Republic of Iraq

Ministry of Higher Education and Scientific Research

University of Babylon

College of Science for Women

Department of Computer Science

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

A Thesis

Submitted to the council of College of Science for woman,

University of Babylon in Partial Fulfillment of the Requirement

For Degree of Master of Science in Computer Science

By

Eman Hamid Hadi

Supervised by:

Prof. Dr. Hussein A. Lafta

Asst . Prof .Dr. Sura Zaki Alrashid

2022 A.D. *1443 A.H.*

قَالُوا سُبْحَانَكَ لَا عِلْمَ لَنَا إِ لا مَا

عَلمْتَنَا إِنكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ

سورة البقرة الآية ٣٢

Declaration

I hereby declare that this dissertation entitled **“ Classification of COVID-19 Disease Using Genes Expression and a Deep Learning Technique ”**, submitted to the University of Babylon in partial fulfillment of requirements for the degree of Master collage of science for women\department of computer science , has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source is appropriately cited in the references.

Signature:

Name: Eman Hamid Hadi

Date: / / 2022

Supervisor's Certification

We certify that this thesis entitled **"Classification of Covid -19 Disease Using Genes Expression and aDeep Learning Technique"**, is completed by the student **"Eman Hamid Hadi"** under my supervision at the Department of Computer Science of University of Babylon in a partial fulfillment of the requirements for MSc Degree in Computer Science.

Signature:

Name: Prof. Dr. Hussein A. Lafta

Date: / / 2022

Address: University of Babylon , College of Information Technology

Signature:

Name: Asst . Prof .Dr. Sura Zaki Alrashid

Date: / / 2022

Address: University of Babylon , College of Information Technology

The Head of the Department Certification

In view of the available recommendations, I forward the research entitled **"Classification of COVID -19 Disease Using Genes Expression and a Deep Learning Technique"** for debate by the examination committee.

Signature:

Name: Asst . Prof .Dr. Saif AL-alak

Date: / / 2022

Address: University of Babylon/College of Science for Women

Dedication

To my father, to the one dearest in existence after God and His Prophet Muhammad

To my dear mother

To my dear husband and children

To my dear brothers and sisters

To my loyal friends

Acknowledgments

First of all, I am grateful Allah the Almighty, the owner of grace and favor who was and is still my support in preparing and achieving this thesis. My deep love and massive thanks go to the masters of creatures, the greatest prophet Mohammed and his progeny (peace be upon them).

Praise is to Allah, who illuminated me the way of science and gave me the patience to continue.

Also sincere appreciation and love go to all my family members whose have encouraged me to go through this study and provides me with love and pure affection.

I would like to express my deepest appreciation and gratitude to my supervisors **Professor Dr. Hussain A. Lafta** and **Asst . Prof .Dr. Sura Zaki Alrashid** for their guidance, valuable advice, helpful discussions, thoughtful comments, continuous encouragement and support throughout the year of study and research at every step of this thesis. Without their support and invaluable feedback, I could not be able to complete it. I acknowledge their contribution to enhance our knowledge on the subject.

I would also like to acknowledge University of Babylon - Collage of Science for Women - Department of Computer Science for providing support in conducting this thesis.

Finally, I would like to thank the people who gave me help and advice.

Abstract

Millions of people are impacted by the coronavirus illness . It should also be highlighted that the cytokine storm has grown to be a significant factor in the high death rate. However, because of our limited understanding of the host defense mechanism and the emergence of a cytokine storm to combat this viral infection, efforts to create medicines, vaccines, and treatments have been unsuccessful. Therefore, a greater comprehension of the processes causing immunological dysregulation and the emergence of cytokine storms may provide us with insights into the clinical treatment of severe instances. Infection with COVID-19 illness may be influenced by genetic factors . therefore ,identifying the genes that influence this disease can help to increase the good therapy response. The proposed system consists of two main stages: the feature selection and the prediction stage. Feature selection is performed using the feature sequential selection (FSS) method to select a subset of important genes and improve the prediction accuracy of the proposed model. In general, the proposed system implements the FSS method, which recognizes the most advantageous features at each step, then they are entered into the model to ensure the importance of these features and the accuracy that can be obtained. Moreover, this thesis sought to provide a prediction model based on artificial neural network (ANN) and a convolutional neural network (CNN) to identify genes related to patients with COVID-19 disease (DORF6), genes related to people with mild symptoms (WT) or genes related to people without infection (MOCK), the available dataset was used to achieve the goals of the current thesis is: the COVID-19 dataset. The evaluation was made based (accuracy).The results obtained showed that the performance of the proposed system is effective, as the prediction accuracy of the original and selected genes was compared before and after applying the FSS method. Using all genes, the prediction accuracy obtained was (30%) with ANN and (30%) with CNN, while the prediction accuracy after applying the FSS method was (97%) with ANN and (82%) with CNN with only 198 genes.

Table of contents

Title No.	Title	Page No.
Chapter One : Introduction		
1.1	Overview	2
1.2	Problem Statement	4
1.3	Aim of Thesis	4
1.4	Thesis motivations	5
1.5	Thesis Challenges	5
1.6	Related Works	5
1.7	Thesis Outline	9
Chapter Two : Theoretical Background		
2.1	Introduction	12
2.2	COVID-19 Disease	12
2.2.1	Genetic Variants That Influence on Susceptibility and Severity To COVID-19	14
2.3	Microarray Technology and Expression Matrix	15
2.4	Dataset	16
2.5	Data Preprocessing	16
2.5.1	Missing value	17
2.5.2	Date Normalization	18
2.6	Feature selection (FS)	19
2.6.1	Pearson correlation coefficient Method (P)	21
2.6.2	Mutual Information Method (MI)	22
2.6.3	Principal Component Analysis (PCA) Method	24
2.7	Deep Learning (DL)	25

2.7.1	Deep Neural Networks (DNN)	27
2.7.2	Types of Learning Algorithms	28
2.7.3	Artificial Neural Network (ANN)	29
2.7.3.1	Multi-Layer Perceptron (MLP) Structure	30
2.7.3.2	The activation functions	31
2.7.3.3	Loss Function	32
2.7.3.4	Optimizer	33
2.7.3.5	Back-propagation Algorithm	34
2.7.4	Convolutional Neural Network (CNN)	36
2.7.4.1	Basic Components of CNN Architecture	37
2.7.4.2	CNN Architecture	38
2.8	Performance Measures	46
2.9	Data Augmentation (DA)	48
Chapter Three :The Proposed System		
3.1	Introduction	51
3.2	The proposed system structure	51
3.2.1	Preprocessing the COVID-19 Dataset	53
3.2.2	Feature Selection	54
3.2.3	Building Prediction Model Using Reduced Features Set	57
3.2.4	Evaluation of the proposed model	61
Chapter four: Experimental Results and Discussion		
4.1	Introduction	63
4.2	Requirements for Hardware and Software	63
4.3	Description of COVID-19 Dataset	64

4.4	Data Preprocessing Results	66
4.4.1	Missing values processing Results	66
4.4.2	Normalization Process Results	67
4.5	Results of Selected Features Methods	68
4.5.1	Result of the Feature Sequential Selection (FSS) Method	68
4.5.1.1	Results of the Pearson Correlation Coefficient Method	68
4.5.1.2	Results of the Mutual Information Method	68
4.5.1.3	Results of the Principal Component Analysis Method	69
4.6	Results of the Prediction Model	73
4.6.1	Results of Artificial Neural Network model (ANN)	73
4.6.2	Results of Convolution Neural Network (CNN) model	74
4.7	Evaluating the Proposed Model	74
Chapter five : Conclusions and future works		
5.1	Conclusions	82
5.2	The Future Works	83
References		

List of Tables

Table No.	Title	Page No.
1.1	Summary of the Related Works	8
2.1	Methods for Pre-processing The Data	17
2.2	Two-Dimensional Confusion Matrix of Classifier System	46

3.1	The proposed structure of the Artificial Neural Network (ANN) model	57
3.2	The CNN Model Proposed Structure	59
4.1	A Brief Description of COVID-19 Dataset	64
4.2	The Top Selected Genes from FSS Method	71
4.3	The Comparative Results of Accuracy and Loss for the ANN and CNN Models	75
4.4	The Comparative Results of the feature Selection Methods for the ANN Model	76
4.5	The Comparative Results of the feature Selection Methods for the CNN Model	77
4.6	The evaluation of the results of the ANN and CNN model with different scales (accuracy, loss, sensitivity, specificity, and area under the curve	78
4.7	The Comparative Results of the feature Selection Methods for the ANN Model using only 5000 genes	79
4.8	The Comparative Results of the feature Selection Methods for the ANN Model using only 500 genes	79
4.9	The Comparative Results of the feature Selection Methods for the CNN Model using only 5000 genes	79
4.10	The Comparative Results of the feature Selection Methods for the CNN Model using only 500 genes	80

List of Figures

Figure No.	Title	Page No.
2.1	The Structure of The COVID-19 Virus	13
2.2	Architecture of the deep network	27

2.3	Topological illustration of a simple neural network	29
2.4	The basic structure of an MLP	30
2.5	Sigmoid Activation Function	31
2.6	ReLU Activation Function	31
2.7	Back-propagation	35
2.8	Architecture of a CNN network	39
2.9	Sliding Receptive Field From The First Location and Continue Moving by One Slide Across The Entire Data	40
2.10	The Convolution Layer	41
2.11	The Convolution Layer Received N Feature Maps and Produced M Feature Maps	42
2.12	Max-Pooling Operation with a 4X4 block size	43
2.13	Connection Between convolution layer and Fully Connected Layer	45
2.14	The Dropout Layer	46
2.15	AUC curve plotted on graph	48
3.1	The Proposed System	52
3.2	Structure of the Used CNN Model	59
4.1	The Class Labels in the COVID-19 Dataset	65
4.2	A sample of COVID-19 Dataset Values	66
4.3	Missing Value Result	66
4.4	The Normalization Result	67
4.5	The Reduction Levels for the COVID-19 Dataset	70
4.6	The gene (A_32_P99744) with class DORF6 for three different replicates during different hours, where x-axis is the hours and y-axis is gene value	72

4.7	The gene (A_32_P99744) with class MOCK for three different replicates during different hours, where x-axis is the hours and y-axis is gene value	72
4.8	The gene (A_32_P99744) with class WT for three different replicates during different hours, where x-axis is the hours and y-axis is gene value	73
4.9	Accuracy, and Loss across the ANN and CNN Models	76
4.10	Accuracy, and Loss across the ANN Model Based on different feature selection Methods	77
4.11	Accuracy, and Loss across the CNN Model Based on different feature selection Methods	78

List of Abbreviations

Abbreviation	*Meaning*
Acc	Accuracy
Adam	Adaptive Moment Estimation
ANN	Artificial Neural Network model
AUC	Area Under Curve
CNN	Convolutional Neural Network
COVID -19	Coronavirus Disease-2019
DA	Data Augmentation
DAEs	Deep Automatic Encoders
DBN	Deep Belief Network
DL	Deep Learning
DNA	Deoxyribonucleic Acid

DNN	Deep Neural Networks
FC	Fully Connected
FN	False Negative
FP	False Positive
FS	Features Selection
FSS	Feature Sequential Selection
GANS	Generative Adversarial Networks
GANs	Generative Adversarial Networks
GEO	Gene Expression Omnibus
LSTMs	Long Short Term Memory Networks
MI	Mutual Information
ML	Machine Learning
MLP	Multi-Layer Perceptron
mRNA	Messenger Ribonucleic Acid
NCBI	National Center for Bioinformatics Information
PCA	Principal Component Analysis
PCs	Principal Components
Pxy	Pearson correlation coefficient
RBFNs	Radial Basis Function Networks
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RNNs	Recurrent Neural Networks
SGD	Stochastic Gradient Descent
SOMs	Self Organizing Maps
SVM	Support Vector Machine
TN	True Negative

TP	True Positive
WHO	World Health Organization

List of Algorithms

Algorithm No.	Title of Algorithm	Page No.
2.1	Normalization	19
2.2	Mutual Information Method	23
2.3	Principal Component Analysis Method	25
3.1	Missing Value	53
3.2	Pearson correlation coefficient Method	55
3.3	The Artificial Neural Network	58
3.4	Convolution Neural Network	60

CHAPTER ONE

INTRODUCTION

Chapter One

Introduction

1.1 Overview

The current development of the novel coronavirus disease known as COVID-19 or SARS-CoV-2, originating from China [Albert Whata and Charles Chimedza,2021] , it is causing a global health emergency with the rapid spread of this epidemic around the world. On March 11, 2020, the World Health Organization (WHO) publicly declared COVID-19 a global pandemic [Raju Babukarthik et al.,2022]. The COVID-19 pandemic has affected more than 200 regions and countries [Hua Ye and et al.,2021] with more than 188,655,968 verified cases including 4,067,517 losses worldwide scientific response to a deadly infection or virus that has killed thousands of people. This latest epidemic of COVID-19 (Coronavirus) caused by severe acute respiratory syndrome is growing lethally at an extraordinary speed. It has infected millions of people and continues to have a horrific impact on the health and well-being of the world's population [Nahida Habib ,and Mohammad Motiur,2021]. COVID-19 attacks the immune system, causing a cellular storm that leads to the death of people with COVID-19 [Priyanka Ramesh and et al,2021].

In general, the field of computer technology and the field of bioinformatics have been used in disease research. Through bioinformatics analysis [Pushpa Singh and Narendra Singh, 2021] , the causes of COVID-19 disease were discovered and an early and appropriate prediction program developed. The analysis and interpretation of enormous datasets produced by numerous biological studies is one of the goals of the interdisciplinary field of research known as bioinformatics. One such biological experiment is the simultaneous measurement of the expression levels for tens of thousands of genes in a given environment

[Priyanka Ramesh and et al,2021]. One of the fundamental tools used by biologists to assess the degree of gene expression in a given organism is the microarray. The difficulty of using microarray technologies to analyze data to gain a better understanding of biological processes has substantially grown. But the microarray method creates a lot of data that includes the expression levels of numerous thousands of genes. This information is typically laid out as a matrix, with the rows denoting samples and the columns denoting genes. Each cell of the matrix contains a numeric value denoting the amount of a gene's expression in a specific sample, is called a gene expression matrix .In reality, the value of microarray gene expression data for illness diagnosis and patient treatment plan selection has increased significantly [Manish Babu and Kamal Sarkar ,2016] . Therefore, advanced computational methods are required, which is the most advanced method for finding the complex code of proteins and genes responsible for stimulating diseases as gene expression data sets are widely used in disease prediction and diagnosis [Heba Abusamra,2013], particularly in the treatment of COVID-19. There are several computational techniques available for gene expression analysis. Deep learning methods are part of machine learning techniques based on artificial neural networks. These include artificial neural networks and convolutional neural networks, which are the most important deep learning algorithms designed for data.

1.2 Problem Statement

The problem could include that sometimes the previously proposed system is not able to recognize the genes, because of the noise and the large number of genes. This is captured by the poor classification accuracy.

Also, the problem of restrictions distinguishing between cases of infection with Covid-19 disease. Which although genes are expected to vary greatly, there may be significant similarities in the gene sets of these data. Thus, each algorithm has an upper theoretical limit in terms of its ability to distinguish between cases of infection with Covid-19 disease and why some people have severe infection, others have a slight infection, and others do not appear to be infected.

1.3 Aims of Thesis

The aim of this thesis is to identify the genes that influence the severity of COVID-19 disease and to develop a prediction model that could assist medical professionals in correctly diagnosing and treating COVID-19 patients. The following objectives will be accomplished in order to accomplish these aims:
1- Deciding which subset of genes (informative genes) is best for the prediction task using gene expression data.

2. Used a prediction model with minimal error and high accuracy based on the selected genes. In order to distinguish between patients with COVID-19 disease (DORF6), genes related to people with mild symptoms (WT) or genes related to people without infection (MOCK).

1.4 Thesis Motivations

There are some motives that were the main reason for researchers' interest in studying the problem of COVID-19, including:

1-To diagnose the genes associated with the severity of COVID-19 disease and the genes that cause death, as ANN and CNN networks were used to diagnose the disease and classify cases to take the necessary measures regarding each case and deal with serious and medium cases.

2- Early diagnosis for accurate treatment, which in turn reduces pressure on the health care system.

1.5 Thesis Challenges

This thesis faced various challenges, including:

1- The curse of dimensionality: the number of genes in the gene expression data is very huge (containing thousands of genes). It is common that not all genes are useful , some genes are irrelevant and redundant information in the dataset. Therefore, working on this huge number of genes is difficult.

2- The application of the prediction model: It presents another difficulty as it should be implemented with the least possible error and highest achievable accuracy.

1.6 Related Works

A number of studies are discussed to cover the most relevant works and provide a summary of the work for the classification of COVID-19. The following is a description of these studies:

• (**Jayadeep Pati,**2018) , the researcher suggested a way to analyze gene expression data to identify and predict the optimal subset of lung cancer-causing genes. The researcher used three models to predict lung cancer-causing genes and

these models are(Multilayer Perceptron, Random Subspace, Saquential Minimal Optimization (SMO)) The classification methods used achieved different accuracy (86%, 68%,91%) respectively.

• **(Nahida Habib and Mohammad Motiur,**2021 **)**, adopted two different methods of detection that 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 that was done using Deep-CNN technology, achieving a test accuracy of 99 % .

• **(Hua Ye and et al.,**2021 **)** , proposed an intelligent prediction model useful for distinguishing between COVID-19 severity, to aid clinical diagnosis for decision-making. In this paper, they proposed a prediction model using Harris hawks optimization (HHO) to improve Fuzzy K-near neighbor (FKNN), which is called HHO-FKNN. This model is used to characterize the severity of COVID-19. In HHO-FKNN, the best classification performance distinguishing severe COVID-19 from mild COVID-19 was obtained with an accuracy of 94%

• **(Nour Eldeen M. Khalifa et al,** 2020), propose a new and improved approach to deep learning that is based on optimizing binary particle swarm, with decision tree (BPSO-DT) and convolutional neural network (CNN) to classify different types of cancer based on gene expression data for renal cell carcinoma, breast cancer and lung adenocarcinoma. The model has been trained and overcome the problem of overfitting and training the model to achieve better accuracy. The proposed approach achieved an overall test accuracy of 96%

• **(Naiyar Iqbal and Pradeep Kumar,** 2022 **) ,** propose a system method for detecting genes associated with corona disease. Their study revealed that differentially expressed genes are used as inputs to implement a machine learning model using SVM to predict genes associated with corona disease severity. The proposed system achieved a high accuracy of 99%.

• (**pranamita nanda and N. Duraipandian**, 2020), propose to apply a number of algorithms to a gene expression data set to predict lung cancer-related genes. The predictive ability of a number of different algorithms was compared. The obtained results showed that "optimized random forest" achieved the highest accuracy of 98%.

• **(Albert Whata and Chartes Chimedza,** 2021 **)** , proposed a deep learning algorithm that uses a convolutional neural network (CNN) as well as a bidirectional long-term memory neural network (Bi-LSTM), to classify SARS-CoV-2, they classified whether the genome sequence contains regulatory motifs or other motives. The proposed bidirectional convolutional neural network for the long-term memory model (CNN-Bi-LSTM) has achieved a classification accuracy of 99%.

• (**Lei Chen and et al.,**2021 **)** , Suggest a method to detect COVID-19 using Boruta, Max-Fit and Min-Redundancy feature selection methods one by one . The mRMR feature list was entered into the IFS method, which included one of the four classification algorithms (RF, SVM, KNN and DT). The SVM classifier is considered the best classifier. This classifier achieved an accuracy of 0.93 .

• **(Faraz Khan and et al.,**2021]**)** , proposed a system for identifying the genes responsible for causing lung cancer in the human race. In this paper, they used the Kruskal-Walli test. Determination of gene expression data. Finally, 12 influencer genes responsible for the pathogenesis of lung cancer have been identified. The accuracy of the model is 84% using the Random Forest algorithm.

• **(Raju Babukarthik et al.,2022**), proposed a technique that aims to provide a solution to identify pneumonia caused by COVID-19 and healthy lung using CXR images. In their research, they used the latest deep learning technology represented by the (GDCNN) algorithm. This algorithm was trained to extract classification features between COVID-19 and normal images, this algorithm achieved an accuracy of 98% for predicting COVID-19.

Table (1.1): Summary of the Related Works

Author Name	Dataset	Disease type	Method	Accuracy
Jayadeep Pati,2018	Gene Expression	lung cancer	MLP Random Subspace SMO	ACC 86% ACC 68% ACC 91%
Nahida Habib ,and Mohammad Motiur,2021.	Gene expression And Image	COVID-19	SVM l& deep CNN	ACC 93% ACC 99%
Hua Ye and et al.,2021	genome sequences	COVID-19	HHO-FKNN	ACC 94%
Nour Eldeen M. Khalifa et al, 2020	Gene Expression	Cancer	CNN	ACC.96%
Naiyar Iqbal and Pradeep Kumar, 2022	Gene Expression	COVID-19	SVM	ACC 99%

pranamita nanda and N. Duraipandian, 2020	Gene Expression	Lung cancer	Random Forest	ACC 98%
Albert whata and chartes chimedza ,2021	DNA sequences	SARS-CoV-2	CNN-Bi-STM	ACC 99%
Lei Chen and et al.,2021	Gene Expression	COVID-19	SVM DT KNN RF	ACC 93% ACC 86% ACC 88% ACC 0.92%
Faraz Khan and et al.,2021	Gene Expression	Lung cancer	Random Forest	ACC 84%
Raju Babukarthik et al.,2022 .	Images	COVID-19	GDCNN	ACC 98%

1.7 Thesis Outline

The thesis included five chapters arranged as follows:

Chapter One: This chapter included the introduction, the problem of the study, the aim of the study, the motives for the study, thesis contributions, thesis challenges , and related works.

Chapter Two : This chapter included a comprehensive description of the main biological concepts, and preprocessing techniques that included both the handling of missing values and normalization process, in addition, this chapter included the methods used to features selection to reduce data dimensions , the deep learning algorithms, the proposed prediction model, and the evaluation methods used.

Chapter Three : This chapter describes the proposed system. In this chapter, the preprocessing stages are explained, and the following stages select important genes in the COVID-19 dataset (features selection). After that, a prediction model is created to predict the genes related to the severity of COVID-19 disease.

Chapter Four : The fourth chapter explains and discusses the implementation the proposed system on the COVID-19 dataset, and illustrates of the experimental results obtained after applying the proposed model.

Chapter Five : The fifth chapter represents the most important conclusions this study reached based on the results of the thesis . In addition to that, this chapter highlights possible future works.

CHAPTER TWO

THEORETICAL BACKGROUND

Chapter Two
Theoretical Background

2.1 Introduction

This chapter clarifies the basics of gene expression as well as the basic information and concepts used in this research that are introduced to detect genes associated with COVID-19 disease. In addition, the basic concepts will be explained by providing a brief overview of the four stages: data pre-processing, selection of the most important features, as well as the prediction techniques used and evaluation of the model used. This chapter mainly focuses on the main methods and techniques adopted in this thesis.

