

# A Deep Learning Model for Prediction Alzheimer's disease Based on Microarray Gene Expression Data

Suhaam Adnan Abdul kareem

Postgraduate Affairs

University of Baghdad

Baghdad, Iraq

Email: [sehamadnan2@gmail.com](mailto:sehamadnan2@gmail.com)

DOI: <http://dx.doi.org/10.31642/JoKMC/2018/090204>

Received Apr. 5, 2022. Accepted for publication Aug. 1, 2022

**Abstract**—Alzheimer's disease (AD) is a major productive neurological illness with complicated genetic architecture. One of the main aims of biomedical research is to identify risk genes and then explain how these genes contribute to disease development. As a result, it is required to increase the list of genes linked to Alzheimer's disease. Genes play a crucial role in every biological activity. Microarray technology has given genes access to a large number of genes, allowing them to evaluate several levels of expression at the same time. Microarray datasets are categorized by a huge number of genes and small sample sizes. This reality is referred to as a multidimensional curse with a difficult task. A promising technology known as gene selection is addressing this issue and has the potential to revolutionize Alzheimer's disease diagnosis. In this work, gene selection approaches such as Singular Value Decomposition (SVD) and Principle Component Analysis (PCA) were used. Techniques can help to minimize an amount from trivial and redundant gene in the unique datasets. Then, using the Convolutional Neural Network (CNN) as a classifier, deep learning (DL) is used to predict AD. The dataset was processed using a CNN with seven layers and varied settings. With the AD dataset, the empirical findings reveal that the PCA-CNN model has 96.60 percent accuracy and a loss of 0.3503, while the SVD-CNN model has 97.08% accuracy and a loss of 0.2466. As a result, the suggested approach is suited for reducing gene dimensions and improving classification accuracy by choosing a subset of relevant genes.

**Keywords**— Deep Learning, Alzheimer's disease; Gene Expression data; Microarray Technology; Classification, Gene Selection; (key words)

## I. INTRODUCTION

Alzheimer's disease is a degenerative memory and cognition impairment that affects millions of people worldwide. The language and memory-related nerve cells in the brain are damaged as a result of this. The symptoms begin to show after 65 years, and as people get older, the prevalence increases dramatically. This is a frequent dementia form[1][2]. AD is the leading cause for 60-80% of all illnesses. By 2050, the number of persons with Alzheimer's disease in the United States is expected to increase by 2030; dementia is expected to cost \$2 trillion globally, rising from 5.4 million to between 11 and 16 million people. Despite this startling figures, there isn't any reliable means for detecting illness before it becomes symptomatic, it may be the sole opportunity to interfere in the disease's progression

[3]. Because of a significance of AD and the lack of a precise treatment, the genes that cause the illness have been identified using a novel approach called microarray. Biologists employ microarray technology to measure gene expression levels in specific organisms. The focus of microarray data analysis is on identifying the optimum treatment for a variety of illnesses and precise medical diagnosis for a variety of genes through a variety of practical instances[4][5]. However, the great complexity of gene expression data obtained by microarray methods is problematic. Genes that are redundant or unnecessary can be deleted without causing serious data loss. A biggest issue in analyzing microarray data is the large number of genes and samples. It might lead to a decrease in predictive performance as well as an

increase in over-fitting problems [6]. A strategy to solve this challenge known as a "gene selection approach" isolates the optimal set of traits (genes) for building classification models. Gene Selection (GS) is the process for choosing a small group from the large group of genes a larger collection of genes that only contains informative genes. Researchers can obtain a lot of insight into the genetic nature of disease by looking at this subset of genes. This strategy has the potential to minimize cost computation and improve classification efficiency for AD[7] . Various techniques, like as PCA and SVD, can be used to pick genes. These algorithms are typical unsupervised approaches for analyzing gene expression microarray data, and they give details on the overall framework of the dataset being studied. They've recently been used to synthesize low-dimensional gene expression data before categorization on very big datasets [8]. Microarray data classification is a difficult task. The bioinformatics community is employing a variety of methods to diagnose and categorize microarray data using machine learning algorithms[9] .

The study uses the Deep Learning (DL) system for detect Alzheimer's disease (AD) utilizing genes expression data. Machine learning has a subset called deep learning. A deep learning algorithm, such as CNN, uses a large quantity of data to learn identify the unknown class label based on the behavior of genes using training set. Furthermore, using CNN architecture to increase predictive accuracy is a possibility. We also place a premium on the accuracy of the categorization after using gene selection, rather than only the methods of gene selection.

