Classification of COVID -19 Disease Using Genes Expression and Deep Learning Technique

Eman Hamid Hadi,[1], Hussein Attya Lafta[2], Sura Z. Al_Rashid[3]

[1]College for Women, University of Babylon, Babylon, Iraq. Email aeminhmed@gmail.com[2]Hussein Attya Lafta works at College of Information Technology, University of Babylon, Babylon, Iraq. Email: hzazmk@yahoo.com [3]Sura Z. Al_Rashid works at College of Information Technology, University of Babylon, Babylon, Iraq. Email:sura_os@itnet.uobabylon.edu.iq

University of Babylon, Iraq

Article Info

Page Number: 3721 - 3731

Publication Issue:

Vol 71 No. 4 (2022)

Article History

***Article Received**: 25 March 2022*

***Revised**: 30 April 2022*

***Accepted**: 15 June 2022*

***Publication**: 19 August 2022*

Abstract

Coronavirus disease 2019 (COVID-19) has spread very quickly among individuals all over the world. And because of the increasing number of cases every day compared to the small quantities of ready-made tests in hospitals. Therefore, it has become necessary to introduce different systems to detect and diagnose this disease to prevent its spread among people. The purpose of this study is to propose a new method using gene expression and deep learning methods to identify patients with COVID-19. Several preprocessing methods have been applied as a method for feature extraction and identification of genes associated with COVID-19 disease severity. Artificial Neural Networks and Convolutional Neural Networks were applied to the COVID-19 dataset. The highest classification accuracy was (91%) using ANN, and the highest classification accuracy using CNN was (87%).

Keywords: Covid-19 , Gene Expression , Corona Disease ,Deep Learning , CNN

I. Introduction

The current development of the novel coronavirus disease, which originated in China, is caused by a global health crisis[1] [2]. To date, the COVID-19 pandemic has affected more than 200 regions and countries with more than 188,655,968 verified cases including 4,067,517 global response loss to a deadly infection or virus that has killed thousands of people [3] [4] This latest epidemic is growing extraordinarily fast, [5]. Therefore, in this study, gene expression analysis will be performed to discover the genes responsible for the cytokine storm that are directly related to the severity of corona disease. [6] . The aim of this work is to build a prediction model based on deep learning represented by ANN, CNN. This paper is divided into five parts, the literature review in section 2, section 3 describing the classification technique used, the classification results in section 4, and the conclusion in section 5

II. Related Works

Nahida Habib et al, in 2021 [4] ,adopted two different methods of detection were proposed the first depends on genes and the screening method for detecting corona diseases, where a random model based on the rules of the random forest was trained with an accuracy close to 93%. The second diagnostic technique proposed is image classification using chest x-rays to classify normal images against COVID-19 and pneumonia and was done using Deep-CNN technology, achieving a test accuracy of 99% .

Ahmed and Jeon, in 2021[2] ,have presented work on genome sequencing analysis of COVID-19 and similar viruses such as SARS, MERS and Ebola. They used different visualization methods to analyze the genome sequences of these viruses, and applied the Support Vector Machine machine learning algorithm to classify the different genome sequences. 97% overall classification accuracy for COVID-19, 96%, SARS, and 95% for MERS and Ebola genome sequencing, respectively.

Hilal Arslan , in 2020 [3] , adoptd a research paper, which uses different machine learning based on classification methods to classify COVID-19 against other types of coronavirus, using features extracted from genome sequences. These classification methods are support vector machines, naive Bayes, K-nearest neighbor, random forest, and decision tree. Use the 2019 Novel Coronavirus Resource Database. The results shown showed that the decision tree achieved a classification accuracy of 93%.

Milad Mostavi1 and et al , in 2021 [7] , they suggested a model for building machine learning models to predict the inference of important cancer genes. Several Convolutional Neural Network (CNN) models were presented that take unregulated gene expression inputs in order to classify tumor and non-tumor samples into specific or normal types of cancer. Based on different designs of gene inserts and convolutional plots, three CNN models were implemented: 1D-CNN, 2D-Vanilla-CNN and 2D-Hybrid-CNN. The models were trained and tested on gene expression profiles from 10,340 samples from 33 cancer types and 713 normal tissues matched with the Cancer Genome Atlas (TCGA). These models achieved prediction accuracy (95%) .

