REPUBLIC OF TÜRKİYE
KIRŞEHİR AHİ EVRAN UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ADVANCED TECHNOLOGIES

MACHINE LEARNING-BASED COVID-19 DIAGNOSIS AND PREDICTION SYSTEM: PERFORMANCE ANALYSIS OF VARIOUS LEARNING ALGORITHMS AND CLASSIFICATION OF RELATED DISEASES

AHMED JADDOA ENAD Al-MAMOORI

MSc THESIS

KIRŞEHİR
2024

REPUBLIC OF TÜRKİYE
KIRŞEHİR AHİ EVRAN UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ADVANCED TECHNOLOGIES

MACHINE LEARNING-BASED COVID-19 DIAGNOSIS AND PREDICTION SYSTEM: PERFORMANCE ANALYSIS OF VARIOUS LEARNING ALGORITHMS AND CLASSIFICATION OF RELATED DISEASES

AHMED JADDOA ENAD Al-MAMOORI

MSc THESIS

SUPERVISOR
ASST. PROF. DR. MUSTAFA AKSU

KIRŞEHİR
2024

KIRŞEHİR AHİ EVRAN UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

ETHICS DECLARATION

In this thesis study, which I have read and understood the Kırşehir Ahi Evran University Scientific Research and Publication Ethics Directive and which I have prepared in accordance with the Kırşehir Ahi Evran University Institute of Natural and Applied Science Thesis Writing Rules;

- I have obtained the data, information and documents I have presented in the thesis within the framework of academic and ethical rules,
- I present all information, documents, evaluations and results in accordance with scientific and ethical rules,
- I have cited all the works I have benefited from in the thesis by making appropriate references,
- I have not made any changes in the data used and the results,
- This study, which I have presented as a thesis, is original,

Otherwise, I declare that I accept all legal actions to be taken against me in this regard and all loss of rights that may arise against me. 25/01/2024

Student
AHMED JADDOA ENAD
Al-MAMOORI

LIST OF CONTENTS

Page No

ACKNOWLEDGEMENTS

For his continual encouragement and counsel during my master's program, Asst. Prof. Dr. Mustafa AKSU, my supervisor, has my sincere gratitude. His knowledge and tolerance have been a great help to me and were essential to the achievement of this thesis.

I want to convey my most profound appreciation to the Advanced Technologies Department in Institute of Natural and Applied Science at Kırşehir Ahi Evran University for giving me the chance to pursue my master's degree. Their support and assistance throughout this research journey have been invaluable.

I also extend my gratitude to Assoc. Prof. Dr. Mustafa YAĞCI and Assoc. Prof. Dr. Murat CANAYAZ for all the information and assistance they gave me throughout my research, not to mention their technical aid.

I appreciate my friends and family's affection and assistance throughout this journey. I would not have finished my adventure if it were not for their support and inspiration.

Last, I thank everyone who participated in my study and was willing to share their knowledge. With their help, this work was completed.

January 2024 AHMED JADDOA ENAD AL- MAMOORI

GENİŞLETİLMİŞ ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENİMİ TABANLI COVID-19 TEŞHİS VE TAHMİN SİSTEMİ: ÇEŞİTLİ ÖĞRENME ALGORİTMALARININ PERFORMANS ANALİZİ VE İLGİLİ HASTALIKLARIN SINIFLANDIRILMASI

AHMED JADDOA ENAD AL-MAMOORI

KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İLERİ TEKNOLOJİLER ANABİLİM DALI

Danışman : **Dr. Öğr. Üyesi Mustafa AKSU**
Yıl: 2024 Sayfa: 85
Jüri: **Dr. Öğr. Üyesi Mustafa AKSU**
Doç. Dr. Mustafa YAĞCI
Doç. Dr. Murat CANAYAZ

Sanal Zeka ve makine öğreniminin alanları, özellikle geçen birkaç yılda çeşitli endüstrilerden önemli ilgi ve yatırım çekmiştir. COVID-19'un son zamanlarda keşfi, sağlık sektöründe yapay zeka yöntemlerinin yaygın kullanımına ve sıkı testlere rağmen, hastalığı teşhis etme, tahmin etme ve önleme amacıyla bu yöntemlerin uygulanmasını gerektirmiştir. Önerilen sistem, COVID-19 enfeksiyonlarını tahmin etmek için Random Forest (RF), Naive Bayes (NB), Destek Vektör Makinesi (SVM), Karar Ağacı (DT), Çok Katmanlı Algılayıcı (MLP) ve K-En Yakın Komşular (KNN) dahil altı makine öğrenme algoritması temelinde bulunmaktadır. Önerilen modelin kullanılan veri setine uygulandığında iyi performans sergilediği tespit edilmiştir. Çalışma, iki aşamayı içermiştir: ilk olarak, modeli eğitmek için veri setini yüklemek ve ikinci olarak, modeli bu vakalarda doğrudan test etmek, otomatik COVID-19 tahminini sağlayarak bir hastanın şüpheli olup olmadığını belirlemek. Çalışmanın amacı, öğrenme tekniklerine dayalı COVID-19 ve ilgili hastalıkların erken teşhisini bulmaktır. Ana araştırma katkıları, kullanılan sınıflandırma sisteminin tespit edilebilir kullanılabilirliğini belirlemek ve EG.5, Eris ve Pirola BA.2.86 gibi diğer ilgili hastalıkları tahmin etmek için kullanılan öğrenme tekniklerine dayalı olarak bir hastanın şüpheli olup olmadığını sınıflandırmaktır. Uygulanan sistem, Java Eclipse programlama ortamını kullanarak Java'da makine öğrenme algoritmalarını uygular. Temel aşamalar, COVID-19 veri seti için ilk aşamada veri madenciliği ön işleme, ham veriyi etkili ve verimli bir formata dönüştürme işlemidir. İkinci aşamada ise önceden işlenmiş eğitim veri seti, normalleştirme, özellik seçimi, eksik veri işleme ve veri dönüşüm yöntemleri kullanılarak özellik değerleri üretmek için kullanılır. Doğruluk sonuçları, kullanılan ilk (Covid Data 1) veri setinin 70 Eğitim ve 30 testleme ile MLP'nin %99.5300'lik yüksek doğruluğa sahip olduğunu ve modelin oluşturulma süresinin 1639469 ms olduğunu gösterdi, SVM'nin %99.4991'lik ikinci doğruluğa sahip olduğunu ve modelin oluşturulma süresinin 1639469 ms olduğunu gösterdi. DT'nin %99.4364'lük üçüncü doğruluğa sahip olduğunu ve modelin oluşturulma süresinin 130785 ms olduğunu gösterdi. Ayrıca, RF'nin

hemen hemen düşük Hata Oranı (MAE) değeri olarak 0.0039'a sahip olması, diğerleriyle karşılaştırıldığında daha iyidir. DT'nin Kök Ortalama Kare Hatası (RMSE) istatistiği sonuçları, diğer algoritmalarla karşılaştırıldığında daha iyi olan 0.0462'dir. MLP algoritmasının hata oranı sonuçları, karşılaştırılan algoritmalar için daha iyidir olarak 0.00469'dur. Ayrıca, 0.99920 olan KNN hassasiyeti, daha fazla ilgili sonuçları döndürme açısından yüksek kalite ölçüsü olarak görülebilir. KNN'nin AUC değeri = 0.998116 daha yüksektir, bu nedenle konumlandırma koordinat sınıfları arasındaki ayrımı belirlemede daha iyidir. MLP'nin DR'si, doğru bir şekilde tespit edilen tüm örneklemin en iyisidir. RF'nin FAR'si en iyisi çünkü kullanılan parametrelerin daha az yanlış alarmı göstermesini ifade eder.

Anahtar kelimeler: Yapay zeka, COVID-19, veri madenciliği, makine öğrenmesi (ML), tahmin.

ABSTRACT

MASTER'S THESIS

MACHINE LEARNING-BASED COVID-19 DIAGNOSIS AND PREDICTION SYSTEM: PERFORMANCE ANALYSIS OF VARIOUS LEARNING ALGORITHMS AND CLASSIFICATION OF RELATED DISEASES

AHMED JADDOA ENAD AL-MAMOORI

KIRŞEHİR AHİ EVRAN UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ADVANCED TECHNOLOGIES

Supervisor : **Asst. Prof. Dr. Mustafa AKSU**
Year: 2024 Pages: 85
Jury : **Asst. Prof. Dr. Mustafa AKSU**
Assoc. Prof. Dr. Mustafa YAĞCI
Assoc. Prof. Dr. Murat CANAYAZ

The fields of Artificial Intelligence and machine learning have attracted significant interest and investment, especially from various industries in the past few years. Despite the widespread use of AI methods in the healthcare sector and rigorous testing, the recent discovery of COVID-19 has necessitated the application of these methods for diagnosing, predicting, and preventing the disease. The used system is based on six machine learning algorithms, including Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN), to predict COVID-19 infections. The used model has been found to perform well when applied to the utilized dataset. The study involved two steps: firstly, loading the dataset to train the model, and secondly, directly testing the model on these cases, enabling automatic COVID-19 prediction to detect whether a patient is suspicious or not. The study's aim is to find an early diagnosis of COVID-19 and related diseases based on learning techniques. Key research contributions include determining the detectable usability of the classification system used to predict other relevant diseases, such as EG.5, Eris, and Pirola BA.2.86, and classifying whether a patient is suspicious based on learning techniques in the used dataset. The implemented system, using the Java Eclipse programming environment, applies machine learning algorithms in Java. The fundamental stages include data mining preprocessing in the first stage for the entire COVID-19 dataset, transforming raw data into an effective and efficient format. In the second stage, the preprocessed training dataset is used to generate feature values, employing normalization, feature selection, handling missing data, and data transformation methods.The accuracy results showed the first used (Covid Data 1) dataset with 70 Training and 30 testings showed the MLP a high accuracy as 99.5300 %, and the time to build model is 1639469 ms, while SVM is the second accuracy as 99.4991 % and the time take to build model is 1639469 ms. While DT is the third accuracy as 99.4364 %, and the time taken to build the model is 130785 ms. Additionally, the RF as 0.0039 almost lower the Mean Absolute Error (MAE) value, so it is better compared with others. DT results of the Root Mean Squared Error (RMSE) statistic the

lower as the better 0.0462 compared with other algorithms. MLP algorithm results of error rate is 0.00469 as the better for the compared algorithms. Besides, the KNN precision as 0.99920 can be seen as a measure of high quality to return more relevant results than irrelevant ones. The AUC of KNN = 0.998116 is higher, so it is better to distinguish between positioning coordination classes. DR of MLP is the best of the whole sample, which was detected correctly. FAR of RF is best because it indicates fewer false alarms of the used parameters.

Keywords: Artificial intelligence, COVID-19, data mining, machine learning (ML), prediction.

LIST OF TABLES

Page No

LIST OF FIGURES

Page No

LIST OF ICONS AND ABBREVIATIONS

Abbreviation		Described
ACE2	:	Angiotensin-Converting Enzyme 2
AI	:	Artificial Intelligence
ANN	:	Artificial Neural Network
CNN	:	Convolutional Neural Network
COVID-19	:	Corona Virus Disease of 2019
DNN	:	Deep Neural Network
DR	:	Detection Rate
DT	:	Decision Tree
FAR	:	False Alert Rate
FN	:	False Negative
FP	:	False Positives
GDCNN	:	Genetic Deep Learning Convolutional Neural Network
GPU	:	Graphics Processing Unit
KDD	:	Knowledge Discovery of Databases
KNN	:	K-Nearest Neighbor
MAE	:	Mean Absolute Error
ML	:	Machine Learning
MLP	:	Multilayer Perceptron
NB	:	Naive Bayes
RF	:	Random Forest
RMSE	:	Root Mean Squared Error
ROC	:	Receiver Operating Characteristic
SVM	:	Support Vector Machine
TN	:	True Negative
TP	:	True Positive
VGG16	:	Visual Geometry Group
WHO	:	World Health Organization

1. INTRODUCTION

Since March 2019, a pandemic caused by the COVID-19 virus has been ongoing for about four years. Despite the vaccination programs prevalent in various nations, there is still an ongoing growth in the number of persons who are sick. Due to the many distinct forms and mutations, the COVID-19 virus has become very contagious, fatal, and sometimes undetected, leading to an increase in the number of persons infected with the virus. The Omicron form has been readily apparent and has spread to a variety of nations, resulting in everyone being afflicted (Morand et al., 2020).

A new COVID-19 variant, known as EG.5.1 or Eris, surfaced in the UK during the summer. The World Health Organization (WHO) officially categorized it as a variant on August 9th. Eris is derived from the Omicron variant that initially emerged in November 2021 and has spawned numerous sub-variants. Currently, Eris is the second most common variant in the UK and the most widespread in the United States, as reported by the Centers for Disease Control and Prevention (CDC) (Wiwoho et al., 2023).

Another COVID-19 variant, BA.2.86 or Pirola, is also spreading. Like Eris, Pirola is a sub-variant of Omicron. It was first detected in the UK on August 18th and has been identified in Denmark and the US. Experts have observed that Pirola possesses many genetic differences compared to earlier versions of COVID-19. Due to its higher number of mutations, Pirola might be more likely to lead to 'breakthrough infections,' where fully vaccinated individuals can still contract the virus. However, vaccination remains effective in preventing severe illness if someone does contract COVID-19. This underscores the ongoing emergence of COVID-19 variants, emphasizing the importance of early diagnosis and appropriate treatment for this disease (Wiwoho et al., 2023).

Many workers, particularly healthcare workers, are experiencing burnout due to the notable rise. In order to lessen people's exposure to the COVID-19 virus and put a stop to its further spread, it is vital to keep track on and monitor individuals. There is a minor link between the two variables despite the fact that they substantiated that aware individuals would practice preventative actions. Therefore, it is necessary to investigate the need to mitigate the effects of the COVID-19 virus by monitoring and tracking (Meo et al., 2023).

Several nations make COVID-19 monitoring and tracking applications and technology accessible to their citizens. According to the findings of (Sepehrinezhad et al., 2020), evidence indicates a high functionality, information quality, and aesthetics level

in Europe. It provides evidence of favorable aesthetics. Conversely, the degree of engagement orientation exhibited a relatively weak level of quality, demonstrating that perceived ease of use is positively associated with perceived value and a mindset in the context of the United Kingdom. In addition, they demonstrated in Germany that the application had to have an agile setup and be capable of providing rapid updates in response to changes (Sepehrinezhad et al., 2020). However, to advertise and make the program useable among individuals from other nations, it was necessary to consider many aspects, although other contact tracing apps are accessible globally. Even though several types of literature about tracing applications are available, There is a dearth of information regarding the "Thai Chana" tracking application originating from Thailand (Toquero et al., 2020).

Thailand's primary contact tracing application is Thai Chana, a web-based platform designed to facilitate contact tracing among Thai individuals. It operates on a self-reporting mechanism from its users. Thailand examined various options, including but not limited to surveillance, laboratory testing, case-control and management, communicating risks, and preparedness of health care providers, facilities, and medical supplies. Thailand extended substantial aid in the form of Thai Chana to support the advancement of its diverse undertakings. For a considerable duration, Thai individuals have been required to utilize it upon entering unfamiliar territories. According to reports, the Thai government has implemented rigorous measures for registration with the Thai Chana mobile application, mandating compliance for all individuals in Thailand, including foreign tourists. Thai Chana can gather personal information, including but not limited to the user's name, age, address, and phone number (Toquero et al., 2020).

Furthermore, it has the potential to indicate and communicate data regarding the existence of ill individuals who have frequented the area. Individuals are provided with information regarding the safety of a particular area, including the necessity of self-isolation, and they undergo testing as a preventative measure against the transmission of the virus. However, Bangkok, Thailand's primary city, continues to be regarded as one of the most infectious urban centers worldwide, being ranked second only to China. Hence, it is imperative to examine Thai chana to enhance its utilization, thereby mitigating the prevalence of communicable ailments in the country (Chuenyindee et al., 2022).

Moreover, AI and machine learning fields have attracted significant interest and investment from a diverse range of industries, especially during the last several years. Although AI methods have been used extensively and put through extensive testing in the

healthcare industry, the recently discovered COVID-19 necessitates the use of these methods to diagnose, forecast, and prevent the emergence of the disease. It has been hypothesized that using AI methods would bring about a paradigm change in healthcare and necessitate the application of these techniques to the ongoing COVID-19 pandemic. Improving the precision of COVID-19 diagnosis is imperative to expeditiously detect affirmative cases, thereby mitigating additional transmissions and guaranteeing prompt medical attention for patients (Bai et al., 2019).

Hence, the availability of large amounts of data in today's world has led to the widespread use of the algorithm for machine learning. It is a tool that may be used for forecasting, categorizing, and identifying patterns within various datasets. Studies have utilized diverse machine learning algorithms, including NN and DT, featuring RF classifiers. Furthermore, a random forest classifier is used to classify the factors that impact the decision to remove a child from their home based on parental factors. The results indicate that the random forest classifier has the potential to examine the factors that impact human behavior. Additionally, using an RF classifier was considered to forecast the risk assessment of flood calamities in China (Ong et al., 2022).

Furthermore, the results of the various investigations have shown that the random forest classifier provides superior classification accuracy compared to the traditional decision tree (Swapnarekha et al., 2020). Conversely, NNs have been employed as a technique for uncovering pattern recognition. An algorithm replicating the process by which neurons transmit information to the brain has been used in developing neural networks. One argument favoring its usefulness is that the findings generated from enormous datasets are considered state-of-the-art (Nagi et al., 2022).

An Artificial Neural Network was used, primarily emphasizing risk assessment in Iran. In addition, they considered neural networks while trying to forecast the number of individuals who will be injured or killed in Indonesia. However, basic neural networks like artificial neural networks have limited capacities to forecast with greater accuracy because they only evaluate a limited number of parameters. Hence, a deep learning neural network would be useful in this context since it takes into account additional hidden layers for the subsequent processing and computation of output. The optimization procedure, encompassing identifying the suitable activation function, optimizer, and node count, is a drawback of utilizing artificial neural networks, primarily because these networks are commonly perceived as black boxes (Jamshidi et al., 2021).

Research on machine learning can only proceed with first preprocessing the data.

As a result, it used semiautomated methods to verify the data for any experimental contaminants that could have been present. There were a few individuals for whom clinical and proteomic data needed to be included. In the context of clinical data, the practice of imputing missing values involved the utilization of the numerical placeholder "-1." The classification algorithms for days 0-7 based on "Clinical Information" were trained using clinical data from 306 individuals, comprising 42 deceased and 264 survivors (Whole dataset I) (Guo et al., 2022).

In addition, many COVID-19 patients quickly deteriorate their condition after experiencing relatively minor symptoms, highlighting the need for more sophisticated risk stratification models. The utilization of predictive models facilitates the detection of patients who exhibit a heightened susceptibility to mortality and the delivery of aid to promptly mitigate the incidence of fatalities. Hence, accurate prognosis forecasting and appropriate triage of critically ill patients are imperative to alleviate the burden on the healthcare system and provide optimal patient care. Moreover, due to a significant level of ambiguity surrounding its definitive influence, medical professionals and policymakers have frequently resorted to the prognostications furnished by diverse computational and statistical models (Mayr et al., 2020).

The proteomics information-based classification model on Day 0 was constructed solely using data related to proteomics. The study found that protein expression values were absent for a single COVID-19-positive patient who passed away within 28 days of hospitalization, while 15 patients among the survivors had missing protein expression values for a small subset of the 1,428 proteins examined. Due to this rationale, the records above were omitted from the dataset, as depicted in Figure 1.1 (Jewell et al., 2020).

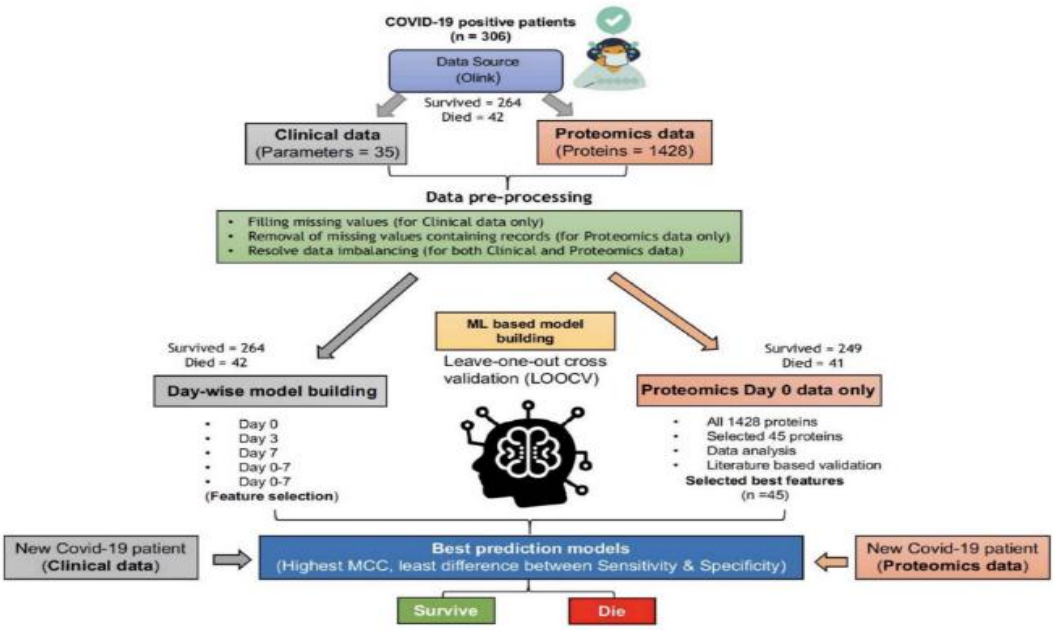


Figure 1.1. Clinical and proteomic data-driven machine learning pipeline for identifying survival-related characteristics (Jewell et al., 2020).