2.2 COVID-19 Disease

COVID-19 is a highly contagious epidemic disease that first appeared in December 2019 in China in the city of Wuhan. The World Health Organization (WHO) declared the outbreak of this disease in 2020 a global health emergency, raising the epidemic to the level of severe severity. The clinical presentation of corona disease varies from person to person, and these clinical presentations are variable, fever represents the main symptom of this disease, in addition to other symptoms of fatigue, cough, diarrhea and nausea, and some cases may be asymptomatic. The cases of elderly people with other diseases, including (heart, diabetes and pressure) are more likely to suffer from respiratory failure [Tamer Ali and et al.,2020]. The severity of corona disease varies from person to person, and the reason for this is due to the presence of genes closely related to the increased severity of COVID-19 disease, as these genes play a vital role in the production of cytokines, which is the main reason for increasing the severity of

infection and increasing death rates [Qiurong Ruan and et al.,2022]. Therefore, the need to identify new targets and effective treatments for COVID-19 disease is of great importance to overcome this epidemic that has swept the entire world. [Lei Chen and et al.,2021]. Therefore, the investigation of this study focuses on identifying genes associated with COVID-19 disease severity using gene expression analysis. In general, the severity of COVID-19 is related to the patient's genetics and an inaccurate diagnosis often delays beneficial treatments. Thus, the discovery of functional genes that can establish correct matches and relationships with clinical symptoms has become urgent. For this purpose, it is necessary to know the set of genes associated with the disease of COVID-19. However, using rapidly developing microarray technology, large amounts of data can be analyzed in order to extract disease-related genes using novel computational methods in order to mitigate disease severity and identify potential therapeutic targets [Chaolin Huang and et al.,2020] . In general, the Corona virus is active inside the living organism because it depends on the cell components for reproduction and is weak outside the living organism [Jenny Oh and Laura Klivans, 2020]. Figure (2.1) shows COVID-19 virus.

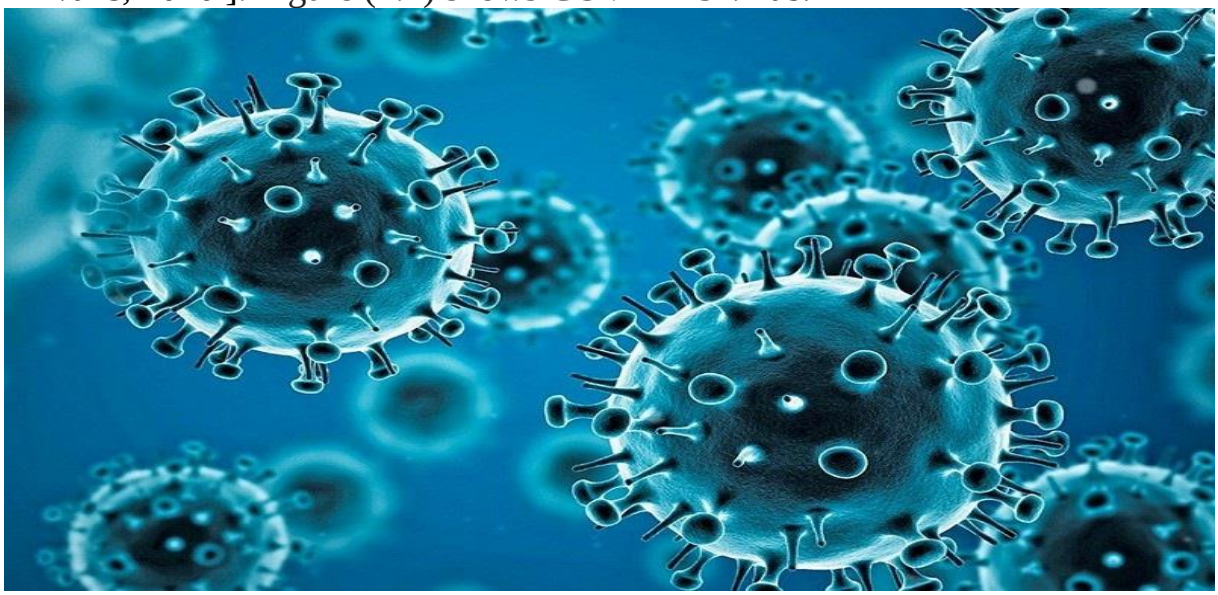


Figure (2.1) :The Structure of the COVID-19 Virus.

2.2.1 Genetic Variants that Influence on Susceptibility and Severity to COVID-19

Certain genetic factors of an individual can influence the risk and severity of infection and symptoms of the disease. Much of the international study has identified large parts of the human genome that can influence the risk and severity of COVID-19 infection [Zhicheng Wang and Kun Tang,2020] .

From the beginning of the Corona virus until now, doctors and scientists are trying to understand why some people contract severe COVID-19 while others do not show any symptoms. Risk factors such as age and underlying medical conditions as well as environmental factors of social and economic determinants have significant roles in determining disease severity. However, the differences in the human genome is the source of the variance in the severity of the disease [Zhicheng Wang and Kun Tang,2020]. Scientists already know that genetic variants influence the severity of human infectious diseases including infection with COVID-19 .Genetic factors may range from factors that are rare, to high-impact mutations that can make the difference between individuals with the disease from mild symptoms and a life-threatening illness, to the most common genetic variants that only moderately affect the severity of the disease [Erola Pairo-Castineira and et al.,2021]**.** However, human genome studies specialized in infection diseases are still rare and limited compared to other immune conditions, including autoimmune disorders for many reasons . The most important of them is that infectious diseases are usually studied by focusing on the microorganisms that cause diseases instead of studying the host. Plus, human genes usually have small effects on the outcome of infection compared to the effects resulting from social and demographic factors, especially age or access to health care. Determining this requires studying large and distinct groups of people in order to

produce sufficient statistical power to detect relevant genetic factors. It could be human genomics **a**n effective tool through which to understand biological mechanisms that determine the immune response to a specific disease infection, in order to identify individuals who are at risk and to develop new drugs and vaccines [Samira Asgari and Lionel Pousaz,2021]

2.3 Microarray Technology and Expression Matrix

Microarray technology has enabled many scientists to measure the expression levels of large numbers of genes in an experiment that takes a very short time, it may take a few days compared to the time it takes to analyze genes in a biological laboratory, which may take months. Usually, it takes several months to diagnose a specific disease based on genetic analysis in a biological laboratory. However, when using the expression technique, the diagnosis of the disease may take much less time, and when the disease is diagnosed early, the severity of the disease can be controlled and the appropriate medications are used. In general, when using microarray data for gene expression as well as performing computational procedures, the disease will be diagnosed in a few days. And therefore , Microarray data can be used to explore the genes responsible for the severity and progression of diseases. It can also help in finding the right treatment [Hiba muthanaa ,2021].

Generally ,the microarray is a glass slide on which DNA molecules are fixed at specific sites called spots. A microarray may contain hundreds of thousands of spots, and each of these spots contains several copies of a DNA molecule that corresponds uniquely to a gene. Then the microarray chip will be laser stimulated with appropriate wavelengths. The final result is stored as an image file for further analysis. A special image analysis program is used to capture the microarray images. As manufacturers of microarray scanners provide various software such as Limma package which is a tool that is used to extract data from

raw image files. In general, microarray technology stores the data of thousands of individual gene expressions in a matrix called the gene expression matrix, where n rows correspond to a specific individual called samples and m columns represent different genes. In addition, the number of samples may reach hundreds, while the number of genes may reach tens of thousands [22].

2.4 Dataset

There are a lot of biological data sets published through the National Center for Biological Information (NCBI). In this thesis, the data set used was taken from the publicly available data source Gene Expression Omnibus (GEO). It was published in November 23, 2020 by the National Center for Biological Information (NCBI). The entry number for data sets is (GSE33267) provided by Gene Reports Journal. In this thesis, a prediction model was applied to the data set (GSE33267), which contains (33630) genes, to find out the genes associated with the severity of COVID-19 disease [Priyanka Ramesh and et al,2021].

2.5 Data Preprocessing

Today's real-world databases are highly vulnerable to missing and cluttered data because of their huge volume and the origin of this data is likely to be from multiple heterogeneous sources. Low quality data will lead to lower quality mining results. Data preprocessing helps improve data quality and thus improves the efficiency and ease of the mining process. Data preprocessing is an important step in the data mining process and data collection methods are often controlled loosely, resulting in out-of-band values, or missing values. Analyzing data that has not been carefully examined to address these problems can lead to illogical and misleading results [Mehdi Toloo and et al.,2008].

There are many data pre-processing techniques. Including data cleaning is applied to remove inconsistencies in the data and remove noise. Data integration is where data from multiple sources are combined into a cohesive data store. Data

reduction can reduce data size through aggregation and deletion of redundant features. Normalization can be applied where data is scaled to fall within a smaller range (0.0 and 1.0). Thus, the accuracy and efficiency of mining can be improved, and the missing values can be estimated [Mehdi Toloo and et al.,2008] . 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 features. Table (2.1) shows the different stages of data pre-processing [Mehdi Toloo and et al.,2008].

Table (2.1) Methods for Pre-processing the Data [Mehdi Toloo and et al.,2008].

Data Preprocessing		
Data Cleaning	Data Transformation	Data Reduction
1-Missing Data a- Ignore the tuple b- Fill the missing value 2- Noisy Data a- Regression b- Clustering	1- Normalization 2- Attribute selection 3- Discretization 4- Concept Hierarchy Generation	1-Data Cube Aggregation 2- Attribute subset selection 3- Numerosity Reduction 4- Dimensionality Reduction

2.5.1 Missing Value

Oftentimes the data contains a lot of missing values in the real world. The main reason for missing values could be the failure to record data or data corruption, and the processing of the missing data is very important during the pre-processing of the dataset used since many deep learning algorithms do not support missing values [Marianne Riksheim and et al.,2019]. There are many ways in which missing values can be handled : either by deleting missing and irrelevant values [Najah Abed, (2021).]or by replacing the missing values with

the value (0) or by replacing the missing value with the mean [Marianne Riksheim and et al.,2019].

2.5.2 Data Normalization

The unit of measure used may affect the analysis of the data and in many cases this data is in a format that is not suitable for use by data mining techniques. So the data is normalized by converting the raw data values into another form with properties that fit the model used . Normalization aims to ensure that all genes used are in the same unit of measurement, i.e. between [0 ,1] or between [1,-1] [Luai A. Al Shalabi and Z. Shaaban,2006] . Therefore, the normalization process is used to eliminate 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 . There are many ways to normalize data , including min-max and z-score normalization . Equation (2.1) is used in the minimum and maximum normalization [Hiba muthanaa ,2021].

$$v' = \frac{v - min\ a}{max\ a - min\ a} * (new\_maxa - new\_mina) + new\_min\ a \qquad (2.1)$$

Where:

v : represents the feature value

$min\ a$: is the minimum original value for any feature

$max\ a$: is the maximum original value for any feature

$new\ -maxa$ and $new\ -mina$ are the maximum and minimum interval of new values **.** for more information on the min-max normalizing method , see algorithm (2.1) .

Algorithm (2.1): Normalization Steps [Hiba muthanaa ,2021].

Input : *Two dimensional array MV* $[n * m]$*in which n is the number of samples and m is number of genes*

Output : *Tow dimensional array DN*$[n * m]$ after normalization process

Begin

1. *Let Max* $[j]$ *and Min* $[j]$ *two arrays represent maximum and minimum value for each gene*

2. *For* $j = 1$ *to* m

3. *Set max and min to the first value of gene j*

4. *For* $i = 1$ *to* n

5. *if* $vji < min$ *then*

6. Min $=$ $v$ji

7. else if $v$ji $>$ max then

8 . Max $=$ $v$ji

9. end if

10. end if

11. *end for*

12. *Apply equation* (2.1)*to modify the every value v in gene j*

13. *end for*

14. *return DN*

End .

2.6 Feature Selection (FS)

Feature selection methods as a preprocessing step in predictive modeling have many important advantages. It can reduce model complexity, enhance learning efficiency, and increase predictive power by reducing noise. Often, in a

high-dimensional dataset, some completely irrelevant and insignificant features remain. It has been observed that the contribution of these types of features is often less towards predictive modeling than towards critical features. These features cause a number of problems that in turn prevent effective predictive modeling [S. Vanjimalar and et al.,2018].

Feature selection technology is a preprocessing technique that aims to select the most advantageous genes, it can differentiate between groups, i.e. subtypes associated with severe COVID-19 disease, or normal versus infected samples. The feature selection method reduces the dimensions of the original feature space to a lower dimensional space by selecting a subset of genes [S. Vanjimalar and et al.,2018].

The methods of selecting features are divided into four categories: (filter, embedded , wrapper and Hybrid methods) [S. Vanjimalar and et al.,2018]. The selection of the feature is done by selecting the features or genes separate or independent of the model or classifier used [Omar Al-Harbi ,2019] . Adequate selection of features may improve the accuracy and efficiency of the model [Younes Bouchlaghem and et al.,]

feature selection has a number of advantages, including:

- enhancing the machine learning algorithm's performance.
- Understanding data, learning about the process, and possibly assisting with visualization.
- data minimization, lowering storage requirements, and possibly assisting in cost reduction.
- simplicity, the ability to use simpler models, and increased speed
- Make the model used easy to building.
- Remove irrelevant features [Younes Bouchlaghem and et al.,].

In these methods, all the features that give optimum results are selected for the deep learning algorithm. The data used contain a lot of genes that are less important and irrelevant, so their unnecessary use in the prediction model leads to

the complexity of the model and the difficulty of its interpretation , it also makes the model work with less accuracy, so the features selection process is critical , as the relevant genes are selected that are more important from a large set of data, which leads to train deep learning algorithms more quickly, facilitates model interpretation and reduces its complexity, in addition to building a good model with better predictive capabilities. Determining the appropriate set of features reduces excessive training time and this results in a small set of features [Nivedhitha Mahendran and et l.,2020].

2.6.1 Pearson Correlation Coefficient (*P*) Method

The Pearson correlation coefficient (*P*) is a measure that determines the degree to which two different variables are related. The Pearson correlation is one of the most common correlations to measure the linear relationship between two variables. When the relationship is nonlinear , this correlation coefficient may not be an appropriate measure of dependence. The range of correlation coefficient values is -1 to 1. In other words, values cannot exceed 1 or be less than -1. Correlation -1 indicates a negative correlation and correlation 1 indicates a positive correlation. If the correlation coefficient is greater than zero , then this is a positive relationship. On the contrary , if the value is less than zero , it is a negative relationship . A zero value indicates that there is no relationship between the two variables [Alexander Ly and et al.,2018] . The related genes are determined by the Pearson correlation coefficient .The association between two objects X = ($x1, x2, x3, \ldots xn$) and Y = ($y1, y2, y3, \ldots yn$)) is described as follows:

$$P\mathrm{xy} = \frac{\Sigma_{i=1}^{n} \left(\mathrm{X}\, i - \bar{x}\right) \left(\mathrm{Y}\, i - \bar{\mathrm{y}}\right)}{\sqrt{\Sigma_{i=1}^{n} \left(\mathrm{X}\, i - \bar{x}\right)^2} \sqrt{\Sigma_{i=1}^{n} \left(\mathrm{Y}\, i - \bar{\mathrm{y}}\right)^2}} \qquad (2.2)$$

Where:

n is the number of samples.

xi is the value of item i .

Y i is the value of class i.

$\bar{x}$ is the mean data of each dimension.

$\bar{y}$ is the mean data of each class label.

The Pearson correlation between variables X and Y is calculated. The formula is basically just dividing the covariance by the product of standard deviations [Alexander Ly and et al.,2018].

2.6.2 Mutual Information (MI) Method

This method is a term from information theory, which shows the strength of the statistical correlation of two random variables, one of which is dependent and the other is independent [Sebastian Wallot and Dan Monster,2018]. In gene expression data, information exchanged between a gene and a class marker is used to identify important and related genes. The genes are selected according to the computed mutual information after the redundant genes are removed [Devi Arockia Vanitha and et al.,2015] . The mutual information estimator is used based on the minimum entropy as a classifier to detect different types of genes related to the severity of corona disease. The exchanged information is calculated as shown by the following equations.

$$H(G) = -\Sigma_{g\in G}\ P(G)\ \ log_2\ P(G) \qquad (2.3)$$

$$H(C) = -\Sigma_{c\in C}\ P(C)\ \ log_2\ P(C) \qquad (2.4)$$

Where:

P(G) and P(C) are probability (G) and (C). Then the cross-entropy of H (G, C) will be calculated in the following form:

$$H(G,C) = -\Sigma_{g\in G}\ \Sigma_{c\in C}\ P(G,C)\ \ log_2\ (G,C) \qquad (2.5)$$

Probability P (G, C) represents the values that are common to both random variables (G) and (C) that occur together. In the last stage, the mutual information

MI is calculated for the variables (C) and (G) using the following equation [Néstor Barraza and et al.,2019].

$$MI(G,C) = H(G) + H(C) - H(G,C) \quad (2.6)$$

Algorithm (2.2) shows the method of selecting genes by mutual information method.

Algorithm (2.2): Mutual Information Method [Hiba muthanaa ,2021] .

Input : *Two dimensional array RDP* $[n * m]$*in which n is the number of the sample and m is the number of features*

Output: *Reduced Dataset* (RDM): *feature's subset that have thehighest of Mutual* Information values

Begin

1. ***S****et features _subset to NULL. // Initialize the features_ subset is empty .*

2. $for\ i\ =\ 1\ to\ m$ // where m: number of features

3 . $for\ j\ =\ 1\ to\ y$ //where y: number of classes

4. *Compute the Mutual Information between feature Fi and class label Yi according to the equations* (2.6)

5. *end for j*

6. *end for i*

7. *Select the features Fi with maximum MI* $(Fi\ .Yi)$

8. *The selected feature is sorted by the Mutual Information value in descending order*

End

2.6.3 Principal Component Analysis (PCA) Method

Principal component analysis is one of the most important ways to reduce dimensions. It is often used to reduce the dimensions of large datasets by converting a large set of variables into a smaller set without losing the information in the large set. The main idea in this method is to reduce the dimensions of the dataset that consists of a large number of variables related to each other while preserving the variance in the dataset, and the same can be done by analyzing the variables into a new set of variables, which are known as the main components [an T. Jolliffe and Jorge Cadima,2016].The principal component analysis method is described in the following equations:

$$\bar{x}= \frac{1}{n} \Sigma_{i=1}^{n} \ xi \qquad (2.7)$$

Where:

$\bar{x}$ is the mean data of each dimension.

n is the number of samples.

xi is the value of item i . and the covariance matrix is calculated by using the equation (2.8)

$$C_x = \frac{1}{n+1} \Sigma_{i=1}^{n} (xi - \bar{x})(xi - \bar{x})^T \qquad (2.8)$$

Then the eigenvectors ($\boldsymbol{v_m}$) and eigenvalues (λ_m) of the covariance matrix will be calculated using the following equation :

$$C_x \ v_m = \lambda_m \ v_m \qquad (2.9)$$

Where:

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 [Ian T. Jolliffe and Jorge Cadima,2016]. See Algorithm (2.3).

Algorithm (2.3): Principal Component Analysis Method [Hiba muthanaa ,2021]

Input: *Two dimensional array RDM* $[n * m]$*in which n is the number samples and m is the number of genes*. // Output of Algorithm

Output : *Reduced set of genes* (PCA)

Begin

*1. X ← Create N * d data matrix, with one row vector* $\boldsymbol{xn}$ *per data point*

2. $\Sigma \leftarrow$ Covariance Matrix of X

3. *Find eigenvectors and eigenvalues of* Σ

4. PCs ← the *K* eigenvectors with largest eigenvalues

5 . *Top PCs.*

End

2.7 Deep Learning (DL)

The field of deep learning is defined as a new research area that deals with finding theories and algorithms that have the ability to allow a machine to learn on its own by simulating neurons in the human body [M. Arif Wani and et al,2020].

Deep learning is a branch of machine learning (ML) and is originally a branch of artificial intelligence (AI) that imitates the way humans acquire certain types of knowledge [Veeru Talreja and et al.,2017] . Traditional machine learning algorithms are linear, while deep learning algorithms are stacked in a complex and abstract hierarchy consisting of layers, each higher layer is based on the previous lower layer which is why it is known as hierarchical learning [Andreas Maier and et al.,2019] . Deep learning first appeared in 2007 [M. Arif Wani and et al,2020]. It has the ability to handle huge amount of data to solve complex problems, which is why it is widely used on a daily basis by tech giants such as Microsoft , Facebook , Amazon, Google , Baidu [M. Arif Wani and et al,2020]. Deep learning consists of processing layers that create computational models that

are used to define complex structures in large and massive datasets by representing data with different levels of abstraction . Deep learning is a new approach to machine learning, as it brings machine learning and artificial intelligence closer together. Deep learning is used in many areas such as object detection and speech recognition as well as in the medical field [Amir Mosavi and et al.,2020] . When the features are useful and describe the database, learning is facilitated . The step of extracting useful features is important in machine learning. But traditional machine learning may be limited in processing natural information and dealing with it in its raw form , this limited ability makes it difficult to solve problems using artificial intelligence technology [Mustafa Khalaf,(2021).], and to solve these problems, deep learning is designed [Ke-Lin Du and M.N.S.Swamy,2013]. DL is also particularly useful in a lot of areas with large, high-dimensional data , which is why deep neural networks outperform shallow ML algorithms in most applications where text, image , video, speech, and audio data need to be processed [Yann LeCun Facebook AI Rese and et al.,2015] . As for low-dimensional data entry , especially in cases where limited training data is available , shallow ML still produces superior results [Ying Zhang1 and Chen Ling,2013] which tend to be better than those generated by deep neural networks [Cynthia Rudin,2019] . One of the most important reasons for the popularity and success of deep learning is the rapid progress in machine learning algorithms and processing [S.H. Shabbeer Basha and et al,2022]. Figure (2.2) shows a deep structure of the nervous system consisting of several layers that make it deep [Noor Fahem,2021] .

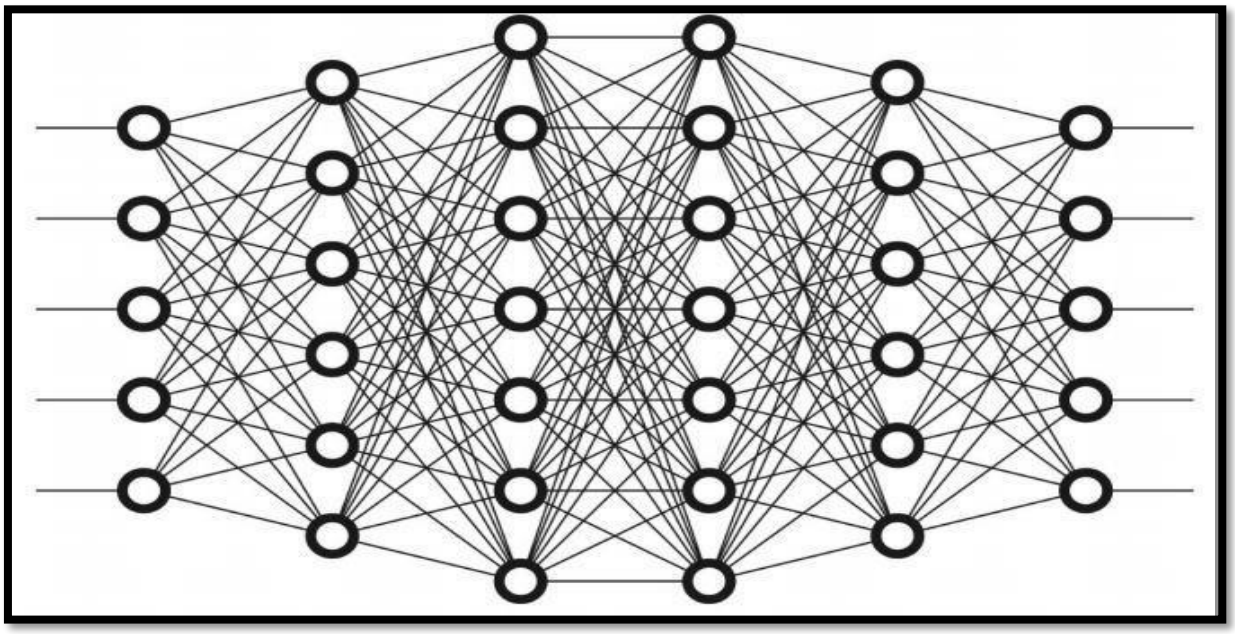


Figure (2.2): Architecture of the deep network.