Babu Karthik and et al., in 2020 [8] , this research aimed To provide the optimal solution for identifying COVID-19 pneumonia and healthy lung using Image (CXR). Deep learning is one of the great techniques used to extract high-dimensional features in medicine. In this research, the latest technology, Genetic Deep, was used Learning Convolutional Neural Network (GDCNN), it is trained from scratch to extract features to classify them between COVID-19 and normal images. GDCNN training begins From the outset that the proposed method performs better compared to other learning techniques, this classification achieved 98% accuracy, 100% sensitivity, and 97% specificity, respectively.

III.Methodology

The proposed system consists of two stages: the first stage is the preprocessing stage, which consists of a number of techniques, which is to initially process missing values and normalization, and then apply the feature selection technique, which consists of a number of ways to reduce high-dimensional data to select genes associated with corona disease from the Covid -19 dataset. . The second stage was the application of the proposed model represented by the Convolutional Neural Network (CNN) to identify the genes associated with the severity of Covid-19 disease and the cause of death. Figure (1) represents the proposed system for identifying genes associated with the severity of COVID-19 disease and that causing death.

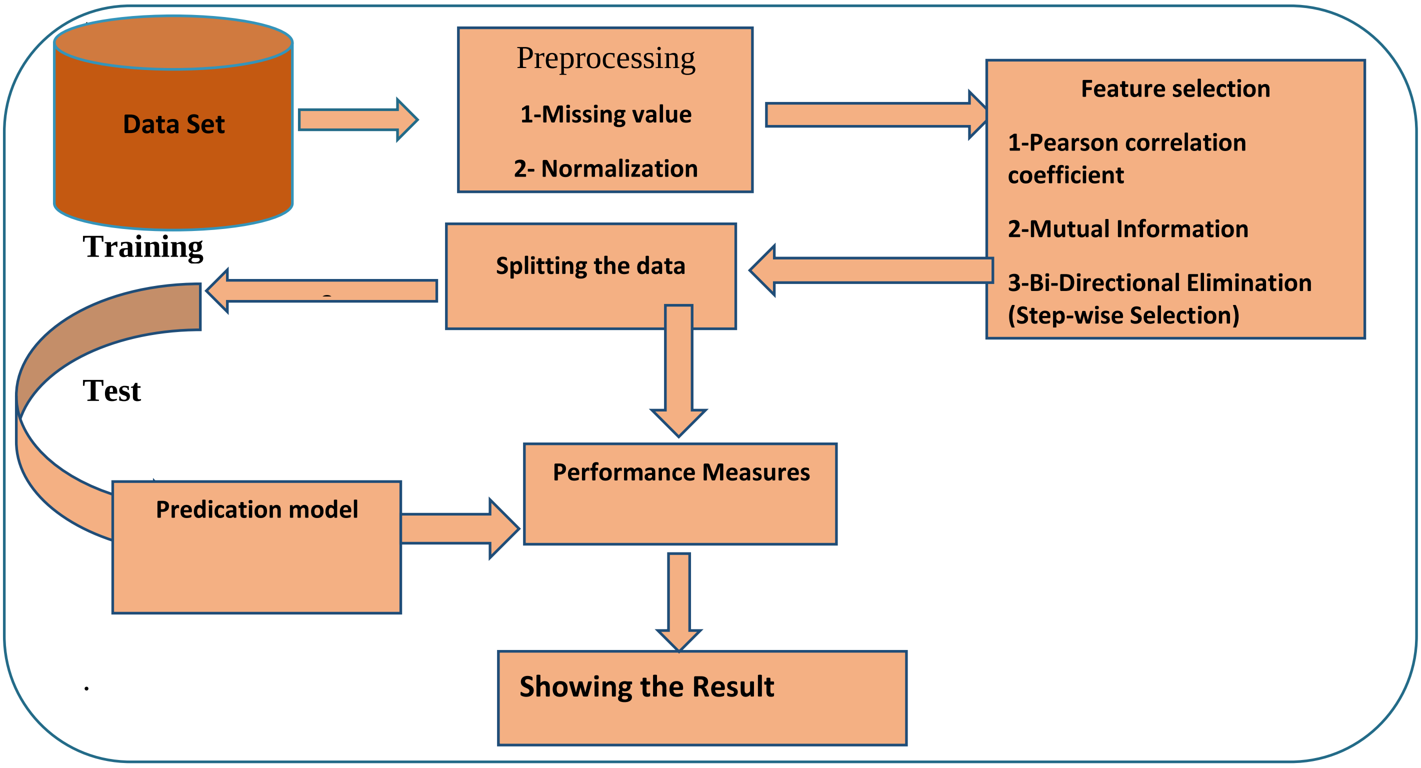


Figure (1) The proposed system

A.Data Set

In this study, the data set was collected from the National Center for Biological Information (NCBI), the data set was taken from the publicly available data source Gene Expression Omnibus (GEO). Which was published on November 23, 2020 by the National Center for Biological Information (NCBI), the data access number is (SE33267) provided by Gene Reports Journal. [5].