It is clear from the retrieved data that there was an uneven distribution of survivors and fatalities in both the clinical and proteomics data. The data collected from the survivor, including clinical and proteomics data, were divided into five groups that were almost equivalent in size. In addition, they educated and verified the models by using each of the five departments and the dataset of deceased patients (Khan et al., 2021).

Personalized protection tactics stand to gain a great deal from accurate population classifications based on categorized COVID-19 susceptibilities. Recent research has challenged the notion that advanced age is an important risk factor for COVID-19. This research suggests that many young adults have developed severe symptoms related to the disease. This discovery suggests a pressing necessity for a thorough assessment of risks grounded in individualized genetic and physiological traits (Frater et al., 2020).

The ACE2 receptor in humans serves as an entry point for the spike glycoprotein of SARS-CoV-2. The enzyme in question is found to be expressed within the epithelial cells of various organs, including the lungs, small intestines, heart, and kidneys. The authors postulated that the augmented expression of ACE2, which could result from administering ACE2-stimulating drugs to manage hypertension and diabetes, could potentially lead to unfavorable clinical consequences in the context of COVID-19 infection. This theory warrants further investigation through rigorous experimental techniques and extensive clinical studies (Russell et al., 2020).

Moreover, it is viable to employ machine learning techniques to investigate the

biochemistry, including ACE2 expression level, and clinical data, such as age, respiratory pattern, viral load, and survival, of COVID-19 patients with pre-existing medical conditions. The previously mentioned methodology enables researchers to discern dependable risk factors, such as ACE2, to predict risk. Furthermore, it enables the implementation of risk stratification and forecasting (MacGowan et al., 2020).

Studies have demonstrated that ACE2 genetic polymorphism, which encompasses diverse variations in the human genome, could influence virus-binding activity, which implies the possibility of a genetic predisposition to acquiring COVID-19. Hence, it is possible to conduct machine learning analysis on genetic variations in asymptomatic, mild, or severe COVID-19 patients to classify and predict individuals based on their vulnerability or immunity to potential COVID-19 infection. The decision-making process of the machine learning model can incorporate prioritized genetic variants, such as ACE2 variation, as significant features for functional and mechanistic research, as illustrated in Figure 1.2. (MacGowan et al., 2020).

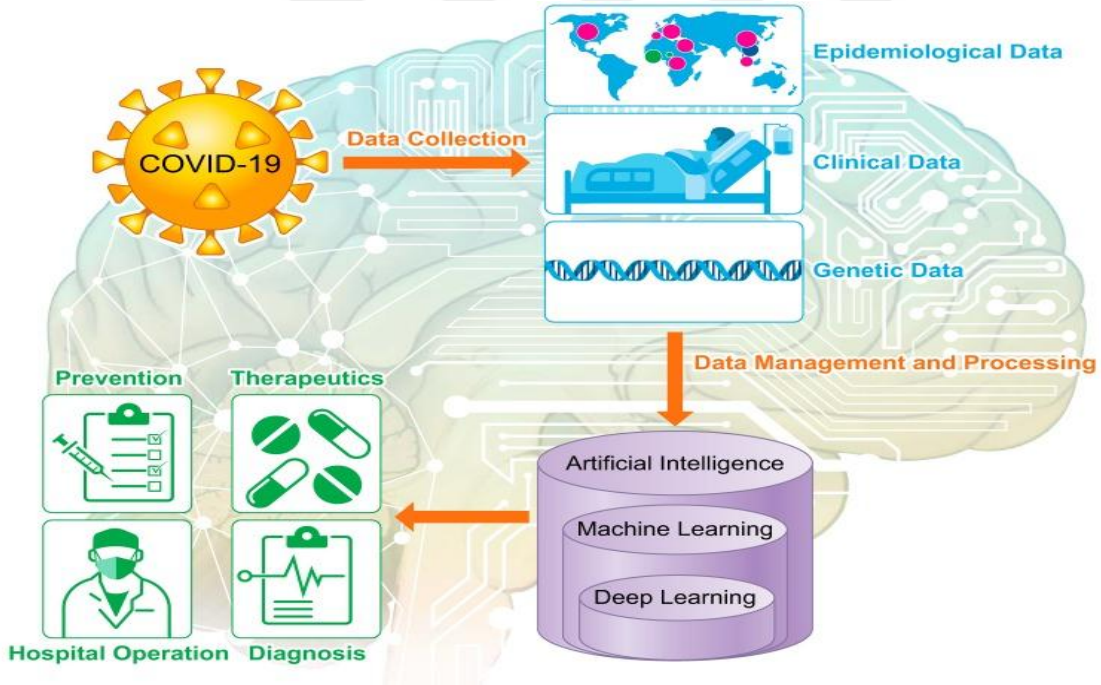


Figure 1.2. In the battle against COVID-19, artificial intelligence and machine learning (MacGowan et al., 2020).

Image classification strategies that use the deep learning algorithm are among the most cutting-edge practices in computer vision research. When a deep learning model is being used, the algorithm will take care of developing and extracting object characteristics independently without the need for human intervention. The engineering process of

extracting the object's characteristics requires a significant investment of both time and effort. An operator who is an expert in the relevant topic is required in order to design and assess the feature extraction technique. When it comes to managing duties, including classification, the deep learning model is the method. It has been determined that the deep learning algorithm is the most important method for categorizing images because of its ability to handle large groups of photographs quickly (Ma et al., 2022).

The deep learning technique is considered the most significant approach for addressing classification challenges due to its adaptability and applicability to various data sets. The time and money required for training deep learning algorithms is the primary challenge presented by these systems. Training a deep learning model might take several weeks and need an expensive graphics processing unit (GPU). Conducting empirical investigations with the primary focus being on the effectiveness of the deep learning algorithm is the most essential component of the evaluation process for the deep learning technique. This study reveals that the deep learning strategy is helpful when it comes to finishing an assignment that requires the classification of pictures (Akhtar et al., 2021).

Deep learning is a kind of artificial intelligence that can mimic human behavior, including expressions and mental processes. 'Training,' a machine learning system, often involves feeding hundreds or thousands of data points as input to complete the process more quickly. The moment the first batch of information is collected, it is put into "training" (Crespo et al., 2022).

When trying to categorize images, computer vision runs into a fundamental obstacle. A picture may be broken down into its parts, known as pixels, which individually have their value when seen through the lens of a computer. Every picture comprises individual elements called pixels, called pixel grids, arranged in a grid fashion. If we consider this, we may deduce that they consider pixels to be the essential building components of digital pictures. The use of color enables us to divide visual material into two separate categories (Crespo et al., 2022).

- In both shades of black and white. The image is a matrix in which the values for the various shades of grey range from 0 to 255. There are 256 possible values, translating to 8 for images with 8 bits of resolution.

- Colour version of the image. It consists of three separate matrices, each with values ranging from 0 to 255. The colors red, green, and blue make up the RGB color

model, and these matrices correspond to those colors in that order. The whole image comprises these three color channels being put together.

There are several approaches to classifying photographs, but each supervised learning image classifier requires three datasets to train on. These datasets comprise pictures and the labels that correspond to them.

- Training set: As its name suggests, this data set is used during the training phase of the learning process for the algorithm.

- Development set: sometimes called the dev set, the validation dataset on which the model will be tested while being trained to identify possible problems such as overfitting.

- This is the test set, which aims to ascertain the classifier's overall accuracy. It is necessary to separate the data into training, validation, and test sets before determining which hyperparameter is the most appropriate for the classifier (Marques et al., 2020). The term "data processing" refers to a series of extremely significant procedures used to change data from its source into a usable format free from mistakes that lead to inaccuracies in the system. These techniques are utilized in the transformation process known as "data transformation." In the context of search and data mining, data preprocessing is divided into two categories; the first is data preparation. Preprocessing of data is employed in a wide variety of contexts, including classic data extraction methods and search and data mining, which comprises converting the data, cleaning the data, and normalizing the data as well as the second category is data reduction, which is the integration of these data and the potential of decreasing it (Albahri et al., 2020).

Processing vast amounts of complicated data, such as the data created from health care in the area of medical and biological sciences, is one of the primary applications for machine learning methods, which are utilized extensively in various sectors. If it is required to discover ways to cope with them, start with the standard approaches. Because standard techniques are not equipped to cope with large data, processing this huge data requires methods and algorithms that are different from those traditionally used. These data come from various origins, such as using methods from artificial intelligence, statistics, cognitive sciences, and many other fields of study that fall under the umbrella of mathematics and engineering (Farid et al., 2020).

This study undertook a bibliometric analysis of the existing literature on utilizing machine learning algorithms in the COVID-19 context, utilizing the VOSviewer software. This analysis aimed to ascertain the lacuna in research that required attention.

The task above was executed by utilizing the Web of Science database in August 2020. As illustrated in Figure 1.3. the bibliometric study's findings suggest that implementing machine learning techniques for COVID-19 is limited to only two subject areas. Upon examining the papers associated with each cluster (theme), it becomes evident that a significant portion of the research has focused on predicting Covid-19 transmission through the utilization of meteorological data (Farid et al., 2020).

The observation above is evident within the initial cluster. Per the second cluster, subsequent inquiries have focused on chest CT scans and chest X-Ray images by utilizing deep learning algorithms. Despite the comprehensive characterization of the COVID-19 diagnosis's high sensitivity through the use of CT and X-Ray imaging, the application of such tests for patient screening may present challenges due to factors such as high radiation doses, elevated costs, and limited equipment availability (Chiroma et al., 2020).

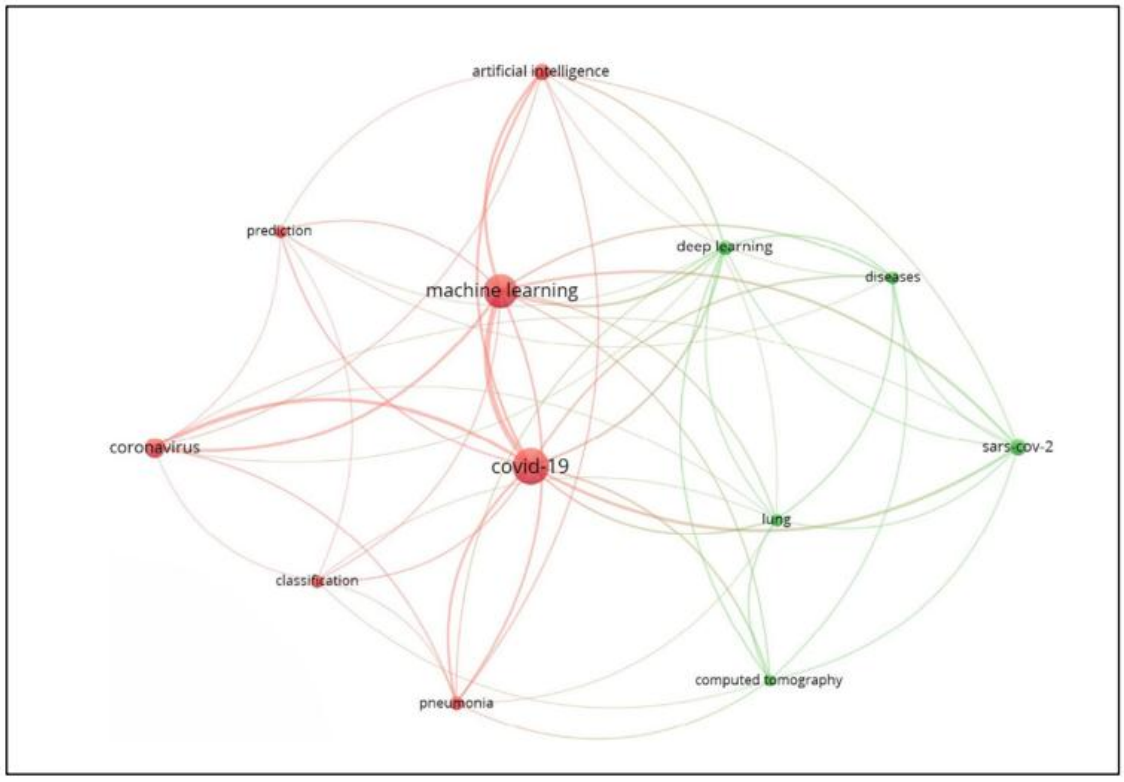


Figure 1.3. On the COVID-19 platform, applications of machine learning techniques(Chiroma et al., 2020).

1.1. Aim of Research

The current medical system has a problem with proper illness diagnosis, which results in a significant economic loss for society. The primary reason for this is that medical data is a composite of many different types of information. However, the proliferation of large amounts of medical data, in conjunction with the advancement of computational methods in the healthcare field, has made it a significant challenge to anticipate illnesses accurately. Moreover, many crucial medical imaging modalities need significant time to rebuild a picture from the raw data samples. The main goals of COVID-19 data classification are to :

1. Determine the qualities of the item or kind of land cover that these characteristics genuinely signal on the ground as different data types, and then display those characteristics in a dataset.
2. Finding the early detection of diseases depending on deep learning will help relieve the pressure on the healthcare systems. It is based on different attributes of deep learning in the used dataset, which helps to predict different diseases before they happen and decrease side effects.

1.2. Problem Statement

Presently medical field suffers from an accurate diagnosis of diseases, which creates a huge loss to society. The prime factor for this is the nature of medical data; it combines all types of data.

- Increasing medical big data, alongside the development of computational techniques in healthcare, has enabled a big problem faced in predicting diseases.
- Many important medical imaging modalities require much time to reconstruct an image from the raw data samples.

1.3. Research Contribution

The main research contributions of the used classification system are to :

- Predict factors affecting COVID-19 detection and other related disease such as EG.5, Eris, and Pirola BA.2.86 with different datasets entered into the trained model.
- Determine the perceived usability of COVID-19 to classify the patient's state as suspected or not.

1.4. Research Novelty

This novel research fulfills the gap of managing the development of an expert system for detecting and classifying the COVID-19 disease by:

- Designing an appropriate data preprocessing approach provides access to discover the causes of specific COVID-19-related diseases such as Omicron-Pirola and EG.5.1 or Eris and knows the proper management.
- Reducing data acquisition time and storage space.

2. LITERATURE REVIEW

2.1. Related Work

There are different research directions related to the used method sorted as follows:

According to (Ezat et al., 2020), a CNN model pre-trained on the Image-Net was suggested for image categorization using the PASCAL VOC 2007 dataset. The deep CNN model uses a modest amount of compute time and machine resources for classification. Hence, the transfer learning technique is employed to increase the model's performance. The final findings assess deep learning as a cutting-edge technique for a COVID-19 classification job.

In reference (Wang et al., 2021), a comparison and contrast of image classification methodologies utilizing deep learning and traditional machine learning were conducted. The study's results indicate that the SVM achieved an accuracy of 0.88, while the CNN achieved an accuracy of 0.98 when utilizing the most extensive sample dataset. Notably, in the case of utilizing the limited sample size of the COREL1000 dataset, the SVM method exhibits an accuracy of 0.86, while the CNN approach yields an accuracy of 0.83. The findings indicate that using a deep learning framework yields superior outcomes compared to conventional methods of pattern recognition, particularly when dealing with extensive datasets.

ResNet-18 and ResNet-50 have been trained on colon gland pictures (Sarwinda et al., 2021). This study used colorectal cancer models trained to discriminate between benign and malignant tumors. Using three distinct forms of test data, it evaluated prototypes (20, 25, and 40 % of whole datasets). Results from three different types of testing data show that ResNet-50 outperforms ResNet-18 in terms of accuracy, sensitivity, and specificity. Its greatest performance across all three test sets was attained with a classification accuracy of 80% or higher, a sensitivity of 87%, and a specificity of 83% on the 20% and 25% test sets.

An implementation of a GDCNN is provided by (Babukarthik et al., 2020). The prediction of COVID-19 is accurate to the tune of 98.84%, with a 93% degree of precision, a 100% sensitivity, and a 97% specificity. This study's high rate of correct illness identification in identifying COVID-19 in an imbalanced setting demonstrates the highest nominal rate in classification. The unique classification model outperforms state-of-the-art methods like ReseNet18, ReseNet50, Squeezenet, DenseNet-121, and the Visual Geometry Group (VGG16).

According to (Alakus et al., 2020), Using a convolutional neural network, they demonstrated a technique for selecting and extracting features from images for further classification. Convolution neural networks may provide superior accuracy over other classifiers. The efficiency and precision are tested on a regular CPU and a GPU. So, CNNs are a great choice for picture categorization. Biometric features might be added to this system in the future.

As suggested in a citation (Díaz-Pernas et al., 2021), the neural model can potentially analyze meningioma, glioma, and pituitary cancers in MRI scans from sagittal, coronal, and axial viewpoints. Obviously, this model does not necessitate preliminary processing of the input images to eliminate segments of the skull or vertebral column. The present study conducts a comparative analysis of the efficacy of the used methodology in processing a publicly available MRI imaging dataset comprising 3064 slices obtained from 233 patients, vis-à-vis the performance of conventional machine learning and deep learning techniques as reported in prior literature. Remarkably, they devised an approach that exhibited superior performance to rival algorithms on identical datasets, achieving a tumor classification accuracy of 0.973.

A sparse representation-based deep learning model is suggested by (Liu et al., 2020), which uses sparse representation extensively to obtain high multidimensional data linear decomposition ability and deep structural benefits of multilayer nonlinear mapping. The experimental findings demonstrate that the suggested technique is more accurate than the state-of-the-art methods and can be effectively applied to other types of picture databases. In order to further enhance the accuracy of image classification, it is superior to other deep learning approaches in resolving the issues of complicated function approximation and poor classifier effect.

According to (Ong et al., 2021), they spoke about how they got the images, what they did with the data before they were processed, what features were extracted from the images, and how those images were categorized and recognized. Studying the input and output points, as well as the number of neurons in a convolutional neural network, led to the conclusion that the deep learning approach can reduce training time and improve identification results while also reducing the need for images in an image library and requiring only moderate amounts of hardware.

The current framework has been developed to systematically collect, organize, and assess information related to image processing by utilizing sandbox simulation operations, modeling techniques, and a wide range of sophisticated algorithms. Reel

neural networks were introduced in reference (Moulaei et al., 2022) to enhance computational capabilities by incorporating deep learning, following a similar approach to the refractostatic neural network model. The primary experimental approach utilized in this study is the target identification algorithm. According to the findings of multiple experiments, utilizing a coproduct neural network model for deep learning could potentially enhance the efficacy of image processing.

Deep learning architectures are employed in the study referenced in (Ali et al., 2022) to assess whether or not an MRI of the brain shows abnormal findings. They also provide a deep learning CNN-based solution for efficient classification in addition to this DNN. Traditional approaches such as SVM are used with various deep learning architectures such as LeNet, AlexNet, and ResNet to analyze and contrast the outcomes. In contrast to SVM (82%) and AlexNet (64%), the LeNet-inspired model achieves an overall accuracy of 88%. In comparison, the CNN-DNN model achieves an accuracy of 80%, with the top accuracy being 100, 92, 94, and 81%, respectively.

The Authors (Lunagaria et al., 2022) were able to effectively recognize image characteristics by making use of a huge dataset and deep convolutional neural networks. It has been shown that the majority of pre-trained networks found in the research literature and the majority of fundamental models put excessive importance on qualities that are not essential when it comes to decision-making. Compiling a large number of chest X-ray images from various sources resulted in the creation of one of the most extensive databases that are open to the general public. To summarise, the CheXNet model, which is commonly utilized, has been adapted to the COVID-CXNet framework through the implementation of transfer learning techniques. This robust approach, which uses significant traits and precise localization, can identify the newly discovered coronavirus pneumonia. With the assistance of COVID-CXNet, it is possible to create a COVID-19 detection system that is fully automated and dependable.

The authors (Alakus et al., 2020) conducted clinical predictive models utilizing deep learning techniques and laboratory data to determine patients' likelihood of COVID-19 illness. The precision, F1-score, recall, area under the curve (AUC), and accuracy scores were produced to evaluate the models' predictive performance. The models underwent verification through 10-fold cross-validation and train-test split methodologies. The models underwent evaluation based on 18 laboratory outcomes obtained from 600 patients. Based on the results of the conducted experiments, it was observed that the employed predictive models exhibited an accuracy rate of 86.66 %, an

F1-score of 91.89%, a precision rate of 86.75 %, a recall rate of 99.42 percent, and an area under the curve (AUC) of 62.5 %. Utilizing laboratory data to train prediction models for COVID-19 infection has been observed as a potential benefit for medical practitioners in effectively allocating available resources.

The research conducted by (Arpaci et al., 2021) resulted in the development of six distinct prediction models for COVID-19 diagnosis, utilizing six distinct classifiers, including BayesNet, Logistic, IBk, CR, PART, and J48. The classifiers were developed using a collection of 14 clinical features. This study conducted a retrospective analysis of the medical records of 114 patients admitted to a hospital in Taizhou, which is situated in the Zhejiang Province of China. According to the findings, the CR meta-classifier demonstrates a notable degree of precision, particularly 84.21%, in its ability to forecast affirmative and negative instances of COVID-19. This statement implies that the CR meta-classifier exhibits the highest level of efficacy among classifiers utilized for this particular objective. The findings have the potential to serve as a valuable resource for expeditiously detecting COVID-19, particularly in situations where RT-PCR kits are insufficient in verifying the existence of the infection. Furthermore, these results could benefit countries, particularly those with limited resources, that encounter difficulties obtaining RT-PCR assays and specialized facilities.