2.7.1 Deep Neural Networks (DNN)

DNN is defined as an artificial neural network that contains more than one hidden layer. MLP are the widely used ANN structures in DNN . Neural networks consist of multiple layers of neurons that are interconnected with each other . Since the number of weights in this network is huge , it can exceed thousands or even millions, which is why DNN requires long computation times and data feeds in the training stages of the samples [Saja Mahdi,(2021)] **.** There are many architectures for deep learning networks , the most important of which are : Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [Pariwat Ongsulee,2017] . During the past decade , many factors have emerged that have allowed training of this type of network [Frank Emmert-Streib and et al.,2020] , such as :

• The significant increase in computing power for both computers and graphics processing units.

• The emergence of a new data-based culture, such as data mining and machine learning that made it possible to collect and analyze a large amount of data in usable databases.

• The emergence of many new methods of training DNNs. Examples include the fast learning algorithm, ReLU layers, and dropout regulation [Peng Tang and et al.,2018].

2.7.2 Types of Learning Algorithms

Deep learning plays an important role in our lives . It has great influence in many fields , including disease diagnosis , precision medicine , speech recognition . and extract features from large size datasets . DL can also overcome the limitations of previous shallow networks . A deep neural network (DNN) uses multiple (deep) layers of modules with highly optimized algorithms and architectures [Ajay Shrestha and Ausif Mahmood,2019] . The following list shows the most widely used types of deep learning algorithms:

- Artificial Neural Networks (ANN)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Generative Adversarial Networks (GANs)
- Long Short Term Memory Networks (LSTMs)
- Radial Basis Function Networks (RBFNs)
- Multilayer Perceptrons (MLPs)
- Self Organizing Maps (SOMs)
- Restricted Boltzmann Machines(RBMs)
- Deep Belief Networks (DBNs)

- Autoencoders

These algorithms are among the most important types of algorithms used in deep learning that work with almost any type of data and require large amounts of computing power and information to solve various complex problems [Ajay Shrestha and Ausif Mahmood,2019] **.**

2.7.3 Artificial Neural Network (ANN)

ANN is one of the most supervised and widely used methods of deep learning [Rene Y. Choi and et al., 2020] . ANNs are collections of connected computational nodes that are used as a computational approach to issues whose solutions are challenging for conventional computer programs to find a suitable representation for. ANN sometimes called as "black boxes" since it is so difficult to comprehend how they work [Serra Xavier,(2017)] . Figure (2.3) shows the architecture of the neural network . Deep learning neural networks are made up of numerous hidden layers of neurons. It takes at least two hidden layers for something to be deep. The majority of deep learning networks, however, have many more hidden layers than just two. The crucial distinction is that a network's depth is determined by the sum of its hidden and output layers . There are numerous varieties of ANN, such as Recurrent Neural Networks (RNN) or Radial Basis Function Network (RBF).

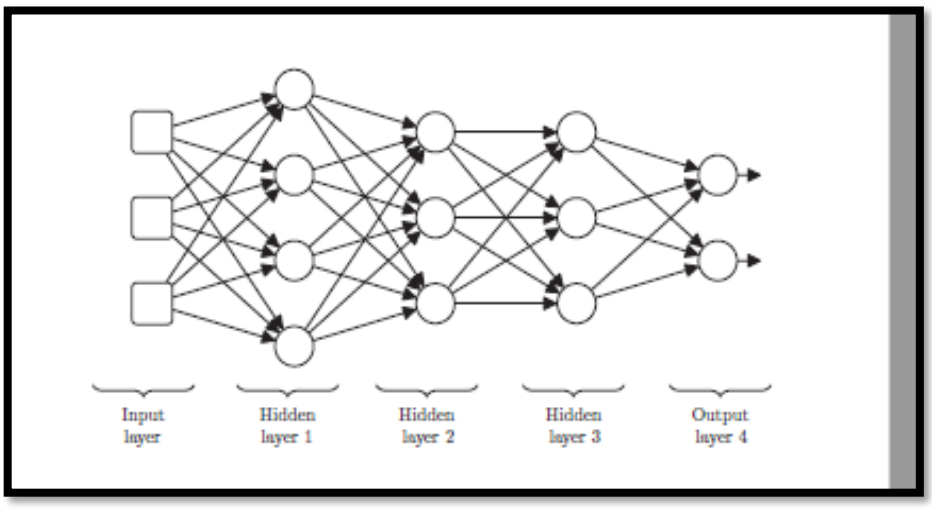


Figure (2.3) Topological illustration of a simple neural network [Rene Y. Choi and et al.,2022]

2.7.3.1 Multi-Layer Perceptron (MLP) Structure

MLPs represent a feed-forward to an ANN with one or more hidden layers, each consisting of a set of computationally interconnected nodes called neurons , as shown in figure (2.4) . Usually all neurons in the hidden layers and the output layer are connected to all neurons in the previous layer, and each of these connections has an associated weight, which represents the strength of connections to the two neurons [Muhind Salim ,(2015).] .

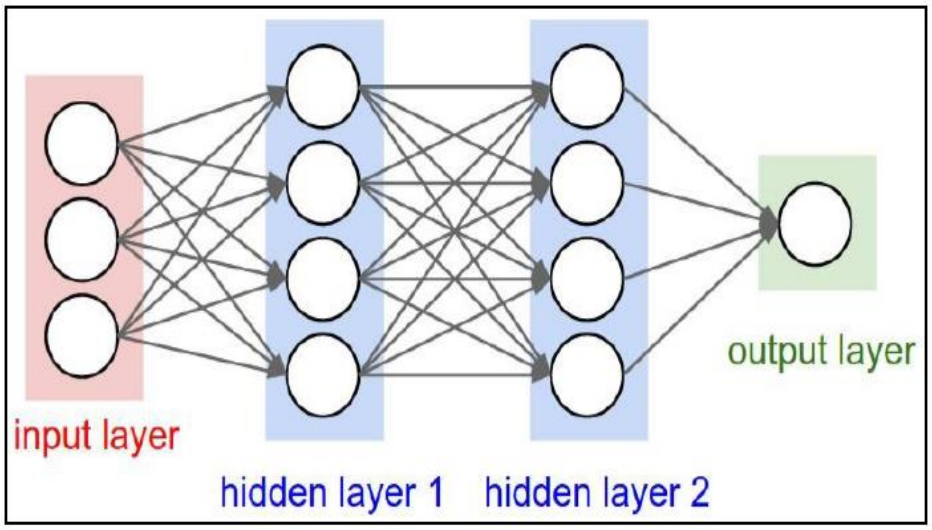


Figure (2.4):The basic structure of an MLP [Husam Imad , 2018]

The output value of each neuron ***j*** in the layer ***l*** is described in equation (2.10) [Ján Vojt,2016]

$$Z_i = \sum_i^N (W_{i\,j} * x_i) + b_j \qquad (2.10)$$

Where

N is the number of inputs

$x\,i$ is the inputs which passed from $(l - 1)$ layer,

(Wi, j) is the weights of all edges that are connecting the neuron j with the neurons of previous layer

And (bj) as bias, equation (2. 10) could be written in a matrices simplified to be as following [Muhind Salim ,(2015)] .

$$Z = W^T x + b \qquad (2.11)$$

2.7.3.2 The Activation Functions

This function is either a linear function or a non-linear function to represent the degree of activation of neurons, and in most cases, the range of this function is between (-1,1) or between (0,1).

Rectified Linear Unit (ReLU) as in equation (2.12):

$g(z)=max\ (z,0)$ (2.12)

The sigmoid function, as shown in equation (2.13) and figure (2.5), is a popular option for the activation function in hidden layers [Stephen moore,2018].

$g(z) = \frac{1}{1+e^{-z}}$ (2.13)

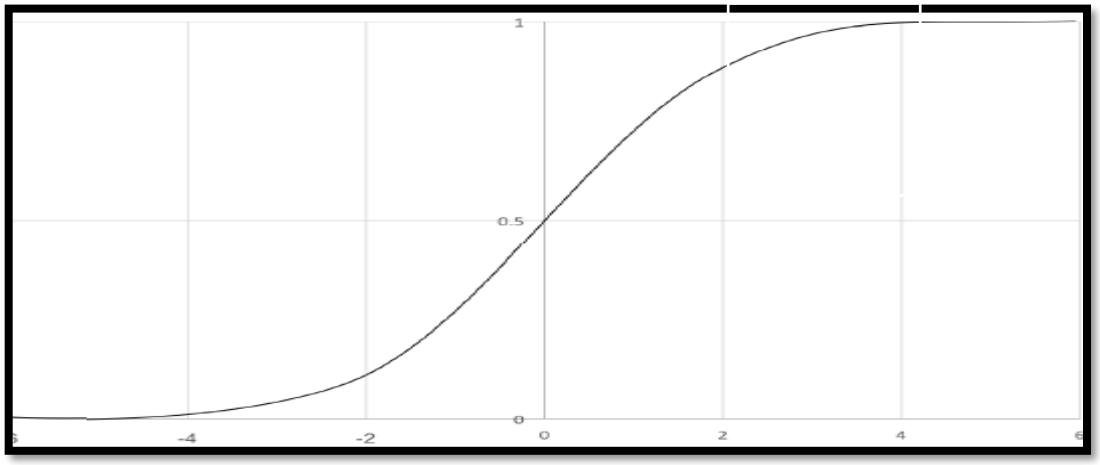


Figure (2.5) Sigmoid Activation Function[Stephen moore,2018] .

figure (2.6) show ReLU Activation Function

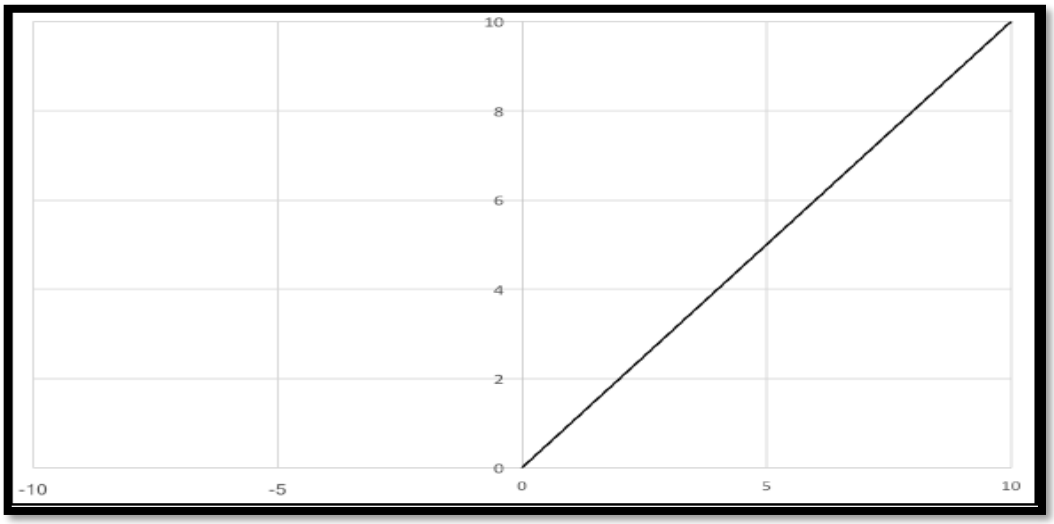


Figure (2.6) ReLU Activation Function [Stephen moore,2018] .

There are two benefits to the Rectified Linear Unit's (ReLU) activation function. The ReLU function is straightforward to assess, and this criterion offers a benefit in terms of computing speed. The second is that the ReLU gradient is constant, which means that the ReLU output value is zero or in the positive region for any numerical input [57]

The probability of each class must be estimated by the output layer. As a result, SoftMax is the most frequently used activation function in the output layer, as shown in equation (2. 14).

$$g(z)_j = \frac{e^{z\,j}}{\sum_{k}^{k=1} e^{zk}} \qquad (2.14)$$

where

zj is the production corresponding to class j, *K* is the total number of classes **[**Stephen moore,2018]

2.7.3.3 Loss Function

Error predictions in neural networks (NNs) are estimated by using a loss function that sometimes called (objective function) , (error function) or (cost function). The loss function is used to measure the quality of our expected probabilities [Aston Zhang,2020].

Since the main goal of machine learning and deep learning is to create a mechanism for converting input dataset values into accurate output, the major method for achieving this goal is through setting the appropriate weights in the artificial neural network. The precise correct figure for each weight cannot be determined in practice due to the numerous unknown values. As a result, the issue of determining the appropriate weight values is turned into an optimization problem. In this way, the algorithm will look for potential weight sets that will produce results that are both predicted and gratifying. These anticipated outcomes are noted with the minimum amount of error. Therefore, every component of the

model's performance should be combined into single value by using the loss function [Aston Zhang,2020].

There are a lot of loss functions, such as mean square error (MSE), cross-entropy, and categorical cross-entropy used in a neural network (NN) to classify into multiple classes using the SoftMax activation function .

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} y^i \ \ln \ g(z^i) \ + (1 - y^i \ ) \ln (1 - g(z^i) \ ) \qquad (2.15)$$

Where

n is number of classes

y i is the desired output for the *i'th* class

$\boldsymbol{g}(\boldsymbol{zi})$ is the estimated probability of class *i* which is calculated using SoftMax activation function in equation (2.14) [Husam Imad, 2018]

In general, the weights in the MLP layers are adjusted during the training process. The back propagation algorithm is the most widely used training algorithm.

2.7.3.4 Optimizer

Optimization refers to the development of an appropriate procedure that will reduce or increase the output of a job, giving it the same input used and that by adjusting some of the variables. This task is very difficult because there is a large and high dimensional space to perform the search for the optimal values of these variables to reduce or increase the output of the function. The variables that the optimizer controls are the weights and biases of the neural network. There are many different types of enhancers[Saja Mahdi,2021]. The most famous enhancer are: Batch Gradient Descent ,Stochastic Gradient Descent SGD [59]. Adaptive Moment Estimation Adam [Saja Mahdi,2021].

1- Adaptive Moment Estimation (Adam) : Adam is one of the most popular enhancer as it gives good results compared to other enhancers. A first-order gradient with a random cost function is its basis. Adam is of great importance in terms of being motivated by the need not to lose a minimum , the speed of the optimizer is reduced as needed during the search to avoid minimal loss. It won't take long to locate the bottom line because this reduction does not occur throughout the entire search process. This is achieved by preserving mean squared gradient values, such as AdaDelta and RMSprop. Also, ADEM keeps the previous decreasing mean of the previous gradients. This brings the benefit of speed convergence. In addition to fixing the problem of fading learning rate, however, this process is computationally expensive [Saja Mahdi,2021].

2- Stochastic Gradient Descent (SGD): The Stochastic Gradient Descent only takes into account one observation before applying the adjustments each time, in contrast to Batch Gradient Descent. Therefore, it will perform the computations once each observation is chosen rather than going through the entire dataset before determining how to adjust the parameters. Since only one observation is chosen to determine how the modifications will be, less memory will be needed, and frequent updates will reduce the likelihood that the optimizer will become stuck at a local minima or saddle point. Frequent modifications are computationally expensive, they can cause an object to depart from the global minima [Nikhil Ketkar,2017].

2.7.3.5 Back-propagation Algorithm

The back propagation is a supervised learning algorithm . Artificial neural networks are frequently trained using a backpropagation algorithm. Based on the error , as seen in the following figure, the weights are adjusted going backward. [57].

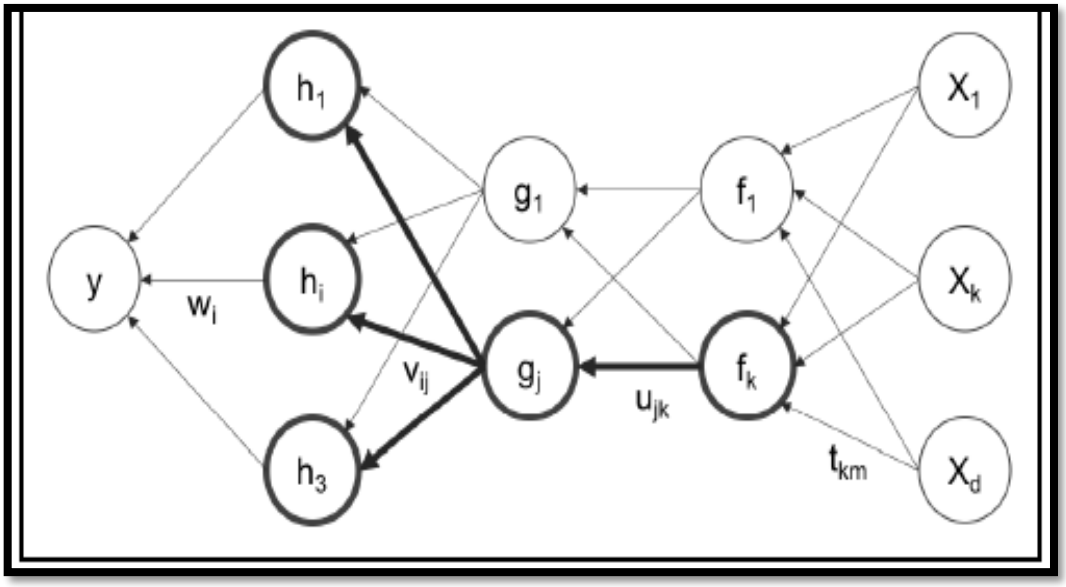


Figure (2.7) Back-propagation [Stephen moore,2018]

Initializing all weights and biases in the network with a random number at the beginning of the training is followed by two steps [Stephen moore,2018].

1- Forward Propagation :

In this stage, the input states are propagated across all layers of the neural network layer by layer from input to output by using equations (2.11) and (2.13) to produce a prediction value [Husam Imad, 2018] .

2- Backward Propagation

Using the gradient descent learning technique starts with the output and works backwards to the input .When utilizing the loss equation (2.15) to calculate the error of the output layer, the delta (δ_j^l) is computed, where l is the layer and j is the index of the neuron in that layer. The delta equation is represented by the following equation [Husam Imad, 2018]:

$$\delta^L = \nabla_x \mathcal{L} \odot \acute{g}(z^L) \qquad (2.16)$$

Where

δ^L A matrix of deltas represents all neurons in L.

L represents last layer in the network.

$\nabla x \mathcal{L}$ is gradient of Loss with respect to x .

x is the output of activation function.

$\odot$ is the Hadamard product (element-wise product of matrices).

Equation (2.17) is used to compute the deltas in lower hidden layers:

$$\delta^{l} = ((w^{l+1})^{T}\ \delta^{l+1})\ \odot \acute{g}(z^{l}) \quad (2.17)$$

Where

l is the number of hidden layer

$(w^{l+1})^{T}$ is weight matrix of next layer.

All of the weights wi, j of the NN must be modified in order to learn it. This is done by computing the weight adjustment wi, j, which is stated in the following equation (2.18) [Husam Imad, 2018].

$$\Delta w_{ij} = -\eta \frac{\partial \mathcal{L}}{\partial w_{ij}} = -\eta \delta_j Z_i \quad (2.18)$$

Where η represent learning rate, and δj represent the delta of neuron j, which computed for output layer using equation (2.16) and for each of rest hidden layers using equation (2.17).

These phases are repeated for all input states, this is called an epoch. The neural network can run for as many epochs needed to find the best solution [Husam Imad, 2018]

2.7.4 Convolutional Neural Network (CNN)

Accurate prediction of corona disease is vital to the diagnosis and treatment of the disease . Through a predictive model , genes related to the severity of COVID-19 disease can be inferred . Several studies have attempted to build deep learning models for this task .

convolutional neural network (CNN) is one of the most supervised and widely used types of deep neural networks in the fields of machine vision [M. Arif Wani and et al,2020] . CNN (also known as ConvNet) is an example of a deep learning strategy that simulates the brain's function in processing information [Sakshi Indolia and et al.,2018] . It is a specific type of feed-forward neural network in artificial intelligence . A CNN network is similar to a multi-

layer network (Perceptron) , it differs only in its ability to combine many locally connected networks. The layers used for feature extraction are accompanied by some layers that are fully connected and used for classification [Alaa Al-Waisy and et al.,2018] **.** CNN is a promising tool for improving automated diagnostic systems and achieving high accuracy for disease prediction [Diyar Qader and et al.,2018] . CNNs are among 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 [Jawid Heidari, 2019] .

A CNN consists of a number of layers that pass information across layers where the output of the previous layer is input to the next layer . The first layer of the network is the input layer , while the last layer of the network is the output layer. The layers between the input and output layers are hidden layers of the network. Each layer is a simple algorithm that contains one type of activation function . In general , the CNN model benefited from capturing high-level interactions between genes and the target set to make predictions with high accuracy [Milad Mostavi and et al.,2020].

2.7.4.1 Basic Components of CNN Architecture

A convolutional neural network has the ability to mimic the human brain . In general , prediction is one of the most important benefits of this network .

A characteristic neural network (CNN) consists of two main components [Y Huang and et al.,2018] :

A) Feature Extractor

It is the first stage of CNN's operation; it involves extracting features and converting them into feature maps. CNN is made up of numerous filters, each with its own task. As a result, multiple feature maps were created, each corresponding to multiple filters. The features extraction procedure includes several phases that result in a low-dimensional feature vector that is then input

into a classifier. Multiple layers make up the feature extractor (multiple convolution layers with optional pooling layers). It passes via a convolution layer in the first step to convolve the filter with the input and produce feature maps, which are then reduced with a pooling layer. Then it uses the previously generated feature maps as input feature maps and repeats the process on them, extracting powerful features layer by layer and producing smaller feature maps. Finally, reduced dimension feature maps are flattened to form a low-dimensional feature vector that can be input into the classifier [Y Huang and et al.,2018].