Table (1) presents a brief description of the COVID-19 dataset

Title of The Dataset:	***Covid -19 dataset***
1-Dataset characteristics	Multivariate
2-Number of Samples	99
3-Number of gene	33629
4-Number of Class Labels	3

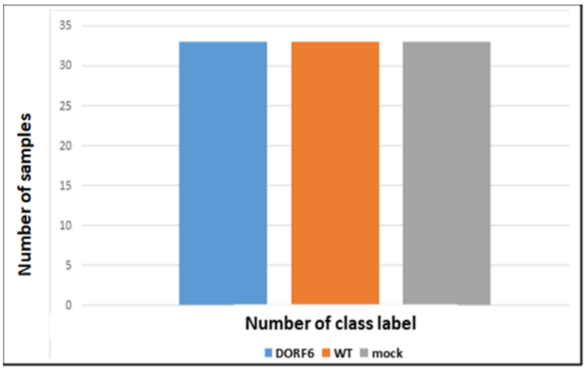


Figure (2): The Number of Class Labels in the COVID-19 Dataset

B . Preprocessing

Data pre-processing is an important step in the data mining process. There are many data pre-processing techniques. Including handling missing values and reducing the size of the data by collecting and deleting redundant features, and normalization can be applied where the data is scaled to fall within a smaller range (0.0 and 1.0) [9]. After applying the pre-processing of the data and obtaining the appropriate results, the final obtained data set can be considered as a reliable source and can be used in any algorithm that is applied to extract the data.

1- Normalization

Normalization aims to ensure that all data used are in the same unit of measure, i.e. between [0, 1] or between [1, -1] [10] . Therefore, a normalization process is used to remove the difference between the influence of small and large values that dominate the results. In general, normalization methods are applied to the data to reduce the error and increase the accuracy of the model used. Equation (1) [11] is used in the minimum and maximum normalization so that all data used are within the range (0 and 1) .

$$v' = \frac{v - \min a}{\max a - \min a}(\text{new\_ max a} - \text{new\_ min a}) + \text{new\_min a} \qquad (1)$$

2-Missing Value

The missing data processing is very important during the preprocessing of the used data set because many deep learning algorithms do not support missing values. There are many ways in which missing values can be handled: either by deleting rows or columns containing null values or by replacing the missing values with (0) and other methods [12].