The remainder of a research project's components is outlined: The relevant work is shown in Section 2. A history of microarray technology is covered in section 3. A materials and procedures utilized in our research are detailed in Section 4. The dataset, as well as gene selection and classification, are discussed in this section. Section 5 contains a detailed description of the suggested technique. Section 6 describes our simulation and outcomes. The conclusion and recommendations for further study are discussed in the last section.

## II. RELATED WORK

P,Zahra, et al.(2016) [10]. The study applied a novel technique to select disease-relevant genes from a certain microarray data set. To eliminate the noisy and redundant genes within the high dimensional microarray data, the study used the Fisher method. When the number of times that each gene appeared in various gene subsets was analyzed, the result was the last subset with very informative genes. The study found that the suggested method had good classification and selection performance that can achieve 94.55 classification precision using 44 genes only. At least 24 (55%) of the Alzheimer's disease is linked to certain genes. When genes were analyzed using GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes), AD-related terms and pathways were identified. The genes may serve as predictors of the disease and for finding new candidate genes.

H. Li et al .(2017) [11]. Presented a classification system for predicting AD from the dataset GSE5281, which is referred to it as the AD dataset. A Wrapper of Genetic Algorithm and Support Vector Machine (GA/SVM) is used as a feature selection method to select a subset of relevant genes that improves the performance of classification. Six different classification methods: Naive Bayes (NB), C4.5 (decision tree), K-Nearest neighbor (KNN), Random Forest (RF), SVM with Gaussian kernel and SVM with linear kernel have been used. The results indicated the accuracy of the above models as follows: (81.4), (78.9), (87.0), (87.0), (85.7) and (91.9) respectively.

M, Balamurugan, et al.(2017) [12]. The KNN Classification Algorithms that has been proposed for diagnosing and classifying Alzheimer's disease (AD), (MCI), in datasets, in accordance with dimensionality reduction. The (NACC) has made available a dataset called the (RDD-UDS) that allows academics to examine clinical and statistical datasets. The KNN approach has limitations based on the feature in the data; in the case of large amounts of information, the timing

and sensitivity of the prediction step depend on the amount and applicability of the data.

K, Sekaran, et al.(2019) [13] . In this study, numerical approaches and machine learning (ML) techniques are used to compare the gene expression profiles of people with Alzheimer's disease with healthy people. Finding genes with variable gene expression helps to identify the most useful genes in a big way. A method called Rhinoceros Search Technique, which is based on a global optimization meta-heuristic (RSA). Researchers have found 24 novel gene biomarkers as a result of RSA. Support Vector Machines, Random Forest, Nave Bayes, and (MLP-NN) are four supervised ML approaches that are used to categorize two different groups of samples. One of these models, the RSA-MLP-NN, proved to be extremely effective at differentiating between genes associated with Alzheimer's disease and healthy genes. The training set could potentially have a lot of noise, which would be a flaw in the study and could affect the model's performance.

C, Park, et al.(2020) [14]. The study recommended using deep learning to predict AD using large-scale gene expression (GE) and DNA methylation data. It is challenging to model Alzheimer's disease using a multi-omics dataset since it calls for combining various omics data and managing a significant amount of small-sample data. To solve this problem, we developed a novel yet straightforward method to reduce the number of features in the multi-omics dataset based on differentially expressed genes and differentially methylated positions. (AUC = 0.79%, 0.75%, 0.77%, and 0.77%, respectively) .Highest feasible computing speed a list of the paper's restrictions.

M. Karaglani ,et al. (2020) [15]. Presented a classification system for predicting AD from the dataset (GEO: GSE63060 and GSE63061), which is referred to it as the AD dataset. We utilized k-nearest neighbor (KNN) classification methods because they capture data features that share non-linear interactions and have robust performance using methods Bayesian statistic. The results indicated the accuracy of the

above models as follows: AUC: 0.73% (ANM1) AUC: 0.66% (ANM2).

N, Le, et al.(2021) [16]. In this study, gene expression microarray data were used to train our machine learning model to use 35 expression features. Classifier performance was outperformed by the 35-feature model on average (AUC 98.3 %). The approach used, which is inadequate for forecasting survival outcomes and even produces a prognosis that is completely at odds with the actual occurrence, is to blame for the paper's inadequacies.