B) classifier

Following the extraction of feature maps and the reduction of dimensions by picking the best features between them, a low-dimensional feature vector is input into a classifier. The classifier returns the likelihood of the input belonging to that class. To do this, the classifier consists of one or more completely connected layers [Y Huang and et al.,2018] **.**

2.7.4.2 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 [Alexander LeNail,2019], as shown in figure (2.8). These layers are:

1. Convolutional layer.
2. Max pooling layer (or Sub Sampling layer).
3. Fully Connected Layer (Classification layer).
4. Dropout Layer

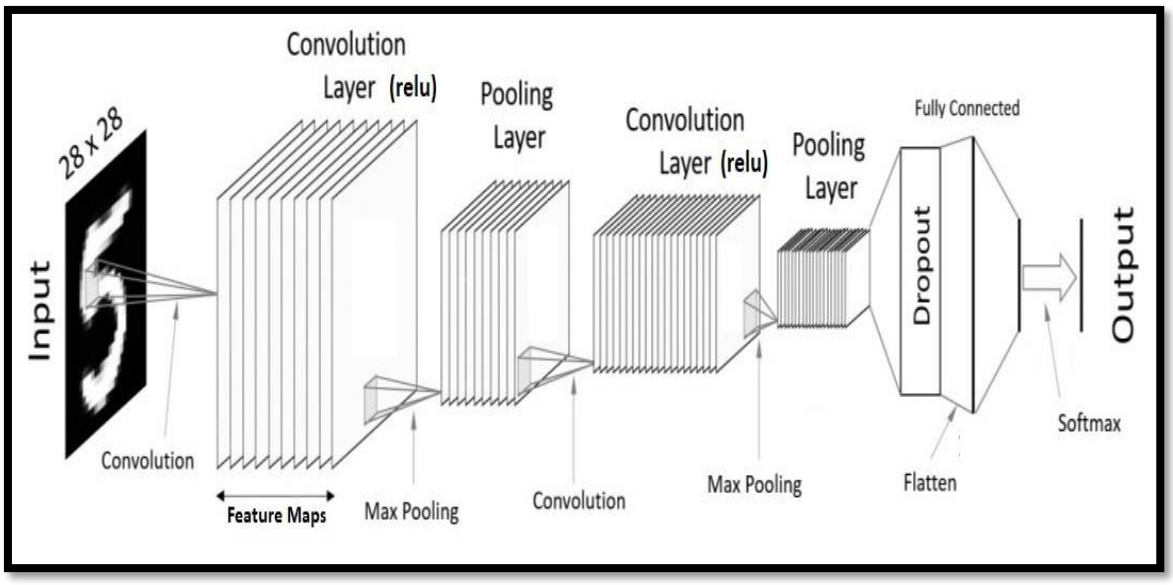
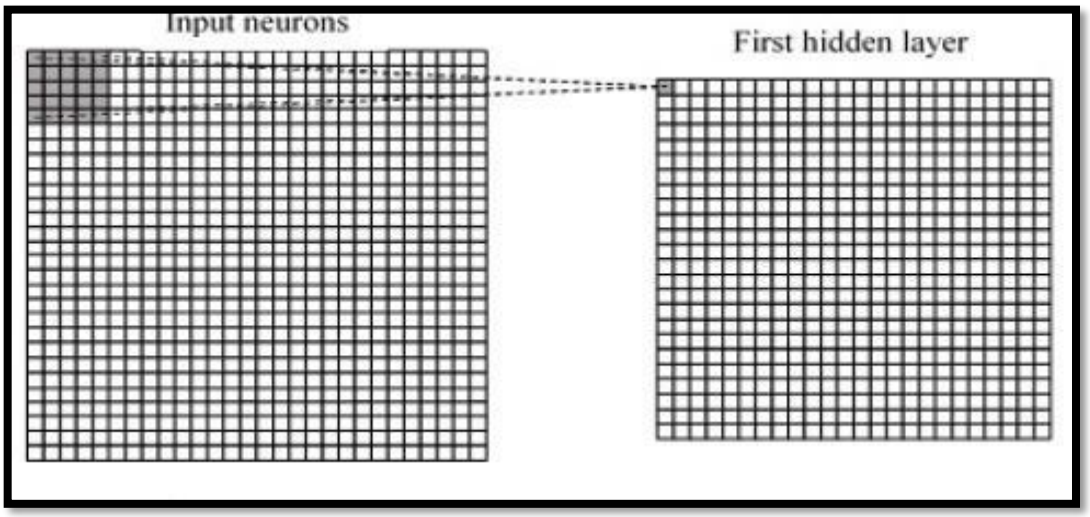


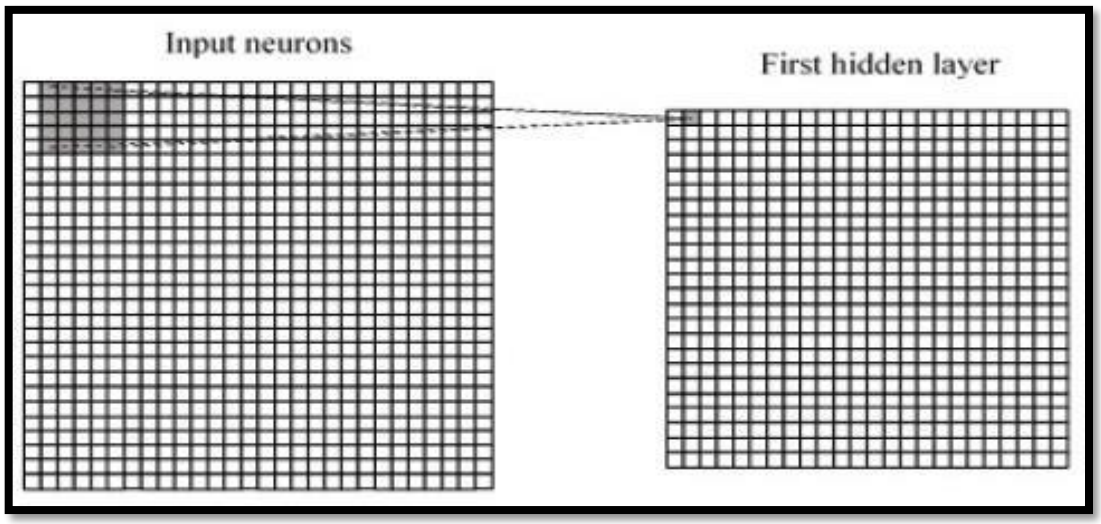
Figure (2.8): Architecture of a CNN network [Alexander LeNail,2019].

1- Convolutional layer

The basic layer in the CNN architecture is the convolution layer, and this layer has the ability to handle high-dimensional data [48] . The convolution layer is the first layer of the convolutional neural network which is partially connected to the next layer (the pooling layer), there is a small window of input neurons, for example (3x3), will be connected to the pooling layer, this window starts from the upper left corner as shown in Figure (2.9) (a) to the bottom corner of the data and keep moving forward in the data by moving its position each time according to the stride value, which is usually (1) as shown in Figure (2.9) (b) . The kernel moves one cell to the right direction until it reaches the end of the columns, and one cell below until it reaches the end of the rows and this process continues until all the data are complete [Rikiya Yamashita and et al.,2018].



(a): Beginning Receptive Field Location



(b) Sliding Receptive Field by One Cell in the Right Direction.

Figure (2.9): Sliding Receptive Field From The First Location and Continue Moving by One Slide Across The Entire Data [Hiba muthanaa ,2021].

The receptive field is a tiny window region created from the input data. To extract features, a tiny section of the input data will be convolved with a shared weights window called kernel or filter [Muhind Salim ,(2015)] . Figure (2.10) depicts the convolution process.

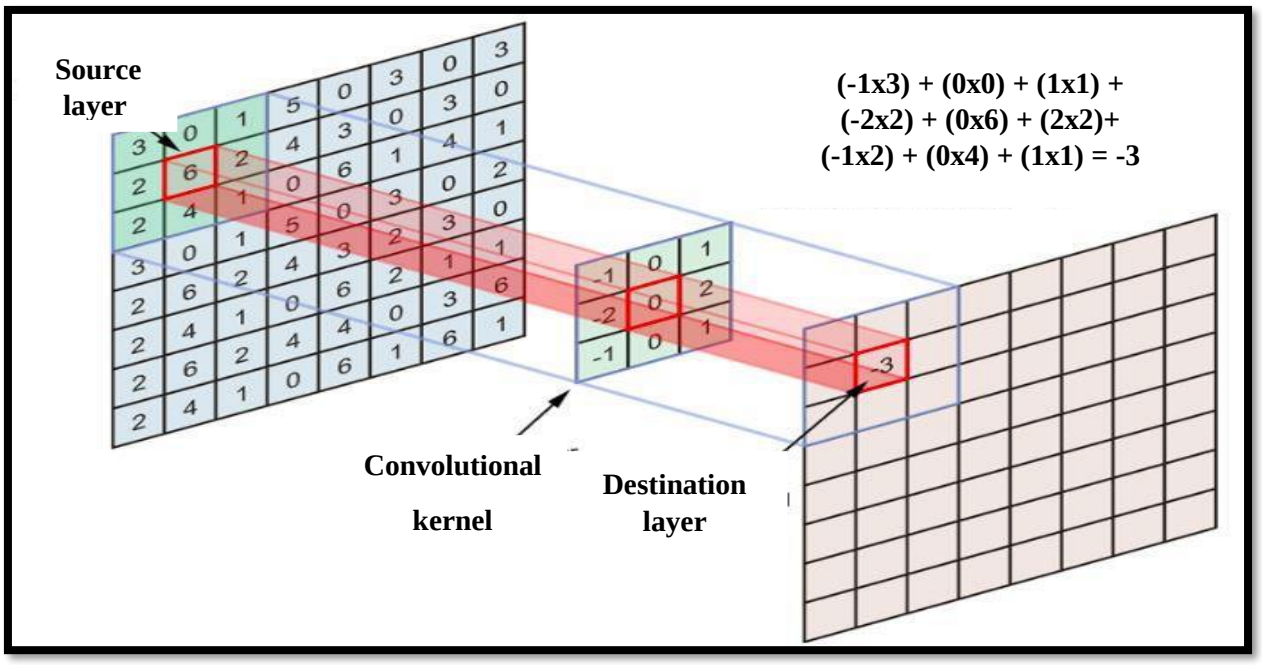


Figure (2.10): The Convolution Layer[Bashkatova Daria,2018].

Lots of wrap filters are applied to a single entry. The activation maps are then combined in order to obtain the final result of one file in the convolutional layer, where this final result represents the input data to the next layer. The values in the filter matrix are represented by the weights given by default. These values must be different from one filter to another to give different properties from each other or different features to each feature map matrices [Min Xia and et al.,2017]. In addition, the architecture of a Convolutional Neural Network (CNN) contains many parameters that are used to control the output size, the overall behavior of the model, or the runtime, these controlled parameters are known as (hyperparameters). These parameters are very important hyperparameters for CNN:

1- **Number of filters**: A reasonable number of different filters can be used, these filters are of different sizes.

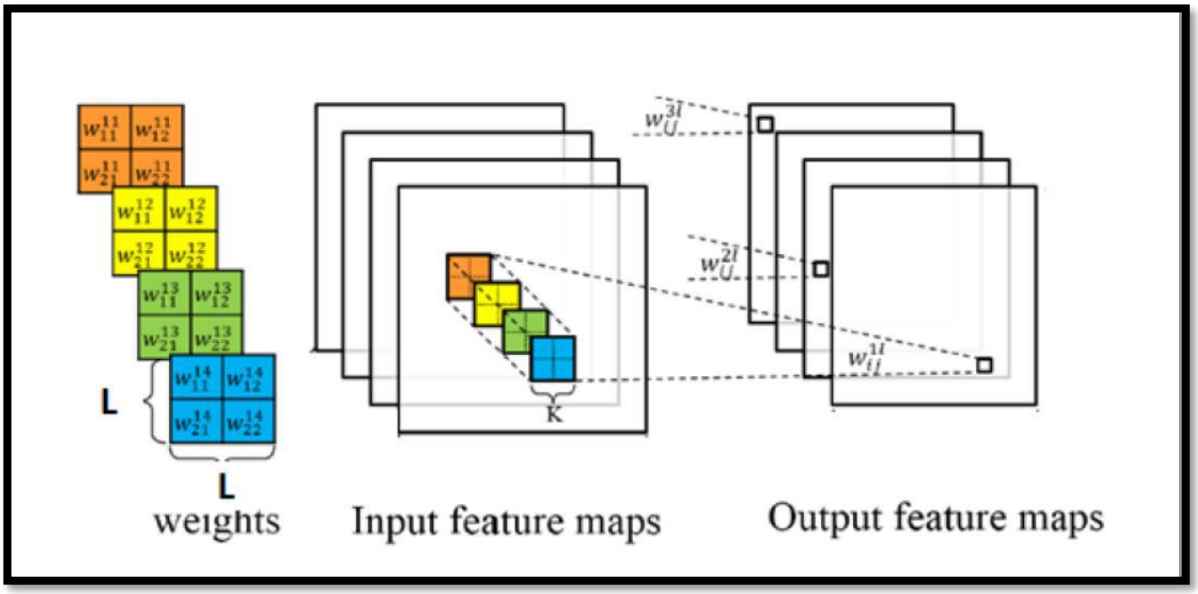
2- **Filter size**: The size of the filter or kernel must be squared L*L, lower than the input data but greater than 2X2.

3- **Stride**: The number of cells that must be moved at once to generate the local receptive field of a filter. A single cell is moved across and down in a single stride. If the stride is too short, there will be overlap, and vice versa.

4- **padding**: The number of pixels to pad the input data is specified by this hyper-parameter. The values "valid" and "same" are the two options. "Valid" denotes the absence of padding, whereas " same " denotes padding with zeros [Lee, Hagyeong and Song, Jongwoo,2019] .In the end, the output of the total convolution layer contains many feature maps, each wrapped with a specific kernel in order to describe certain features. Suppose the input source data is called I, the convolution kernel or weight is called W, the bias is called b. Next, suppose that filters k, W, of size L x L are transformed with the input data in order to produce k feature maps called Ck each containing indices i, j. The convolution operation is found mathematically in equation (2.19)

$$C_{i.j}^{k} = \Sigma_{m=0}^{l-1} \Sigma_{n=0}^{l-1} W_{m.n}^{k} * I_{i+m.j+n} + b^{k} \qquad (2.19)$$

Figure (2.11) shows the map display of the resulting multiple features.



Figure(2.11) : The Convolution Layer Received N Feature Maps and Produced M Feature Maps [71] .

The activation function's mathematical action clarifies the following equation (2.20), which substitutes equation (2.19), which is the linear output of the convolution layer, With the equation for the ReLU activation function described earlier in equation (2.12) in section (2.6.3.2)

$$C_{i.j}^{k} = \max\left(0, \Sigma_{m=0}^{l-1} \Sigma_{n=0}^{l-1} W_{m.n}^{k} * I_{i+m.j+n} + b^{k}\right) \quad (2.20)$$

2- Max pooling layer or Sub Sampling layer

Not only does CNN have convolution layers, but it also has certain pooling layers. The primary goal of this layer is to reduce the dimensionality of its input in order to provide reduced-dimension output by retaining just the most important data. Maximum and average pooling are the two methods used by this layer to reduce dimensionality. In either case, the pooling layer divides the input feature map into non-overlapping blocks, keeping the maximum number in max pooling and the average of block numbers in average pooling, and finally returning only a single value for each block [Satheshkumar Kaliyugarsan,2019] . Figure (2.12) shows the max pooling procedure.

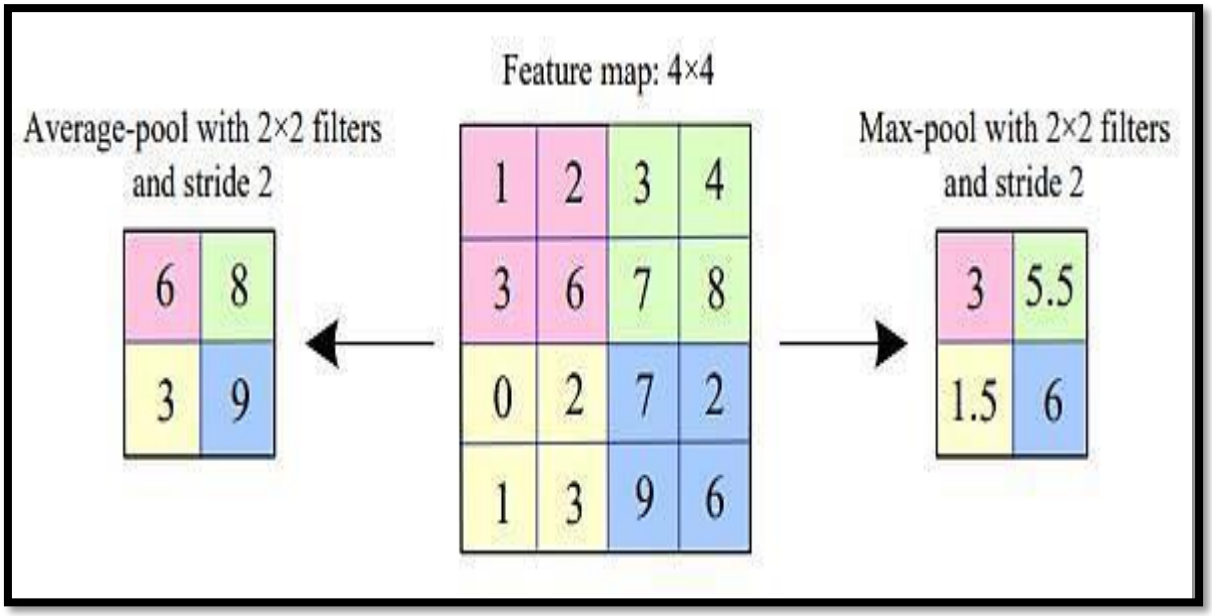


Figure (2.12) : Max-Pooling Operation with a 4X4 block size [Satheshkumar Kaliyugarsan,2019].

The maximum pooling layer does the same for each of the feature maps that were previously obtained from the previous convolution layer. Equation (2.21) shows the mathematical operation of the maximum pooling layer, P, on feature maps previously activated by the convolution layer.

$$P_{i.j}^{k} = \max \begin{pmatrix} C_{(2i.2j)}^{k} & C_{(2i+1\ .2j)}^{k} \\ C_{(2i.2j+1)}^{k} & C_{(2i+1.2j+1)}^{k} \end{pmatrix} \quad (2.21)$$

Equations (2.19), (2.20) and (2.21) keep iterating with each new layer of the model used until all model layers are finished, and all features previously generated from the feature extractor characterized (convolution, activation, and pooling layers) are fed into the classifier, but This classifier must be completely connected to all the probabilistic neurons of the output classes, which in turn is a fully connected layer[M. Arif Wani and et al,2020] .

3- Fully Connected Layer (Classification layer) (FC).

A fully connected layer (FC), often known as a dense layer, is the network's final layer. Every neuron from the previous layer is connected to every neuron in the next layer in a fully connected layer. The resulted feature map from the previous layer must first be flattened to become a feature vector before being fully connected with the output layer, which consists of neurons equal to the number of classes according to the softmax or sigmoid activation functions used in the final CNN layer to classify the trained data which are specialized for multi-class and binary class classification respectively [Martin Thoma,2017] . The connection between final feature maps and a fully connected layer is depicted in figure (2.13).

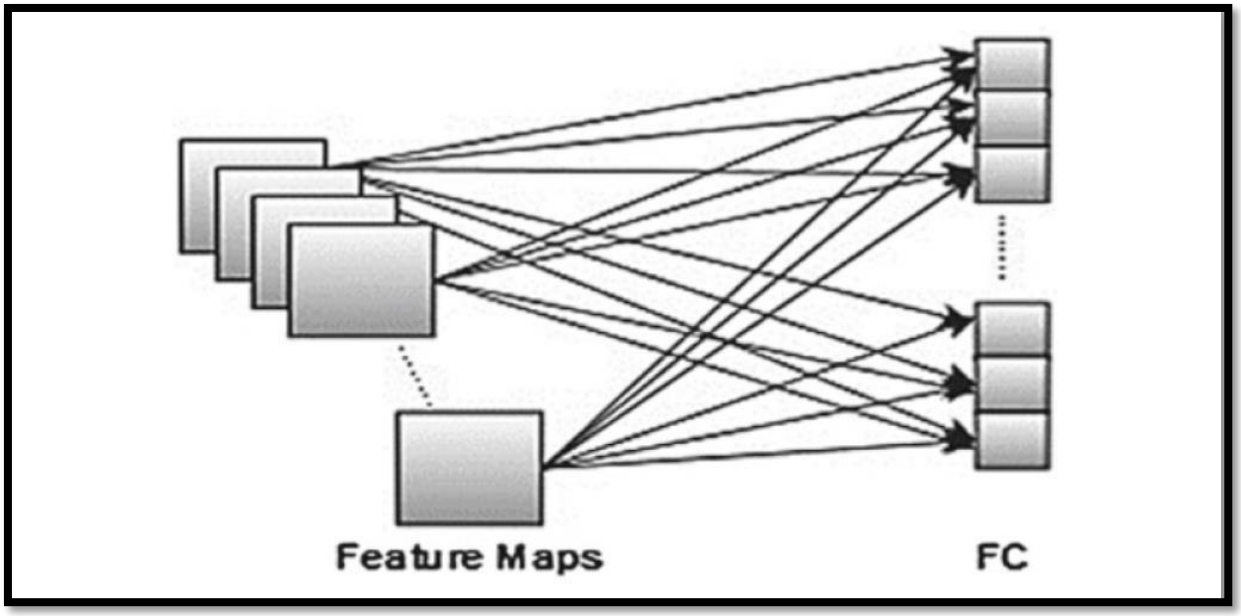


Figure (2.13): Connection Between convolution layer and Fully Connected Layer [Alexander LeNail,2019] .

The calculation for the full connection F that links each of the previously generated feature maps with all output classes is shown in equation (2.22).

$$F^K = \Sigma_{i=1}^{R} \quad W_i^k \ * C^i \qquad (2.22)$$

4- Dropout Layer

When all of the features are connected to the fully connected layer, the training dataset is prone to overfitting. Overfitting happens when a model performs so well on training data that it has a negative impact on its performance when applied to new data. A dropout layer is utilized to solve this problem in which some neurons and their connections are randomly discarded from the network during the training process.Figure (2.14) depicts the dropout process [M. Arif Wani and et al,2020].