3- Feature selection

Feature selection methods as a preprocessing step in predictive modeling have several important advantages. It can reduce model complexity, enhance learning efficiency, and increase predictive power by reducing noise. One of the methods of selecting the feature that was used in this research is:

1- Mutual Information Method (MI)

This method demonstrates the strength of the statistical association of two random variables, one dependent and one independent [13] . A mutual information estimator based on minimum entropy is used as a classifier to detect different types of genes related to corona disease severity. The mutual information is calculated as shown by the following equation:

$$MI(G,C) = H(G) + H(C) - H(G,C) \qquad (2)$$

2- Principal Component Analysis method

Principal component analysis is one of the most important ways to reduce dimensionality. It is often used to reduce the dimensionality of large data sets, by converting a large set of variables into a smaller set without losing information in the large set. The main idea of this method is to reduce the dimensions of a data set consisting of a large number of variables related to each other, while maintaining the variance in the data set, and the same can be done by analyzing the variables into a new set of variables known as principal components [14]. :

- $\boldsymbol{C_x v_m = \lambda_m v_m}$ **(3)**
- **Where:**
- $\boldsymbol{v_m}$is the eigenvectors and λ_m is the eigenvalue of the covariance matrix.

By relying on eigenvalues, the dimensions of the Principal components (PCs) will be reduced

3- Pearson correlation coefficient

The correlation coefficient (ρ) is a measure that determines the degree to which two different variables are related. Pearson's correlation is one of the most common correlations for measuring the linear relationship between two variables. The range of correlation coefficient values is -1 to 1. A correlation of -1 indicates a negative correlation and a correlation of 1 indicates a positive correlation. . A value of zero indicates that there is no relationship between the two variables. Related genes are determined by the Pearson correlation coefficient. The association between two objects X = (1, $x2$, $x3$, ... xn) and Y = (1, $y2$, $y3$, ... yn)) [15] is described as follow

$$\mathbf{Pxy} = \frac{\Sigma_{i=1}^{n}(X\,i - \bar{x})(Y\,i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(X\,i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(Y\,i - \bar{y})^2}} \quad \mathbf{(4)}$$

C- Predication Model

1-Artificial Neural Network (ANN)

An artificial neural network (ANN) is represented by a number of computational nodes that are interconnected with each other, and ANN represents a distinctive computational method for problems in which it is difficult to find a suitable solution [16].

Multi-Layer Perceptron (MLP) Structure

Multi-Layer Perceptron’s (MLPs) is a feed forward architecture of ANN that have one or more than one hidden layer, each of them consists of a set of computationally simple interconnected nodes which called the neurons as shown in Figure (3)