To predict COVID-19 using a given dataset, machine learning techniques are provided (Podder et al., 2021). According to the findings of the experiments, the blood glucose level is the factor that has the most impact on one's ability to predict COVID-19 in this specific dataset. According to the findings, XGBoost has the greatest accuracy value for the case of cv, with a value of 92.67%, while LR has the second-best accuracy value, 92.58%. On the other hand, the precision, recall, and F1 scores for both XGBoost and LR are the same, at 93%. LR demonstrates the maximum level of testing accuracy, which is 94.06%, when the holdout technique is used with 20% of the testing data samples. As a result, XGBoost and LR are both viable options for predicting COVID-19. The authors aimed to predict the factors that influence the perceived usability of Thai Chana. The researchers integrated the Protection Motivation Theory and the Technology Acceptance Theory while also incorporating the System Usability Scale to achieve their objective. The authors utilized a deep-learning neural network and a random forest classifier in their study, as stated in reference (Ong et al., 2022). The study used convenience sampling to gather information from a cohort of 800 respondents. The principal aim of the investigation was to assess a range of factors, encompassing

knowledge about COVID-19, perceptions of its severity and vulnerability, the inclination to utilize it, the factual employment of the system, and perceptions of its usability. Based on the analysis conducted using a deep learning neural network, it was found that a substantial majority of 97.32% of the participants attributed the perceived usefulness of COVID-19 to their understanding of the disease. Furthermore, the results indicate that the random forest classifier achieved a precision rate of 92%, accompanied by a standard deviation of 0.00. The findings of this investigation suggest that a favorable association exists between possessing knowledge regarding COVID-19 and the perception of vulnerability to it, as well as an augmented perception of the efficacy of measures implemented to hinder its transmission.

Additionally, a positive correlation was observed between the perceived severity and perceived ease of use with the perceived usability. The results suggest a favorable association exists between the perceived usability and comprehension of COVID-19, as well as the perceived susceptibility. The results of this investigation could be employed by governmental bodies to encourage the implementation of contact tracing technology in the United States and other nations. To summarise, deep learning neural networks and RF classifiers represent two viable machine learning algorithms for forecasting the determinants that impact human conduct about deploying technologies or systems globally.

According to authors (Moulaei et al., 2022), they aimed to assess multiple machine learning (ML) algorithms to predict the COVID-19 mortality rate based on patient data collected during their initial hospital admission. Finally, the metrics derived from the confusion matrix were calculated to assess the efficacy of the models. The study involved a sample size of 1500 participants, with a notable gender disparity favoring males (836) over females (664). The participants' median age was 57.25 years, with an interquartile range spanning from 18 to 100 years. After conducting the feature selection process, it was found that the three most significant predictors were dyspnea, hospitalization in the intensive care unit, and treatment with oxygen. The analysis encompassed a total of 38 distinct characteristics. The study revealed that smoking, alanine aminotransferase, and platelet count exhibited the lowest precision in forecasting mortality due to COVID-19. The experimental findings indicate that the random forest (RF) technique outperformed other machine learning (ML) algorithms regarding accuracy, sensitivity, precision, and specificity, with 95.03%, 90.70%, 94.23%, and 95.10%, respectively. Additionally, the receiver operating characteristic (ROC) score was

99.02%.

2.2. COVID-19

The disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is commonly called coronavirus. This viral infection is known to primarily affect the respiratory system and body parts. The emergence of this disease can be traced back to the city of Wuhan in China in 2019. Subsequently, it gained momentum and spread rapidly at the onset of 2020, prompting the World Health Organisation to declare it a global pandemic (Booth et al., 2021).

The variability in the severity of Covid-19 is evident among individuals; some experienced mild symptoms while others succumbed to the disease. Individuals who have contracted this virus may exhibit a range of symptoms, such as elevated body temperature, respiratory distress, anosmia and ageusia, myalgia, rhinorrhea, cephalalgia, and thoracic discomfort. The viral outbreak under discussion proved a formidable challenge, causing a staggering 2 million fatalities and infecting a global population of 100 million as of January 25th, 2021. The epidemic's ramifications were far-reaching, with numerous sectors, including industry, commerce, tourism, economics, healthcare, and experiencing significant disruptions (Shinde et al., 2020).

Scholars have endeavoured to assist the healthcare industry by utilizing artificial intelligence methodologies and machine learning algorithms to differentiate individuals who have contracted an infection. The researchers investigate the manifestations exhibited by individuals and develop software applications, such as the "AMAN" app, that transmit notifications to individuals who have been near a Covid-19 patient. The AMAN application is a mobile software designed to notify users of potential exposure to the Coronavirus through GPS technology in cases where contact with an infected individual is suspected (Shinde et al., 2020).

2.3. Medical image classification

COVID-19, a member of the coronavirus family, shares similarities with other viruses such as SARS and ARDS. The World Health Organisation (W.H.O.) has officially declared the current outbreak as a public health emergency. According to their statement, the virus is primarily transmitted through the respiratory tract via contact with an infected individual. The COVID-19 outbreak was first identified in December 2019 in Wuhan, located in the Hubei region of China. After three months, the outbreak above was officially declared a global pandemic by the World Health Organisation (WHO). As of November 16th, 2020, over 54.40 million confirmed cases of COVID-19 and over 1.32

million deaths worldwide have been reported. As indicated in reference (Alazab et al., 2020), the present situation has been deemed the most pressing global emergency since the conclusion of World War II.

The COVID-19 pandemic has affected Kurdistan, Iraq, in other parts of the world, with a rapid spread observed in Sulaymaniyah City. The fatality rate of this ailment is progressively increasing with each passing day, posing a significant threat to the global population. In addition to clinical investigations, the analysis of relevant data will provide valuable support for the human population. According to recent research, machine learning (ML) and artificial intelligence (AI) have emerged as promising technologies utilized by diverse healthcare providers due to their ability to enhance scale-up, expedite processing power, ensure reliability, and even surpass human performance in specific healthcare tasks (Estiri et al., 2021).

Health care and its applications as medical image classification have significantly impacted human life concerning big data technology because the sources of formation of this data in health care are characterized by large size, heterogeneous complexity, and high dynamic. Besides, within the context of big medical data, the success of these applications that depend on health care data depends on the infrastructure and the use of appropriate methods to deal with this big data, for example, predictive models, clinical decision-making, disease control, and public health safety methods, where big data and its analysis play an important role in dealing with heterogeneous medical data, electronic E-health records (Zimmerman et al., 2020).

Some challenges must be dealt with in processing medical data. Interested programmers must make decisions due to the availability of large amounts of primary and complex data, which allows the organization to store and analyze it after the loss. The organization can use effective tools to deal with this data, including the structured and unstructured ones that have been collected from different data generation sources and from these challenges mentioned as follows (Zimmerman et al., 2020):

A. Handling a Large Amount of Data

In order to process large amounts of data, it is considered a challenge to make an appropriate decision to deal with the increase in data access to data from the past years, as these institutions contain everything that the consumer needs and how he interacts, and specific characteristics are very accurate (Zimmerman et al., 2020).

The extensive medical data exceeds the amount of stored data that has been calculated by traditional processing methods and is challenging in the ability to provide

and manage this data, where the high number of unstructured data represented by video, audio, social media, smart device data, which requires advanced methods for data management through a combination of From relational databases (Zimmerman et al., 2020).

B. Data Complexity

With the massive update of medical data generation every second, organizations need to be familiar with the processing of this data; for example, it can help in analyzing data at present for current purchases when selling retail to a company that wants to be able to analyze customer behavior, which contributes to adding accuracy and speed to the processing of this temporary data (Zimmerman et al., 2020).

C. Shortage of Skilled Resources

There is a law that needs for be moreional expertise for those interested in the field of big medical data processing at this time. The challenges facing organizations that seek to use big data in better ways and build data analysis in a way that ensures the effectiveness of these methods have been mentioned (Zimmerman et al., 2020).

2.4. Data Mining

The data mining goal is to find patterns known after the mining. Once these patterns are found, they can be used more and more effectively to make business development decisions. Considerable massive amounts of data in our daily life are generated from different sources, as mentioned. It is worth noting that this data is difficult and almost impossible to handle with manual cases, and it is not easy to obtain knowledge directly as a result of this amount of data that is difficult to understand, as it became necessary to find ways to deal with this data and extract knowledge from it (Bhattacharya et al., 2021).

Knowledge and its quality extracted after data mining depends not only on the design of the method but on the quality of this data, its coherence, and suitability. Unfortunately, it is worth mentioning several negative factors that affect the data, such as noise, missing values, unformatted data, unimportant data, redundant data, and large volumes. This data that is used to extract knowledge and is pain from low-quality data is also low knowledge, and the teacher is a primary, essential, and very important stage aimed at obtaining the final identifier for more data mining algorithms or machine learning. The KDD process (Rahman et al., 2021).

2.4.1. Data preprocessing

Pre-processing big data is an important concept that precedes the data mining process to extract knowledge and clean and refine the data from problems (Rahman et al., 2021).

Often, the discovering extensive data knowledge process includes seven stages from the beginning of discovery to the end:

- Phase 1: Data Integration: it is the process of collecting data from the sources
- Phase 2: Data Selection: It represents selecting the valuable data
- Phase 3: involves the process of data cleaning, which aims to remove any inaccuracies, omissions, or discrepancies in the data, which includes identifying and addressing data errors, values that are absent, and inconsistent data.
- Phase 4: Data Transformation: This is represented by a set of data normalization and smoothing tools, and other forms are considered suitable for data mining
- Phase 5: Data Mining: The application of various modification techniques to detect patterns.
- Phase 6: Evaluation and Presentation of Patterns: It consists of removing various types of redundant patterns through visualization.
- Phase 7: The process of uncovering patterns, trends, and insights from large datasets, commonly called Knowledge Discovery, is a crucial aspect of data analysis in various fields.
- The preprocessing tasks (Rahman et al., 2021).

After the successful data pre-processing stage is applied, the final data is reliable and more suitable for data mining situations and machine learning algorithms. Furthermore, preprocessing is not limited to data mining but rather to other ways to improve and adapt data models to build new proposals (Rahman et al., 2021).

A. Imperfect data

Data mining techniques rely on the assumption that the data is free of errors, but unfortunately, the data generated from reality is almost far from clean and without errors [48].

B. Missing values

One of the assumptions reached by data mining methods and that the set of these data may be complete. However, commonly, the missing values in these data happened since these missing data are data that have yet to be stored or collected due to a defect in the process of taking and calculating samples or related processes. Cost, operations, or

restrictions are affected by purchases (Rahman et al., 2021).

C. Feature indexers and encoders

These properties transform the features of the data from one type without understanding its usefulness to another more understandable type depending on the techniques of indexing and encoding techniques.

- StringIndexer: It is based on converting the text into a series of numeric indicators, depending on the arrangement of the indicators according to the order of the naming frequencies.
- OneHotEncoder: Depends on mapping text columns to columns of unique binary vectors and allows representing the best speakers in strong because it contributes to removing the numerical order dependent on the previous method.
- VectorIndexer: Depends on the automatic report of categorical features and also converts these features into a category index (Rahman et al., 2021).

D. TF-IDF

It aims to determine and know the relevance of each term to a document based on a complete set of different documents. Term Frequency (TF) Depends on measuring how many times a term is found within documents. In contrast, Inverse Document Frequency (IDF) Measures the amount of information provided by the term based on the frequency of this term depending on the download feature to get better performance as it depends on mapping all raw features in a specific index (Solayman et al., 2023).

E. Other pre-processing methods for text mining

They are based on attempts to search for the text after structuring the input, which generates organized information patterns (Solayman et al., 2023).

2.4.2. Deep Learning

It is a way of teaching machines to do activities like humans. By stimulating neurons in the human brain, this strategy aims to discover theories and methods that enable devices to learn by themselves and find ways to extract features from large data sets using linear and nonlinear variables. The basic idea of deep learning is that any object in an image can be described in several ways, such as using the brightness vector for each pixel or the sum of the edges and areas that make up the image, in addition to many additional ways that can be used to describe these images, which is the essence of deep learning. In machine learning, some strategies (such as studying a face or noticing expressions) outperform others. Because of this, people who study deep learning want to eliminate the need for human intervention in feature elicitation and replace it with

algorithms that generate features automatically or almost automatically (Omran et al., 2021).

2.5. Machine Learning Algorithms Taxonomy

Machine learning (ML) is a subfield of artificial intelligence (AI) with developing applications that can learn from data and enhance their accuracy without explicit programming. Using training algorithms, ML models can identify patterns and features in data, enabling them to make informed decisions and predictions based on new data. The ultimate goal of ML is to achieve optimal performance in handling complex and dynamic real-world problems. Machine learning algorithms typically operate through a structured sequence of stages, commencing with identifying and preparing the dataset. Subsequently, an appropriate algorithm is selected to be applied to the training dataset, followed by the algorithm's training to generate the desired model. Ultimately, the model is utilized and refined to enhance its performance. Various categories of machine learning algorithms can be discerned, (Kumari et al., 2021).

ML is an implementation of AI to teach machines how to manage data efficiently. ML aims to learn from the data. In recent times, machine learning (ML) methodologies have been employed for medical prognostication. Diverse ML algorithms can be utilized for distinct applications across various domains. Numerous research studies have indicated that machine learning algorithms have provided superior assistance to clinical support systems, particularly in utilizing patient data for decision-making purposes. Utilizing machine learning (ML) predictive algorithms for illness prognosis represents a precious and robust application within medical services. Commonly, the machine learning algorithm is designed to analyze atypical datasets related to the COVID-19 disease.

Numerous scholarly investigations have been conducted on the topic of autonomous machine learning. Among the most prevalent approaches to machine learning are supervised learning, unsupervised learning, and semi-supervised learning (Abdulkareem et al., 2021).

2.5.1. The Supervised Learning/Predictive Models

Algorithms of this type are based on labeled or untitled training documents and include different methodological methods. Depending on the ML algorithm is trained to work in a system addressed to a specific field. Many algorithms of this type are used for a pre-sorted data set and are considered to be very accurate, and when this data is domain-specific, the model works only for this setting[54].

In data mining, supervised learning can be categorized into two distinct problem types, classification, and regression, as outlined in the reference (Kang et al., 2021).

Classification employs an algorithmic approach to allocate test data into distinct categories precisely. The algorithm identifies distinct entities in the dataset and attempts to formulate definitive labels or definitions for said entities. The prevalent classification algorithms include linear classifiers, SVM, decision trees, k-nearest neighbors, and random forests. These algorithms will be elaborated upon in the subsequent sections.

The statistical technique of regression is commonly employed to comprehend the association between dependent and independent variables. Projections, such as those for sales revenue in a particular business, are frequently generated through its common usage. The regression algorithms that are commonly utilized include linear regression, logistical regression, and polynomial regression (Kang et al., 2021).

Supervised learning models have the potential to develop and enhance various business applications, such as those listed below:

Supervised learning algorithms can identify, segregate, and classify objects from images or videos, rendering them valuable for diverse computer vision techniques and imagery analysis, particularly in image- and object recognition.

Supervised learning models are commonly employed in developing predictive analytics systems, which offer comprehensive insights into diverse business data points. Enterprises can utilize this approach to predict specific outcomes by considering a given output variable, which can assist business leaders in rationalizing their decisions or making necessary adjustments for the betterment of the organization (Kang et al., 2021).

The analysis of customer sentiment can be conducted using supervised machine-learning algorithms. This approach enables organizations to extract and categorize significant information from vast amounts of data, encompassing contextual, emotional, and intentional aspects, with minimal human involvement. This approach is highly advantageous in enhancing comprehension of customer interactions, thereby facilitating the refinement of brand engagement endeavours (Kang et al., 2021).

The identification of spam is an instance of a supervised learning model. By employing supervised classification algorithms, entities can instruct databases to identify regularities or deviations in fresh data, thereby efficiently categorizing spam and non-spam-related communications (Kang et al., 2021).

While supervised learning can benefit businesses, such as enhanced automation and profound data insights, certain obstacles exist in constructing sustainable supervised

learning models. Several challenges exist, including the following:

Accurately structuring supervised learning models may necessitate a certain level of expertise. The process of training models with supervision can be significantly time-consuming.

Datasets may exhibit a greater propensity for human error, leading to erroneous learning of algorithms; in contrast to unsupervised learning models, supervised learning cannot autonomously cluster or classify data, as noted in (Kang et al., 2021).

Inductive machine learning refers to acquiring a set of rules from instances, commonly known as examples in a training set. The objective is to develop a classifier that can be utilized to generalize from new instances. (Kang et al., 2021).

The first phase entails the procurement of the dataset. If a qualified expert is available, they may suggest the attributes or features that are most indicative. In cases where alternative methods are not feasible, the most direct approach involves the implementation of a "brute-force" methodology, which entails measuring all available variables to identify the pertinent characteristics that are both informative and relevant. However, a dataset acquired via the "brute-force" method is unsuitable for induction. The dataset often exhibits noise and incomplete feature values, requiring comprehensive pre-processing methodologies (Altini et al., 2021).

The subsequent phase entails the preparation and pre-processing of data. Researchers can use Various methodologies to handle missing data, depending on the situation. The researchers have identified both the benefits and drawbacks of the techniques above. The application of instance selection is not restricted to mitigating the noise problem but also functions as a strategy to cope with the infeasibility of obtaining knowledge from extensive datasets. Selecting instances from these datasets can be conceptualized as an optimization problem, where the objective is to maintain the integrity of data mining results while concurrently decreasing the sample size. The utilization of data reduction techniques is a method that enhances the efficacy of data mining algorithms when dealing with voluminous datasets (Altini et al., 2021).

There are various methodologies available to select representative samples from large datasets. These methodologies are diverse. The feature subset selection procedure entails identifying and removing superfluous and duplicative features to enhance the efficacy of a model, and decreasing the number of dimensions in data aids in the optimization and efficacy of data mining algorithms. The precision interdependence of multiple features can significantly affect the n of supervised machine learning

classification models previously mentioned problem can be solved by generating innovative characteristics derived from the core set of features. Feature construction or transformation is a widely used term for generating or modifying dataset features. The newly generated features can yield the advancement of more concise and accurate classifiers. Moreover, recognizing noteworthy attributes amplifies the comprehensibility of the resulting classifier and enables a more thorough understanding of the acquired concept (Karthikeyan et al., 2021).

Supervised learning encompasses various algorithms such as Decision Trees (DT), Naive Bayes (NB), K-Nearest Neighbours (KNN), Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) (Karthikeyan et al., 2021). The system under consideration is predicated on the application of machine learning algorithms, which are presented below:

A- Naïve Bayes

Naïve Bayes is the most popular method of classification algorithms that uses filtering applications, and this popularity is back to the quick training speeds they attain and high accuracy despite their relative simplicity to implement. Also, it is one of the simplest classification methods in machine learning. It depends on Bayes' theory with some independent assumptions between the predictors (Arshed et al., 2021).

Naïve Bayes is the most popular method of classification algorithms that uses filtering applications, and this popularity is back to the quick training speeds they attain and high accuracy despite their relative simplicity to implement. Also, it is one of the simplest classification methods in machine learning. It depends on Bayes' theory with some independent assumptions between the predictors (Arshed et al., 2021). The Bayes Theorem equation is (Arshed et al., 2021).

$$P(A|B) = \frac{P(A)*P(B|A)}{P(B)}$$

For explanations :

- A& B: Events.
- P(A), P(B): the (A, B) probabilities.
- P(A|B): conditional probability.
- P(B|A): the probability of B given A (Arshed et al., 2021).

The Algorithm (2.1) (Arshed et al., 2021).

Algorithm (2.1): The used Naive Bayes algorithm
Begin 1- Training the Naïve Bayes document, category (D, C) 2- For every one of the classes c found in category (C), compute the probability of each class (c) 3- compute log-priority [c] for log (Number of Category divided on number of documents) 4- Compute the probability for everyone in the documents. 5- Return Log-prior and Log probability. 6- Removes a final -e. 7- After affixes are omitted (prefixes and suffixes), the word length is determined.
End of algorithm

B- Random Forest (RF) Algorithm

It is a supervised and collective learning algorithm for building a decision tree where the standard network of all variables separates each node. At the same time, the main set is divided into random subtotals (Rasheed et al., 2021).

This algorithm can deal with classification and regression problems and is based on various classification mechanisms, including the random tree classifier based on the input property and the classifier based on the RF. The classifier with the most votes is chosen and represents the average of the responses across all subtrees within RF [58], as shown in Algorithm (2.2) (Rasheed et al., 2021).

Algorithm (2.2):Random Forest Algorithm

Input : dataset = pd.read_csv(path, names = headernames)
Output : Subset of features

```
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 4].values
dataset.head()
sklearn.ensemble import RandomForestClassifier
    classifier = RandomForestClassifier(n_estimators = 50)
     classifier.fit(X_train, y_train)
  from sklearn.metrics
    import classification_report, confusion_matrix, accuracy_score
    result = confusion_matrix(y_test, y_pred)
          print("Confusion Matrix:")
          print(result)
    result1 = classification_report(y_test, y_pred)
          print("Classification Report:",)
          print (result1)
Output: result2 = accuracy_score(y_test,y_pred)
print("Accuracy:",result2)
```

End Algorithm

C- K Nearest Neighbor (KNN)

The K-Nearest Neighbour (KNN) algorithm is a supervised learning approach that involves classifying the outcome of a new instance query based on the majority category of its k-nearest neighbors. The fundamental purpose of an algorithm is to categorize a novel entity by utilizing features and instructional examples. The classification process involves employing a majority vote approach based on the classification outcomes of the k objects. An illustrative instance involves the implementation of a survey to assess the market value of a specific item by examining its consumption patterns. Presented below is an exemplar training Table 2.1. (Jamshidi et al., 2022).

Table 2.1. Training sample

X1	X2	Result
8	8	No
8	5	No
4	5	Yes
1	5	Yes

Determining the binary outcome, "Yes" or "No," is contingent upon the variable values of X1 and X2. When the combination of X1 = 4 and X2 = 8 is absent in the data table, the kNN classification method can predict the results, circumventing the need for extensive surveying procedures. The following Pseudocode (Algorithm (2.3)) serves as

an illustration of instance-based learning techniques (Jamshidi et al., 2022).

Algorithm (2.3): KNN algorithm

Procedure InstanceBaseLearner (Testing Instances) for each testing instance
{
find the k most nearest instances of the training set according to a distance metric
Resulting Class: most frequent class label of the k nearest instances
}
End of algorithm

The main feature of this algorithm is a simple and valuable method in the classification process. It involves classifying a sample based on the majority vote of its neighbors (Jamshidi et al., 2022).