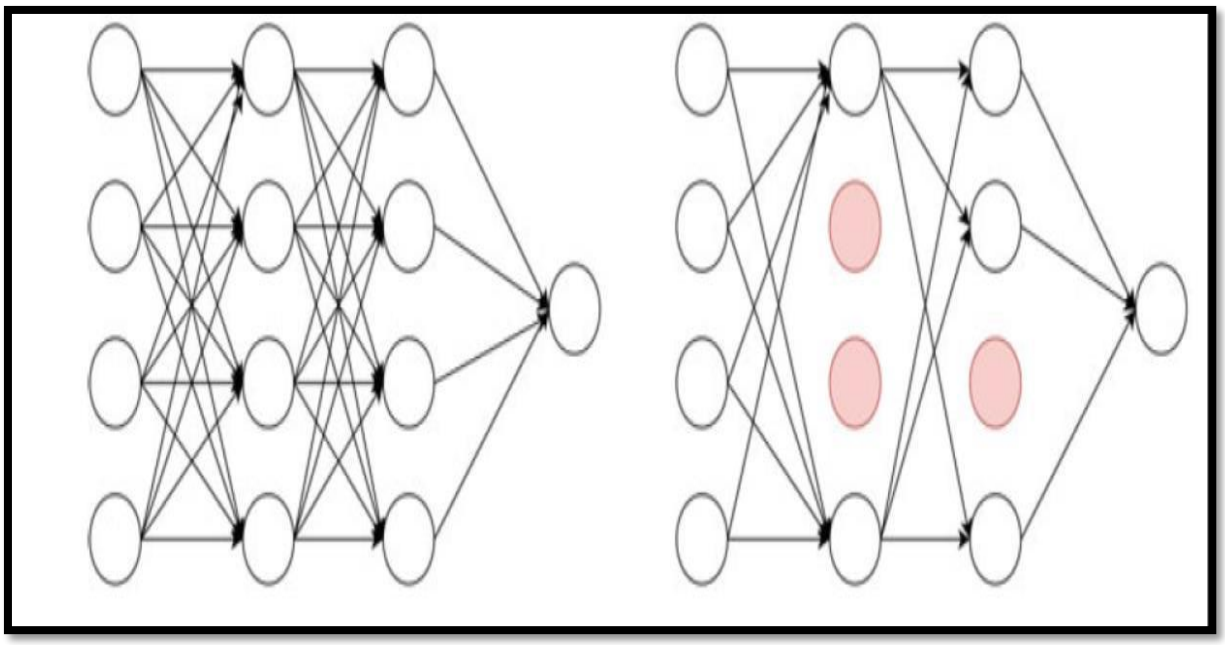


Figure (2.14) : The Dropout Layer[M. Arif Wani and et al,2020]

2.8 Performance Measures

There are many criteria that can be used to evaluate performance classification algorithms . Accuracy and loss are the most used performance metrics in the field of deep learning. The accuracy is a common metric for assessing the model's power of generalization. The trained model is assessed for accuracy based on the total cases that it correctly predicts when the unknown data is tested . This measure's calculation is based on calculating the confusion matrix. This matrix lists how many occurrences a prediction model correctly or incorrectly predicted [74]. More details in the table (2.2) .

Table (2.2) :Two-Dimensional Confusion Matrix of Classifier System.

Confusion Matrix		**Predicted**	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

The prediction error is recorded by four parameters:

- True Positive (TP) is the positive states that are correctly labeled as positive states.
- False Positive (FP) denotes the negative states that are incorrectly labeled as positive states.
- True Negative (TN) represents the right classification of negative diagnosis.
- False Negative (FN) indicates the positive cases that are incorrectly classified as negative [Ali Narin and et al.,2021]] .

Measures accuracy and loss will be discussed as follows:

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 (2.23) [Hema Shekar Basavegowda and Guesh Dagnew,2020] .

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP + FP + FN + TN} \qquad (2.23)$$

2- Precision: is the percentage of relevant instances in the collection of received instances. The precision measure is computed in equation (2.24)

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2.24)$$

3-Recall: Recall is also called sensitivity, and it represents the percentage of relevant cases that have been received. Equation (2.25) shows how recall can be calculated.

$$\text{Recall} = \frac{TP}{TN+FN} \qquad (2.25)$$

4- Specificity: the true negative rate. It represents the true negative values in the model. Equation (2.26) shows how Specificity can be calculated.

$$\text{Specificity} = \frac{TN}{TP+FP} \qquad (2.26)$$

5- Area Under the ROC Curve (AUC): This curve calculates the area under the ROC curve. The amount of instances that are successfully identified based on various threshold values will be determined, in other words, this refers to the

model's capacity to distinguish between various classes and classify objects. AUC measures separability. It shows how well the model is able to distinguish across classes. Between 0 and 1, the AUC value is rated. The lowest number, 0, stands for a misclassification rate of 100%. And 1 signifies the model's highest ranking and its complete accuracy in categorization. It shows how well the model is able to distinguish across classes. The more accurate the model is in predicting zero classes as zero and one classes as the bigger the AUC, the better the model is at differentiating between healthy and unwell patients [Saja Mahdi,2021]. A graph with the AUC is displayed in figure (2.15).

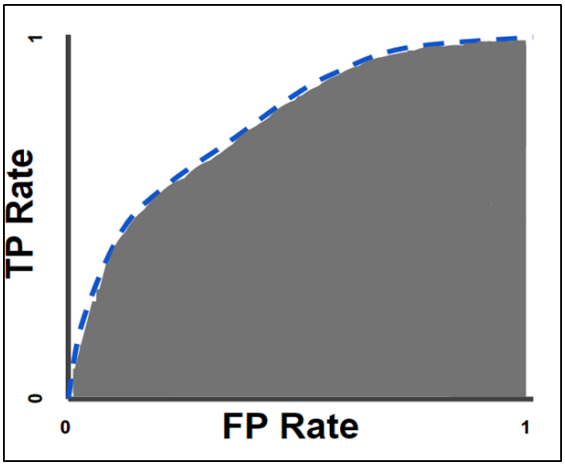


Figure (2.15): AUC curve plotted on graph [Saja Mahdi,2021]

The AUC curve will be employed as the main classification performance parameter in this study because the challenge involves multiclass classification. To obtain a thorough understanding of the model performance, however, the Specificity, Sensitivity, and Confusion Matrix will be used [Saja Mahdi,2021] .

2.9 Data Augmentation (DA)

Unbalanced datasets are a frequent problem that prevent classification in machine learning . The performance of the deep model is significantly impacted by under- and over-fitting which might result from a shortage of genes in each class. This problem is addressed by presenting a data augmentation approach within the (GEO) datasets to enhance classifier performance [Connor Shorten and Taghi M. Khoshgoftaar ,2019].

Data augmentation is a method that increases the number of genes used to train a neural network. By creating new data for categories that have fewer data in the dataset .The data augmentation method is successful in preventing unequal representation and successfully avoiding the problems associated with over-allocation by having a limiting effect on the data. The effectiveness of a deep learning model can be increased with the use of suitable data augmentation techniques [Connor Shorten and Taghi M. Khoshgoftaar ,2019].

Deep neural networks excel at a wide variety of tasks. However, these networks rely to a large extent on big data in order to avoid overfitting.

Due to the enormous complexity of deep neural networks, millions of weights must be tuned and matched to the training set. As a result, having a good model for the issue at hand is essential, as is ensuring that the data collection in is of a suitable scale and quality. To feed and set up a larger, more complicated model that can handle more challenging and real-world challenges, this complexity will demand a larger dataset. However, in the majority of deep learning applications, a neural network's number of parameters frequently outnumbers the dataset's number of data points by a significant amount [Yann LeCun Facebook AI Rese and et al.,2015].Data augmentation means artificially increasing the sample size by modifying the original data instances. [Magdalena Kircher and et al.,2022].

CHAPTER THREE

THE PROPOSED SYSTEM

Chapter Three

The Proposed System

3.1 Introduction

The steps involved to achieve the main objectives of this thesis are discussed in this chapter. This involves formulating a strategy to predict COVID-19 using a number of genes. First , The structure of this system is depicted, After that, the data preprocessing is discussed to prepare the data for subsequent operations, then , the features selection are used to select a subset of the original features, the prediction model used and evaluation methods are discussed subsequently in detail .

3.2 The Proposed System Structure

In this thesis, the proposed system includes four main phases (data processing , feature selection , prediction model , and evaluation of the model used) in order to achieve the basic objectives. In this chapter, an overview of all these phases are given , while detailed description is provided in the following subsections. First, the preprocessing phase of the data includes handling missing data and normalization process, second, feature selection phase, this phase includes identifying the important features by applying a number of methods (Pearson correlation coefficient, mutual information and Principal component analysis) . These phases will help us reduce the large number of genes as well as identifying the most important of these genes that are used in the prediction phase, third , the prediction model was built and implemented using the artificial neural network (ANN) and convolutional neural network (CNN). Finally, the suggested model's findings were assessed using several approaches to determine the projected class. The suggested system's general structure is depicted in figure (3.1).

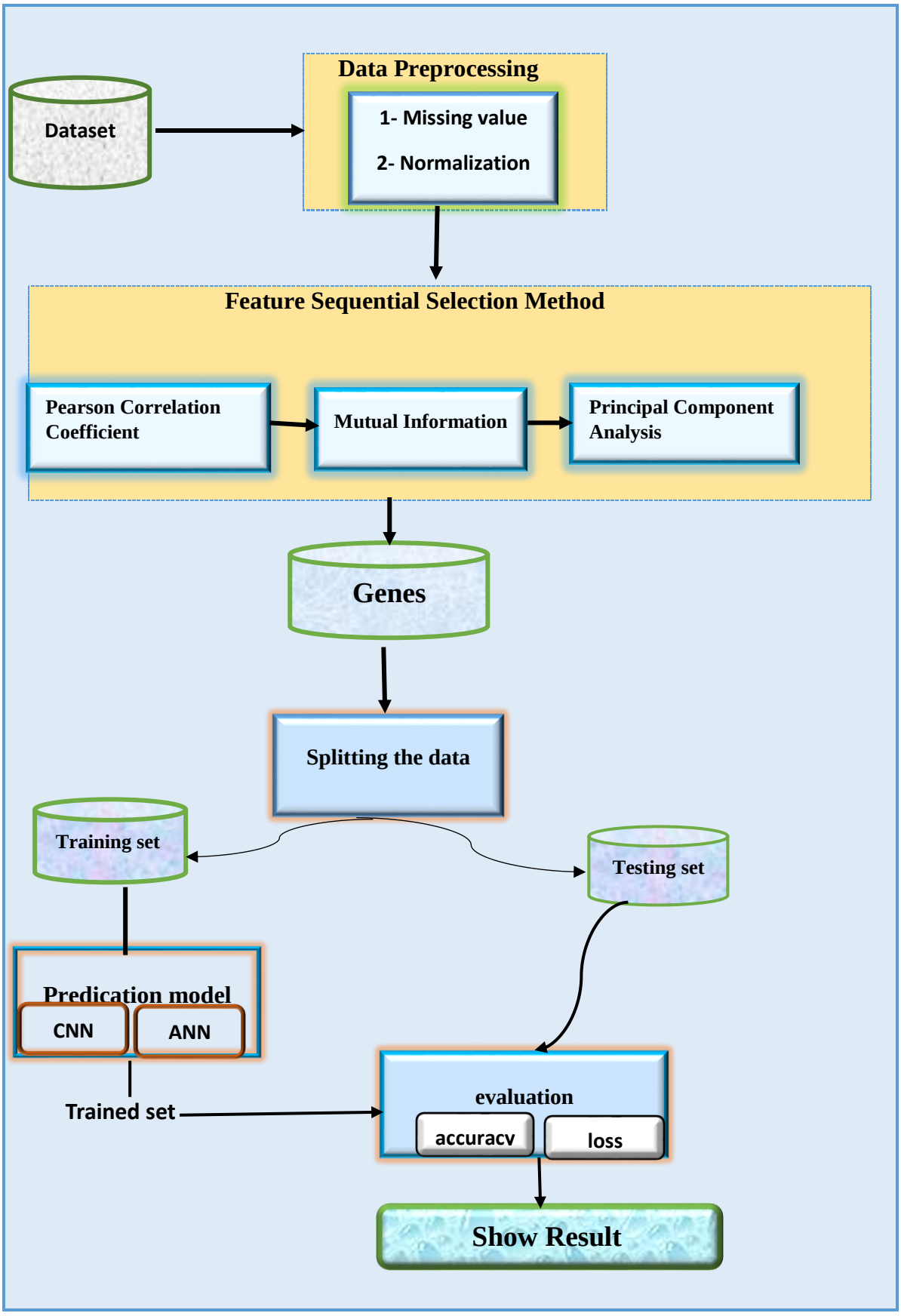


Figure (3.1) The Proposed System

3.2.1 Preprocessing the COVID-19 Dataset

The proposed system consists of a number of phases. The Preprocessing is the first phase of the system. At this phase, the raw data is converted into an efficient and easy-to-understand format. It is a very important process to obtain a reliable dataset that is used in deep learning techniques (prediction models). Preprocessing consists of a number of stages:

a) Missing Value Stage

In this Stage, the missing value is an empty cell for the missing gene in the COVID-19 dataset or this cell may contain the letter N , or it may contain the word nuall which indicates a missing value which controls the results of the calculation and gives inaccurate results . There are different ways to deal with missing values. In this thesis, the missing values were dealt with by replacing the missing value with the mean, depending on the values of other features within the same column. See algorithm (3.1)

Algorithm (3.1): missing value

Input : *two dimensional array* $A[n * m]$ *in which n is the n number of samples and m is number of genes*

Output : *Two dimensional array MV* $[n * m]$*after handling missing value process*

Begin

1. $for\ i\ = 1\ to\ n$

2. $df\ =\ True$

3. $for\ j\ =\ 1\ to\ m$

4. $If\ DNij\ not\ find\ in\ Dataset$

5. df = False

6. $End\ for$

7. $If\ df\ =\ True$

8. *End for*

9. *Replace the missing value with the mean based on the values of other features within the column containing the missing values*

End

b) Normalization Stage

The normalization stage was used for all numerical genetic values within the COVID-19 dataset , which are inputs for deep learning algorithms . This process was implemented to avoid genes with large values controlling the calculation results. All values are normalized to be within the constant range between zero and one using the min-max normalization method mentioned in the second chapter in equation (2.1) and algorithm (2.1) in section (2.4.2)

3.2.2 Feature Selection

In the COVID-19 dataset, a feature selection approach was used to reduce feature space dimensions, select the important features for the prediction process, and improve prediction accuracy. A subset of the most useful features are selected as a result of this phase. Feature selection techniques were used first. However, a feature sequential selection (FSS) approach has been proposed to reduce the number of features and increase prediction accuracy because a very high degree of accuracy cannot be achieved. Since the COVID-19 dataset includes a huge number of genes, the main stage in this thesis is the selection of appropriate features to reduce the curse of dimensionality.

In the feature sequential selection phase different stages were used for feature selection initially. Prior to the prediction model, these techniques are used to select feature subsets based on how well they perform in the prediction model. The goal of the FSS technique is to find a subset of the most informative genes. three strategies for feature selection were used in this thesis because the COVID-19 dataset includes a large number of genes. To select the best subset of features,

these techniques were applied sequentially. The subset with the highest resolution at the end of this stage indicates the ideal feature set for the prediction.

a) Pearson Correlation Coefficient Stage

The Pearson correlation method is implemented to select genes of interest and eliminate unlinked genes. The Pearson correlation coefficient between genes and class label in the dataset is calculated according to equation (2.2) which was explained in the previous chapter. The data are arranged in descending order from the genes of greatest importance to the genes of least importance, then a subset of the important genes is selected to be used in other processes. for more details, see Algorithm (3.2).

Algorithm (3.2): Pearson correlation coefficient Method .

Input : *Two dimensional array DN* $[n * m]$*in which n is the number Sample and m is the number of features*

Output : *Reduced Dataset* (RDP): *feature's subset that have the highest of pearson correlation coefficient values*

Begin

1. *Set features _subset to NULL. // Initialize the features_ subset is empty .*

2. *for* $i = 1$ *to* m // where m: number of features

3. *for* $j = 1$ *to* y //where y: number of classes

4. *Compute the Pearson correlation coefficient betweenfeature Fi and class label Yi according to the equations in section* (2.5.1)

5. *end for j*

6. *end for i*

7. *Select the features Fi with maximum PXY* $(Fi . Yi)$

8. *The selected features are sorted by the Pearson correlation coefficient values in descending order*

End

b) Mutual Information Stage

A mutual information filtering approach is used to identify important genes while removing unwanted genes from the dataset, the mutual information between genes and class naming is determined using equation (2.6) which is explained in the second chapter. The gene with the highest weight is selected for the mutual information, and the gene with the lowest weight is filtered out from the gene pool. As a result, the genes are sorted in descending order from maximum to minimum based on weights of mutual information. Then, a subset of previously selected genes is passed and used as input to the next process. call the algorithm (2.2) mentioned in Chapter second in Section (2.5.2)

c) Principal Component Analysis stage

The primary purpose of principal component analysis is to reduce the data's dimensionality, resulting in the creation of a new collection of variables while maintaining as much of the original data as possible. This method, in general, converts a set of correlated variables into a set of linearly non-correlated variables known as principal components (PCs). The PCs are calculated using the covariance matrix's eigenvalues and eigenvectors. The biggest eigenvalue (i.e., maximum variance) is assigned to the first principal component (PC 1), while the smallest eigenvalue is assigned to the last principal component (PC 2). As a result, PCA has the capacity to maximize the dataset's variability while minimizing its dimensionality. As a result, the data size can be lowered by removing the weaker PCs with low variance. The strongest PCs are then sorted in order of strength and used as inputs for further analysis. call the algorithm (2.3) mentioned in chapter two in Section (2.5.3)

3.2.3 Building Prediction Model Using Reduced Features Set

Before starting the application of the prediction model, the dataset resulting from the previous steps is divided into two phases called the training set and the test set. Usually, when the dataset is divided into a training set and a test set, the largest part of the data is used for training, and a small part is used for testing. In this thesis, (65%) of the dataset was used to train the model through passing in the three phases of the above , as for the rest of the dataset, (35%) was used to test the model after fully trained . In general, the most important stage in the proposed system is the construction of a prediction model. In this thesis, two deep learning algorithms are applied which are artificial neural networks and convolutional neural networks.

a) An Artificial Neural Network Model

ANN is used in this thesis because it is easy to examine while also being able to manage the COVID-19 dataset. It can also achieve the maximum level of prediction accuracy. An ANN model is used with well-chosen input data as it has been shown to provide satisfactory results and improve prediction performance. As a result, the proposed system used the ANN model to predict class labels. Table (3.1) shows the structure of the ANN as follows:

Table (3.1) The proposed structure of the Artificial Neural Network (ANN) model

Name layer	Number of node	Activation function
1- Dense	20 nodes	Relu
2-Dense	40 nodes	Relu

3-Dense	20 nodes	Relu
4- Dense	40 nodes	Relu
5-Dense	3 nodes	Sigmoid

Algorithm (3.3) shows Structure of ANN model used

Algorithm (3.3): The Artificial Neural Network

Input: *Two dimensional array PCA* $[n * m]$ *in which n is the number of Samples and m is number of genes* // Output of Algorithm (2.3)

Output : *Optimized ANN*

Begin

1. Assign Random weights to all the linkages in ANN layers

2. *Start the forward propagation of the input data*

3. Each node output is calculated by multiplying the inputs by the weights according to the equations(2.10)

4. *The node output is passed through ReLU activation function according to the equations* (2.12)

5. *The output of final layer with three nodes are passed through Sigmoind activation function according to the equations* (2.13)

6. *The loss function is calculated by using sparse categorical crossentropyaccording to the equations* (2.15)

7. *Optimizer Adam will update the weights depending on the loss function*

8. *Steps* 3: 8 *are repeated until the maximum number of epochs are met* (150 *epoch*)

End

b) Convolution Neural Network Model (CNN)

In this thesis, another prediction model is used which is the convolutional neural network. The proposed system uses the CNN model to predict class labels. Table (3.2) shows the structure of the convolutional neural network as follows

Table (3.2): The CNN Model Proposed Structure.

Layer (type)	Layer information	Output shape	Params number
layer_1 (Conv1D)	No. of filter =16 ,filter size (3,3), activation = RelU	(None, 198, 16)	64
Flatten (Flatten)	/	(None, 3168)	0
layer_3 (Dense)	64	(None, 64)	202816
layer_4 (Dense)	Number of Classes	(None, 3)	195
Total Parameters:	203,075		

Figure (3.2) shows the structure of the CNN model used which is built to extract features according to pre-defined criteria and is proposed to predict COVID-19 indicators to provide the desired results.

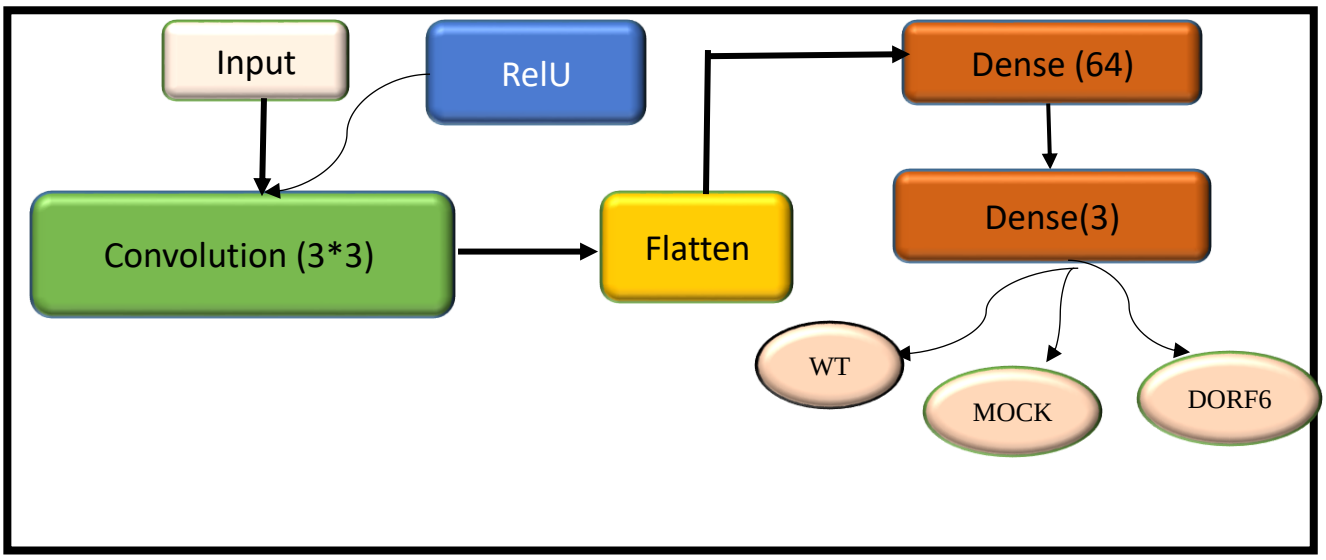


Figure (3.2): Structure of the Used CNN Model

The following algorithm (3.4) shows Structure of CNN model used

Algorithm (3.4): Convolution Neural Network

Input: *Two dimensional array PCA* $[n * m]$ *in which n is the number of Samples and m is number of genes* // Output of Algorithm (2.3)

Output : *Optimized CNN*

Begin

1. set random values for the CNN parameters (convolutional kernels. and CNN weights)

2. *Pass forward the input data through the CNN model for feature extraction according to the equations* (2.19)

3. *Pass the output of the convolutional layer though ReLU activation function according to the equations* (2.12)

4. *pass the extracted features through Flatten layer*

5. *Pass the flattened data through Fully Connected layers according to the equations* (2.22)

6. *Pass the output of the Fully Connected layer though ReLU activation function according to the equations* (2.13)

7. *pass the extracted features through classification layer. by using SoftMax activation function according to the equations* (2.14)

8. *Calculate the loss function by using sparse categorical crossentropy according to the equations* (2.15)

9. *Use SGDM optimizer to update the parameters of the model (kernels. and CNN weights)*

10. *Calculate the validation value for each iteration*

11. *Repeat the steps* 3: 11 *until the max epochs are reached* (300)

End

3.2.4 Evaluation of the Proposed Model

The model evaluation phase is the last phase of the proposed system. In this phase , the accuracy of the ANN and CNN model is evaluated depending on the test set , in order to obtain a more reliable evaluation. The performance of the adopted model is evaluated using training and testing on the dataset . The proposed model was trained several times while still controlling for loss and accuracy in the training and validation dataset. After each epoch, the proposed model results in a loss of training and accuracy, as well as a loss of validation and accuracy. In general, there are a lot of other scales used such as accuracy, recall, Precision and loss function, these measurements are of great importance in supporting Performing the model used, in addition to verifying the validity of the obtained result.