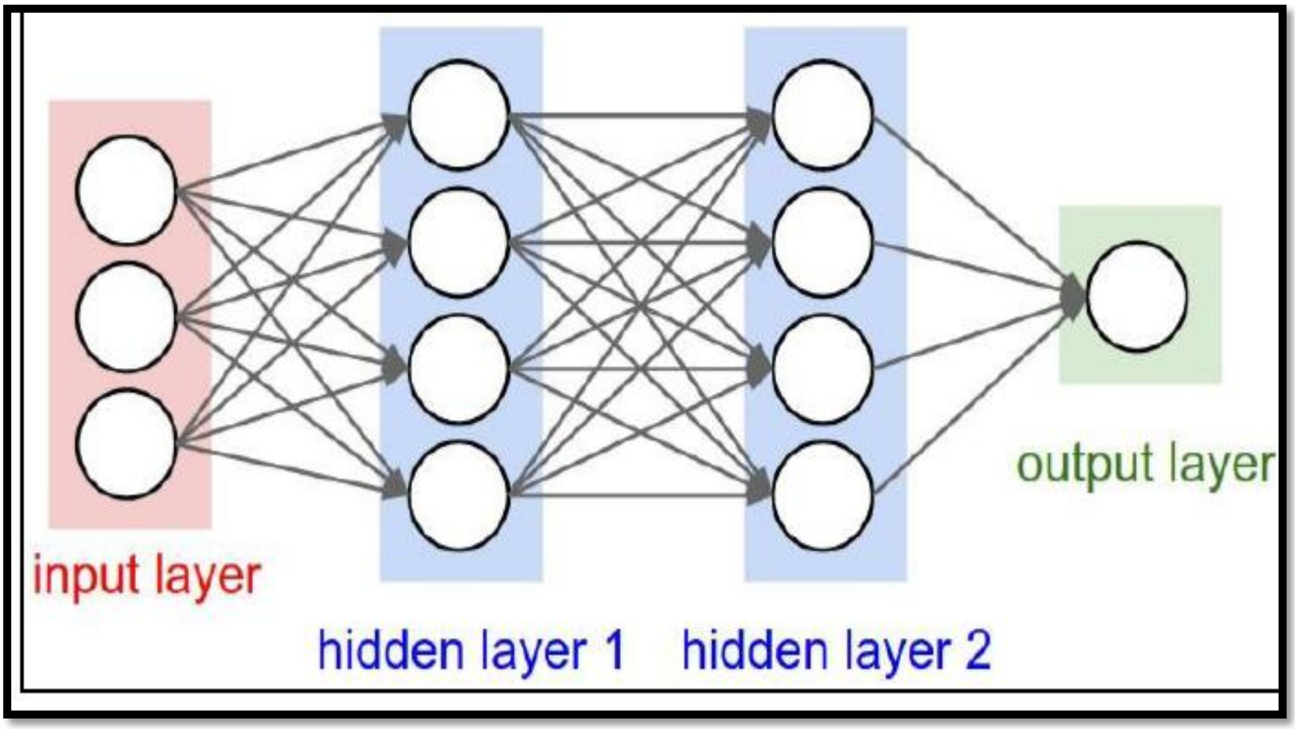


Figure (3):The basic structure of an MLP [16]

In this paper, the Artificial Neural Network (CNN) model includes two stages. The first is the design of the form. The second is the training parameters**.**

1-ANN design

The ANN design include arranging the different layers in order to get the best feature extraction process.. The layers of the ANN are as follow: the first layer is dense : these layer with 20 node and activation function is Relu, the second layer is dense :these layer with 40 node and activation function is Relu, the 3 layers is dense : these layer with 20 node and activation function is Relu, the 4 layer is dense : these layer with 40 node and activation function is Relu, - the 5 layers is dense : these layer with 3 node that represent three class ,with activation function is sigmoid.

2-Training hyper-parameters

Splitting the data set: In this step, after the dataset pre-processing process is finished, the data set is uploaded to the proposed system. The data set is divided into 65% training and 35% testing groups ,Optimizer: The optimization process is about finding the best parameters and values for the kernel and biases, to fine tune the training process by comparing all the available optimizers, an optimizer was selected (Adam),Epochs:150,Batch size:15

2-Convolutional Neural Network (CNN)

CNN is one of the most important and widely used types of deep neural networks in the fields of machine vision. [17]. CNN is a promising tool for improving automated diagnostic systems and achieving high accuracy for disease.It is one of the most widely used neural networks in the field of artificial intelligence, because it has a large capacity to process a large amount of data and does not need to extract features manually, and does not need the complexity of hashing A CNN consists of a number of layers that pass information through layers as the output of the previous layer is fed to the next layer. The first layer of the network is called the input layer, while the last layer of the network is called the output layer. There are also hidden layers of the network between the layers of input and output. Each layer is a simple algorithm containing one type of activation function. [18].

CNN Architecture

The first layer of a convolutional neural network is the input layer that reflects the model's input (selected features), but this layer does not compute with the number of CNN layers. In general, when studying gene expression data that needs to be analyzed using a CNN, the input layer is a two-dimensional matrix of size (n × m) where n is the number of samples and m is the number of features. A distinct neural network consists of a number of layers, as shown in Figure (2). These classes are:

1.Convolutional layer.
2. Max pooling layer (or Sub Sampling layer).
3. Activation Layer

