification," BMC Bioinformatics, vol. 20, no. 1, pp. 1-15, 2019.
20. D. H. Lim, "Principal Component Analysis using Singular Value Decomposition of Microarray Data," International Journal of Mathematical and Computational Sciences, vol. 7, no. 9, pp. 1390-1392, 2013.
21. M. U. Ali, S. Ahmed, J. Ferzund, A. Mehmood, and A. Rehman, "Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data," International Journal of Advanced Computer Science and Applications, vol. 8, no. 5, pp. 415-426, 2017. N. Parveen1, H. H.Inbarani, and E.N. Sathish, " Performance analysis of unsupervised feature selection methods," International Conference on Computing, Communication and Applications (ICCCA), 2012.
22. H. Vural, and A. Subaşı, "Data-Mining Techniques to Classify Microarray Gene Expression Data Using Gene Selection by SVD and Information Gain," Modeling of Artificial Intelligence, vol. 6, no. 2, pp. 171-182, 2015.
23. N. Varghese, V. Verghese, "Asurvey of dimensionality reduction and classification methods," International Journal of Computer Science & Engineering Survey (IJCSES), vol. 3, no. 3, pp. 45-54, 2012. Mirzal, "SVD based gene selection algorithm," vol. 285 LNEE, no. December 2013, pp. 223-230, 2014.
24. N. M. Khalifa et al., "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. 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.
26. S. Kiranyaz et al., "1D Convolutional Neural Networks and Applications: A Survey," pp. 1-20, 2019.
27. S. Sakib et al., "An Overview of Convolutional Neural Network: Its Architecture and Applications," no. November, 2018.
28. M. Mostavi et al., "Convolutional neural network models for cancer type prediction based on gene expression," Studies in Health Technology and Informatics, vol. 267, pp. 181-186, 2019.
29. 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.
30. H. S. Basavegowda, and G. Dagnew, "Deep learning approach for microarray cancer data classification," CAAI Transactions on Intelligence Technology, vol. 5, no. 1, pp. 22-33, 2020.