In addition, some experiments were conducted on this data, where ANN and CNN were trained on another data set, and different results were obtained, which will be explained in the fourth chapter.

CHAPTER FOUR

EXPERIMENTAL RESULTS AND DISCUSSION

Chapter four

Experimental Results and Discussion

4.1 Introduction

The activities and processes of the proposed system for identifying genes associated with the severity of COVID-19 disease were explained in the previous chapter. As for the fourth chapter, the results of the performance of the system will be discussed, which is divided into two phases: training and testing. This chapter contains four parts: the preprocessing results, the feature selection results, the results of the steps of the structure of each of the artificial neural network (ANN) and the convolutional neural network (CNN) and the total outputs for each step in the proposed technique, , as well as the evaluation of the proposed system by calculating the performance using gene expression , and these are the main topics covered in this chapter.

4.2 Hardware and Software Requirements

Specific software and hardware requirements are necessary to carry out the process of developing, training, and evaluating the system.

4.2.1 Hardware Requirements:

The proposed COVID-19 classification system operates by employing personal computer hp that have parameters such us Intel(R) Core i7- 10510U @ 1.80 GHz 2.30 GHz for CPU, 8 GB, Windows10 of RAM and 64-bit Operating System.

4.2.2 Software Requirements

The suggested system was developed using Google's Colab notebook and Python version 3.6.9. ANN coding is simply implemented when used with Python. The system depends on open source libraries including Open, Scikit

Learn, and Pandas. It also uses Google's TensorFlow framework (an open-source software library designed for efficient work with tensors) and Keras (a neural network library based on TensorFlow built in Python and also open-source). These libraries are specialize in handling data analysis and machine learning.

4.3 Description of COVID-19 Dataset

The COVID-19 dataset used in this thesis includes many genes and samples that indicate characteristics of an individuals (such as histology, condition, race, age, and gender). It contains (33630) genes, 99 samples consisting (33) controls , which are represented by non-infected people (MOCK) , (33) samples representing people with corona disease (DORF6) , and (33) samples representing a mutation (WT). The first column (sample) contains an identifier for each gene. Each column contains a label describing the patient's status, whether the patient is (DORF6, MOCK, or WT). The remaining values in the dataset represent gene expression levels. For additional details, see Table (4.1) which shows the summary of the dataset used.

Table (4.1) A Brief Description of COVID-19 Dataset

Title of The Dataset:	COVID **-19 dataset**
Dataset characteristics	Multivariate
Number of Samples	99
Number of genes	33629
Number of Class Labels	3

Figure (4.1) shows class labels in the COVID-19 dataset, and figure (4.2) displays an example of COVID-19 dataset values in Excel format, for further information.

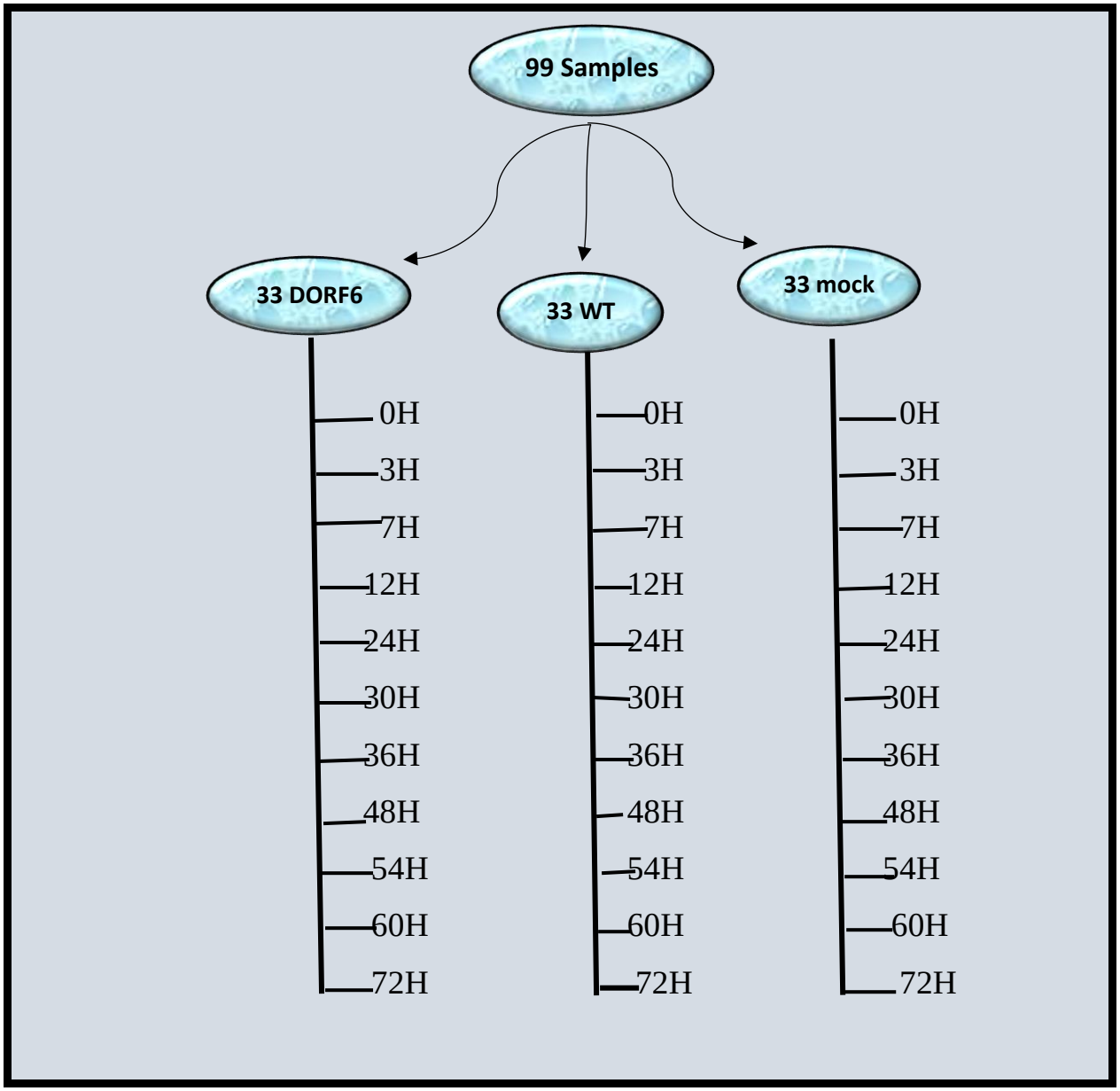
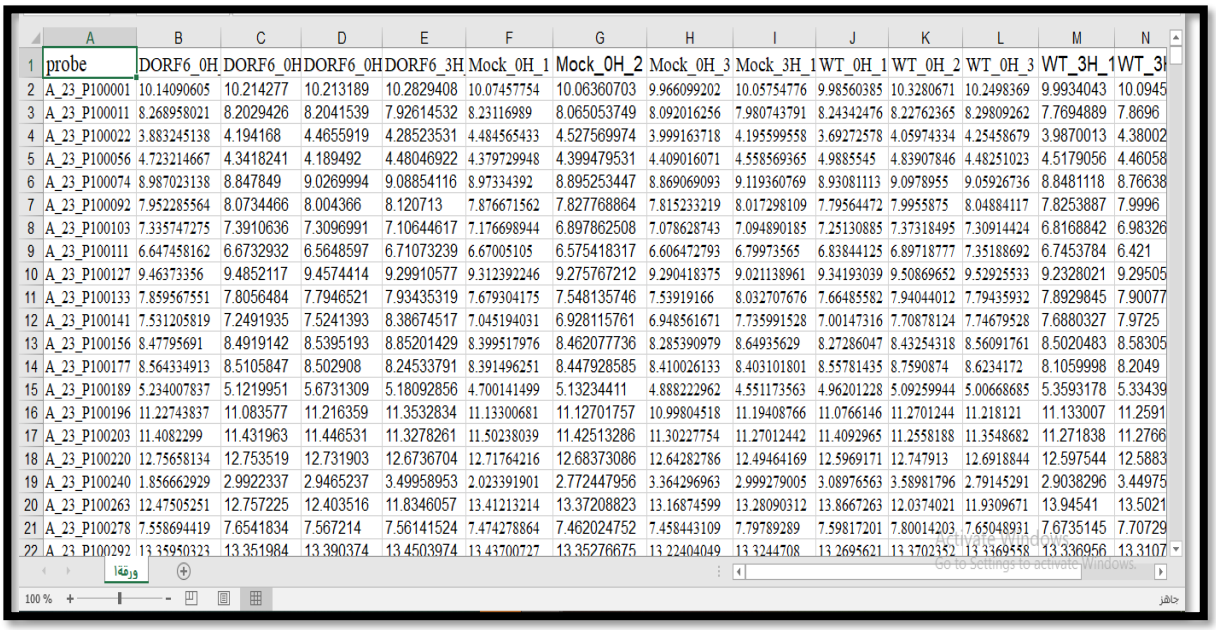


Figure (4.1): The Class Labels in the COVID-19 Dataset



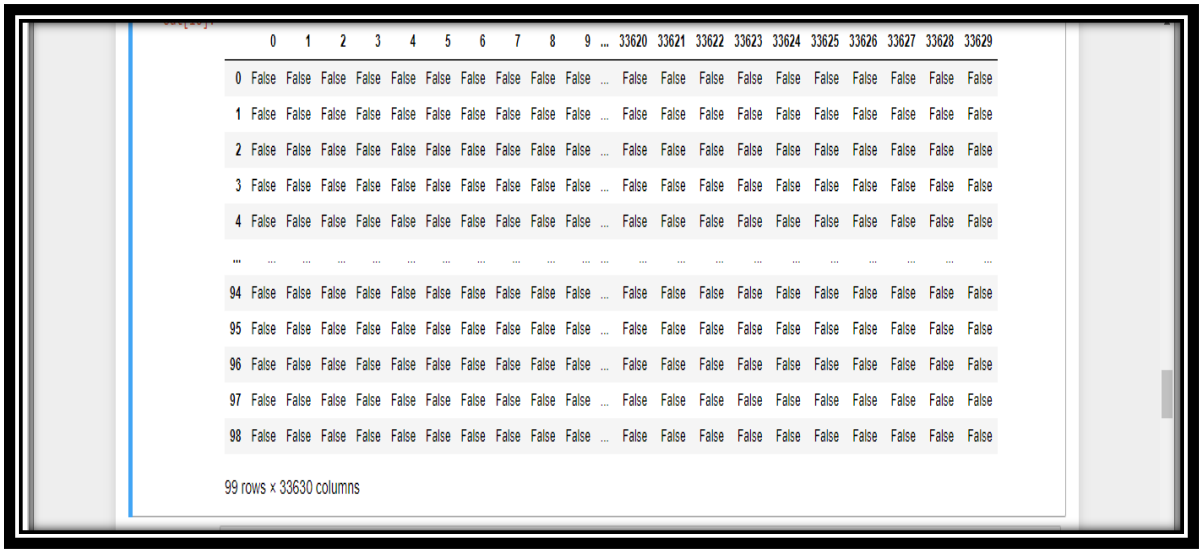
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	probe	DORF6_0H	DORF6_0H	DORF6_0H	DORF6_3H	Mock_0H_1	Mock_0H_2	Mock_0H_3	Mock_3H_1	WT_0H_1	WT_0H_2	WT_0H_3	WT_3H_1	WT_3H
2	A_23_P100001	10.14090605	10.214277	10.213189	10.2829408	10.07457754	10.06360703	9.966099202	10.05754776	9.98560385	10.3280671	10.2498369	9.9934043	10.0945
3	A_23_P100011	8.268958021	8.2029426	8.2041539	7.92614532	8.23116989	8.065053749	8.092016256	7.980743791	8.24342476	8.22762365	8.2980926	7.7694889	7.8696
4	A_23_P100022	3.883245138	4.194168	4.4655919	4.28523531	4.484565433	4.527569974	3.999163718	4.195599558	3.69272578	4.05974334	4.2545867	3.9870013	4.38002
5	A_23_P100056	4.723214667	4.3418241	4.189492	4.48046922	4.379729948	4.399479531	4.409016071	4.558569365	4.9885545	4.83907846	4.4825102	4.5179056	4.46058
6	A_23_P100074	8.987023138	8.847849	9.0269994	9.08854116	8.97334392	8.895253447	8.869069093	9.119360769	8.93081113	9.0978955	9.0592673	8.8481118	8.76638
7	A_23_P100092	7.952285564	8.0734466	8.004366	8.120713	7.876671562	7.827768864	7.815233219	8.017298109	7.79564472	7.9955875	8.0488411	7.8253887	7.9996
8	A_23_P100103	7.335747275	7.3910636	7.3096991	7.10644617	7.176698944	6.897862508	7.078628743	7.094890185	7.25130885	7.37318495	7.3091442	6.8168842	6.98326
9	A_23_P100111	6.647458162	6.6732932	6.5648597	6.71073239	6.67005105	6.575418317	6.606472793	6.79973565	6.83844125	6.89718777	7.3518869	6.7453784	6.421
10	A_23_P100127	9.46373356	9.4852117	9.4574414	9.29910577	9.312392246	9.275767212	9.290418375	9.021138961	9.34193039	9.50869652	9.5292553	9.2328021	9.29505
11	A_23_P100133	7.859567551	7.8056484	7.7946521	7.93435319	7.679304175	7.548135746	7.53919166	8.032707676	7.66485582	7.94044012	7.7943593	7.8929845	7.90077
12	A_23_P100141	7.531205819	7.2491935	7.5241393	8.38674517	7.045194031	6.928115761	6.948561671	7.735991528	7.00147316	7.70878124	7.7467952	7.6880327	7.9725
13	A_23_P100156	8.47795691	8.4919142	8.5395193	8.85201429	8.399517976	8.462077736	8.285390979	8.64935629	8.27286047	8.43254318	8.5609176	8.5020483	8.58305
14	A_23_P100177	8.564334913	8.5105847	8.502908	8.24533791	8.391496251	8.447928585	8.410026133	8.403101801	8.55781435	8.7590874	8.6234172	8.1059998	8.2049
15	A_23_P100189	5.234007837	5.1219951	5.6731309	5.18092856	4.700141499	5.13234411	4.888222962	4.551173563	4.96201228	5.09259944	5.0066868	5.3593178	5.33439
16	A_23_P100196	11.22743837	11.083577	11.216359	11.3532834	11.13300681	11.12701757	10.99804518	11.19408766	11.0766146	11.2701244	11.218121	11.133007	11.2591
17	A_23_P100203	11.4082299	11.431963	11.446531	11.3278261	11.50238039	11.42513286	11.30227754	11.27012442	11.4092965	11.2558188	11.3548682	11.271838	11.2766
18	A_23_P100220	12.75658134	12.753519	12.731903	12.6736704	12.71764216	12.68373086	12.64282786	12.49464169	12.5969171	12.747913	12.6918844	12.597544	12.5883
19	A_23_P100240	1.856662929	2.9922337	2.9465237	3.49958953	2.023391901	2.772447956	3.364296963	2.999279005	3.08976563	3.58981796	2.7914529	2.9038296	3.44975
20	A_23_P100263	12.47505251	12.757225	12.403516	11.8346057	13.41213214	13.37208823	13.16874599	13.28090312	13.8667263	12.0374021	11.9309671	13.94541	13.5021
21	A_23_P100278	7.558694419	7.6541834	7.567214	7.56141524	7.474278864	7.462024752	7.458443109	7.79789289	7.59817201	7.80014203	7.6504893	7.6735145	7.70729
22	A_23_P100292	13.35950323	13.351984	13.390374	13.4503974	13.43700727	13.35276675	13.22404049	13.3244708	13.2695621	13.3702352	13.3369558	13.336956	13.3107

Figure (4.2): A sample of COVID-19 Dataset Values

4.4 Data Preprocessing Results

4.4.1 Missing Values Processing Results

In this stage, the COVID-19 data will be tested to see if this data contains missing values (NaN) or not. Where missing values are processed by replacing the mean of the column containing missing values, depending on the values of other features . After a number of tests, we noticed that the COVID-19 data is free from missing values, as shown in figure (4.3)



	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 (4.3) Missing Value Result

4.4.2 Normalization Process Results

Normalization is an important step that has been conducted on the COVID -19 dataset to minimize substantial value disparities from dominating the results. The gene values ranged between zero and one by using the min-max normalization approach. Figure (4.4) shows the data normalization on a tiny sample of the COVID -19 dataset .

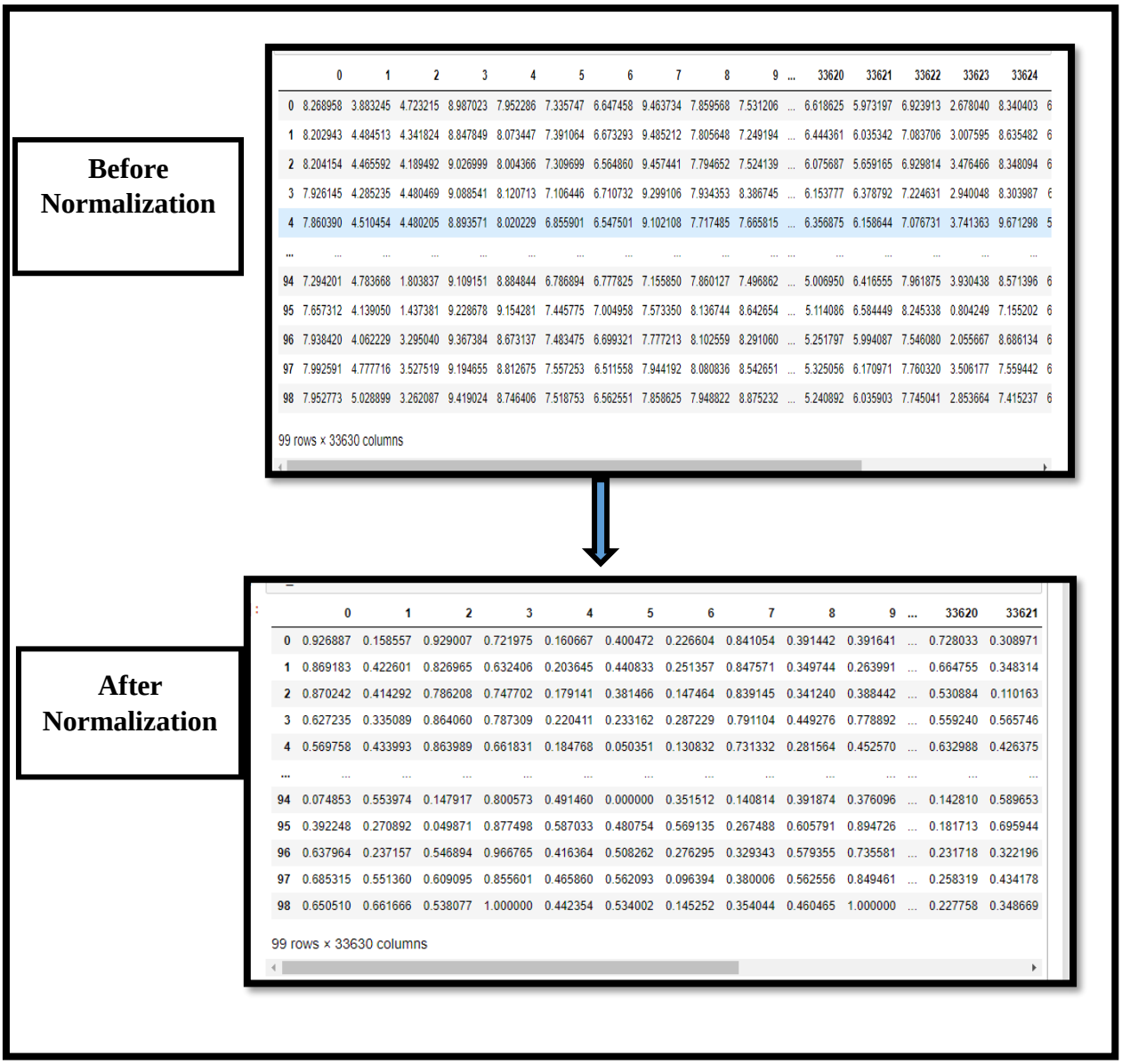


Figure (4.4): The Normalization Result

4.5 Results of Selected Features Methods

The goal of this study is to use the reduction levels for COVID-19 dataset to identify the most informative features and investigate their impact on the prediction model. Although the COVID-19 dataset contains a large number of features , not all of them are suitable for the prediction process, as some of them reduced the prediction model's accuracy and increased the time complexity. For this reason, the feature selection system is employed in this thesis , in order to extract the most important genes that greatly affect COVID-19 .

4.5.1 Result of the Feature Sequential Selection Method

Three feature selection methods were used to better understand the COVID-19 dataset. These methods are used to reduce the dimensions of the dataset and to identify relevant genes associated with COVID-19 disease. The FSS approach was suggested because it significantly greatly improved the accuracy of the prediction model.

4.5.1.1 Results of the Pearson Correlation Coefficient Method

The main goal of this method is to reduce the dimensions of the dataset by identifying relevant genes. The Pearson correlation coefficient method is calculated between a gene and a class label . According to this method, the genes with the highest value were selected. The first (15000) genes out of (33630) genes were selected as informational genes to be used for further analysis.

4.5.1.2 Results of the Mutual Information Method

By locating the pertinent genes, this technique seeks to minimize the dataset's dimensionality calculated using the mutual information approach between the gene and the class label. This approach has resulted in the selection of genes with the highest mutual information value. Out of (15000) genes, the first

(10000) are chosen as the informative genes that will be used in further investigation.

4.5.1.3 Results of the Principal Component Analysis Method

Principle component analysis which identifies a new and condensed group of genes as the principal components is another crucial technique. These major components are arranged in order of maximum inter-component variation. The eigenvectors and eigenvalues derived from the covariance matrix serve as the foundation for the gene selection process in PCA. The threshold determines the number of eigenvectors. As a result, only a small number of eigenvectors with the highest eigenvalues are chosen. Only (198) of the (10000) genes remain after doing the PCA, therefore these chosen genes are then utilized as input for another procedure.

In this thesis, the most informative genes that can aid in COVID-19 prediction have been found using the FSS technique to determine the optimum subset of the original dataset. The FSS technique aims to minimize the number of genes while keeping high accuracy and minimal error before building the prediction model. In general, by selecting the pertinent genes with a reduction rate. Figure (4.5) explains the levels that have been accomplished to reduce the dimensions of the dataset.