4. Fully Connected Layer (Classification layer).
5. Dropout Layer

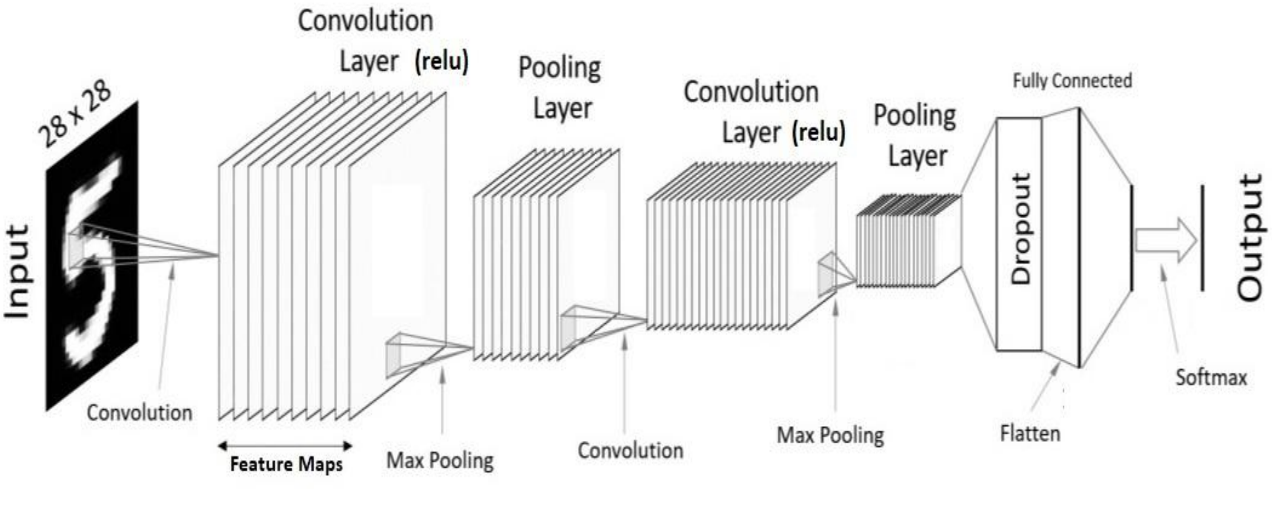


Figure (4) shows the architecture of a CNN network [19].

In this paper, the Convolutional Neural Network (CNN) model includes two stages. The first is the design of the form. The second is the training parameters.

1-CNN design

the first layer is convolutional layer (1D): With 16 filters the layer will learn. With 3*3 kernel size, padding (same), and the used activation function is ReLU, Layer flatten: this layer will transform the previous feature map matrix into single column of values. So it will be appropriate to be entered to the next fully connected layer, Layer 3 Dense layer : In the Three layer, the number of nodes is 64 node and the activation function is Relu,Layer 4 Dense layer: In the Three layer, the number of nodes is set to 3 (to represent the three classes of DORF, mock,wt. And the activation function is Softmax.

2-**Training hyper-parameters:**

Splitting the data set: In this step, after the dataset pre-processing process is finished, the data set is uploaded to the proposed system. The data set is divided into 65% training and 35% testing groups,**Optimizer**: The optimization process is about finding the best parameters and values for the kernel and biases, to fine tune the training process by comparing all the available optimizers, an optimizer was selected (stochastic gradient descent(SGD)),Initial learning rate: 0.001

,Epochs:300 ,Batch size:20, Learning rate reduction will depend on the value of validation loss too. With patience of 2, verbose of 1, and changing factor of 0.1, and minimum learning rate of 0.00001.

D-Performance Measures

There are many criteria that can be used to evaluate performance Classification Algorithms .

1- Accuracy is achieved by the number of correctly classified instances, whether they are positive or negative states. The accuracy can be calculated using equation (5) [20]

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

2- Precision: is the percentage of relevant instances in the collection of received instances. The precision measure is computed in equation(6)

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (6)$$

3-Recall: Recall is also called sensitivity, and it represents the percentage of relevant cases that have been received. Equation (7) shows how can be calculated Recall

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (7)$$

IV.classification results

1- COVID-19 data preprocessing results

The normalization process is an important step that has been implemented on the COVID-19 dataset to avoid differences in large values that dominate the results. The data preprocessing step on a small sample of the COVID-19 dataset is shown in Figure (5)