### III. Microarray Technology

Microarrays are sometimes called DNA chips. They are made up of thousands of small patches of DNA that have been attached to a solid surface. Many copies of the same DNA sequence can be found in each place, a single gene in an organism is represented by this symbol. The spots are arranged in regular pen groups [17]. For each gene, the amount of expression is saved as a picture (CEL File). The data is then extracted from the image using specialized software[18]. Fig. 1 depicts the surfaces of a DNA microarray.

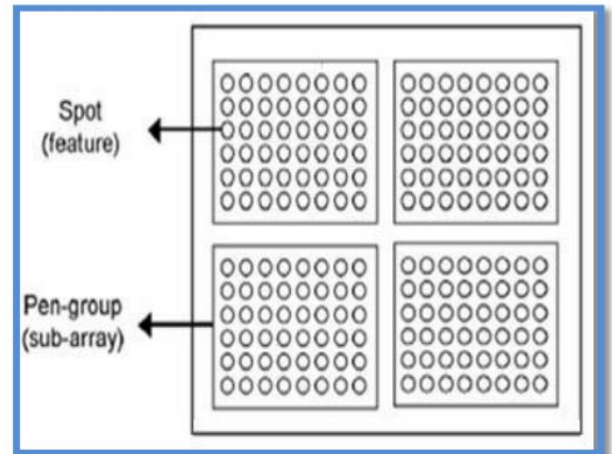


Fig 1. (A DNA Microarray's Surface) [18]

Several microarray companies provide their own software. For example, the most often used software is Limma. It is a component of analysis tools for raw CEL microarray data [19]. Genes responsible for various illnesses might be found

via analyzing and measuring the amount of gene expression in diseased and healthy cells [20].

Data from hundreds of distinct gene expressions is often stored in DNA microarrays. The dataset obtained by DNA microarray studies is often represented as a matrix, sometimes called a gene expression matrix, the sample is represented by a row, and the gene expression level are shown by the column. A collecting of this data serves as the foundation for any analysis[21].

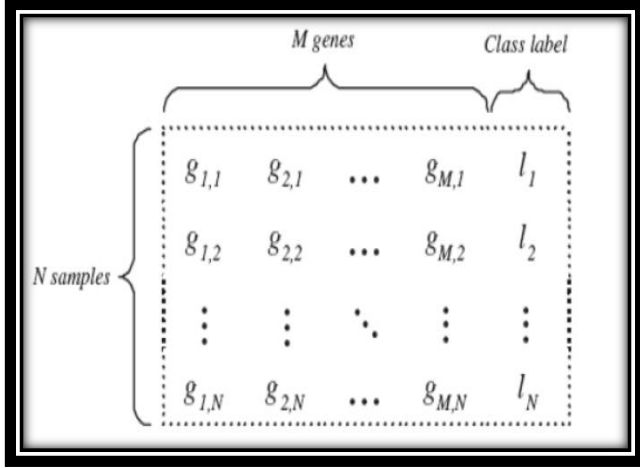


Fig 2. (The gene expression data matrix)

#### IV. Materials and Methods

##### A. Dataset

A dataset was collected of National Center for Bioinformatics Data's Gene Expression Omnibus (GEO: GSE63060 and GSE63061), which is a publicly available data source (NCBI). Then we combined the two datasets into one. The AD datasets comprises 16383 gene and 569 samples, including 245 individuals with Alzheimer's disease, 142 patients with mild cognitive impairment (MCI), and 182 healthy controls (CTL).

##### B. Gene Selection Methods

The expression of differences across numerous situations in many genes may be observed using a series of microarray experiments. Because the majority of genes are unrelated to the classification process, microarray data has a high dimensionality issue. As a result, gene selection strategies are

excellent in removing duplicated and unnecessary genes while also reducing data dimensionality [22]. The purpose of gene selection method is to locate a limited collection of genes that gets a high result [23], which can minimize computing reduces expenses and improves AD classifier accuracy. Several genes selection methods, such as Principal Component Analysis (PCA) and Singular value decomposition (SVD), have been used, were utilized in this work to identify useful genes that are directly linked to illness diagnosis.

- Principal Component Analysis (PCA)

Principal component analysis, a popular unsupervised technique for analyzing gene expression data, reveals details about the entire shape of the data. PCA is one of the most effective gene-selection methods[24]. PCA aims to transform high-dimensional data into a new, lower-dimensional subset of the original data. In advance of further research, significant gene information is extracted from a huge dataset using a principal component analysis testing [24]. It is good to know that using PCA to select genes aids in avoiding over-fitting, improving accuracy, and maintaining model simplicity while enhancing classification accuracy.