D- Decision Tree Algorithm (DT)

The structure comprises a cluster of interconnected points situated within the arboreal environment. The items are categorized in advance according to their respective branches and the scale of their weight assessment. An integrated text document can be classified by traversing the document's hierarchy, commencing from the root, and following the query structure until a particular page is located within the system. It is worth noting that most training data do not fit with the construction of the decision tree memory, which is often ineffective because of switching a set of training sets (Arista et al., 2022).

This algorithm is capable of processing dataset types that are nominal, ordinal, interval, and ratio. (Arista et al., 2022).

F. Support Vector Machine (SVM)

Supervised machine learning techniques are frequently employed for classification-related issues and can be defined as such. SVM aims to identify the optimal hyperplane that separates the training data with a maximum margin. The classification algorithm is

utilized to forecast whether an item falls under a specific group, as demonstrated in Algorithm (2.4) (Villavicencio et al., 2021).

Algorithm (2.4): Support vector machine
Input: Determine the various training and testing data **Output**: Predicated Class Y

candidateSV = {closest pair from opposite classes }
while there are violating points do
Find a violator
candidateSV = candidateSV ∪violator
if any αp < 0 due to addition of c to S then
candidateSV = candidateSV \p
repeat till all such points are pruned
end if
End Algorithm

Most practical issues entail non-separable data, wherein the training set lacks a hyperplane capable of effectively distinguishing between positive and negative instances. A viable approach to address the issue of inseparability involves projecting data onto a space of higher dimensionality, followed by establishing a separating hyperplane within that space. The space of higher dimensions distinct from the input space where the training instances are located is called the transformed feature space (Villavicencio et al., 2021).

Selecting an appropriate kernel function is crucial for achieving improved results, as the kernel function plays a pivotal role in defining the transformed feature space that will be used to classify the training set instances (Villavicencio et al., 2021).

G. Neural Network Classifier

It is considered one of the multi-layer classifiers used in this research, as it classifies the neural network for the foreground feed to assign a set of data in an input format to find the outputs based on several layers of multiple nodes (Alves et al., 2021).

The Multilayer Perceptron (MLP) has been selected as a classifier for implementation in this study. A classifier based on a feedforward neural network is capable of mapping input data sets to corresponding output sets. The Multi-Layer Perceptron (MLP) architecture comprises numerous layers, each comprising multiple nodes. MLP exhibits primary characteristics, as outlined in the reference (Alves et al., 2021):

The concept of layers.

The Multilayer Perceptron (MLP) comprises a minimum of three layers. Multi-layer Perceptron (MLP) is commonly employed as a fully connected neural network. Each node within the system applies a specific weight to the input data and transmits the resulting output to the subsequent layer. The number of nodes should be determined on each occasion when the experiment is conducted. The MLP classifier notation (x,y,z) denotes the number of neurons in each layer of a three-layered structure. In the context of representing the structure of a Multi-Layer Perceptron (MLP), it is common to use the notation MLP (10, 10, 20) to denote a three-layer architecture with 10 nodes in the first and second layers and 20 nodes in the third layer. (Alves et al., 2021) .

The input layers are situated on the left-hand side, where features are incorporated into the classifier. The hidden layer is located in the middle, while the output of the classifier is situated on the right-hand side.

b. Weight

The weights in a neural network are randomly assigned and subsequently adjusted through exposure to the training set as it propagates from the input layers to the output layer. The adaptations persisted until a particular threshold of error was attained. The MLP classifier is influenced by various parameters, including but not limited to gradient, momentum (with a default value of 0.2 and a recommended range of 0 to 1), and learning rate (with a default value of 0.3 and a recommended range of 0 to 1), as stated in reference (Le et al., 2021).

2.5.2. Unsupervised Learning

These algorithms depend on developing descriptive models in an unsupervised learning manner, where it is possible to know that the inputs between the outputs are unknown. It includes transaction data. For example, these algorithms are k-Means and k-Medians clustering (Ong et al., 2022).

2.5.3. Semi-supervised Learning

It is based on the unlabeled and unlabeled data in the training data set. For instance LR method (Levin et al., 2022).

2.6. The used Dataset

The study utilized a COVID-19 database with medical data to develop and evaluate predictive models for identifying COVID-19 patients. The models were based on ten independent variables. The attributes above are depicted in the database. The data set is being updated regularly. The variables under consideration are sex, age, classification, patient type, pneumonia, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, cardiovascular, renal chronic, other disease, obesity, tobacco, usmr, medical unit, intubed, icu and death. (Matsuda et al., 2023).

2.7. Evaluation Metric

When assessing medical data, such as COVID-19 samples, more than just o accuracy is needed for the effectiveness of a machine learning algorithm. Moreover, it is crucial to accurately diagnose the patient as misidentifying a COVID-19-afflicted patient as a non-infected individual can have significant consequences. This chapter employs various widely used metrics to facilitate the data-based diagnosis of patients with COVID-19. In the context of performance evaluation, TP denotes COVID-19 samples that have been accurately classified as originating from patients affected by SARS-CoV-2e. The variable TN represents the count of individuals within the normative population who accurately receive negative prognostications, which implies that they are categorized as typical individuals seeking medical treatment.

The notation FN represents the count of COVID-19-positive patients who remain undetected. At the same time, FP denotes the count of samples erroneously classified as COVID-19 positive despite being negative. This information has been reported in reference (Xiao et al., 2022).

The study utilized a series of assessment scales founded on the confusion matrix framework. Specifically, a set of equations with distinct nomenclature was employed, as exemplified in Equations 2.1 through 2.8. [68].

- **Precision**

The metric referred to is the ratio of true positives (TP) to the sum of true positives and false positives (TP+FP), commonly known as the TP rate or precision. The computation was performed using Equation 2.1 (Fuhrman et al., 2022).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \textbf{Equation (2.1)}$$

- **Accuracy**

The accuracy of a prediction model is determined by dividing the number of correct predictions by the total number of predictions. The calculation used Equation 2.2, as referenced in (Dellière et al., 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \qquad \textbf{Equation 2.2}$$

- **Recall**

The expression denotes the ratio of true positives to the sum of true positives and false negatives. The computation of this metric can be derived from Equation 2.3, as stated in reference (Agrawal et al., 2021).

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \textbf{Equation 2.3}$$

- **F1-score**

The formula represents the outcome of doubling the product of precision and recall divided by the sum of precision and recall. The metric's equation can be calculated utilizing Equation 2.4, as explained (Yao et al., 2020).

$$F1-\text{score} = \frac{(2*TP)}{(2*TP+FN+FP)} \qquad \textbf{Equation 2.4}$$

- **Detection Rate (DR)**

It is the proportion of correctly identified positive (anomaly) instances; it is calculated by dividing the number of valid positive instances by the total number of actual positive instances. The computation of this metric is feasible by utilizing Equation 2.5. (Rahaman et al., 2020).

$$\text{Detection Rate(DR)} = \frac{TP}{TP+FN} \qquad \textbf{Equation 2.5}$$

False Alert Rate (FAR)

The metric denotes the ratio of negative predictions erroneously classified as positive (anomalies). A lower value is considered to be more desirable. The computation of this metric can be derived from Equation 2.6 (Hemdan et al., 2020).

$$\text{False Alert Rate(FAR)} = \frac{FP}{FP + TN} \qquad \textbf{Equation 2.6}$$

Error rate

The operational definition of "accuracy" can be expressed as the proportion of incorrect predictions to the overall number of predictions conducted on a specific dataset, is shown in Equation 2.7 (Shan et al., 2020).

$$ERR = \frac{b+c}{a+b+c+d} \qquad \textbf{Equation 2.7}$$

Besides, in some cases, it is calculated as follows (Roberts et al., 2021):

Error Rate = Incorrect Predictions / Total Predictions **Equation 2.8**

Error Rate = 1 – Accuracy **Equation 2.9**

- **Mean Absolute Error**

One of the evaluated parameters that are used to measure the closeness of the prediction or the exception to the values of the result, where the mean absolute error is adopted, is shown in Equation 2.10 (Fayyoumi et al., 2020):

$$\textbf{MAE} = \frac{SAE}{N} = \frac{\sum_{i-1}^{n} |x_i - \hat{x}_i|}{N} \qquad \textbf{Equation 2.10}$$

That is, MEN as absolute errors (or deviations), while the N value is a non-missing number of the data points; besides, the x_i is the observations time series, and $\hat{x}_i$ is the forecasted time set (Fayyoumi et al., 2020).

- **Root Mean Squared Error**

It is represented by the square root of the average values, as it squares the Weight of the high errors before calculating the average. It also gives exceptions to this algorithm concerning significant errors, it shown in Equation 2.11 (He et al., 2020):

$$\textbf{RMSE} = \sqrt{\frac{\sum(x_{obs,i} - x_{model,i})}{N}} \qquad \textbf{Equation 2.11}$$

Where X_{obs} has observed values, and X_{model} has modeled values at time/place i. Relative and absolute values are different, and the absolute error measures result in deviation from the actual value. At the same time, the relative error is a percentage measure compared to the actual value (Kumar et al., 2022).

2.8. Confusion Matrix

The evaluation of various models and algorithms is commonly conducted to determine their performance, with metrics such as accuracy, recall, and F-measure being utilized for this purpose. The matrix above depicts the efficacy of a given model on a diverse array of test data. The outcome manifests as a pair of accurate predictive categories and a duo of erroneous prognostications for the employed classifier. Table 2.2. displays the confusion matrix.

There exist multiple measures that can be employed to assess the effectiveness of a classification scheme. In addition to utilizing the Confusion Matrix, alternative methods for assessing performance measures, including Accuracy and Error Rate, have been identified and organized accordingly (Islam et al., 2021).

- **A True Positive (TP)**: It refers to values that have been correctly classified.
- **False Negative (FN)**: The classification needed to be more accurate.
- **False Positive (FP)**: The results indicated that the negative values were inaccurately predicted and classified.
- **True Negative (TN)**: The classification model accurately predicted negative instances, as evidenced by the findings reported in reference (Andreu-Perez et al., 2021). Shows Table 2.2. Confusion Matrixes .

Table 2.2. Confusion Matrixes.

Confusion Matrix		Predicated Class	
		Positive +	**Negative -**
Actual Class	**Positive +**	**TP**	**FN**
	Negative -	**FP**	**TN**

3. MATERIAL AND METHOD

3.1. Work Description

The system under consideration has been executed using the Java Eclipse programming environment. Java is utilized for the implementation of machine learning algorithms. The process comprised three primary phases, namely:

The initial stage involves the pre-processing of data mining on the complete COVID-19 dataset to convert the raw data into an effective and efficient format.

In the second stage, the pre-processed training dataset generates value attributes.

Phase three involves the utilization of machine learning algorithms to obtain outcomes.

The system model depicted in Figure 3.1 is presented as used.

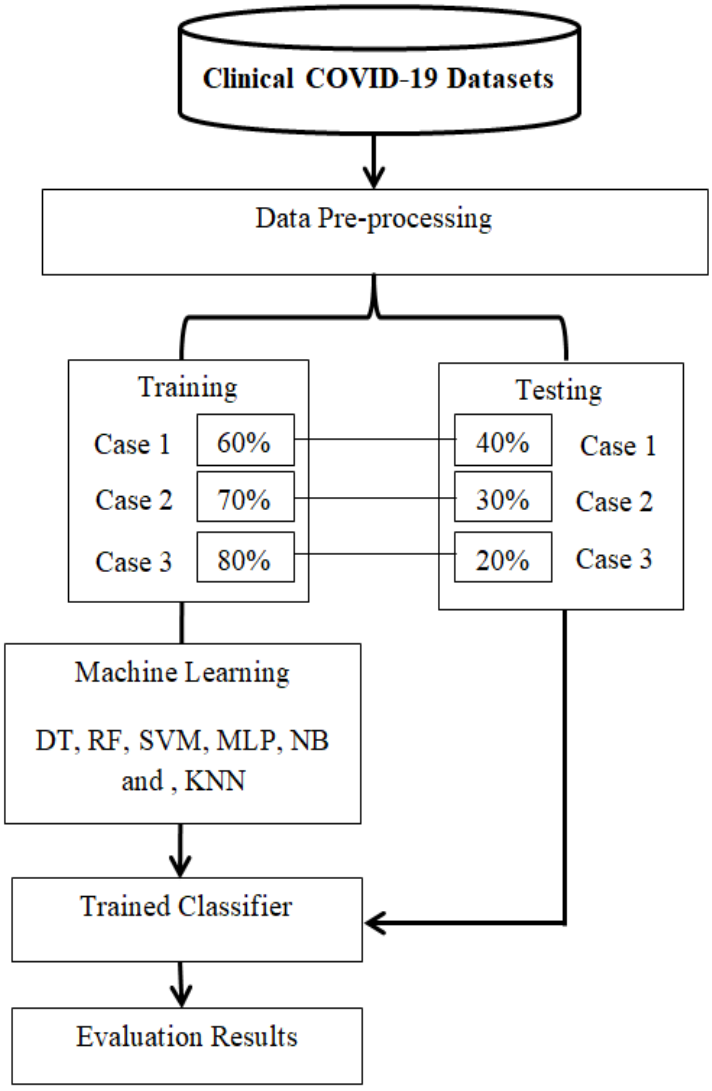


Figure 3.1. The used system Model.

3.1.1. Data Mining Pre-processing

It is considered one of the important processes in extracting unstructured data and the task of converting it into a meaningful and effective format on the other jobs, alongside, it based on the useful data from preprocessing to evaluated with machine learning classifiers as it showed in Figure 3.2 as the main steps of data preprocessing.

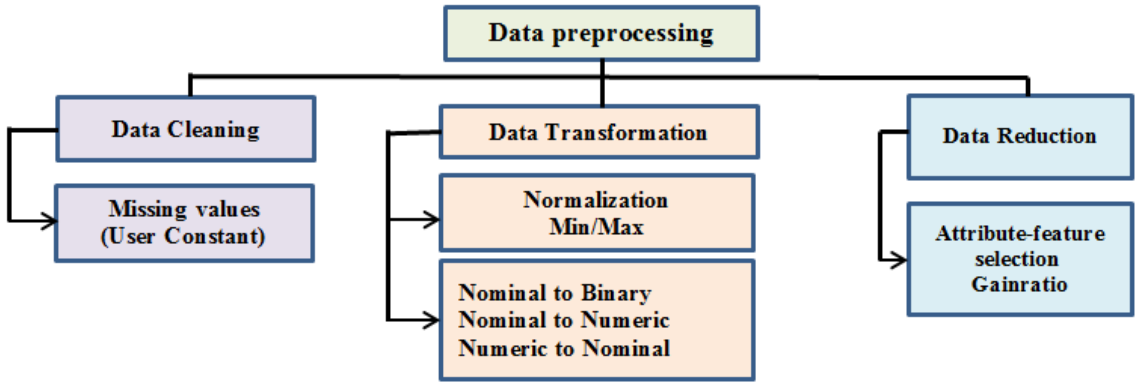


Figure 3.2. The used Data Mining Pre-processing Methods.

A- Normalization: This is done to scale the data values in a specified range.

B- Attribute-feature-selection: It is very important to use attributes that are related and interrelated with each other, and it is possible to get rid of other characteristics as it has become necessary to use a high level of importance of characteristics and ignore other characteristics of little importance is very important in the implementation process. The used method used a gain ratio to determine the splits and to select the most important features.

C- The missing values: This value can be exchanged on other days, the reason, the maximum value, or the average value. Sometimes, the zero value can be used for the missing values. Moreover, fixed values can be adopted as an alternative to the missing action. The used system method uses user-constant values to replace missed attributes in dataset records.

D- The method employed by the used system for converting nominal attribute values into binary states was Nominal to Binary. The present methodology involves utilizing a system that converts nominal attributes (i.e., string values) within the COVID-19 dataset into binary data (0, 1). This approach is deemed optimal for use in the used machine learning algorithm, as it enhances prediction accuracy.

E- Nominal to Numeric: The Nominal to Binary method was utilized by the used system to transform nominal attribute values into binary states. The current approach entails utilizing a system that functions by transforming nominal

attributes, specifically string values, present in the COVID-19 dataset into binary data, denoted as (0, 1). The methodology above is considered the most advantageous for implementing the used machine learning algorithm, as it effectively improves forecasting precision. In this work, the method converts string values in dataset attributes such as ever-married, smoking-status, and residence-type attributes.

F- Numeric to Nominal: Previously, it was utilized to transform numerical data into categorical data. In this work, the used method applied to class values to deal with them as nominal values by classifying the class state as Patient-Test-Status (class 1) with COVID-19 and non-COVID-19 (class 0).

Furthermore, when gathering the healthcare dataset about COVID-19, it is observed that the data comprises both categorical and numeric variables. As Machine Learning algorithms are designed to comprehend numeric data, it is recommended to transform the categorical data into numeric data through techniques such as Label Encoder or one hot encoding.

The Label Encoder technique, which falls under data mining transformation techniques, involves converting categorical data into numeric data. The process involves converting ascending numerical values into a numeric data range of 0 to n-1. The dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients while the second dataset contains on 199999 unique patients records and the same unique features as 21. In the Boolean features, 1 means "yes" and 2 means "no". Values as 97 and 99 are missing data. The first COVID-19 dataset is uploaded from the link: https://www.kaggle.com/datasets/meirnizri/covid19-dataset

and the second COVID-19 dataset is uploaded from the link :

https://www.kaggle.com/datasets/sajoan/covid19-dataset

The dataset description is explained as follows :

- USMER: Indicates whether the patient treated medical units of the first, second or third level.
- MEDICAL_UNIT: type of institution of the National Health System that provided the care.
- SEX: 1 - female. 2 - male
- PATIENT_TYPE: type of care the patient received in the unit. One for return home and 2 for hospitalization.

- DATE_DIED: If the patient died, indicate the date of death, and 9999-99-99 otherwise.
- INTUBED: whether the patient was connected to the ventilator.
- PNEUMONIA: whether the patient already has air sac inflammation or not.
- AGE: Age of the patient. The histogram of age shows that the majority of patients fall within the age group of 30 to 60 years. There is a relatively smaller number of cases in the age group of 0 to 18 years, suggesting that children and teenagers are less affected. The distribution is slightly right-skewed, with a gradual decline in the number of cases in older age groups.
- PREGNANT: whether the patient is pregnant or not.
- DIABETES: whether the patient has diabetes or not
- COPD: whether the patient has Chronic obstructive pulmonary disease or not
- ASTHMA: whether the patient has asthma or not
- INMSUPR: whether the patient is immunosuppressed or not.
- HYPERTENSION: whether the patient has hypertension or not
- OTHER_DISEASE: whether the patient has another disease or not
- CARDIOVASCULAR: whether the patient has heart or blood vessels-related disease.
- DATE DIED: If the patient died, indicate the date of death, and 9999-99-99 otherwise.
- OBESITY: whether the patient is obese or not
- RENAL_CHRONIC: whether the patient has chronic renal disease or not
- TOBACCO: whether the patient is a tobacco user
- CLASIFFICATION_FINAL: Covid test results. covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- ICU: whether the patient had been admitted to an Intensive Care Unit.

In addition, Histograms for "Hypertension" and "Obesity" indicate the prevalence of these comorbidities among the patients. The majority of patients do not have hypertension or obesity, as evident from the high counts in the "0" category for these features. Scaling normalization of numerical features is a fundamental preprocessing step to standardize their values. Standardization ensures that features have a mean of 0 and a standard deviation of 1, making them compatible for use in various machine learning algorithms. Besides, Table 3.1 shows the used data preprocessing methods for each attribute in the

COVID-19 dataset

Table 3.1. The present study aims to elucidate the data mining methodology employed for each attribute in the COVID-19 dataset.

Column Name	Type	Data Mining (Pre-processing)
USMER	int64	Handling Missing Values
MEDICAL_UNIT	int64	Attribute-feature selection (Gain Ratio)
SEX	int64	One-Hot Encoding for Categorical Variables
PATIENT_TYPE	int64	One-Hot Encoding for Categorical Variables
DATE_DIED	Object	Handling Missing Values
INTUBED	int64	One-Hot Encoding for Categorical Variables
PNEUMONIA	int64	One-Hot Encoding for Categorical Variables
AGE	int64	Feature Engineering (Age Groups)
PREGNANT	int64	Transformation (Nominal to Numeric), Cleaning Replace Missing Value (User Constant)
DIABETES	int64	Transformation (Numeric to Nominal)
COPD	int64	Transformation (Nominal to Numeric), Cleaning Replace Missing Value (User Constant)
ASTHMA	int64	Transformation (Nominal to Numeric), Cleaning Replace Missing Value (User Constant)
INMSUPR	int64	Transformation (Nominal to Numeric), Cleaning Replace Missing Value (User Constant)
HIPERTENSION	int64	Normalization
OTHER_DISEASE	int64	Normalization
CARDIOVASCULAR	int64	Normalization
OBESITY	int64	Normalization
RENAL_CHRONIC	int64	Normalization
TOBACCO	int64	Normalization
CLASIFFICATION_FINAL	int64	Normalization
ICU	int64	One-Hot Encoding for Categorical Variables

3.1.2. Machine Learning Algorithms

The system under consideration employs six distinct machine-learning algorithms and is evaluated based on their respective accuracy and building time. The current investigation involved implementing experiments on various machine learning algorithms: RF, NB, SVM, DT, MLP, and KNN.

3.1.2.1. Random Forest (RF) Algorithm

The ensemble classifier comprises numerous decision trees, which collectively determine the mode of class results. The algorithm can handle many input variables and employs an automated instance selection process without implementing pruning techniques. The decision trees are grown to their maximum depth, and each tree is classified independently. Key Features of Random Forest :

- Ensemble of Decision Trees: Random Forest builds multiple decision trees during the training phase. Each tree is trained on a random subset of the training data and features.
- Random Feature Selection: At each node of the decision tree, a random subset of features is considered for splitting. This randomness helps in decorating the trees and making the ensemble more robust.
- Create and Configure Random Forest: Create an instance of the RandomForest class and set its parameters.

```
RandomForest randomForest = new RandomForest();
randomForest.setOptions(weka.core.Utils.splitOptions("-I 100 -K 0 -S 1"));
```

In this example, -I is the number of trees, -K is the number of features to consider at each split, and -S is the random seed.