Before Normalization

Probe	DORF6_0H_1	mock_0H_1	WT_0H_1
A_23_P100001	10.14090605	10.07458	9.985604
A_23_P100011	8.268958021	8.23117	8.243425
A_23_P100022	3.883245138	4.484565	3.692726
A_23_P100056	4.723214667	4.37973	4.988555
A_23_P100074	8.987023138	8.973344	8.930811
A_23_P100092	7.952285564	7.876672	7.795645

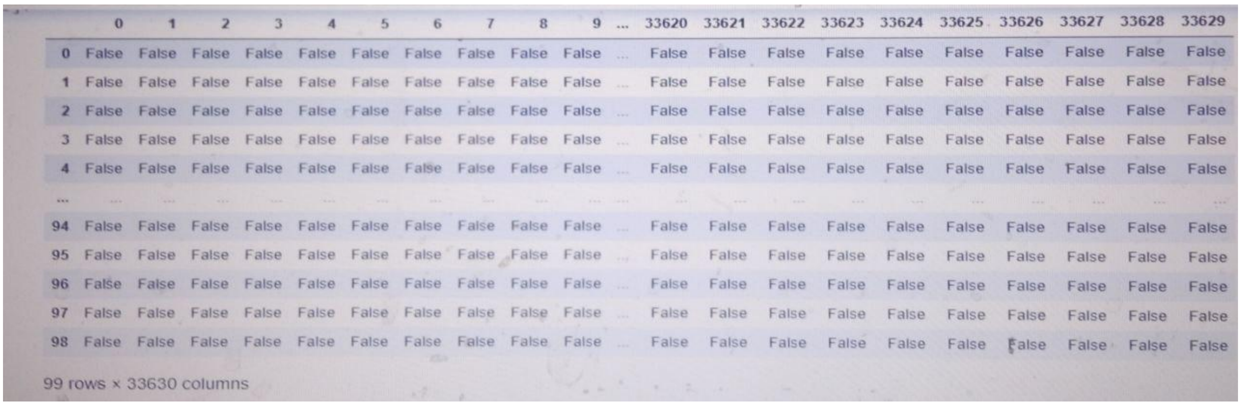
After Normalization

Probe	**DORF6_0H_1**	**mock_0H_1**	**WT_0H_1**
A_23_P100001	**0.424545**	**0.404530**	**0.398985**
A_23_P100011	**0.238897**	**0.309852**	**0.269142**
A_23_P100022	**0.260173**	**0.093088**	**0.191556**
A_23_P100056	**0.470306**	**0.467705**	**0.463207**
A_23_P100074	**0.426656**	**0.549037**	**0.522823**

Figure (5): shows the Normalization step

2- Missing Values Result.

In this stage, the COVID-19 data will be tested to see if this data contains missing values (NaN) or not. After a number of tests, we noticed that the Covid-19 data is free from missing values, as shown in Figure (6)



	0	1	2	3	4	5	6	7	8	9	...	33620	33621	33622	33623	33624	33625	33626	33627	33628	33629
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
94	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
95	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
96	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
97	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
98	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False

99 rows × 33630 columns

Figure (6) shows that the Covid-19 data are free of missing values

3- Results of the Prediction Model

A- Result of Artificial Neural Network (ANN)

After applying a number of feature selection methods to select genes related to ANN in this thesis we obtained different results

Table (2) showing the results obtained from applying feature selection methods with ANN

Method	Number of gene	accuracy	Loss
Pearson +ANN	**15000**	**0.271**	**1.112**
MI+ANN	**10000**	**31.43**	**1.100**
FS(Pearson+MI+PCA)+ANN	**198**	**97.14**	**0.533**

B- Result of Convolution Neural Network (CNN)

After applying a number of feature selection methods to select genes related to CNN in this paper we obtained different results

Table (3) showing the results obtained from applying feature selection methods with ANN

Method	Number of gene	accuracy	Loss
Pearson +CNN	**15000**	0.271	1.098
MI+CNN	**10000**	**0.328**	**1.098**
FS(Pearson+MI+PCA)+CNN	**198**	**0.871**	**0.719**

Conclusions

After a number of experiments using the best deep learning algorithms represented by the artificial neural network algorithm and the convolutional neural network, a prediction model for the COVID-19 data used in this thesis was built and the best results were obtained using the Feature Sequential Selection (FSS) Method. Using FSS with ANN (97%) ,and CNN with FSS(87%)..The results obtained prove that the artificial neural network remains the best deep learning algorithm for dealing with various data, whether it is image data, speech discrimination data, or tabular data.Through the obtained results, it was shown that the convolutional neural network gives high accuracy with the visual data more than the tabular data.