Figure (4.5): The Reduction Levels for the COVID-19 Dataset

Table (4.2) The Top Selected Genes from FSS Method

No	Gene ID	No	Gene ID	No	Gene ID	No	Gene ID
1	A_32_P99744	51	A_32_P9107	101	A_32_P857658	151	A_32_P785717
2	A_32_P99280	52	A_32_P90859	102	A_32_P85732	152	A_32_P78488
3	A_32_P98930	53	A_32_P90812	103	A_32_P85591	153	A_32_P7823
4	A_32_P98927	54	A_32_P90709	104	A_32_P8551	154	A_32_P78101
5	A_32_P98847	55	A_32_P90685	105	A_32_P85500	155	A_32_P77665
6	A_32_P9861	56	A_32_P90551	105	A_32_P85495	156	A_32_P77252
7	A_32_P98534	57	A_32_P90346	107	A_32_P85344	157	A_32_P7721
8	A_32_P98136	58	A_32_P89755	108	A_32_P8529	158	A_32_P77178
9	A_32_P98136	59	A_32_P89730	109	A_32_P85279	159	A_32_P77139
10	A_32_P9753	60	A_32_P89691	110	A_32_P84772	160	A_32_P76853
11	A_32_P97506	61	A_32_P89646	111	A_32_P84580	161	A_32_P76842
12	A_32_P97489	62	A_32_P89169	112	A_32_P84454	162	A_32_P76602
13	A_32_P_97315	63	A_32_P889903	113	A_32_P84580	163	A_32_P76576
14	A_32_P_97243	64	A_32_P88905	114	A_32_P84454	164	A_32_P76526
15	A_32_P_9700	65	A_32_P88764	115	A_32_P84388	165	A_32_P76399
16	A_32_P96858	66	A_32_P886243	116	A_32_P84119	166	A_32_P76025
17	A_32_P96838	67	A_32_P88605	117	A_32_P841049	167	A_32_P75792
18	A_32_P96692	68	A_32_P88603	118	A_32_P840463	168	A_32_P75695
19	A_32_P96579	69	A_32_P88598	119	A_32_P84009	169	A_32_P75661
20	A_32_P963	70	A_32_P88555	120	A_32_P83845	170	A_32_P75563
21	A_32_P95949	71	A_32_P88479	121	A_32_P83579	171	A_32_P755542
22	A_32_P95823	72	A_32_P88349	122	A_32_P835626	172	A_32_P75357
23	A_32_P95757	73	A_32_P88327	123	A_32_P83453	173	A_32_P752551
24	A_32_P95176	74	A_32_P88252	124	A_32_P83305	174	A_32_P75141
25	A_32_P95176	75	A_32_P88231	125	A_32_P832555	175	A_32_P74964
26	A_32_P9491	76	A_32_P88163	126	A_32_P831725	176	A_32_P74932
27	A_32_P_94722	77	A_32_P8806	127	A_32_P83000	178	A_32_P74901
28	A_32_P_94722	78	A_32_P87993	128	A_32_P82907	179	A_32_P74831
29	A_32_P94122	79	A_32_P87872	129	A_32_P82671	180	A_32_P74588
30	A_32_P93868	80	A_32_P8772	130	A_32_P82530	181	A_32_P74366
31	A_32_P93782	81	A_32_P87703	131	A_32_P82515	182	A_32_P743407
32	A_32_P93597	82	A_32_P877	132	A_32_P82251	183	A_32_P74213
33	A_32 _ P935950	83	A_32_P875758	133	A_32_P82179	184	A_32_P73707
34	A_32_P93525	84	A_32_P87568	134	A_32_P82119	185	A_32_P73654
35	A_32_P93442	85	A_32_P874898	135	A_32_P81324	186	A_32_P73535
36	A_32_P_P39209	86	A_32_P87323	136	A_32_P81092	187	A_32_P73452
37	A_32_93144	87	A_32_P8732	137	A_32_P80897	188	A_32_P73413
38	A_32_P92922	88	A_32_P871061	138	A_32_P80697	189	A_32_P73217
39	A_32_P92814	89	A_32_P867789	139	A_32_P80678	190	A_32_P7316
40	A_32_P926336	90	A_32_P86739	140	A_32_P8067	191	A_32_P7308
41	A_32_P92563	91	A_32_P86705	141	A_32_P80413	192	A_32_P72611
42	A_32_P925529	92	A_32_P865613	142	A_32_P80089	193	A_32_P72541
43	A_32_P92281	93	A_32_P86517	143	A_32_P80016	194	A_32_P725218
44	A_32_P92212	94	A_32_P86400	144	A_32_P79190	195	A_32_P72447
45	A_32_P91821	95	A_32_P86318	145	A_32_P79103	196	A_32_P72203
46	A_32_P91633	96	A_3__P86173	146	A_32_P790284	197	A_32_P721983
47	A_32_P91519	97	A_32_P86150	147	A_32_P78876	198	A_32_P72110
48	A_32_P91507	98	A_32_P86135	148	A_32_P78809		
49	A_32_P914221	99	A_32_P86118	149	A_32_P78681		
50	A_32_P91228	100	A_32_P85835	150	A_32_P78649		

4.6 Genes Exploration

The following figures show the traceability of a gene over time and the changes that can occur. In the figures below different plots approaches are used to explore the genes

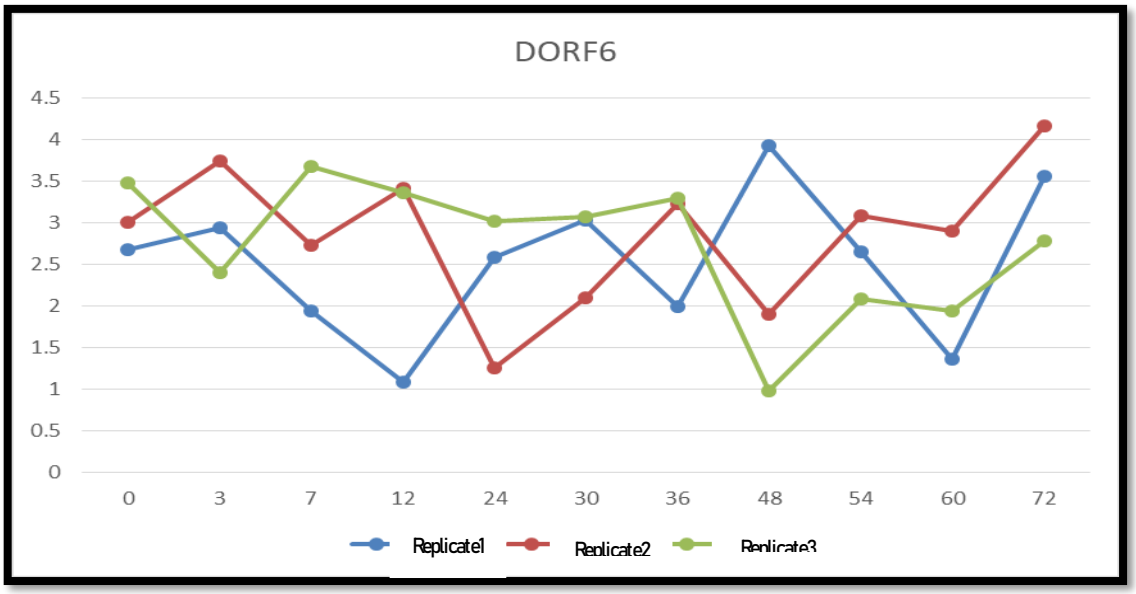


Figure (4.6) The gene (A_32_P99744) with class DORF6 for three different replicates during different hours, where x-axis is the hours and y-axis is gene value.

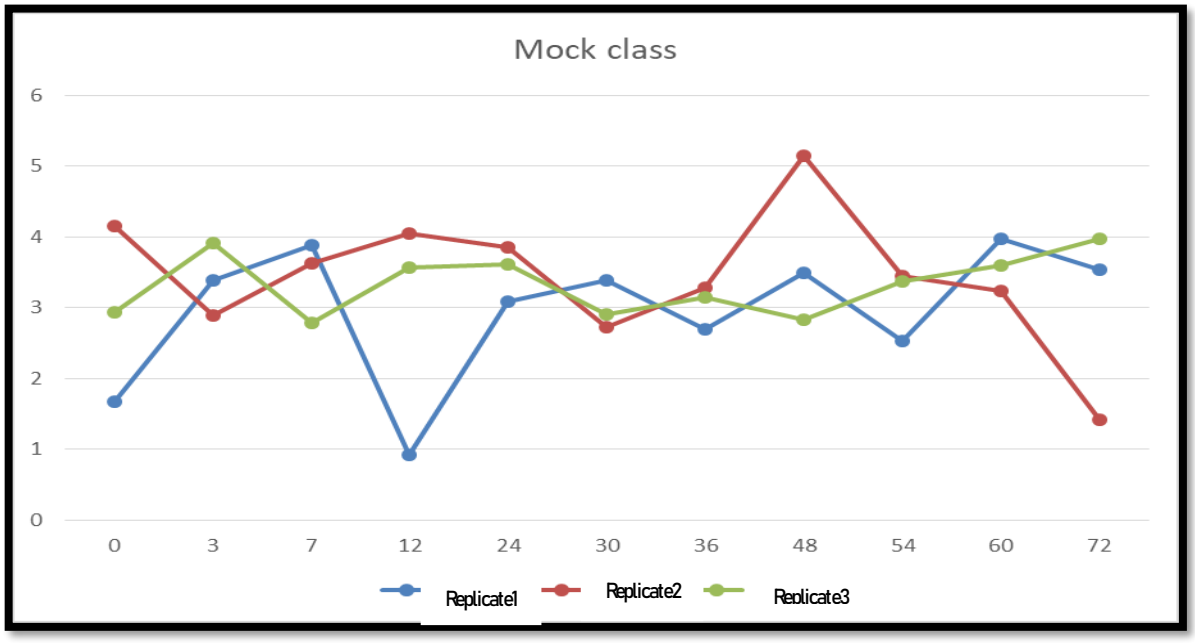


Figure (4.7) The gene (A_32_P99744) with class MOCK for three different replicates during different hours, where x-axis is the hours and y-axis is gene value.

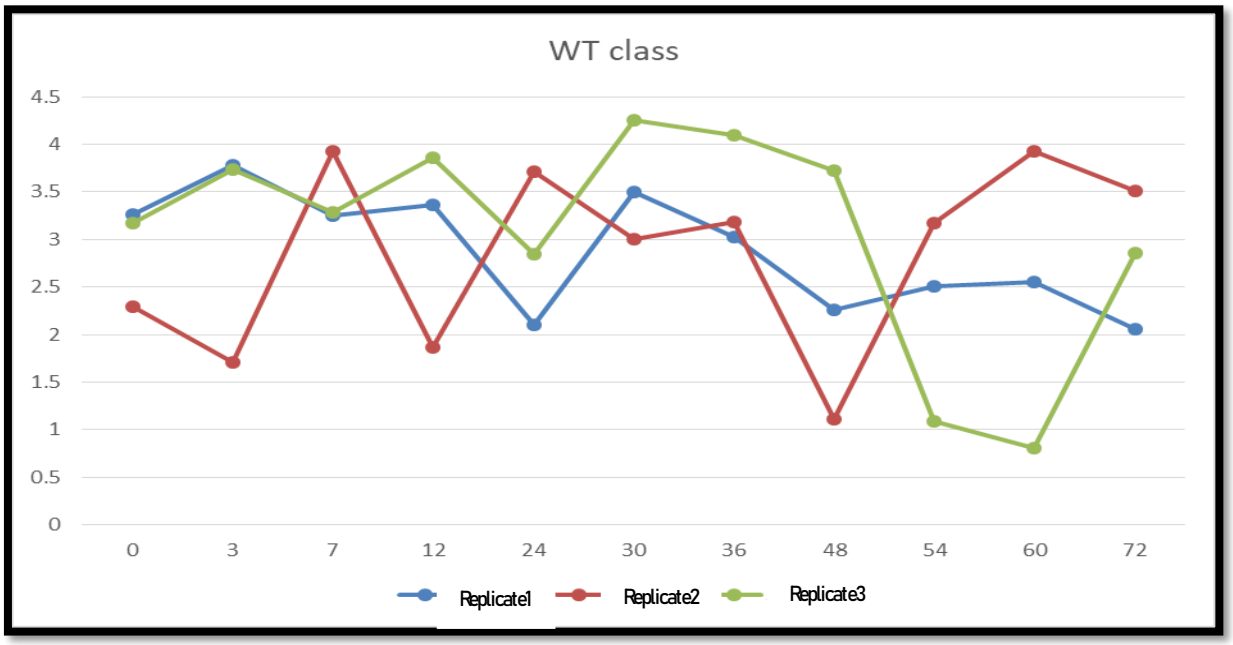


Figure (4.8) The gene (A_32_P99744) with class WT for three different replicates during different hours, where x-axis is the hours and y-axis is gene value.

4.6 Results of the Prediction Model

4.6.1 Results of Artificial Neural Network Model

In order to predict the genes associated with COVID-19 disease, artificial neural network (ANN) model was developed. The gene of interest for the highest level of precision constitutes the best genetic subset of the FSS approach. Generally, an ANN model is generated using Python Keras software. A high-level neural network toolkit called Keras has been created to speed up experimentation. At the top level, passing the output from one layer to another does not require writing code. All need to know is the architecture of the neural network and the input data needed to train it. In particular, the ANN model .It was chosen and its architecture consists of five layers, including three Hidden layers, with an input layer and an output layer.

The first layer is the input layer (Dense) with 20 neurons, and activation layer is RelU. The second is hidden layer (Dense) with 40 neurons and activation

layer is RelU. The third layer is hidden layer (Dense) with 20 neurons and activation layer is RelU .The fourth layer is hidden layer(Dense) with 40 neurons and activation layer is RelU. The fifth layer is output layer (Dense) with 3 neurons and activation layer is sigmoid.

In addition, categorical cross entropy as the loss function to compute the loss, accuracy as a metric and the adaptive moment estimation (ADAM) optimizer with batch_size=15 . This model is trained with 150 epochs.

4.6.2 Results of Convolution Neural Network model

In order to predict the genes associated with COVID-19 disease, a second prediction model, the Convolutional Neural Network (CNN) was developed. Generally, a CNN model is generated using Python Keras software. A high-level neural network toolkit called Keras was created to speed up experimentation, and a four-layer CNN model was chosen. The first layer is convolution layer with 16 different filters , kernel size 3*3 and activation function RelU. The second layer is flatten layer. The thired layer is fully connection layer with 64 neurons The last fully connected layer has 3 neurons to predict 3 classes for the COVID-19 dataset in layer number fourth and uses a softmax layer as a prediction layer. In addition, categorical cross entropy as the loss function to compute the loss, accuracy as a metric and the adaptive moment estimation (SGD) optimizer with learning-rate 0.001 are selected for CNN model. This model is trained with 300 epochs.

4.7 Evaluating the Proposed Model

The performance of the prediction model (artificial neural network and convolutional neural network) was evaluated. Usually all genes that have appeared are taken as influencer genes. Then, they were added to the model to confirm their importance and the level of accuracy that could be achieved by incorporating these genes. In fact, the use of selected genes greatly improves accuracy. The results of the FSS method were satisfactory in the prediction model

because the informative gene selection process required multiple steps and took advantage of the majority of gene selection strategies. The performance of the FSS approach can be demonstrated effectively especially when compared to a raw dataset without feature selection methods techniques. Table (4.3) shows the results of comparing the accuracy and loss of the artificial neural network model and the convolutional neural network before and after applying the FSS method.

Table (4.3): The Comparative Results of Accuracy and Loss for the ANN and CNN Models

The method	Accuracy	Loss	Used genes
Raw Data +ANN	30%	1.068	33630
Raw Data + CNN	30%	1.104	33630
FSS + ANN	97%	0.041	198
FSS+CNN	82%	0.0691	198

By looking at table (4.3), it is found that the best subgroup consists of 198 genes which were previously explained in table (4.2). As these genes were selected when applying the FSS method based on the ANN and CNN model, the highest accuracy was obtained. Therefore, it is recommended that these selected genes are used as the most important genes. Figure (4.9) shows the graph of raw data accuracy and loss , FSS method is based on ANN and CNN model.

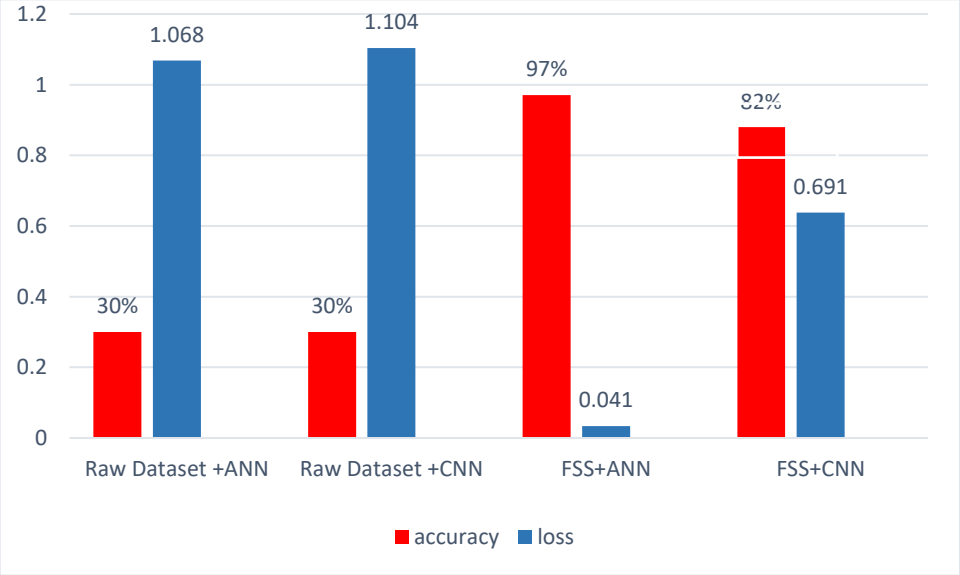


Figure (4.9): Accuracy, and Loss across the ANN and CNN Models

On the other hand, it is taking into account that the outcomes of the gene selection approaches that are used separately are insufficient for the prediction model. The number of genes chosen by each of the three feature selection techniques for the ANN model is compared in table (4.4).

Table (4.4):The Comparative Results of the feature Selection Methods for the ANN Model

The Method	Number of Genes	ANN Model	
		Accuracy	Loss
Pearson	15000	24%	1.135
MI	10000	31%	1.100
PCA	198	97%	0.041

It can be seen that all the three selected feature selection methods can reduce the dimensions of the data, but their results were not satisfactory in the proposed prediction model. Thus, it is clear from the comparison that the proposed method (FSS) outperformed all the methods that were applied in terms of the accuracy of the prediction model. Figure (4.10) shows the accuracy and loss of the ANN model based on the different feature selection methods.

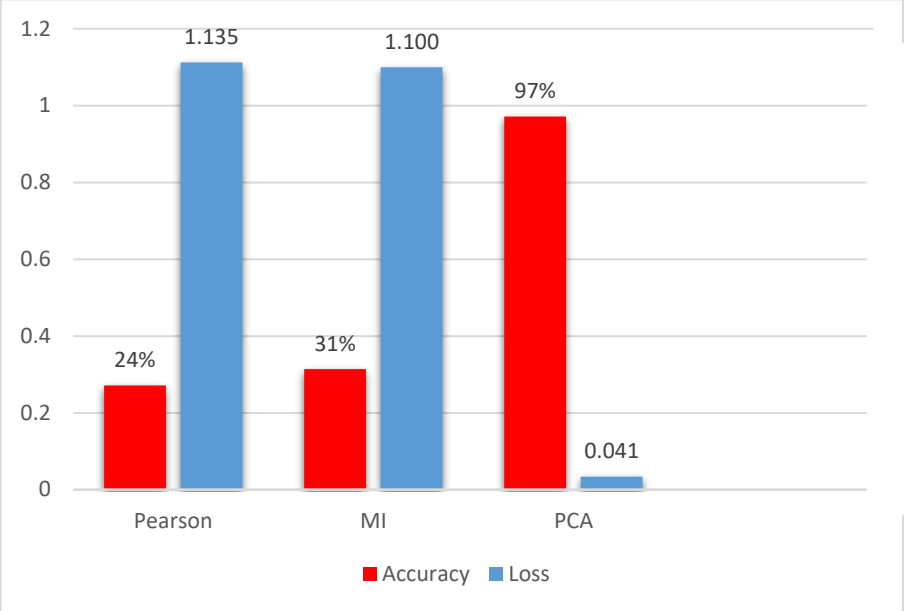


Figure (4.10): Accuracy, and Loss across the ANN Model Based on different feature selection Methods

Table (4.5):The Comparative Results of the feature Selection Methods for the CNN Model

The Method	Number of Genes	CNN Model	
		Accuracy	Loss
Pearson	15000	24%	1.099
MI	10000	32%	1.098
PCA	198	82%	0.069

Through the above table, it is clear from the comparison that the proposed method (FSS) outperformed all the methods that were applied in terms of the accuracy of the prediction model. Figure (4.11) shows the accuracy and loss of the CNN model based on the different feature selection methods.

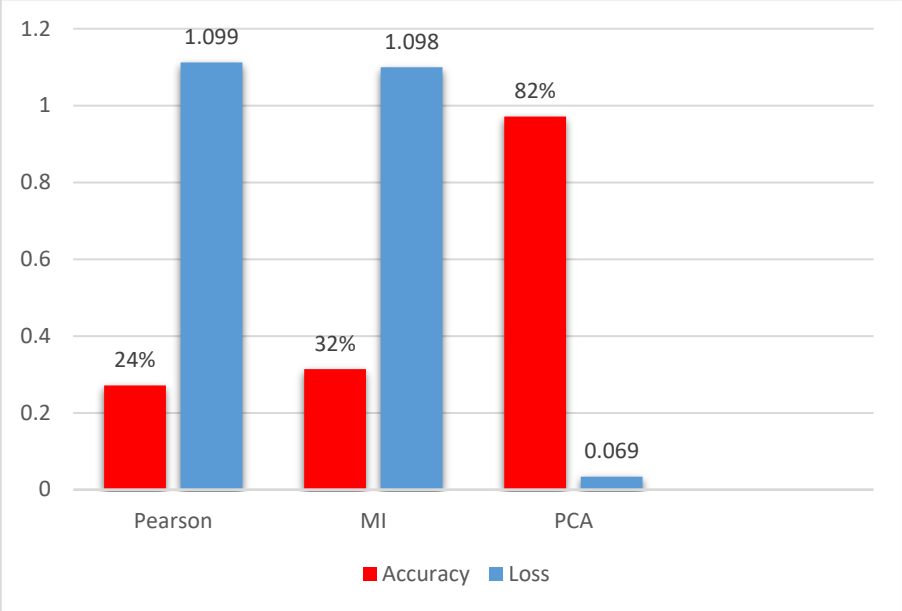


Figure (4.11): Accuracy, and Loss across the CNN Model Based on different feature selection Methods

Table (4.6) The evaluation of the results of the ANN and CNN model with different scales (accuracy, loss, sensitivity, specificity, and area under the curve).