- Evaluate Model: Evaluate the performance of the model using cross-validation with 10 folds is used. The attributes are allocated to individual trees, which subsequently determine the class with the highest number of votes across all trees through either a majority vote or averaging .

The algorithmic procedure. The process of pseudo code can be described as follows: A bootstrap is selected from a set S for any given forest tree. Si labels the i-th bootstrap. Furthermore, an updated decision-tree learning algorithm is utilized to learn a decision tree, which is subsequently modified in the following manner: In the context of tree structures, it is common practice to randomly select subsets of features for each node rather than exhaustively testing all possible feature splits (where F denotes the set of features). The node exhibited superior performance about the optimal feature denoted by

(F), with (f) representing a significantly lower value. The new weight (hi) was computed at each iteration, commencing with an initial value of hi=0.

3.1.2.2. Naive Bayes (NB) Algorithm

The Naive Bayes classifier methodology is predicated upon using the Bayesian theorem. Despite its simplicity, this method can outperform many sophisticated classification techniques. The classifier is a machine learning model utilized to differentiate between objects based on specific features. The Naive Bayes model is a probabilistic approach utilized in machine learning for classification purposes. The main features of the implementation of the NB algorithm are as follows:

- Training and Testing:

In Java, typically load the dataset into an Instances object. It is used to train the Naive Bayes classifier (buildClassifier method) and then apply it to new instances for classification.

- Evaluation: Java provides tools for evaluating the performance of your classifier, such as cross-validation or splitting your dataset into training and testing sets.

The Naive Bayes model is a probabilistic approach utilized in machine learning for classification purposes.

3.1.2.3. Support Vector Machine (SVM) Algorithm

This is based on the non-probability side, as this aspect is the source of power for this algorithm. It contrasts with many algorithms based on probabilistic classifiers, as they include a set of feature vectors, and this set of data used from them is a subset of data based on decision limits appropriately for vector support. It supports different kernel types for SVM. The used SVM is based on Linear Kernel. It is the default kernel and works well for linearly separable data. Cross-validation is essential for evaluating the performance of the SVM model. The number of optimal folds of the used SVM in the Cross-validation is 10 folds.

3.1.2.4. Decision Tree (DT) Algorithm

What distinguishes this algorithm is that it is based on classification systems with common variables and sometimes multiple, where it classifies data into a group of branches similar to an inverted tree depending on the root node and the nodes inside the tree, and the end nodes, and complicated.

The Decision Tree approach under consideration involves partitioning a given dataset into progressively smaller subsets. Simultaneously, a decision tree linked to the process above is gradually constructed. The outcome of employing this methodology is a

hierarchical structure comprising decision nodes and terminal nodes. Decision trees are capable of processing data that is either categorical or numerical. In the context of decision trees, the term "number of trees" typically refers to the number of decision trees in an ensemble. It performs 10-fold cross-validation using the DT decision tree classifier.as shown in the Figure 3.3 .

The used Decision Tree Algorithm (DT) Algorithm
Input : The data-set D is fed Output: Decision tree
Begin: 1. if D is "pure" OR other stopping criteria met then 2. terminate 3. end if 4. for all attribute ϵ D do 5. Compute information gain and entropy as shown in the following **Entropy (Term)=**$-\sum_{j=1}^{n} Pjlog2pj$**,Gain(age) = Entropy(s) – Infoage(s)** 6. abest =Best attribute according to above computed information gain 7. Tree=Create a decision node that tests abest in the root 8. Dv= Induced sub-datasets from D based on abest 9. For all Dv do 10. Treev=j48 (Dv) 11. Attach Treev to the corresponding branch of Tree 12. End for 13. Return Tree **End**
End of algorithm

Figure 3.3. The used Decision Tree Algorithm (DT) Algorithm

3.1.2.5. Multi-Layer perceptron (MLP)

The aforementioned is a supplementary component of a feedforward neural network. The neural network consists of three discrete layers, specifically the input, output, and hidden layers. The used model experiments with different configurations, including varying the number of hidden layers and neurons, to find the best model for your specific task. Consisting of 20 hidden layers was utilized to forecast daily COVID-19 fatalities in the dataset under consideration. MLP performs cross-validation and adjusts the number of folds to 10 folds. This model was identified as the most profound overall.

3.1.2.6. K-Nearest Neighbor(KNN)

The system is engineered to preserve all accessible data and subsequently categorize a novel data point by assessing its resemblance to pre-existing data points. The K-NN algorithm efficiently classifies newly acquired data into appropriate categories. The algorithm in question is classified as lazy due to its methodology of storing the dataset and deferring learning until the classification stage. Rather than immediately learning from the training set, the algorithm performs actions on the dataset during classification.

The prediction of a new instance (X) is accomplished by examining the complete training set to identify the k most comparable states (neighbors) and summarising the output variable for these K cases. The mode, which refers to the class value that occurs most frequently, is utilized in the classification process. In KNN, the 10 represents the number of folds in the cross-validation, and the new Random(1) is a seed for reproducibility.

3.2. System Installation Requirements

The used system is based on the JAVA programming language, and the main configuration for both is shown as follows:

The main class of the code of the used algorithm in Java is as follows in Algorithm (3.1) of the Java machine learning Pseudo code:

Algorithm(3.1): Machine learning main Class Algorithm

Input : COVID-19 Dataset

Output: Machine learning evaluation results

Begin

- converters.ConverterUtils.DataSource.

Step1: For all classifiers do

Begin

DataSource source = new DataSource("C:\\ COVID-19.csv");

Instances dataset = source.getDataSet(); //Getting data source

ML type Rule = new ML type (); // building classifier algorithm

Rule1.buildClassifier(dataset);

End

Step 2: For each Building Evaluators do

Begin

Evaluation eval = new Evaluation(dataset);

Rest Dataset for Evaluation

DataSource source1 = new DataSource("C:\\newCOVID-19.csv ");

Instances testdataset = source1.getDataSet();

Set class index to the last attribute

eval.evaluateModel(Rule1, testdataset);

End

Step 3: For all evaluation results do

Begin

Print evaluation results based on the machine learning type

eval.pctCorrect(),eval.pctIncorrect(),eval.meanAbsoluteError(),eval.relativeAbsoluteError(), eval.rootRelativeSquaredError(), eval.precision(), eval.recall(), eval.fMeasure(), eval.errorRate(), eval.avgCost()

Then Building the confusion Matrix parameters

End

The splitting dataset into training and testing algorithms in Java is explained in Algorithm (3.2) of Pseudocode splitting the dataset into training and testing modules.

Algorithm (3.2): Pseudocode of splitting dataset

Step1: Adding libraries by import classed as :

- Core.Instances;
- java.io.File;
- java. Util. Random;
- ConverterUtils.DataSource;

Step2: Building main function TrainandTest

A- Loading dataset

```
load dataset
DataSource source = new DataSource("C:\ COVID-19.csv");
Instances dataset = source.getDataSet();
```

B- Setting class index to the last attribute

```
dataset.setClassIndex(dataset.numAttributes()-1);
int seed = 1;
int folds = 10;
```

C- Randomize data

```
Random rand = new Random(seed);
Create random dataset
Instances and data = new Instances(dataset);
randData.randomize(rand);
```

D- Stratify dataset processes

```
if
(randData.classAttribute().nominal())
randData.stratify(folds);
```

E- Perform cross-validation

```
for (int n = 0; n < folds; n++)
Evaluation eval = new Evaluation(and data);
```

F- Get the folds

```
Instances train = and data.trainCV(folds, n);
Instances test = and data.testCV(folds, n);
```

Step 3: Split the dataset

A- Training dataset

```
1- CSVSaver saver = new CSVSaver();
2- saver.set instances(train);
3- System. out.println ("No of folds done = " + (n+1));
4- saver.set file(new File("C:\COVID-19\mytrain.csv"));
5- saver.write batch();
```

B- Testing Dataset

```
6- CSVSaver saver2 = new CSVSaver();
7- Saver2.setInstances(test);
8- Saver2.setFile(new File("C:\COVID-19\mytest.arff"));
9- Saver2.writeBatch();
```

Step 4: End Function

End of Algorithm

4. RESULT AND DISCUSSIONS

The present chapter presents an analysis of the outcomes obtained from the system used in the previous section, as outlined in chapter three. The used system was implemented by studying six cases with data mining (data pre-processing) and machine learning classifiers (DT, SVM, RF, NB, MLP, and KNN).

4.1. The used system implementation

The system under consideration is predicated on three distinct case studies that utilize machine learning, as illustrated in Figure 4.1.

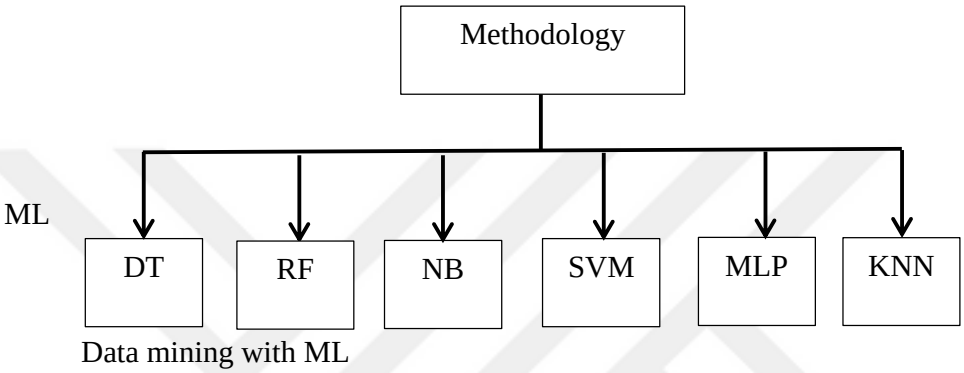


Figure 4.1. The used machine learning algorithms.

4.2. The results of the 1st (Covid Data 1) Dataset

The used system is based on the 1st case study on three splitting of datasets into (60 training, 40 testings, 70 training, and 30 testings, and 80 training and 20 testings) of the first Covid Data 1 dataset, which is evaluated with machine/deep learning classifiers with the maximum accuracy and minimum time required to build the system. The Covid Data 1 dataset contains different columns and rows explained in Table 4.1.

Table 4.1. Number of records and attributed of (Covid Data 1) dataset.

(Covid Data 1) dataset Features	
Number of data instances	1048575
Number of data attributes	21

4.2.1. Results of splitting Covid Data 1 into 60 training and 40 testing

The Covid Data 1 is split into 60 training and 40 testing for the total number of rows used to train the classifier model. Besides, the main accuracy details of the used case study using Data mining/machine learning on the used dataset SVM is the high accuracy as 99.4103 %, and the time to build the model is 1716178 ms. In comparison, MLP has

the second-highest accuracy as 99.4089 %, and the time taken to build the model is 769026 ms. While DT is the third accuracy as 99.3916 %, and the time taken to build the model is 129626 ms. Table 4.2 shows the accuracy and time details with the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset.

Table 4.2. The results of machine Learning for (Covid Data 1) Dataset Analysis of 60 training and 40 of testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**FalseNegative Rate**	
1	Decision Tree (DT)	99.3916 %	0.7702	0.85533	129626 ms
2	Support Vector Machine (SVM)	99.4103 %	0.0	1.0	1716178 ms
3	Random Forest (RF)	99.3541 %	0.00225	0.68133	438407 ms
4	Naïve Bayes (NB)	98.0164 %	0.003501	0.97188	2788 ms
5	Multilayer Perceptron (MLP) Neural	99.4089 %	0.81117	0.999075	769026 ms
6	K-Nearest Neighbor(KNN)	99.2542 %	0.00323	0.483183	1109 ms

Table 4.3 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithm of SVM on the training dataset used ((Covid Data 1) Dataset).

Table 4.3. Correctly / Incorrectly Classified Testing Instances of the data Preprocessing of 60 training and 40 of testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	416878 = 99.3916 %	2552 = 0.6084 %
Support Vector Machine (SVM)	386361 = 99.4103 %	2292 = 0.5897 %
Random Forest (RF)	416721 = 99.3541 %	2709 = 0.6459 %
Naïve Bayes (NB)	411110 = 98.0164 %	8320 = 1.9836 %
Multilayer Perceptron (MLP) Neural	416950 = 99.4089 %	2478 = 0.59101 %
K-Nearest Neighbor(KNN)	246883 = 99.2542 %	1855 = 0.7458 %

Furthermore, the evaluation criteria used MAE, RMSE, and Error Rate shown in Table 4.4 and Figure 4.2 show the prediction of the evaluation criteria of the used algorithms, the KNN as 0.004 almost lower the MAE value, so it is the better compared with others. RF results of the RMSE statistic are lower as the better 0.0522 compared with other algorithms. SVM and MLP algorithms result of error rate is the better for the compared algorithms.

Table 4.4. MAE and RMSE for the (Covid Data 1) machine learning of 60 training and 40 of testing of 60 training and 40 of testing.

Evaluation Criteria	Predication					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.0199	0.2505	0.0055	0.0129	0.0054	0.004
Root Mean Squared Error	0.0582	0.3126	0.0522	0.0976	0.48237	0.0564
Error Rate	0.00608	0.005897	0.006458	0.019836	0.00591	0.00745

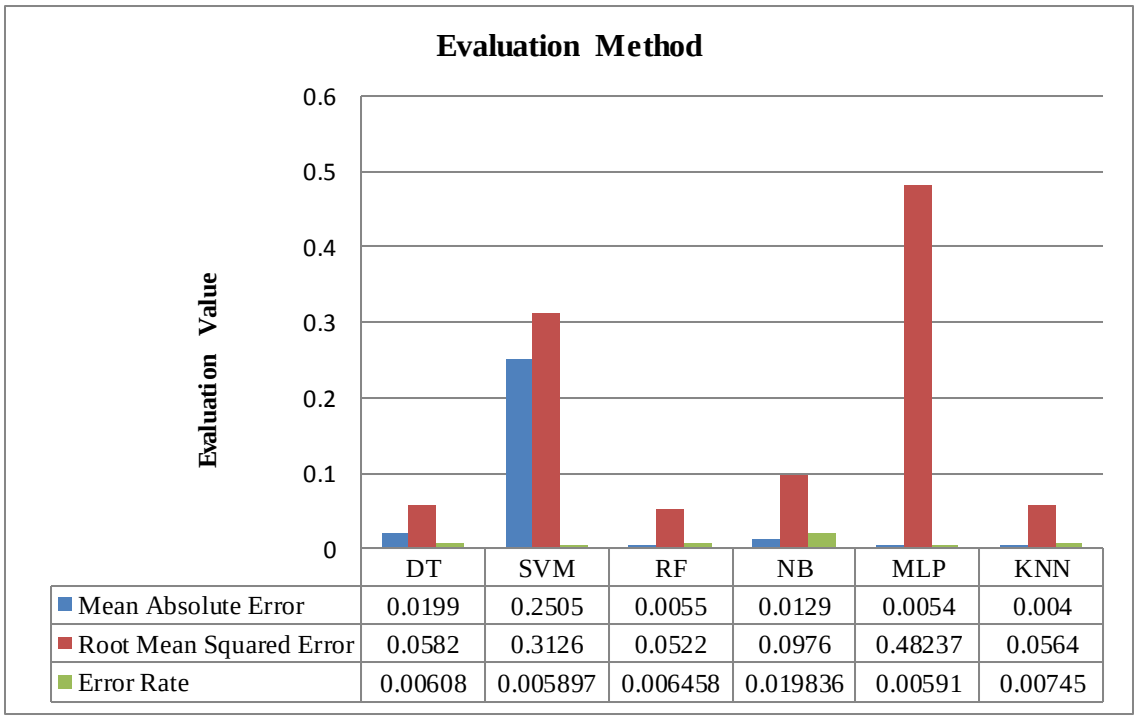
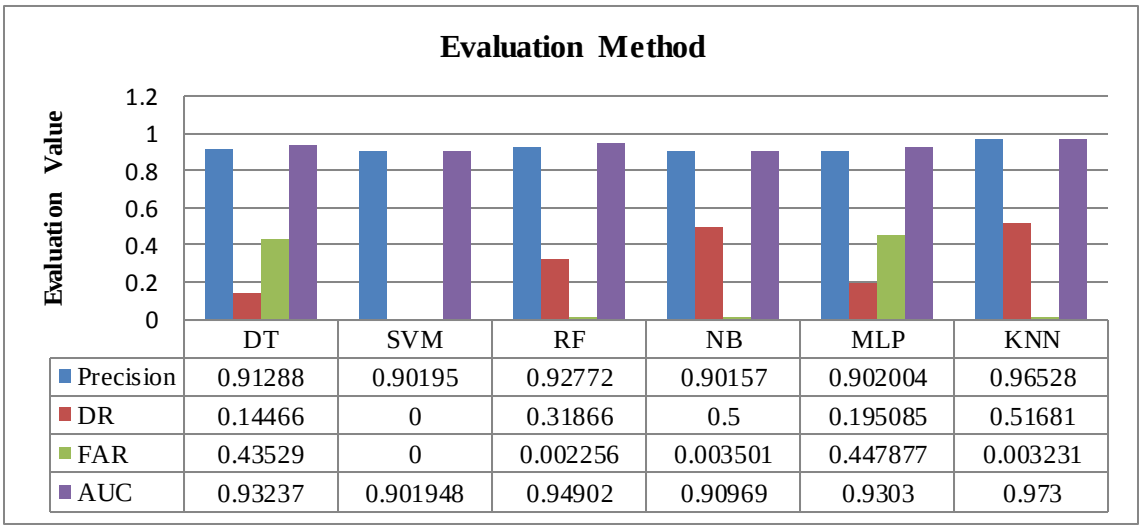


Figure 4.2. The main evaluation Parameters MAE RMSE for the (Covid Data 1) dataset analysis of 60 training 40 and testing.

Besides, there are other evaluation classifiers, as in Table 4.5 of the used system based on the six machine learning classifiers, and they are implemented for both normal cases without (Covid Data 1) (class 0) and with (Covid Data 1) (class 1) as shown in Figure 4.3 of (Covid Data 1) case based on confusion matrix values. In addition, the used system includes various evaluation classifiers, KNN precision as 0.96528 can be seen as a measure of high quality to return more relevant results than irrelevant ones. The AUC of KNN = 0.97300 is higher, so it is better to distinguish between positioning coordination classes. DR of KNN is the best of the whole sample, which was detected correctly. FAR of SVM is best because it indicates fewer false alarms of the used parameters.

Table 4.5. Evaluation (Covid Data 1) of the machine learning for the used data analysis of 60 training and 40 of testing.

Evaluation Parameters	Machine Learning Algorithms					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Precision	0.91288	0.90195	0.92772	0.90157	0.902004	0.96528
Detection Rate (DR)	0.14466	0	0.31866	0.5	0.195085	0.51681
False Alert Rate (FAR)	0.43529	0	0.002256	0.003501	0.447877	0.00323
Area Under Curve (AUC)	0.932370	0.901948	0.94902	0.90969	0.93030	0.97300
True Positive (TP) Rate	0.144661	0.0	0.31866	0.02811	0.24214	0.516816
True Negative (TN) Rate	0.99922	1.0	0.997743	0.99649	0.99998	0.99676



	DT	SVM	RF	NB	MLP	KNN
Precision	0.91288	0.90195	0.92772	0.90157	0.902004	0.96528
DR	0.14466	0	0.31866	0.5	0.195085	0.51681
FAR	0.43529	0	0.002256	0.003501	0.447877	0.003231
AUC	0.93237	0.901948	0.94902	0.90969	0.9303	0.973

Figure 4.3. Precision, DR, FAR and AUS of the used machine learning algorithms of 60 training and 40 of testing.

Table 4.6 shows the SVM is best in recall, which measures the ability of the SVM classifier to detect positive samples; it is ideally the value 0.99990 high for a good classifier. F-Measure showed the KNN is the best model to make a correct prediction across the entire dataset. The Kappa result can be interpreted as follows: Values less than or equal to zero are interpreted as signifying no agreement. Values ranging from 0.01 to 0.20 are considered as indicating none to slight agreement. Values between 0.21 and 0.40 are classified as fair agreement. Values falling within the range of 0.41 to 0.60 are categorized as moderate agreement. Values ranging from 0.61 to 0.80 are considered substantial agreement. Finally, values between 0.81 and 1.00 are interpreted as almost perfect agreement. The Kappa results of KNN and SVM are the almost perfect agreement between dataset attributes.

Table 4.6. Recall the F-Measure and Kappa Coefficient of the (Covid Data 1) of 60 training and 40 of testing.

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.98659	0.94831	0.9526
Support Vector Machine (SVM)	0.99990	0.94840	0.9528
Random Forest (RF)	0.96140	0.94426	0.9498
Naïve Bayes (NB)	0.97395	0.93637	0.8568
Multilayer Perceptron (MLP) Neural	0.99957	0.94828	0.9527
K-Nearest Neighbor(KNN)	0.973038	0.96914	0.96834

Figure 4.4 shows the confusion matrix to describe the performance of a classification model on a set of data for which the true values are known. It allows for the visualization of the performance of an algorithm by breaking down the number of correct and incorrect predictions into various classes.