So it is suggested to discover important original genes for principle components using PCA, an unsupervised gene selection method based on eigenvector analysis. Assume that a dataset  $(X_1, X_2, \dots, X_m)$  has m-dimensional data, PCA projects m-dimensional data into a k-dimensional sub-space ( $k < m$ ). The steps for PCA are described below [39]:

1. Suppose X is an input matrix for PCA made up of an m-dimensional n-vector.
2. Use equation (1) to determine the mean data ( $\bar{X}$ ) for each dimension:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (1)$$

Where:

N: is the number of samples, and  $X_i$ : is the value of item i.

3. Calculate the covariance matrix ( $C_x$ ) using the following equation:

$$C_x = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}) (X_i - \bar{X})^T \quad (2)$$

4. Use equation(3) to determine the eigenvalues ( $\lambda_m$ ) and eigenvectors ( $v_m$ ) of the correlation matrix:

$$C_x v_m = \lambda_m v_m \quad (3)$$

5. Order the eigenvalues decreasing.
6. A group of eigenvectors known as the principal component (PC) corresponds to the sorted eigenvalues in step 5.
7. The eigenvalues will determine how to minimize the PC's dimension [40].

- Singular Value Decomposition (SVD)

SVD from the matrix X is the decomposition of the matrix X into a product of matrices, where X is the gene expression data matrix of size  $d \times n$  [25].

$$X = U \Sigma V^T = \sum_{i=1}^r \lambda_i u_i v_i^T \quad (4)$$

where  $r$  is the rank of a matrix X,  $U = [u_1, u_2, \dots, u_r]$  is a matrix with size  $d \times r$  with orthonormal columns,  $V = [v_1, v_2, \dots, v_r]$  is a matrix with size  $N \times r$  with orthonormal columns, and  $\Sigma$  is a matrix with size  $r \times r$  with the elements  $\lambda_1$  ( $\lambda_1 > 0, 1 \leq i \leq r, \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ ) located in the diagonal of matrix.

Below are some fundamental SVD matrices' mathematical characteristics.

- The square root of the eigenvalues  $s_1, s_2, \dots, s_r$  of the matrix  $X^T X$  are the singular values of the rectangular matrix X.
- The number  $l$  of rank's positive singular values determines the matrix X's rank ( $X$ ) =  $l, l \leq r$ .
- The biggest singular value is the same as the matrix X's Euclidean norm:  $\|X\|_2 = \lambda_1$ .

d. An orthonormal basis for the space occupied by the columns of matrix X is formed by the first  $l$  columns of matrix U.

e. The first  $l$  columns of matrix V serve as an orthonormal basis for the area covered by matrix X's rows.

- Features Modalities

Let matrix  $C_i$  as rows of matrix  $\Sigma V^T$  and from Equation 4 we get.

$$\begin{aligned} X &= \sum_{i=1}^r (u_i) (\lambda_i v_i)^T \\ &= \begin{bmatrix} u_1 \sqrt{\lambda_1} & u_2 \sqrt{\lambda_2} & \dots & u_r \sqrt{\lambda_r} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} v_1 & \sqrt{\lambda_2} v_2 & \dots & \sqrt{\lambda_r} v_r \end{bmatrix}^T \\ &= RC^T \end{aligned} \quad (5)$$

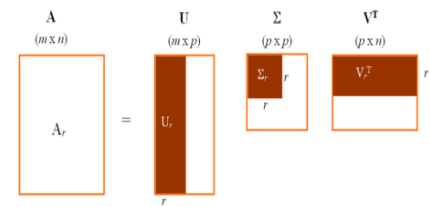


Fig.3. reduced SVD diagram matrix[25]

Characteristics mode connected to matrix X is called orthogonal vector  $C_i$ . It is simple to demonstrate that the expression of the  $j$ -th gene varies among the examined samples, as seen on Equation 3, where the coefficients of the combination are the appropriate entries of matrix X, may be represented precisely as a linear combination of the characteristic modes included in the row  $X_j$  of matrix X. Typically, only some of the characteristic modes are required to reasonably reconstruct the gene expression pattern. As seen in Fig. 3, we can use an incomplete SVD expression.

$$X_j = \sum_{i=1}^l U_{ji} C_i \quad l \leq r \quad (6)$$

Equation 6 gives us information on the analysis of singular values and demonstrates how the dimension of matrix X can be lowered by reducing the dimension of the characteristic

mode matrix. The original pattern of gene expression data can also be described using Ci [25].