References

[1] ALbert Whata and Charles Chimedza, Deep Learning for SARS COV-2 Genome Sequences,IEEE Acces, vol 9,2021, https://creativecommons.org/licenses/by/4.0/

[2] Ahmed and G. Jeon, “Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses,” *Interdiscip. Sci. Comput. Life Sci.*, Aug. 2021, doi: 10.1007/s12539-021-00465-

[3] **HilalArslan ,2020**, “Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data,” *Proceedings*, vol. 74, no. 1, p. 20, Mar. 2021, doi: 10.3390/proceedings2021074020

[4] Nahida Habib and et al. , Diagnosis of corona diseases from associated genes and X-ray images using machine learning algorithms and deep CNN, 2021, https://doi.org/10.1016/j.imu.2021.100621

[5] Priyanka Ramesh and et al,Gene expression profiling of corona virus microarray datasets to identify crucial targets in COVID-19 patients, Gene Reports Journal,vol 22 ,https://doi.org/10.1016/j.genrep.2020.100980.

[6] 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.

[7] Milad Mostavi1and other, Convolutional neural network models for cancer type prediction based on gene expression,BMC Medical Genomics vol 13 , 2020 ,p1.

[8]Babu Karthik and et al, Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN,: International Journal of Grid and High Performance Computing, vol 8,2020 p 177647

[9] Mehdi Toloo and et al, A new method for ranking discovered rules from data mining by DEA, Expert Systems with Applications, journal Expert Systems with Applications,vol 36, doi:10.1016/j.eswa.2008.10.038,p 8504

[10] Fadl Rahman Shamil,Data mining normalization method, *International Journal of Computer Science*, 2020, Vol. 2, pp 111–117

[11] P. Tang, M. Steinbach, and V. Kumar, “Introduction to data mining”, Pearson Education, 2006.

[12] Satyam Kumar,7 Ways to Handle Missing Values in Machine Learning,2020, https://www.linkedin.com/in/satkr7/

[13] Sebastian Wallot and Dan Monster, Calculation of Average Mutual Information (AMI) and False-Nearest Neighbors (FNN) for the Estimation of Embedding Parameters of Multidimensional Time Series in Matlab, TECHNOLOGY REPORT article ,2018 https://doi.org/10.3389/fpsyg.2018.01679

[14] IanT.Jolliffe1 and Jorge Cadima , Principal component analysis : areview and recent developments,2016 , http://dx.doi.org/10.1098/rsta.2015.0202.

[15] Benesty J., Chen J., Huang Y., Cohen I, Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2, (2009), Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

[16] X. SERRA, “Face recognition using Deep Learning,” Polytechnic University of Catalonia, 2017

[17] Alaa Al-Waisi, et al. “Multi-biometric iris recognition system based on Deep learning approach. Pattern Analysis and Applications 21.3 (2018): 783-802. doi.org/10.1007/s10044-017-0656-1.

[18]Jawid Heidari “Classifying Material Defects with Convolutional Neural Networks and Image Processing”. 2019, http://urn.kb.se/ resolve?urn=urn: nbn:se: uu:diva-387797.

[19] LeNail, Alexander. "Nn-svg: Publication-ready neural network architecture schematics." Journal of Open Source Software 4.33 (2019): 747. doi:10.1109/TVCG.2011.185

[20] H. S. Basavegowda and G. Dagnew, “Deep learning approach to classifying cancer microarray data,” CAAI Transactions on Intelligence Technology, vol. 5, no. 1, pp. 22 - 33, 2020.

doi.org/10.3390/s19071693