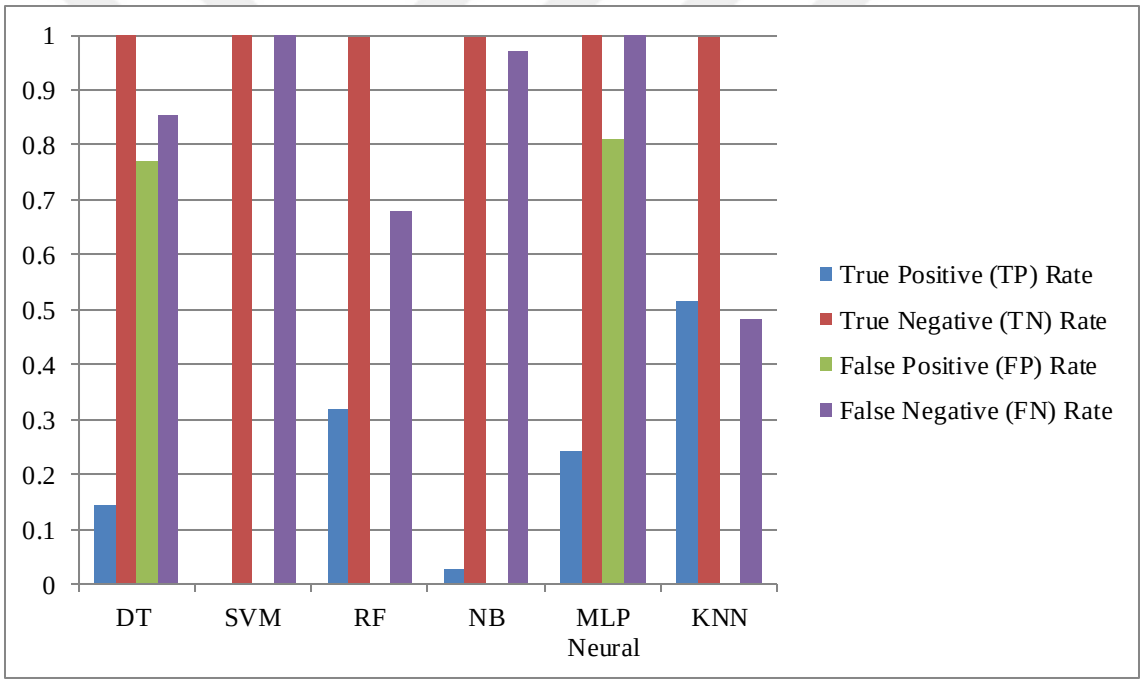


Figure 4.4. Showed the confusion matrix of the used dataset of 60 training and 40 of testing.

4.2.2. Results of splitting Covid Data 1 into 70 Training and 30 testing

Another case study using (the Covid Data 1) dataset with 70 Training and 30 tests showed the MLP has the highest accuracy as 99.5300 %, and time to build model is 1639469 ms, while SVM has the second accuracy as 99.4991 %, and the time take to build model is 1639469 ms. While DT is the third accuracy as 99.4364 %, and the time taken to build the model is 130785 ms. Table 4.7 shows the accuracy and time details with

the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset.

Table 4.7. The results of machine learning for (Covid Data 1) data analysis of 70 Training and 30 testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**False Negative Rate**	
1	Decision Tree (DT)	99.4364 %	3.5165	0.99717	130785 ms
2	Support Vector Machine (SVM)	99.4991 %	0.0	1.0	1639469 ms
3	Random Forest (RF)	99.4087 %	0.00124	0.82273	697883 ms
4	Naïve Bayes (NB)	98.878 %	0.00333	0.96946	4703 ms
5	Multilayer Perceptron (MLP) Neural	99.5300 %	0.00131	0.67739	585496 ms
6	K-Nearest Neighbor(KNN)	99.1939 %	0.02173	4.34216	342007 ms

Table 4.8 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithms (ML, SVM, DT) on the testing dataset used ((Covid Data 1) Dataset).

Table 4.8. Correctly / Incorrectly Classified Testing Instances of the data Preprocessing of 70 Training and 30 testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	312799 = 99.4364 %	1773 = 0.5636 %
Support Vector Machine (SVM)	290030 = 99.4991 %	1460 = 0.5009 %
Random Forest (RF)	208475 = 99.4087 %	1240 = 0.5913 %
Naïve Bayes (NB)	207362 = 98.878 %	2353 = 1.122 %
Multilayer Perceptron (MLP) Neural	309671 = 99.5300 %	1470 = 0.4699 %
K-Nearest Neighbor(KNN)	42699 = 99.1939 %	347 = 0.8061 %

Furthermore, the evaluation criteria used in the used system as Mean Absolute Error (MAE), Root Mean Squared Error(RMSE), and Error Rate shown in Table 4.9 and Figure 4.5 show the prediction of the evaluation criteria of the used algorithms, the RF as 0.0039 almost lower the MAE value, so it is the better compared with others. DT results of the RMSE statistic is the lower as the better 0.0462 compared with other algorithms. MLP algorithm results of error rate is 0.00469 as the better for the compared algorithms.

Table 4.9. MAE and RMSE for the COVID-19 machine learning of 70 Training and 30 testing.

Evaluation Criteria	Predication					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.0048	0.2504	0.0039	0.0075	0.004753	0.0992
Root Mean Squared Error	0.0462	0.3125	0.0492	0.072	0.46531	0.1568
Error Rate	0.00563	0.00500	0.005912	0.01121	0.00469	0.00806

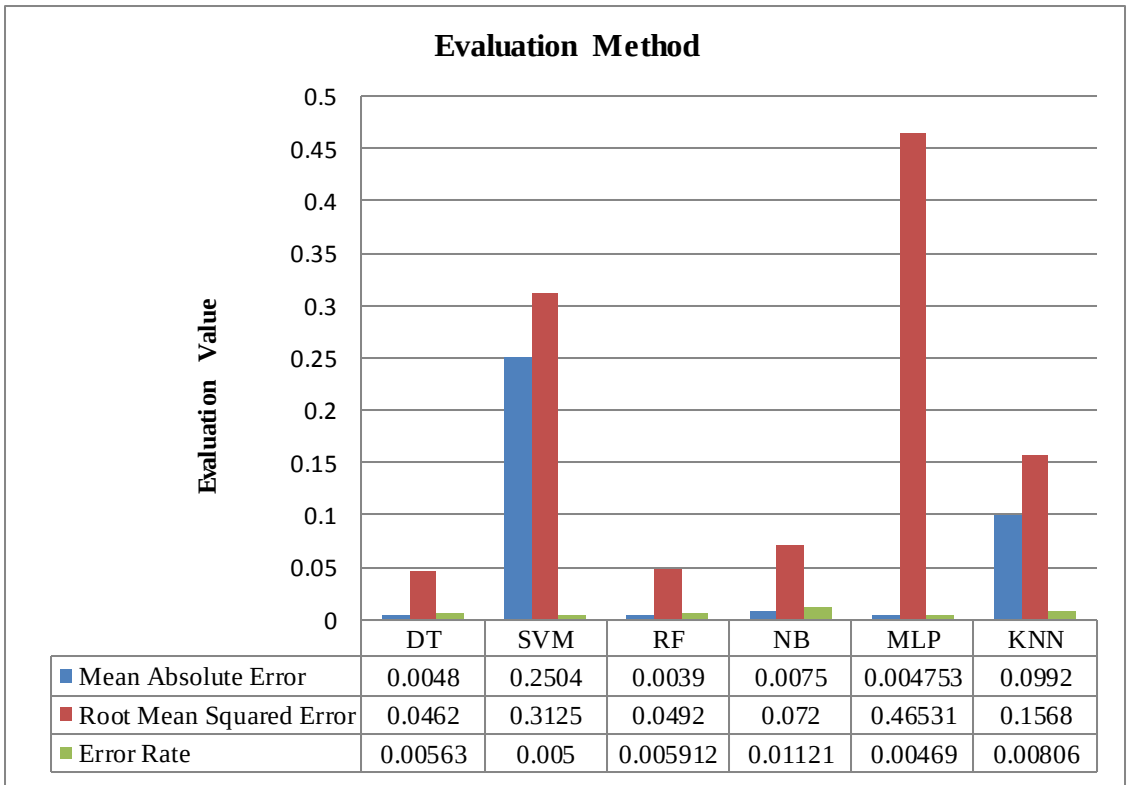


Figure 4.5. The main evaluation Parameters are MAE RMSE for the (Covid Data 1) dataset analysis of 70 Training and 30 tests.

Besides, there are other evaluation classifiers, as in Table 4.10 of the used system based on the six machine learning classifiers. They are implemented for both normal cases without (Covid Data 1) (class 0) and with (Covid Data 1) (class 1) as shown in Figure 4.6 of (Covid Data 1) case based on confusion matrix values. In addition, the used system includes various evaluation classifiers, as outlined in Table 4.49. KNN precision as 0.99920 can be seen as a measure of high quality to return more relevant results than irrelevant ones. The AUC of KNN = 0.998116 is higher, so it is better to distinguish between positioning coordination classes. DR of MLP is the best of the whole sample, which was detected correctly. FAR of RF is best because it indicates fewer false alarms of the used parameters.

Table 4.10. Evaluation (Covid Data 1) of the machine learning for the used data analysis of 70 Training and 30 tests.

Evaluation Parameters	Machine Learning Algorithms					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Precision	0.8952	0.89689	0.90436	0.88950	0.92568	0.99920
Detection Rate (DR)	0.002829	0	0.177268	0.030534	0.322602	0.187124
False Alert Rate (FAR)	3.51657	0	0.001241	0.003337	0.001313	0.021739
Area Under Curve (AUC)	0.929470	0.89689	0.93798	0.907446	0.947943	0.998116
True Positive (TP) Rate	0.00282	0.0	0.177268	0.030534	0.322602	0.999565
True Negative (TN) Rate	0.99996	1.0	0.99875	0.99666	0.998686	0.978260

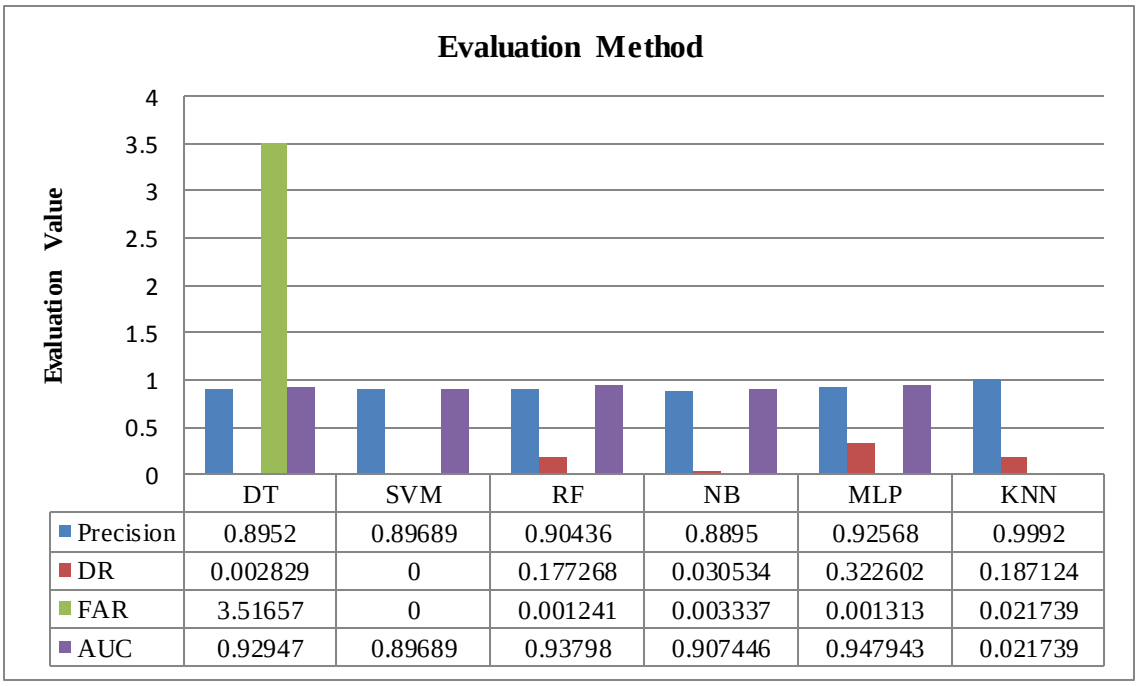


Figure 4.6. Precision, DR, FAR, and AUS of the used machine learning algorithms of 70 Training and 30 testing.

Table 4.11 shows the SVM is best in recall, which measures the ability of the SVM classifier to detect positive samples; it is ideally the value 1 (high) for a good classifier. F-Measure showed the KNN and MLP are the best models to make a correct prediction across the entire dataset. The Kappa result can be interpreted as follows: Values less than or equal to zero are interpreted as signifying no agreement. Values ranging from 0.01 to 0.20 are considered as indicating none to slight agreement. Values between 0.21 and 0.40 are classified as fair agreement. Values falling within the range of 0.41 to 0.60 are categorized as moderate agreement. Values ranging from 0.61 to 0.80 are considered substantial agreement. Finally, values between 0.81 and 1.00 are interpreted as almost

perfect agreement. The Kappa results of KNN and MLP are the almost perfect agreement between dataset attributes. Figure 4.7 shows the confusion matrix to describe the performance of a classification model.

Table 4.11. Recall, F-Measure, and Kappa Coefficient of the (Covid Data 1) of 70 Training and 30 testing

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.999269	0.9443939	0.945475
Support Vector Machine (SVM)	1.0	0.94564	0.9470041
Random Forest (RF)	0.971377	0.936672	0.940000
Naïve Bayes (NB)	0.975405	0.9304747	0.889424
Multilayer Perceptron (MLP) Neural	0.97	0.9473239	0.95033
K-Nearest Neighbor(KNN)	0.9782608	0.9886218	0.982381

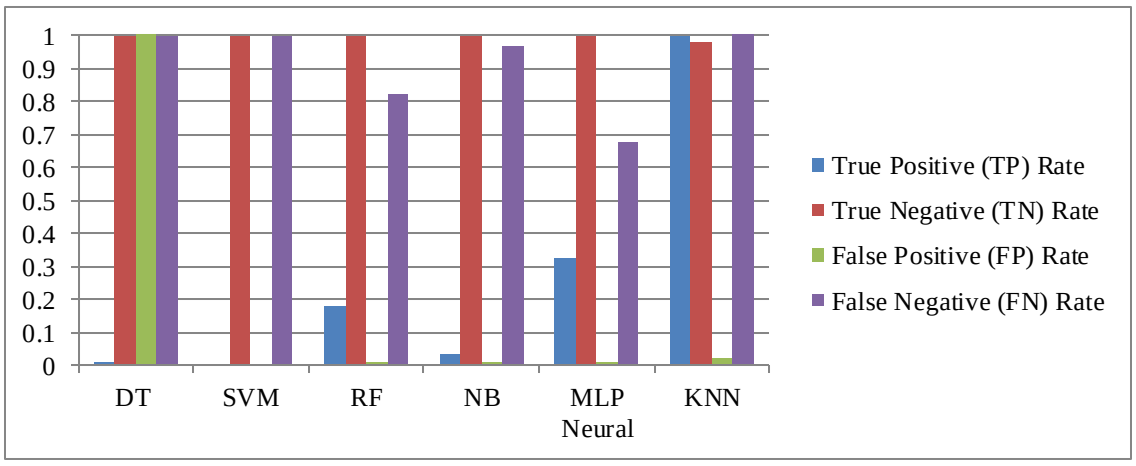


Figure 4.7. Showed the confusion matrix of the used dataset of 70 training and 30 of testing.

4.2.3. Results of splitting Covid Data 1 into 80 Training and 20 testing

The main accuracy of the used splitting dataset into 80 training and 20 testing dataset MLP is the high accuracy as 99.4884 %, and the time to build the model is 760216 ms.

Table 4.12 shows the accuracy and time details with the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset.

Table 4.12. The results of machine learning for (Covid Data 1) Data Analysis of 80 Training and 20 testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**False Negative Rate**	
1	Decision Tree (DT)	99.4349 %	0.52748	0.99575	128822 ms
2	Support Vector Machine(SVM)	99.4566 %	0.0	1.0	2000818 ms
3	Random Forest (RF)	99.4566 %	0.0	1.0	2201277 ms
4	Naïve Bayes (NB)	98.878 %	0.003337	0.96946	2922 ms
5	Multilayer Perceptron (MLP) Neural	99.4884 %	0.5004	0.913825	760216 ms
6	K-Nearest Neighbor(KNN)	99.3728 %	0.0024728	0.450389	1272 ms

Table 4.13 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithms (ML, SVM, RF) on the testing dataset used ((Covid Data 1) Dataset).

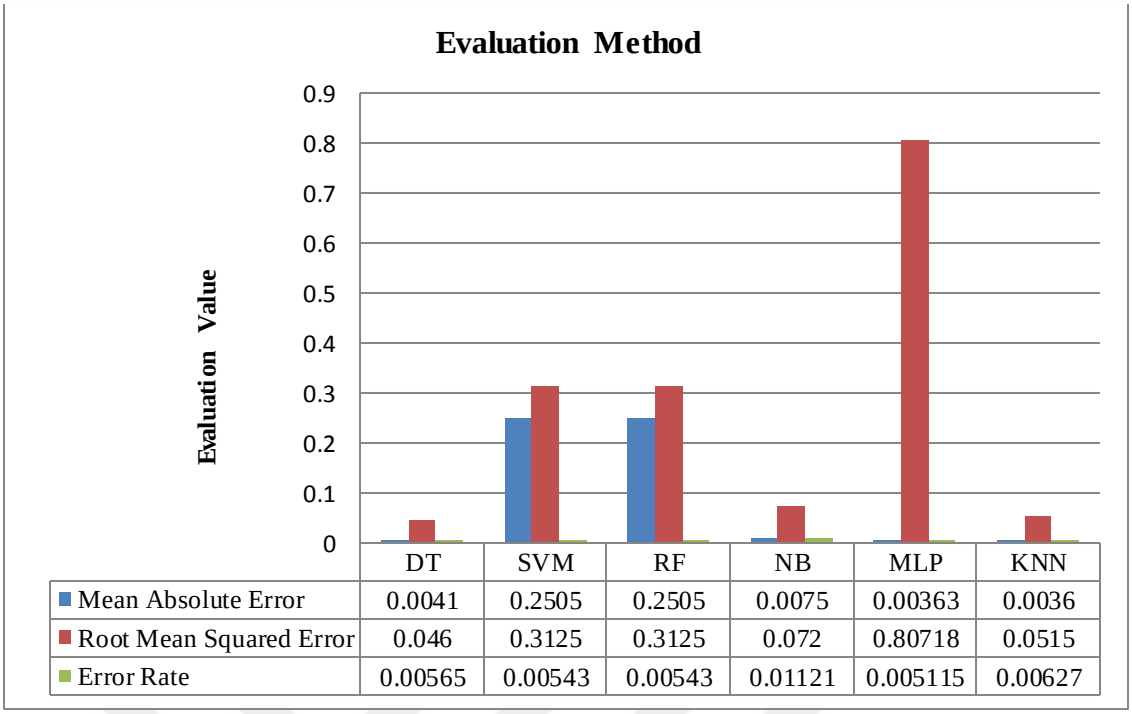
Table 4.13. Correctly / Incorrectly Classified Testing Instances of the data Preprocessing of 80 Training and 20 testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	208530 = 99.4349 %	1185 = 0.5651 %
Support Vector Machine (SVM)	193271 = 99.4566 %	1056 = 0.5434 %
Random Forest (RF)	193271 = 99.4566 %	1056 = 0.5434 %
Naïve Bayes (NB)	207362 = 98.878 %	2353 = 1.1219 %
Multilayer Perceptron (MLP) Neural	210060 = 99.4884 %	1079 = 0.5115 %
K-Nearest Neighbor(KNN)	123589 = 99.3728 %	780 = 0.6272 %

Table 4.14 and Figure 4.8 show the prediction of the evaluation criteria of the used algorithms. The MLP of 0.00363 almost lower the MAE value, so it is better compared with others. DT results of the RMSE statistic are the lower as the better 0.046 compared with other algorithms. MLP algorithm result of error rate is 0.005115 as the better for the compared algorithms.

Table 4.14. MAE and RMSE for the COVID-19 machine learning of 80 Training and 20 testing.

Evaluation Criteria	Predication					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.0041	0.2505	0.2505	0.0075	0.00363	0.0036
Root Mean Squared Error	0.046	0.3125	0.3125	0.072	0.80718	0.0515
Error Rate	0.00565	0.00543	0.00543	0.01121	0.005115	0.00627



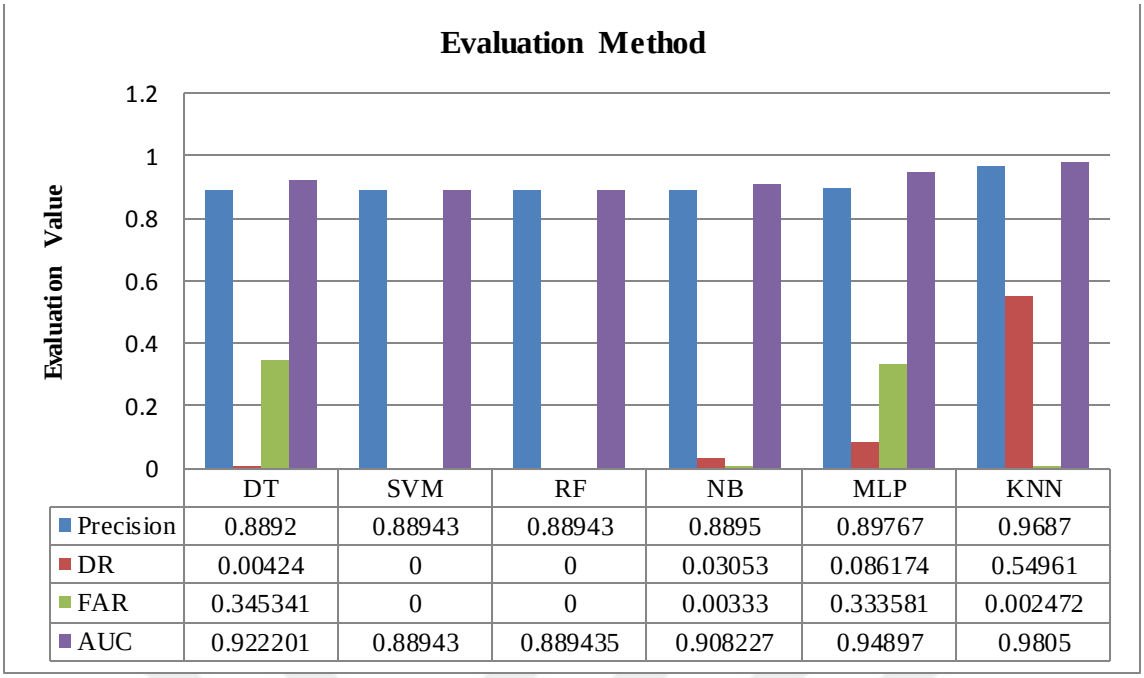
	DT	SVM	RF	NB	MLP	KNN
Mean Absolute Error	0.0041	0.2505	0.2505	0.0075	0.00363	0.0036
Root Mean Squared Error	0.046	0.3125	0.3125	0.072	0.80718	0.0515
Error Rate	0.00565	0.00543	0.00543	0.01121	0.005115	0.00627

Figure 4.8. The main evaluation Parameters MAE RMSE for the (Covid Data 1) dataset Analysis of 80 Training and 20 tests.