### C. A Deep learning model classification for Alzheimer's disease

Deep learning is a type of artificial intelligence uses algorithms for replicate data processing and mental processes, as well as to construct abstractions. Algorithm layers are used by DL to process, analyze, and uncover data's hidden patterns. The deep network's layers send information to each other, as well as the output of a proceeding as the input layer for the next layer. A network's additional levels, hidden layers are those that exist between both the input and output layers. Each layer is usually basic, regular, and contains at least one activation function. Now it's accepted a promising technique for developed automatic detection system that can produce better findings, broaden the spectrum for illness classification systems, and do real-time medical diagnosis [26]. Convolutional Neural Networks are a common architecture for creating deep learning models, and they are described in this article (CNN). CNN is the most widely used supervised DL model for classifying Alzheimer's disease based on information on gene expression.

- Convolutional Neural Networks (CNN)

Convolutional neural networks are deep learning technique the mimics the brain's information processing function. This study proposes using multilayer CNN to identify gene expression patterns from microarrays. The Convolutional Neural Network has been recommended due of its capacity to handle enormous volumes to increase the accuracy of data categorization. Also, the Convolutional Neural Network is good at merging closely related datasets, which improves classification performance. This is due to its capacity to distinguish latent elements of Alzheimer's disease from other diseases [14]. The advantages listed below can be attributed to deep CNN's vast variety of application fields:

- CNN combines the selection and classification operations into a single learning unit. During the training phase, they

learn how to maximize the characteristics straight from a raw input.

- Because CNN neurons are connected to bound weights, it has the potential to process large inputs while still having a high computational efficiency.
- Small input data transformations like as scaling, encoding, distortion, and skewing are resistant to CNN.
- CNN can deal with a wide range of input sizes.

Many applications, as well as early diagnostic and structural health monitoring, and data categorization to tailored biomedicine, have recently been suggested and directly achieved contemporary levels of performance using 2D CNN. Another significant advantage is that the use of real-time data and low-cost technology allows for the 2D CNNs with a simple and compact design that only performs 2D convolution [14]. There's an input and output layer, as well as a slew of other layers that aren't visible, make up a CNN. Convolutional, pool, or thick layers, and short for completely linked layers, are the three types of these layers (FC). Fig 4 illustrates the CNN model.

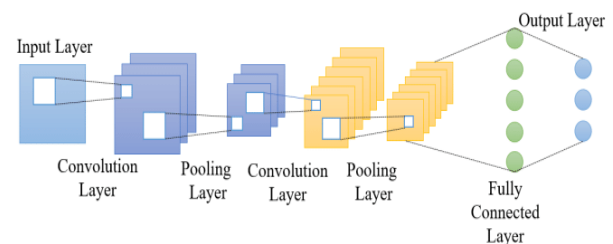


Fig 4. A convolutional neural network's foundational design [14] .

The CNN model in our research uses expression of genes as a vector; it employs 2D kernel at the input vector. Two convolution layers and two thick layers make up this model with a single layer of flattening. We'll refer to this model as 2D CNN for convenience's reason [27].

### V. Methodology

A suggested technique covers essential procedures including importing the raw microarray AD data set in this part .Then, as shown in Figure 5, normalization utilizing the Min – Max

methodology, the gene selection techniques, and classifier using a Convolutional Neural Network .