The Model	accuracy	loss	sensitivity	Specificity	area under the curve
ANN	97%	0.041	91%	100%	95%
CNN	82%	0.691	91%	100%	95%

In addition to the previous results, which represent the best results obtained, in this thesis, feature extraction methods with ANN and CNN were applied using different numbers of genes with different percentages of data partitioning. More details in the following tables.

5,000 genes will be used, which represent the best genes. The artificial neural networks will be trained on this data after dividing it into two groups, the training group and the test group. These data were divided by different proportions, and different precisions were obtained. More details in Table (4.7)

Table (4.7):The Comparative Results of the feature Selection Methods for the ANN Model using only 5000 genes

The method	*Testing set=0.2*		*Testing set=0.1*		*Testing set=0.3*		*Used genes*
	accuracy	*loss*	*accuracy*	*loss*	*accuracy*	*Loss*	
Person +A NN	32%	1.099	25%	1.105	26%	1.131	5000
MI+ANN	30%	1.101	25%	1.106	23%	1.137	5000
PCA+ANN	93%	0.310	90%	0.663	93%	0.239	5000

5,00 genes will be used, which represent the best genes. The artificial neural networks will be trained on this data after dividing it into two groups, the training group and the test group. These data were divided by different proportions, and different precisions were obtained. More details in Table (4.8).

Table (4.8):The Comparative Results of the feature Selection Methods for the ANN Model using only 500 genes

The method	*Testing set=0.2*		*Testing set=0.1*		*Testing set=0.3*		*Used genes*
	accuracy	*Loss*	*accuracy*	*loss*	*accuracy*	*Loss*	
Person + ANN	72%	0.552	55%	0.987	60%	0.794	500
MI+ANN	62%	0.848	80%	0.429	83%	0.430	500
PCA+ANN	90%	0.775	90%	1.053	90%	0.731	500

Also 5,000 genes were used, which represent the best genes. Convolutional neural networks were trained on this data after dividing it into two groups, the training group and the test group. These data were divided by different proportions, and different precisions were obtained. More details in Table (4.9).

Table (4.9):The Comparative Results of the feature Selection Methods for the CNN Model using only 5000 genes

The method	*Testing set=0.2*		*Testing set=0.1*		*Testing set=0.3*		*Used genes*
	accuracy	*loss*	*accuracy*	*loss*	*accuracy*	*loss*	
Person + CNN	32%	1.098	30%	1.098	28%	1.096	5000

MI+CNN	35%	1.092	45%	1.098	31%	1.102	5000
PCA+CNN	75%	0.502	65%	0.730	78%	0.529	5000

Also 5,00 genes were used, which represent the best genes. Convolutional neural networks were trained on this data after dividing it into two groups, the training group and the test group. These data were divided by different proportions, and different precisions were obtained. More details in Table (4.10).

Table (4.10):The Comparative Results of the feature Selection Methods for the CNN Model using only 500 genes

The method	***Testing set=0.2***		***Testing set=0.1***		***Testing set=0.3***		***Used genes***
	accuracy	***loss***	***Accuracy***	***loss***	***accuracy***	***loss***	
Person + CNN	50%	1.062	25%	1.098	43%	1.066	500
MI+CNN	72%	0.997	45%	1.098	31%	1.102	500
PCA+CNN	80%	0.791	70%	0.675	81%	0.649	500

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORKS

Chapter five

Conclusions and future works

5.1 Conclusions

The following are the main conclusions gained from the results obtained utilizing the proposed system for detecting COVID-19 in gene expression:

1- The suggested system has demonstrated its effectiveness in detecting relevant genes (the best genes) and deleting irrelevant or harmful genes. Furthermore, according to all conventional assessment metrics, this proposed system produces good outcomes in the prediction model.

2- The proposed system has successfully proved the feature selection approach according to the COVID-19 dataset's nature, with satisfying results in the prediction model.

3- The results indicate that the best accuracy was obtained with the ANN model using only the best genes. The accuracy (0.971) was obtained from the FSS method.

4- The results also indicate that the best accuracy was obtained with the CNN model using only the best genes. The accuracy (0.828) was obtained from the FSS method.

5- The system has been trained and tested to classify COVID-19 data. The results of this system are as follows: 0.916 % sensitivity, 100% specificity, and 0.958% AUC.

6-This thesis demonstrates how deep learning and neural networks aid and progress biology and medical sciences by demonstrating neural networks' remarkable capacity to categorize and forecast diseases.

5.2 The Future Works

The following is recommended for future works:

1- Classifying the genes related with the severity of different diseases using a big dataset containing thousands of genes

2- Apply the proposed approach to other datasets such as diabetes, breast cancer, Alzheimer's disease , and etc.

3- Examine various models that can be compared to the CNN and ANN models and attempting to reduce prediction error.

4- studding Other feature selection methods, such as Information Gain (IG), Singular Value Decomposition (SVD), and others, are being investigated for picking a best subset of genes and highlighting their impact on the prediction model.

5- Compare the results of other classification approaches, such as SVM (Support Vector Machines), with the results of the CNN algorithm and ANN algorithm in classifying genes associated with COVID-19 disease severity

REFERENCES

References

Whata, A., & Chimedza, C. (2021). Deep learning for sars cov-2 genome sequences. Ieee Access, 9, 59597-59611, https://creativecommons .org/licenses/by /4.0/

Babukarthik , R. ,Ananth ,V. and Sambasiva,G.(2020). Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN)", *Ieee Access*, 8 , 177647-177663,https:// 10.1109/ACCESS .2020.3025164

Ye,H.,Wu,P ,Zhu,T.,Xiao,Z., Zhang,X., Zheng,L.,Sun,Y.Zhou,W., Fu,Q., Yu,X., Chan,A.&(2021). Diagnosing Coronavirus Disease 2019 (COVID-19): Efficient Harris Hawks-Inspired Fuzzy K-Nearest Neighbor Prediction Methods . *Ieee Access*, 9, 17787 - 17797, e https://creativecommons.org/licenses/by/4.0/

Habib,N., & Motiur,M.,(2021),Diagnosis of corona diseases from associated genes and X-ray images using machine learning algorithms and deep CNN, *Informatics in Medicine Unlocked*, 24, 2-12, https://doi.org /10.1016/j.imu.2021.100621.

Ramesh, P., Veerappapillai, S., & Karuppasamy, R. (2021). Gene expression profiling of corona virus microarray datasets to identify crucial targets in COVID-19 patients. Gene reports, 22, 100980. https://doi.org/ 10.1016/j .genrep.2020.100980

Singh, P., & Singh, N. (2021). Role of data mining techniques in bioinformatics. International Journal of Applied Research in Bioinformatics (IJARB), 11(1), 51-60,http doi:10.4018/IJARB.2021010106

Babu, M., & Sarkar, K. (2016). A comparative study of gene selection methods for cancer classification using microarray data. Ieee , 204-211,http://doi:10.1109/ ICRCICN.2016. 7813657

Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. Procedia Computer Science, 23, 5-14,http:// doi:10.1016/j.procs.2013.10.003

Nanda,P. and Duraipandian,N.(2020).Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest", *Ieee Access* ,http:// doi: 10.1109/ICICT48043.2020.9112558

Pati, J. (2018). Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach. IEEE Access, 7, 4232-4238,http:// doi: 10.1109/ACCESS.2018.2886604.

Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A., & Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach. IEEE Access, 8, 22874-22883,http:// doi: 10.1109 /ACCESS.2020.2970210

Iqbal,N. & Kumar,P., (2022).Integrated COVID-19 Predictor: Differential expression analysis to reveal potential biomarkers and prediction of coronavirus using RNA-Seq profile data, *Computers in Biology and Medicine*, 140, 1-9 ,http:doi : 10.1016/j. compbiomed .2022.105684

Chen, L., Li, Z., Zeng, T., Zhang, Y. H., Feng, K., Huang, T., & Cai, Y. D. (2021). Identifying COVID-19-specific transcriptomic biomarkers with machine learning methods. BioMed Research International, 2021. https://doi.org/10.1155/2021/9939134

Khan, F., Pradhan ,K. &Sinha,D.,(2022) ,A Model for Lung Cancer Prediction, *(ICAC3N)IEEE*, 251-255,http://, doi: 10.1109/ICAC3N53548.2021.9725462.

Ali,T. , Tawab,M., & ElHariri,M.,(2020).CT chest of COVID-19 patients: what should a radiologist know?, *Egyptian Journal of Radiology and Nuclear Medicine* 51, 120, 2-6,http:// doi.org/10.1186/s43055-020 00245-8?.

Ruan, Q., Yang, K., Wang, W., Jiang, L., & Song, J. (2020). Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive care medicine, 46(5), 846-848., https://doi.org/10.1007/s00134-020-05991-x.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506., https://doi.org/10.1016/S0140- 736(20)30183-5.

Oh,J., & Klivans,L.,(2020) ,"How the Coronavirus Attacks Your Lungs ". May 5, 2020. Date of retrieval 12-1-2021. https://www.kqed.org/science/1963200/how covid- 19-attacks -
-your-lungs.

Wang, Z., & Tang, K. (2020). Combating COVID-19: health equity matters. Nature medicine, 26(4), 458-458., https://**doi.org/10.1038/s41591-020-0823-6**

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., ... & Baillie, J. K. (2021). Genetic mechanisms of critical illness in COVID-19. Nature, 591(7848), 92-98..

Asgari, S., & Pousaz, L. A. (2021). Human genetic variants identified that affect COVID susceptibility and severity

. ***Hiba muthanaa ,2021*** .Gene Expression Classification of Alzeheimar Disease Stages Using Machine Learning", Master Thesis, University of Babylon, College of Information Technology .

Toloo,M., Sohrabi,B., & .Nalchigar,S.,(2008). A new method for ranking discovered rules from data mining by DEA, *journal Expert Systems with Applications*, 36, 8503–8508,http:// doi:10.1016/j.eswa.2008.10.038.

Riksheim,M., Clausen1,T., & Røislien,J.,(2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data , *SAGE Open Medicine*,.7, 1-12, https://doi.org/10.1177/2050312118822912.

Najah Abed, (2021).An Approach to Bacteria Classification based on Topic Modeling and Naive Bayes", Master Thesis, University of Babylon College of Science for Women .

Al Shalabi, L., & Shaaban, Z. (2006, May). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In 2006 International conference on dependability of computer systems (pp. 207-214). IEEE..

Vanjimalar, S., Ramyachitra, D., & Manikandan, P. (2018, December). A review on feature selection techniques for gene expression data. In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-4). IEEE.http:// doi: 10.1109/ICCIC.2018.8782294

Al-Harbi, O. (2019). A comparative study of feature selection methods for dialectal Arabic sentiment classification using support vector machine. arXiv preprint,https://doi.org/10. arXiv:1902.06242,(.48550/arXiv.1902.06242

Bouchlaghem, Y., Akhiat, Y., & Amjad, S. (2022). Feature Selection: A Review and Comparative Study. In E3S Web of Conferences (Vol. 351, p. 01046). EDP Sciences. https://doi.org/10.1051/e3sconf/202235101046

Mahendran, N., Durai Raj Vincent, P. M., Srinivasan, K., & Chang, C. Y. (2020). Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. Frontiers in genetics, 11, 603808.https://doi.org/10. .3389/fgene.2020.6038

Ly,A., Marsman,M., & Wagenmakers ,J,.,(2018).Analytic posteriors for Pearson'scorrelation coefficient, *Statistica Neerlandica* ,. 72,1. 4–13, http://doi:10.1111/stan.12111

Wallot,S., & Monste,M.,(2018) . 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* , l 9, 1697 , https://doi.org/ 10.3389/fpsyg.2018.01679 .

Vanitha, C. D. A., Devaraj, D., & Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. procedia computer science, 47, 13-21, https://doi.org/10.1016/j.procs.2015.03.178

Barraza, N., Moro, S., Ferreyra, M., & de la Peña, A. (2019). Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. Journal of Information Science, 45(1), 53-67. https://doi.org/10.1177/016555151877069

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202., http://dx.doi.org /10.1098/rsta.2015. 0202
Wani. M. A,. Bhat. F. A, Afzal. S, Khan. A. I ,(2020).Advances in Deep Learning , *Srinagar, India, Springer*.

Talreja, V., Valenti, M. C., & Nasrabadi, N. M. (2017, November). Multibiometric secure system based on deep learning. In 2017 IEEE Global conference on signal and information processing (globalSIP) (pp. 298-302). IEEE.http:// doi:10.1109/GlobalSIP .2017.8308652

Maier, A., Syben, C., Lasser, T., & Riess, C. (2019). A gentle introduction to deep learning in medical image processing. Zeitschrift für Medizinische Physik, 29(2), 86-101.https:// doi.org /10.1016/j.zemedi.2018.12.003

Mosavi. A., Ardabil, S., & Koczy. A. R. V.,(2020) .List of Deep Learning Layers, 2, 202–214, http://doi:10.20944/preprints 201908.0152.v1

Mustafa Khalaf,(2021).COVID-19 diagnosis lung CT images based on convolutional neural networks, Master thesis, University of Babylon, College of Information Technology,15-19

Du, K. L., & Swamy, M. N. (2013). Neural networks and statistical learning. Springer Science & Business Media,http://doi.1010071978-1-4471-5571-3

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444., https://doi.org/10.1038/nature14539.

Zhang,Y., & Ling,C.,(2018). A strategy to apply machine learning to small datasets in materials science, *npj Computational Materials*, 4, 25, , https://doi.org/10.1038/s41524-018-0081-z.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. http:// doi.org/10.1038/ s42256-019-0048-x.

Basha, S. S., Dubey, S. R., Pulabaigari, V., & Mukherjee, S. (2020). Impact of fully connected layers on performance of convolutional neural networks for image classification. Neurocomputing, 378, 112-119, https://doi.org/10.1016/j.neucom.2019.10.008, .

Noor Fahem,(2021). Automated Diagnosis of COVID-19 based on Hybrid Machine Learning Approaches, Master Thesis, University of Babylon, College of Science for Women.

Saja Mahdi,(2021).Classification of Macular Degeneration Based on Deep Learning Method, Master Thesis, University of Babylon, College of Science for Women.

Ongsulee, P. (2017, November). Artificial intelligence, machine learning and deep learning. In 2017 15th international conference on ICT and knowledge engineering (ICT&KE) (pp. 1-6). IEEE.http:// doi: 10.1109/ICTKE.2017.8259629.

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. Frontiers in Artificial Intelligence, 3, 4.,https://doi.org/10.3389/frai.2020.00004

Tang, P., Peng, K., Zhang, K., Chen, Z., Yang, X., & Li, L. (2018). A deep belief network-based fault detection method for nonlinear processes. IFAC-PapersOnLine, 51(24), 9-14., https://doi.org/10.1016/j.ifacol.2018.09.522

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE access, 7, 53040-53065.,http:// doi: 1090.1109/ACCESS .2019.2912200.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology, 9(2), 14-14,https://doi.org/10.1167/tvst.9.2.14,

Xavier,(2017).Face recognition using Deep Learning, Master Thesis Polytechnic University of Catalonia, 2017.

Muhind Salim ,(2015). Robust Classification With Convolutional Neural Network, Master Thesis ,University of Missouri-Columbia.

Husam Imad AbdulRazzaq, 2018. Robust Authorization Based on Multi Biometric Techniques", Master thesis, University of Technology ,2018.

Ján Vojt,(2016) .Deep neural networks and their implementation, Master thesis, Charles University in Prague Faculty of Mathematics and Physics, 2016.

Moor,M.,(2018) . Deep Learning for Computer Vision ,BOOK, Packt Publishing Ltd.

Zhang,A.,(2020) . Zachary C. Lipton, Mu Li, and Alexander J. Smola, "Dive into Deep Learning", BOOK .

Ketkar, N. (2017). Stochastic gradient descent. In Deep learning with Python (pp. 113-132). Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4842-2766-4_8

Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P. (2018). Conceptual understanding of convolutional neural network-a deep learning approach. Procedia computer science, 132, 679-688., http://doi: 10.1016/j.procs.2018.05.069.

Al-Waisy, A. S., Qahwaji, R., Ipson, S., Al-Fahdawi, S., & Nagem, T. A. (2018). A multi-biometric iris recognition system based on a deep learning approach. Pattern Analysis and Applications, 21(3), 783-802.,http:// doi.org/10.1007/s10044-017-0656-1.

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018, October). Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE, http:// doi: 10.1109/ICOASE.2018.8548836

Heidari, J. (2019). Classifying Material Defects with Convolutional Neural Networks and Image Processing., http://urn.kb.se/ resolve?urn=urn: nbn:se: uu:diva-387797.

Mostavi, M., Chiu, Y. C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC medical genomics, 13(5), 1-13, https://github.com/chenlabgccri/CancerTypePrediction.

Huang, Y., Shen, J., Wang, Z., Wen, M., & Zhang, C. (2018, May). A high-efficiency FPGA-based accelerator for convolutional neural networks using Winograd algorithm. In Journal of Physics: Conference Series (Vol. 1026, No. 1, p. 012019). IOP Publishing. ,http:// doi :10.1088/1742-6596/1026/1/012019 .

LeNail, A. (2019). NN-SVG: Publication-Ready Neural Network Architecture Schematics. J. Open Source Softw., 4(33), 747.http:// doi:10.1109/TVCG.2011.185.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9(4), 611-629. https://doi.org/10.1007/s13244-018-0639-9

Bashkatova, D. (2018). "Classification of Images Using Convolutional Neural Networks". In Science. Research. Practice (pp. 37-39),

Xia, M., Li, T., Xu, L., Liu, L., & De Silva, C. W. (2017). Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks. IEEE/ASME transactions on mechatronics, 23(1), 101-110. , http :/ /doi. org/10 .1109/ TMECH.2017.2728371.

Lee, H., & Song, J. (2019). Introduction to convolutional neural network using Keras; an understanding from a statistician. Communications for Statistical Applications and Methods, 26(6), 591-610.. https://doi.org/10.29220/CSAM.2019.26.6.591.

[71]W. Gong, H.Chen, Z.Zhang ,M. Zhang,R.Wang, C.Cuan,Q.Wang. "A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion." *Sensors*, vol. 19, No.7 , P. [1693],2019, doi.org/10.339 0/s19071693.

Kaliyugarsan,S.,(2019). Deep transfer learning in medical imaging, Master thesis. The University of Bergen.

Thoma,M.,(2017) .Analysis and optimization of convolutional neural network architectures. Master Thesis, University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association .

Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Analysis and Applications, 24(3), 1207-1220. https://doi.org/10.1007/s10044-021-00984-y

Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. CAAI Transactions on Intelligence Technology, 5(1), 22-33, https://doi.org/10.1049/trit.2019.0028.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.http://, doi: 10.1186/s40537-019-0197-0.

Kircher, M., Chludzinski, E., Krepel, J., Saremi, B., Beineke, A., & Jung, K. (2022). Augmentation of Transcriptomic Data for Improved Classification of Patients with Respiratory Diseases of Viral Origin. International journal of molecular sciences, 23(5), 2481. https://doi.org/10.3390/ijms23052481.

المستخلص

يتأثر ملايين الأشـخاص بتفشـي مرض فيروس كورونا الحالي (COVID-19) ، والذي أودى بحياة مأسـاوية في جميع أنحاء العالم. وتجدر الإشـارة أيضًـا إلى أن عاصـفة السـيتوكينات قد نمت لتصـبح عاملاً مهمًا في ارتفاع معدل الوفيات. ومع ذلك ، نظرًا لفهمنا المحدود لآلية دفاع المضـيف وظهور عاصـفة خلوية لمكافحة هذه العدوى الفيروسـية ، فقد باءت الجهود المبذولة لابتكار الأدوية واللقاحات والعلاجات بالفشـل. لذلك ، فإن الفهم الأكبر للعمليات التي تسـبب خلل التنظيم المناعي وظهور عواصـف خلوية قد يوفر لنا نظرة ثاقبة في العلاج السـريري للحالات الشـديدة. قد تتأثر الإصـابة بمرض كوفيد -19 بالعوامل الوراثية. وفقًا للدراسـات الحديثة ، يتفوق العلاج الجيني على الأدوية التقليدية الأخرى لمرضـى COVID-19. نتيجة لذلك ، يمكن أن يسـاعد تحديد الجينات التي تؤثر على هذا المرض في زيادة الاسـتجابة العلاجية الجيدة. علاوة على ذلك ، فإن المصـفوفة الدقيقة هي أداة واعدة تسـمح للعلماء بقياس مسـتوى التعبير لمئات الآلاف من الجينات في نفس الوقت. الهدف الرئيسـي من هذه الأطروحة هو تطوير نموذج يمكنه التنبؤ بدقة بتطور مرض COVID-19 بأقل قدر من الخطأ عن طريق اختيار أهم الجينات (الجينات المعلوماتية).

يتكون النظام المقترح من مرحلتين رئيسـيتين: اختيار الميزة ومرحلة التنبؤ. يتم اختيار الميزة باسـتخدام طريقة تسـلسـل الميزات (FSS) لتحديد مجموعة فرعية من الجينات المهمة وتحسـين دقة التنبؤ للنموذج المقترح. تتضـمن طريقة FSS ثلاث طرق: طريقة معامل ارتباط بيرسـون ، طريقة المعلومات المتبادلة وطريقة تحليل المكونات الرئيسـية. بشـكل عام ، يقوم النظام المقترح بتنفيذ طريقة FSS ، والتي تتعرف على الميزات الأكثر فائدة في كل خطوة ، ثم يتم إدخالها في النموذج لضـمان أهمية هذه الميزات والدقة التي يمكن الحصول عليها. علاوة على ذلك ، سعت هذه الأطروحة إلى تقديم نموذج تنبؤ يعتمد على الشـبكة العصـبية الاصـطناعية (ANN) والشـبكة العصـبية التلافيفية (CNN) لتحديد الجينات المتعلقة بالمرضـى المصـابين بمرض COVID-19 (DORF6) ، والجينات المتعلقة بالأشـخاص الذين يعانون من أعراض خفيفة (WT).) أو الجينات المتعلقة بأشـخاص غير مصـابين (MOCK) ، تم اسـتخدام مجموعة البيانات المتاحة لتحقيق أهداف الأطروحة الحالية: مجموعة بيانات COVID-19. تم إجراء التقييم بناءً على مقياسين تنبؤيين (الدقة والخسارة).

أظهرت النتائج التي تم الحصـول عليها أن أداء النظام المقترح فعال حيث تمت مقارنة دقة التنبؤ بالجينات الأصلية والمختارة قبل وبعد تطبيق طريقة FSS. باستخدام جميع الجينات ، كانت دقة التنبؤ التي تم الحصول عليها (30٪) مع ANN و (30٪) مع CNN ، بينما كانت دقة التنبؤ بعد تطبيق طريقة FSS (0.97) مع ANN و (82٪) مع CNN مع 198 جين فقط.

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة بابل-كلية العلوم البنات

قسم علوم الحاسبات

تصنيف مرض كوفيد -19 باستخدام التعبير الجيني وتقنية التعلم العميق

رسالة مقدمة الى

مجلس كلية العلوم للبنات-جامعة بابل

وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسبات

من قبل

ايمان حميد هادي

بإشراف

البرفسور. الدكتور. حسين عطية

مساعد. بروفيسور د. سرى زكي الراشد

1443هـ **2022م**