In addition, the used system includes various evaluation classifiers, as outlined in Table 4.15 and Figure 4.9 KNN precision of 0.96870 can be seen as a measure of high quality to return more relevant results than irrelevant ones. The AUC of KNN = 0.98050 is higher, so it is better to distinguish between positioning coordination classes. DR of KNN is the best of the whole sample, which is detected correctly; it indicates that the KNN model is effective at capturing positive instances. FAR of RF and SVM is best because it indicates fewer false alarms of the used parameters.

Table 4.15. Evaluation (Covid Data 1) of the machine learning for the used data analysis of 80 training and 20 testing.

Evaluation Parameters	Machine Learning Algorithms					
	DT	SVM	RF	NB	MLP	KNN
Precision	0.88920	0.88943	0.88943	0.88950	0.89767	0.96870
Detection Rate (DR)	0.004240	0	0	0.03053	0.086174	0.549610
False Alert Rate (FAR)	0.345341	0	0	0.00333	0.333581	0.002472
Area Under Curve (AUC)	0.922201	0.88943	0.889435	0.908227	0.94897	0.98050
True Positive (TP) Rate	0.004240	0.0	0.0	0.03053	0.08617	0.54961
True Negative (TN) Rate	0.99994	1.0	1.0	0.99666	0.99984	0.99752



	DT	SVM	RF	NB	MLP	KNN
Precision	0.8892	0.88943	0.88943	0.8895	0.89767	0.9687
DR	0.00424	0	0	0.03053	0.086174	0.54961
FAR	0.345341	0	0	0.00333	0.333581	0.002472
AUC	0.922201	0.88943	0.889435	0.908227	0.94897	0.9805

Figure 4.9. Precision, DR, FAR, and AUS of the used machine learning algorithms of 80 Training and 20 testing.

Table 4.16 shows the SVM is best in recall, which measures the ability of the SVM and RF classifier to detect positive samples; it is ideally the value 1 (high) for a good classifier. F-Measure showed the KNN is the best model to make a correct prediction across the entire dataset. The Kappa results of KNN and MLP are the almost perfect agreement between dataset attributes.

Table 4.16. Recall, F-Measure, and Kappa Coefficient of the (Covid Data 1) of 80 training and 20 testing.

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.998833	0.940835	0.942610
Support Vector Machine (SVM)	1.0	0.94148	0.943314
Random Forest (RF)	1.0	0.94148	0.94331
Naïve Bayes (NB)	0.97540	0.93047	0.889424
Multilayer Perceptron (MLP) Neural	0.99658	0.94454	0.94665
K-Nearest Neighbor(KNN)	0.97946	0.974057	0.973361

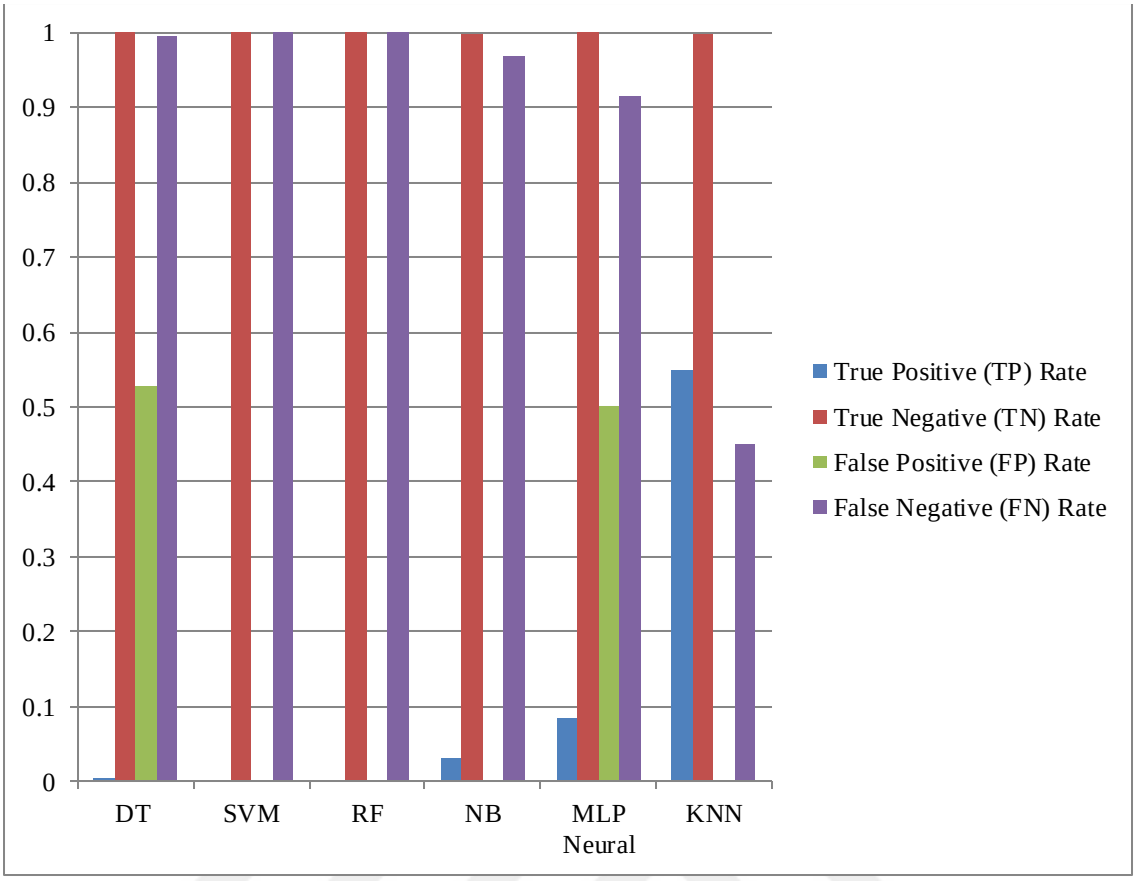


Figure 4.10. Showed the confusion matrix of the used dataset of 80 training and 20 of testing.

4.3. The results of the 2nd (Covid Data 2) Dataset

The used system is based on the 2nd case study preprocessing procedure, which is evaluated with machine/deep learning classifiers with the maximum accuracy and minimum time required to build the system. The used dataset is (Covid Data 2), which contains different columns and rows explained in Table 4.17.

Table 4.17. Number of records and attributed of (Covid Data 2) dataset.

(Covid Data 2) dataset Features	
Number of data instances	199999
Number of data attributes	21

4.3.1. The case of 60 Training and 40 of Testing

The main accuracy details of 60 Training and 40 of Testing dataset DT is 92.12620 %, and MLP is 92.12625 % as the high accuracy and time to build model is 75434 ms and 1488944 ms of DT and MLP algorithms, respectively.

Table 4.18 shows the accuracy and time details with the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset of 60 training and 40 testing.

Table 4.18. The results of machine learning for (Covid Data 2) data analysis of 60 training and 40 of testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**False Negative Rate**	
1	Decision Tree (DT)	92.12620 %	1.0	0.0	75434 ms
2	Support Vector Machine (SVM)	82.4433 %	0.8696	0.0988	2122065 ms
3	Random Forest (RF)	76.16 %	0.7721	0.19277	141583 ms
4	Naïve Bayes (NB)	91.71 %	0.8517	0.0171	468 ms
5	Multilayer Perceptron (MLP) Neural	92.12625 %	1.0	0.0	1488944 ms
6	K-Nearest Neighbor(KNN)	73.4912 %	0.68296	0.22937	163 ms

Table 4.19 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithms (DT and MLP) on the testing dataset used ((Covid Data 2) Dataset).

Table 4.19. Correctly / Incorrectly Classified Testing Instances of the data Preprocessing of 60 training and 40 of testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	73701= 92.1262 %	6299 = 7.8738 %
Support Vector Machine (SVM)	65954 = 82.4433 %	14045 = 17.5566 %
Random Forest (RF)	60928 = 76.16 %	19072 = 23.84 %
Naïve Bayes (NB)	73368 = 91.71 %	6632 = 8.29 %
Multilayer Perceptron (MLP) Neural	73701 = 92.1262 %	62990 = 7.87375 %
K-Nearest Neighbor(KNN)	58793 = 73.4912 %	21207 = 26.5087 %

Table 4.20 and Figure 4.11 show the prediction of the evaluation criteria of the used algorithms. The NB is almost lower than the MAE value, so it is better compared with others. ML results of the RMSE statistic are lower and better compared with other algorithms. DT and MLP algorithm results of error rate is 0 as the better for the compared algorithms.

Table 4.20. MAE and RMSE for the COVID-19 machine learning of 60 training and 40 of testing.

Evaluation Criteria	Predication					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.3945	0.3844	0.4012	0.3301	0.454	0.4164
Root Mean Squared Error	0.4002	0.9870	0.4318	0.348	0.1050	0.5264
Error Rate	0.0787	0.1755	0.2384	0.0829	0.0787	0.2650

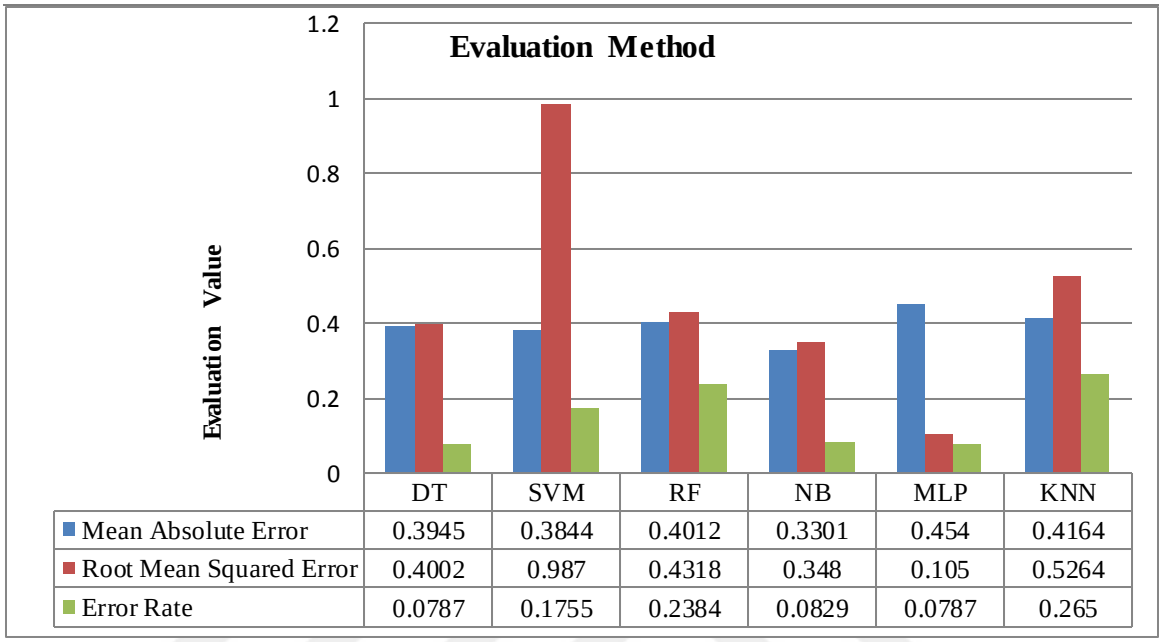
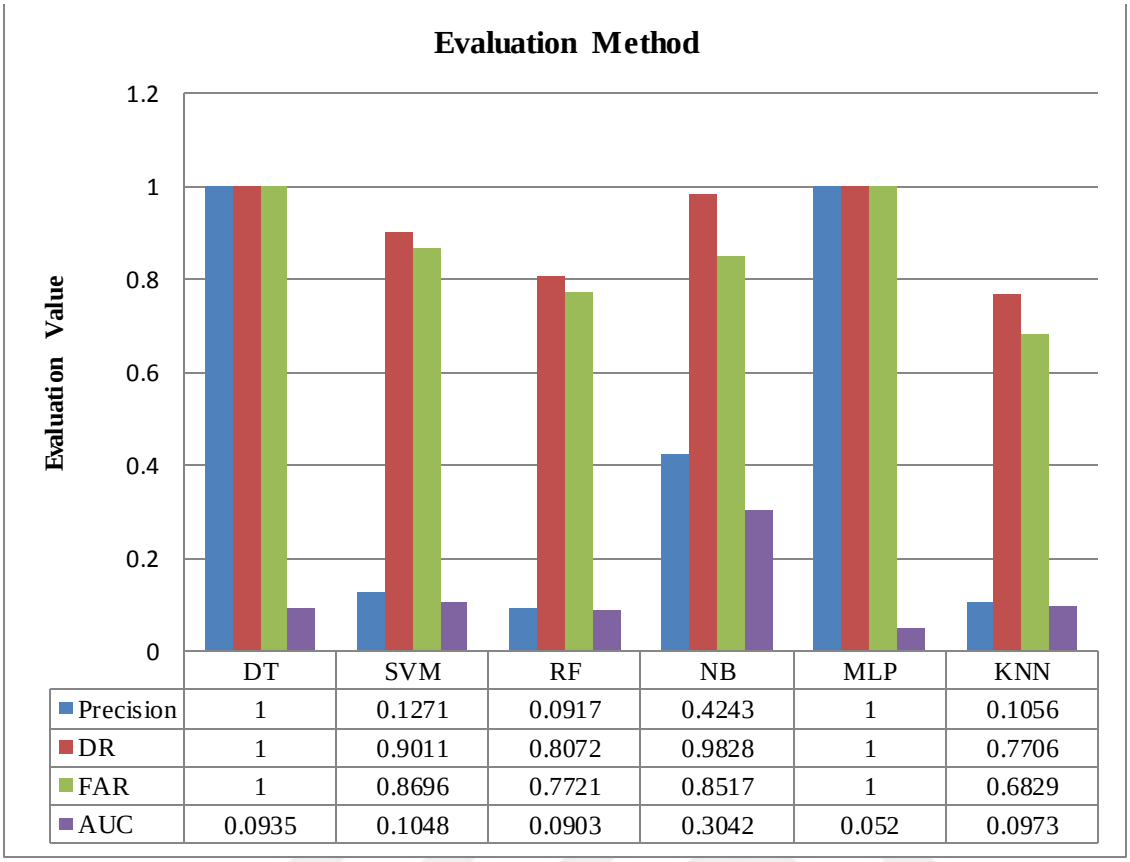


Figure 4.11. The main evaluation Parameters MAE RMSE for the (Covid Data 2) dataset analysis of 60 training and 40 testing.

Besides, there are other evaluation classifiers, as in Table 4.21 of the used system and Figure 4.12 of (Covid Data 2) case based on confusion matrix values as outlined in Table 4.49. DT and MLP precision as 1 can be seen as a measure of high quality to return more relevant results than irrelevant ones. DR of DT MLP is the best of the whole sample, which was detected correctly. FAR of KNN is the best because it indicates fewer false alarms of the used parameters.

Table 4.21 Evaluation (Covid Data 2) of the machine learning for the used data analysis of 60 training and 40 of testing.

Evaluation Parameters	Machine Learning Algorithms					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Precision	1.0	0.1271	0.0917	0.4243	1.0	0.1056
Detection Rate (DR)	1.0	0.9011	0.8072	0.9828	1.0	0.7706
False Alert Rate (FAR)	1.0	0.8696	0.7721	0.8517	1.0	0.6829
Area Under Curve (AUC)	0.0935	0.1048	0.0903	0.3042	0.0520	0.0973
True Positive (TP) Rate	1.0	0.9011	0.8072	0.9828	1.0	0.7706
True Negative (TN) Rate	0.0	0.1303	0.2278	0.1482	0.0	0.3170



	DT	SVM	RF	NB	MLP	KNN
Precision	1	0.1271	0.0917	0.4243	1	0.1056
DR	1	0.9011	0.8072	0.9828	1	0.7706
FAR	1	0.8696	0.7721	0.8517	1	0.6829
AUC	0.0935	0.1048	0.0903	0.3042	0.052	0.0973

Figure 4.12. Precision, DR, FAR and AUS of the used machine learning algorithms of 60 training and 40 of testing.

Table 4.22 shows the DT and MLP are low in recall, which measures the ability of classifiers to detect positive samples. F-Measure showed the SVM is a low model to make a correct prediction across the entire dataset as it is not performing well in the ability of the NB algorithm to capture all the relevant instances. The Kappa results of NB are the slight agreement between dataset attributes.

Table 4.22. Recall, F-Measure and Kappa Coefficient of the (Covid Data 2) of 60 training and 40 of testing.

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.0	1.0	0.0
Support Vector Machine (SVM)	0.1303	0.1287	0.03110
Random Forest (RF)	0.2278	0.1307	0.0209
Naïve Bayes (NB)	0.1482	0.2197	0.1866
Multilayer Perceptron (MLP) Neural	0.0	1.0	0.0
K-Nearest Neighbor(KNN)	0.3170	0.1584	0.0457

Figure 4.13 shows the confusion matrix to describe the performance of a classification model.

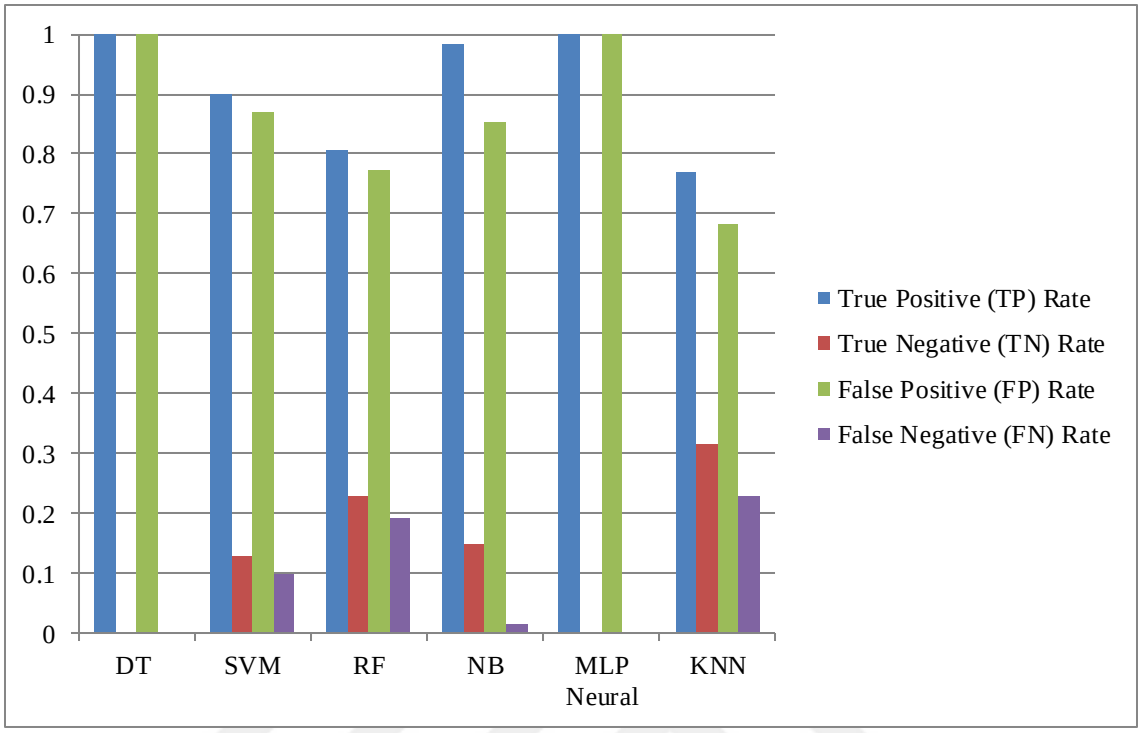


Figure 4.13. Showed the confusion matrix of the used dataset of 60 training and 40 of testing.

4.3.2. The case of 70 Training and 30 of Testing

Besides, the main accuracy details of the used case study using preprocessing of machine learning on the used dataset MLP is the high accuracy as 92.1262 %, and the time to build model is 39027474 ms of DT algorithm. At the same time, DT is the second level of accuracy as 90.0517 %, and the time taken to build the model is 79205 ms. Table 4.23 shows the accuracy and time details with the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset.

Table 4.23. The results of machine learning for (Covid Data 2) Data analysis of 70 training and 30 testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**False Negative Rate**	
1	Decision Tree (DT)	90.0517 %	1.0	0.0	79205 ms
2	Support Vector Machine (SVM)	86.1599 %	1.0	1.0	11689 ms
3	Random Forest (RF)	71.7967 %	0.81990	0.22261	231485 ms
4	Naïve Bayes (NB)	89.8617 %	0.8239	0.02156	921 ms
5	Multilayer Perceptron (MLP) Neural	92.1262 %	1.0	0.0	39027474 ms
6	K-Nearest Neighbor(KNN)	69.515 %	0.71971	0.25901	272 ms

Table 4.24 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithms (MLP and DT) on the testing dataset used ((Covid Data 2) Dataset).