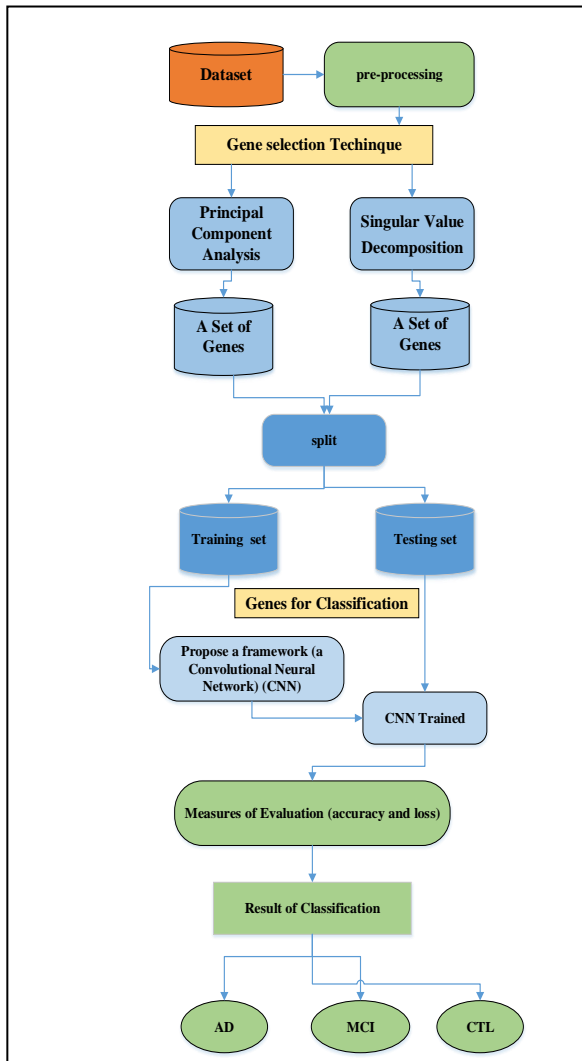


Fig 5: Proposed Method

#### A. Pre-processing Stage

Because the data obtained by microarray technology is noisy, data pretreatment is a must before beginning the studies. To limit the variance in expression measurements, the dataset must be standardized[28]. To normalize the data collection, use the Min-max normalization method. The levels of gene expression are determined such that each gene has a minimum and maximum value of zero and one [29].

#### B. Gene Selection

Gene selection stage procedures are designed to lower the dimensionality of a dataset's computing space. It's a method for selecting a subset of genes from the bigger dataset because genes are frequently uninteresting. Regardless of the machine learning methodologies, these strategies are usually applied prior to the development of algorithms for machine learning (ML) and the selection of genes based on measures. PCA and SVD gene selection methods were utilized to determine a subset of genes that are directly employed in categorization in this study.

#### C. Classification Stage

At this step, the gene expression data is classified deep convolutional neural networks are used in this model. CNN models have been configured once the data gathering, pre - processing, and gene selection processes are now complete. A CNN with multiple layers are selected. The convolutional layer's design was chosen because it can handle vast volumes of multidimensional data, such as gene expression data [14]. This study proposes a novel approach using a 2-Dimensional convolutional. The convolution layer was employed with a filter of 65 kernels of size 4 and non-linearity activation termed Rectified Linear Units. ReLU might be proven using the formula in eq. (7).

$$g(X) = 0 \text{ for } X < 0 \quad (7)$$

X for  $X \geq 0$  for the dataset, they used CNN architecture with six layers, the result can be seen in Table 2. A reality is a following layer's input layers have an identical amount of neurons as previous layer input layer. For the epoch, a size of 100 is selected. The total number of trainable parameters in our study is 1,476,035. At the conclusion of the final layer, the softmax activation function is also applied to increase network performance. Now, as seen in eq., a SoftMax function may be proven (8).

$$\text{SoftMax}(C) j = e^{c_j} / \sum_{k=1}^K e^{c_k} \quad (8)$$

The adaptive moment estimation (ADAM) optimizer is used to quantify the loss in training and testing data using a predetermined objective function, categorical cross-entropy[30]. The ADAM optimizer calculates the learning rate for each parameter. The system was developed with training takes up 80% of the data, while testing takes up 20%.

Table 1: An overview of the structure of a 2D CNN model

Layer type	Output Shape	No. variables
Conv2d-1 (Conv2D)	(15380, 15)	65
Conv1d-2 (Conv2D)	(15378, 33)	1558
Dense-1 (Dense)	(15378, 33)	1046
Dense-2 (Dense)	(15378, 33)	1046
Flatten-1 (Flatten)	(524095)	0
Dense-3 (Dense)	(3)	15722912

The total number of variables that may be trained is 1,676,035

#### D. Evaluation Measures

The definition of a performance metric is the measurement of outcomes. It generates trustworthy information regarding the suggested methodology's efficacy and efficiency. Using performance measurements like as accuracy and loss, the link between the input and output values of the suggested approach may be understood. Also, accuracy is one of the most useful assessment criteria for determining the efficacy of the suggested Alzheimer disease categorization methods. The simplest intuitive performance indicator is accuracy, which is just a ratio of all observations to accurately anticipated observations. Eq. (9) and (10) provide a generic formula for calculating accuracy and loss.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (9)$$

FP stands for false positive, TN stands for true negative, TP stands for true positive, and FN stands for false negative.

A loss function is used to determine the error score in the suggested technique, where N is the number of genes,  $X_i$  ' is

the real class label, and  $X_i$  ' is a projected one. Classifier cross-entropy indicates for loss that occurs when categorical outcomes are non-binary and  $>2$ .

$$Loss = - \sum X_i \log_2 X_i \quad (10)$$

## VII. Discussion and Result

A suggested strategy was utilized to illustrate the value of using PCA and SVD gene selection techniques to identify pertinent genes .Assessing the classification algorithm's performance in identifying the best gene numbers and classifying gene expression data. Reading the gene expression data comes first. The data is then normalized using the Min-Max approach. The number of genes was reduced using gene selection approaches such that it was close to the number of samples. According to the initial gene count and the genes picked, PCA and SVD were used to choose the genes, as shown in Table 2. The suggested gene selection procedures produce fewer informative genes while enhancing classification performance by removing irrelevant genes because the majority of the genes in the first dataset have minimal influence on class label prediction. Table 2: The selected data's summary.

Method	Samples	Genes	Selected Genes
PCA	568	16383	500
SVD			450

In compared to unprocessed datasets and alternative gene selection methods, the suggested methodology employing on the AD dataset, PCA combined with a CNN model can enhance classification accuracy was (96.60%). Table 4 shows the average classification accuracy and loss as a function of time. When compared to other gene selection methods, The

SVD-based gene selection method also works well with the CNN model (97.08 %).

Table 3. We examine average accuracy and loss.

Method	CNN	
	Accuracy	Loss
OriginalDataset		82.921%
	0.5952	
PCA	96.60	
	0.3503	
SVD	97.08	
	0.2466	

## Conclusion

The convolutional neural network (CNN) modeling for multiclass microarray samples is suggested in this article. The min-max technique is used to normalize a set of data. Two gene selection strategies, PCA and SVD, are well adapted to circumvent the dimensionality curse and other data-related issues. To verify the effectiveness of the suggested approach, performance measurements for accuracy and loss have been created. Classification cross-entropy is especially helpful for non-binary categorization issues because it is a loss function that frequently arises. Optimizing is ADAM's major objective. The outcomes of a dataset show that the suggested approach may lessen the issue of dimensional data by creating a subset that contains useful data to increase classification accuracy. The suggested method not only provides a smaller subset for diagnosing Alzheimer's disease, but it also improves categorization efficiency while consuming less processing time. Future study will focus on improving the recommended approach and testing it against datasets that are less accurate than the unprocessed

dataset with no gene selection. Although the recommended technique can reduce data dimensions and hence mitigate the over-fitting problem, it still has to be improved to perform effectively across all datasets.

## REFERENCES

- [1] H. Ahmed, H. Soliman, and M. Elmogy, "Early Detection of Alzheimer's Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, pp. 15–20, 2020, doi: 10.1109/ICDABI51230.2020.9325640.
- [2] S. T. Ahmed and S. M. Kadhem, "Early Alzheimer's Disease Detection Using Different Techniques Based on Microarray Data : A Review," *iJOE – Vol. 18, No. 04, 2022*, no. Mci.
- [3] Inzhong *et al.*, "Systematic Analysis and Biomarker Study for Alzheimer's Disease," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, 2018, doi: 10.1038/s41598-018-35789-3.
- [4] M. S. Othman, S. R. Kumaran, and L. M. Yusuf, "Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data," *IEEE Access*, vol. 8, pp. 186348–186361, 2020, doi: 10.1109/ACCESS.2020.3029890.
- [5] S. T. Ahmed, Q. K. Kadhim, H. S. Mahdi, and W. S. A. Almahdy, "Applying the MCMSI for Online Educational Systems Using the Two-Factor Authentication," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 13, pp. 162–171, 2021, doi: 10.3991/ijim.v15i13.23227.
- [6] W. Zhongxin, S. Gang, Z. Jing, and Z. Jia, "Feature Selection Algorithm Based on Mutual Information and Lasso for Microarray Data," *Open Biotechnol. J.*, vol. 10, no. 1, pp. 278–286, 2016, doi: 10.2174/1874070701610010278.
- [7] J. Zahoor and K. Zafar, "Classification of microarray gene expression data using an infiltration tactics optimization (Ito) algorithm," *Genes (Basel)*, vol. 11, no. 7, pp. 1–28, 2020, doi: 10.3390/genes11070819.
- [8] M. Babu and K. Sarkar, "A comparative study of gene selection methods for cancer classification using microarray data," *Proc. - 2016 2nd IEEE Int. Conf. Res. Comput. Intell. Commun. Networks, ICRCIN 2016*, no. September 2016, pp. 204–211, 2017, doi: 10.1109/ICRCIN.2016.7813657.
- [9] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings, Twent. Int. Conf. Mach. Learn.*, vol. 2, pp. 856–863, 2003.
- [10] S. Zahra Paylakhi, S. Ozgoli, and S. Paylakhi, "Identification of Alzheimer disease-relevant genes using a novel hybrid method," *Prog. Biol. Sci.*, vol. 6, no. 1, pp. 37–46, 2016, [Online]. Available:

- <http://www.ncbi.nlm.nih.gov/pubmed>.
- [11] H. Li *et al.*, "Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017, doi: 10.1038/s41598-017-13700-w.
  - [12] M. Balamurugan, A. Nancy, and S. Vijaykumar, "Alzheimer's disease diagnosis by using dimensionality reduction based on KNN Classifier," *Biomed. Pharmacol. J.*, vol. 10, no. 4, pp. 1823–1830, 2017, doi: 10.13005/bpj/1299.
  - [13] K. Sekaran and M. Sudha, "Diagnostic gene biomarker selection for alzheimer's classification using machine learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 2348–2352, 2019, doi: 10.35940/ijitee.L3372.1081219.
  - [14] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Med. Genomics*, vol. 13, 2020, doi: 10.1186/s12920-020-0677-2.
  - [15] S. Perera, K. Hewage, C. Gunarathne, R. Navarathna, D. Herath, and R. G. Ragel, "Detection of Novel Biomarker Genes of Alzheimer's Disease Using Gene Expression Data," *MERCon 2020 - 6th Int. Multidiscip. Moratuwa Eng. Res. Conf. Proc.*, pp. 1–6, 2020, doi: 10.1109/MERCon50084.2020.9185336.
  - [16] N. Jameel and H. S. Abdullah, "A Proposed Intelligent Features Selection Method Using Meerkat Clan Algorithm," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012061.
  - [17] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Syst. Appl.*, vol. 140, p. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.
  - [18] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen, "Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC Bioinformatics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2105-13-266.
  - [19] K. Nishiwaki, K. Kanamori, and H. Ohwada, "Gene Selection from Microarray Data for Alzheimer's Disease Using Random Forest," *Int. J. Softw. Sci. Comput. Intell.*, vol. 9, no. 2, pp. 14–30, 2017, doi: 10.4018/ijssci.2017040102.
  - [20] S. Z. Paylakhi, S. Ozgoli, and S. H. Paylakhi, "A novel gene selection method using GA/SVM and Fisher criteria in Alzheimer's disease," *ICEE 2015 - Proc. 23rd Iran. Conf. Electr. Eng.*, vol. 10, pp. 956–959, 2015, doi: 10.1109/IranianCEE.2015.7146349.
  - [21] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks," *Adv. Bioinformatics*, vol. 2015, pp. 1–9, 2015, doi: 10.1155/2015/639367.
  - [22] S. Taha Ahmed and S. Malallah Kadhem, "No Title," *Int. J. Interact. Mob. tchnologies(iJIM)*, vol. 15, no. 16, p. 95, 2021.
  - [23] N. Q. K. Le, D. T. Do, T.-T.-D. Nguyen, N. T. K. Nguyen, T. N. K. Hung, and N. T. T. Trang, "Identification of gene expression signatures for psoriasis classification using machine learning techniques," *Med. Omi.*, vol. 1, no. December 2020, p. 100001, 2021, doi: 10.1016/j.meomic.2020.100001.
  - [24] M. Lenz, F. J. Muller, M. Zenke, and A. Schuppert, "Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data," *Sci. Rep.*, vol. 6, no. June, 2016, doi: 10.1038/srep25696.
  - [25] D. H. Lim, "Principal Component Analysis using Singular Value Decomposition of Microarray Data," vol. 7, no. 9, pp. 1390–1392, 2013, [Online]. Available: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.943.4177&rep=rep1&type=pdf>.
  - [26] Y. Zhang, J. M. Gorriz, and Z. Dong, "Deep learning in medical image analysis," *J. Imaging*, vol. 7, no. 4, p. NA, 2021, doi: 10.3390/jimaging7040074.
  - [27] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, p. 107398, 2021, doi: 10.1016/j.ymssp.2020.107398.
  - [28] T. Ragunthar and S. Selvakumar, "Classification of gene expression data with optimized feature selection," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 4763–4769, 2019, doi: 10.35940/ijrte.B1845.078219.
  - [29] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014, doi: 10.1007/s00521-013-1368-0.
  - [30] K. Lunnon *et al.*, "A blood gene expression marker of early Alzheimer's disease," *J. Alzheimer's Dis.*, vol. 33, no. 3, pp. 737–753, 2013, doi: 10.3233/JAD-2012-121363.