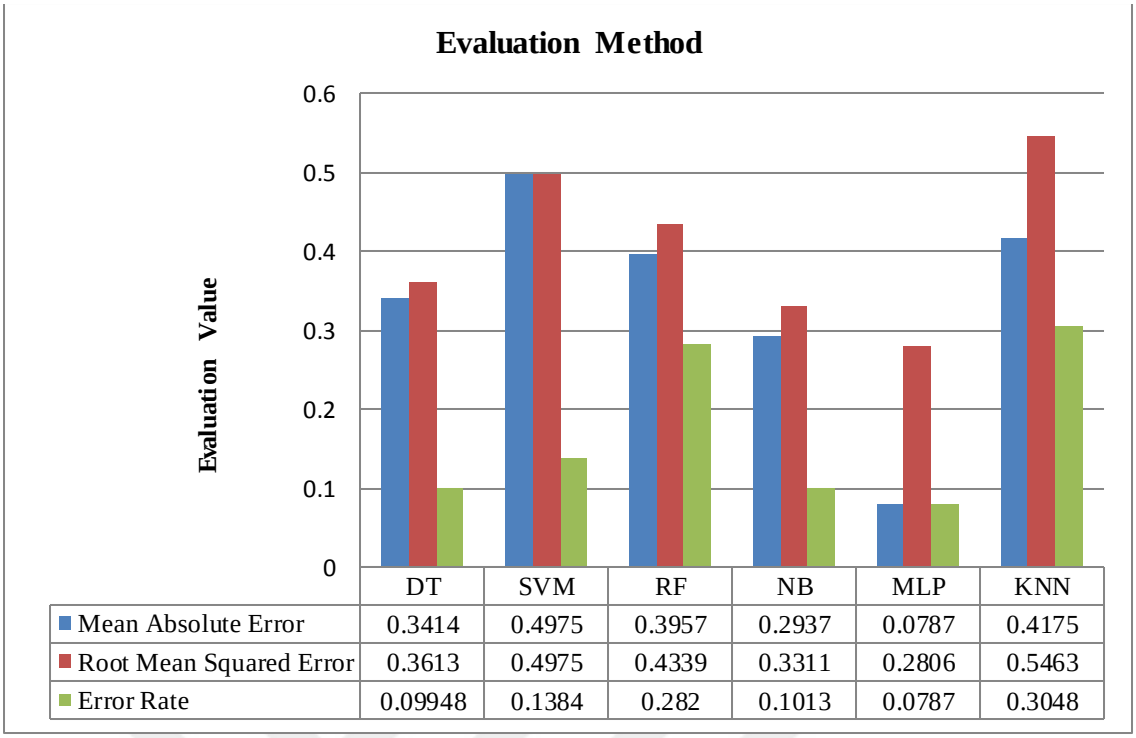
Table 4.24. Correctly / Incorrectly Classified Testing Instances of the data preprocessing of 70 training and 30 of testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	54031 = 90.0517 %	5969 = 9.9483 %
Support Vector Machine (SVM)	44803 = 86.1599 %	7196 = 13.8401 %
Random Forest (RF)	43078 = 71.7967 %	16922 = 28.2033 %
Naïve Bayes (NB)	53917 = 89.8617 %	6083 = 10.1383 %
Multilayer Perceptron (MLP) Neural	73701 = 92.1262 %	6299 = 7.8738 %
K-Nearest Neighbor(KNN)	41709 = 69.515 %	18291 = 30.485 %

Furthermore, the evaluation criteria used in the used system as MAE, RMSE, and Error Rate shown in Table 4.25 and Figure 4.14 show the prediction of the evaluation criteria of the used algorithms, the MLP as 0.0787 almost lower the MAE value, so it is the better compared with others. MLP results of the RMSE statistic are the lower as, the better with 0.2806 compared with other algorithms. MLP algorithm results of error rate is 0.0787 as the better for the compared algorithms.

Table 4.25. Error metrics for the Covid Data 2 of 70 training and 30 of testing.

	Predication					
Evaluation Criteria	**DT**	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.3414	0.4975	0.3957	0.2937	0.0787	0.4175
Root Mean Squared Error	0.3613	0.4975	0.4339	0.3311	0.2806	0.5463
Error Rate	0.09948	0.1384	0.2820	0.1013	0.0787	0.3048



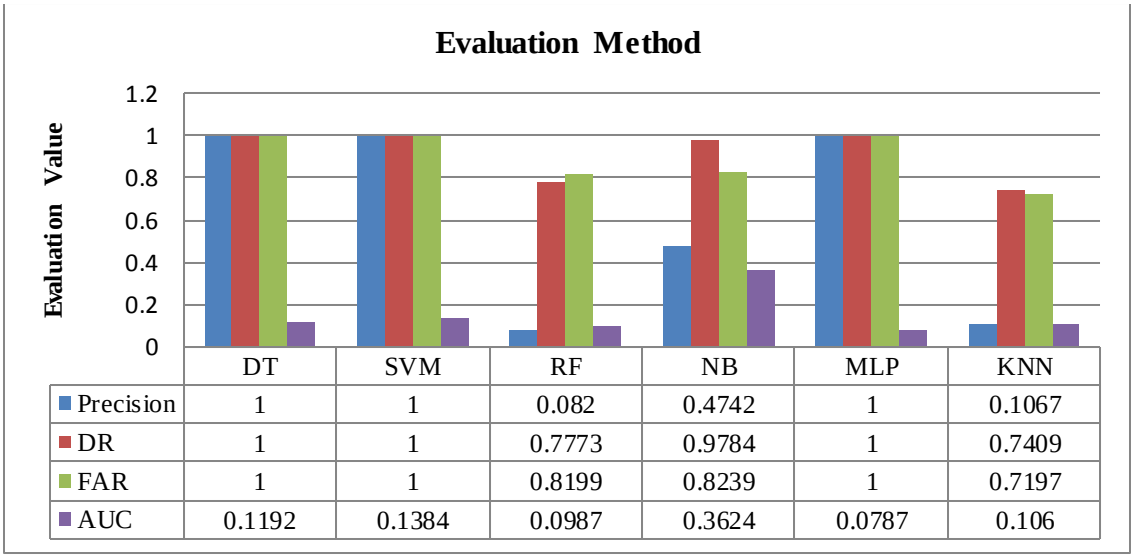
	DT	SVM	RF	NB	MLP	KNN
Mean Absolute Error	0.3414	0.4975	0.3957	0.2937	0.0787	0.4175
Root Mean Squared Error	0.3613	0.4975	0.4339	0.3311	0.2806	0.5463
Error Rate	0.09948	0.1384	0.282	0.1013	0.0787	0.3048

Figure 4.14. The MAE, RMSE evaluation for the (Covid Data 2) dataset analysis.

Besides, there are other evaluation classifiers, as in Table 4.26 of the used system based on the six machine learning classifiers. They are implemented for both normal cases without (Covid Data) (class 0) and with (Covid Data) (class 1), as shown in Figure 4.15 of the (Covid Data) case based on confusion matrix values. In addition, the used system includes various evaluation classifiers, as outlined in Table 4.49. DT, SVM, and MLP precision as 1 can be seen as a measure of high quality to return more relevant results than irrelevant ones. DT, SVM, and MLP are 1 as higher, so it is better to distinguish between positioning coordination classes. DR of DT, SVM, and MLP is the best of the whole sample, which was detected correctly.

Table 4.26. Evaluation (Covid Data 2) of the machine learning for the used data analysis.

Evaluation Parameters	Machine Learning Algorithms					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Precision	1.0	1.0	0.0820	0.4742	1.0	0.1067
Detection Rate (DR)	1	1	0.7773	0.9784	1	0.7409
False Alert Rate (FAR)	1	1	0.8199	0.8239	1	0.7197
Area Under Curve (AUC)	0.1192	0.1384	0.0987	0.3624	0.0787	0.1060
True Positive (TP) Rate	1.0	1.0	0.7773	0.9784	1.0	0.7409
True Negative (TN) Rate	0.0	0.0	0.1800	0.1760	0.0	0.2802



	DT	SVM	RF	NB	MLP	KNN
Precision	1	1	0.082	0.4742	1	0.1067
DR	1	1	0.7773	0.9784	1	0.7409
FAR	1	1	0.8199	0.8239	1	0.7197
AUC	0.1192	0.1384	0.0987	0.3624	0.0787	0.106

Figure 4.15. Precision, DR, FAR, and AUS of the used machine learning algorithms.

Table 4.27 showed the DT, SVM, and MLP are low in recall, which measures the ability of the KNN classifier to detect positive samples. F-Measure showed the RF is a low model to make a correct prediction across the entire dataset as it is not performing well in the ability of the KNN algorithm to capture all the relevant instances. The Kappa results of DT, SVM, RF, and MLP are the almost perfect agreement between dataset attributes.

Table 4.27. Recall, F-Measure and Kappa Coefficient of the (Covid Data 2)

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.0	1.0	0
Support Vector Machine (SVM)	0.0	1.0	0
Random Forest (RF)	0.1800	0.1127	0.0278
Naïve Bayes (NB)	0.1760	0.2568	0.2145
Multilayer Perceptron (MLP) Neural	0.0	1.0	0
K-Nearest Neighbor(KNN)	0.2802	0.1546	0.0123

Figure 4.16 shows the confusion matrix to describe the performance of a classification model.

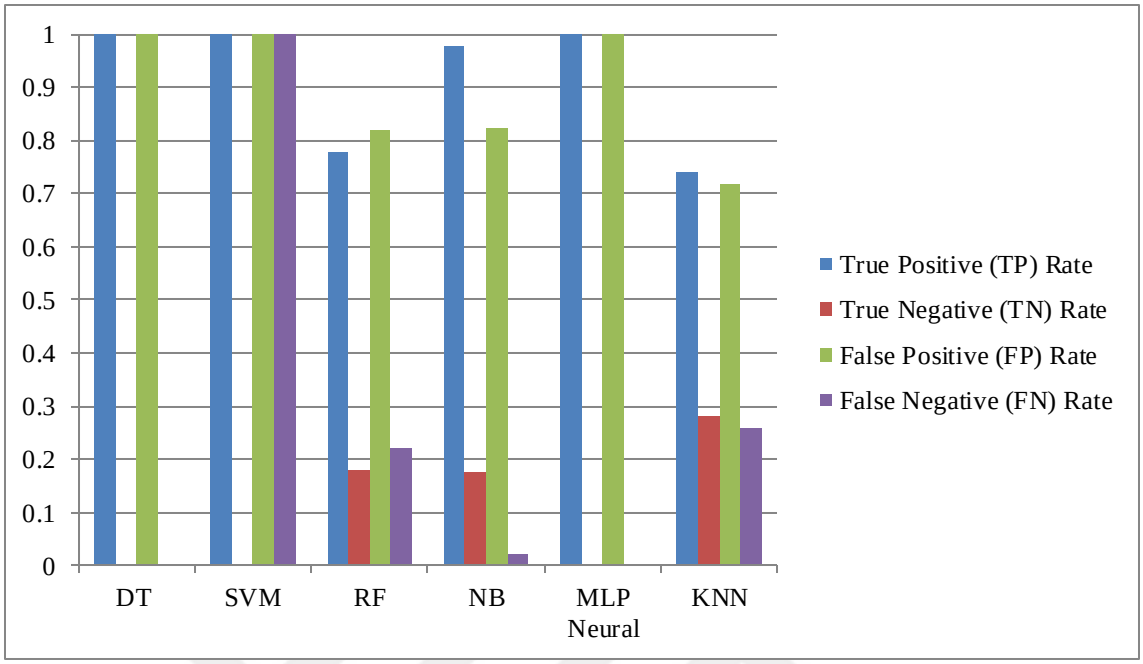


Figure 4.16. Shows the confusion matrix of the used dataset of 70 training and 30 of testing.

4.3.3. The case of 80 Training and 20 of Testing

The main accuracy details of 80 Training and 20 Testing case studies using preprocessing of machine learning on the used dataset KNN is the high accuracy as 96.3525 %, and the time to build the model is 194 ms. Table 4.28 shows the accuracy and time details with the confusion matrix evaluated parameters as False Positive Rate and False Negative Rate of the used dataset.

Table 4.28. The results of machine learning for (Covid Data 2) data analysis of 80 training and 20 of testing.

Item	Method Name	Accuracy	Confusion Matrix		Time
			False Positive Rate	**False Negative Rate**	
1	Decision Tree (DT)	91.9925 %	0.9845	0.2448	110312 ms
2	Support Vector Machine (SVM)	92.6998 %	0.00253	0.09672	1125 ms
3	Random Forest (RF)	92.1925 %	0.0080	0.0842	224431 ms
4	Naïve Bayes (NB)	93.51 %	0.4682	0.0293	625 ms
5	Multilayer Perceptron (MLP) Neural	91.89 %	1.0	0.0	1814378 ms
6	K-Nearest Neighbor(KNN)	96.3525 %	0.2672	0.0161	194 ms

Table 4.29 presents the results of correctly classified with incorrectly classified instances of the used data mining (preprocessing) with the highly accurate machine learning algorithms (KNN and NB) on the testing dataset used ((Covid Data 2) Dataset).

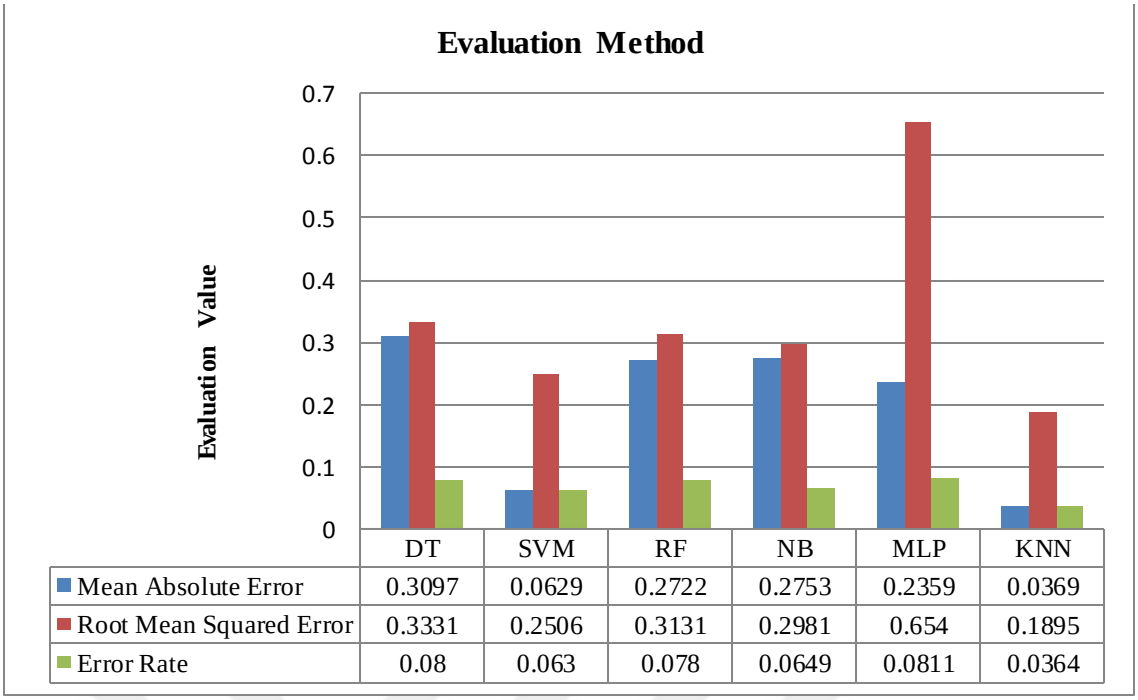
Table 4.29. Correctly / Incorrectly Classified Testing Instances of the data preprocessing of 80 training and 20 of testing.

Machine learning algorithm	Correctly Classified	Incorrectly Classified
Decision Tree (DT)	36797 = 91.9925 %	3203 = 8.0075 %
Support Vector Machine (SVM)	37079 = 92.6998 %	2520 = 6.3002 %
Random Forest (RF)	36877 = 92.1925 %	3123 = 7.8075 %
Naïve Bayes (NB)	37404 = 93.51 %	2596 = 6.49 %
Multilayer Perceptron (MLP) Neural	36756 = 91.89 %	3244 = 8.11 %
K-Nearest Neighbor(KNN)	38541 = 96.3525 %	1459 = 3.6475 %

Furthermore, the evaluation criteria used in the used system as MAE, RMSE, and Error Rate shown in Table 4.30 and Figure 4.17 show the prediction of the evaluation criteria of the used algorithms, the KNN almost lower the MAE value, so it is the better compared with others. KNN results of the RMSE statistic the lower and better compared with other algorithms. KN algorithm results of error rate is 0.0364, so it is better for the compared algorithms.

Table 4.30. MAE and RMSE for the (Covid Data 2) machine learning of 80 training and 20 of testing.

	Prediction					
Evaluation Criteria	**DT**	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Mean Absolute Error	0.3097	0.0629	0.2722	0.2753	0.2359	0.0369
Root Mean Squared Error	0.3331	0.2506	0.3131	0.2981	0.6540	0.1895
Error Rate	0.0800	0.0630	0.0780	0.0649	0.0811	0.0364



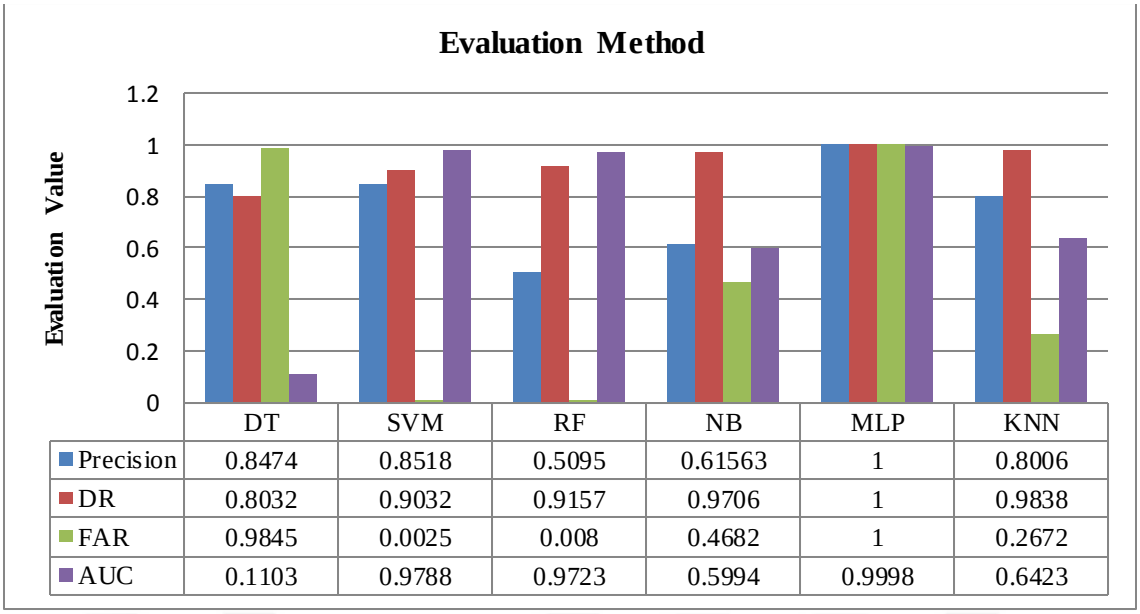
	DT	SVM	RF	NB	MLP	KNN
Mean Absolute Error	0.3097	0.0629	0.2722	0.2753	0.2359	0.0369
Root Mean Squared Error	0.3331	0.2506	0.3131	0.2981	0.654	0.1895
Error Rate	0.08	0.063	0.078	0.0649	0.0811	0.0364

Figure 4.17. The main evaluation Parameters MAE RMSE for the (Covid Data 2) dataset analysis of 80 training and 20 testing.

In addition Table, the used system includes various evaluation classifiers, as outlined in Table 4.31 and Figure 4.18. MLP precision as 1 can be seen as a measure of high quality to return more relevant results than irrelevant ones. MLP is 1 as higher, so it is better to distinguish between positioning coordination classes. MLP is the best of the whole sample which detected correctly. MLP algorithm is best because it indicates fewer false alarms of the used parameters.

Table 4.31. Evaluation (Covid Data 2) of the machine learning for the used data analysis of 80 training and 20 testing

Evaluation Parameters	Machine Learning Algorithms					
	DT	**SVM**	**RF**	**NB**	**MLP**	**KNN**
Precision	0.8474	0.8518	0.5095	0.61563	1.0	0.8006
Detection Rate (DR)	0.8032	0.9032	0.9157	0.9706	1	0.9838
False Alert Rate (FAR)	0.9845	0.0025	0.0080	0.4682	1	0.2672
Area Under Curve (AUC)	0.1103	0.9788	0.9723	0.5994	0.9998	0.6423
True Positive (TP) Rate	0.9997	0.9032	0.9157	0.9706	1.0	0.9838
True Negative (TN) Rate	0.0154	0.9974	0.9919	0.5317	0.0	0.7327



	DT	SVM	RF	NB	MLP	KNN
Precision	0.8474	0.8518	0.5095	0.61563	1	0.8006
DR	0.8032	0.9032	0.9157	0.9706	1	0.9838
FAR	0.9845	0.0025	0.008	0.4682	1	0.2672
AUC	0.1103	0.9788	0.9723	0.5994	0.9998	0.6423

Figure 4.18 Precision, DR, FAR and AUS of the used machine learning algorithms of 80 training and 20 of testing.

Table 4.32 showed RF algorithm recall is 0.9919 which means that the model has successfully identified all relevant instances of the positive class, and there are a few false negatives. F-Measure showed the DT is a low model to make a correct prediction across the entire dataset as it is not performing well in its ability to capture all the relevant instances. The Kappa results of SVM are the almost perfect agreement between dataset attributes.

Table 4.32. Recall, F-Measure and Kappa Coefficient of the (Covid Data 2) of 80 training and 20 of testing.

Machine learning algorithm	Recall	F-Measure	Kappa Coefficient
Decision Tree (DT)	0.0154	0.0302	0.0275
Support Vector Machine (SVM)	0.9904	0.9189	0.8679
Random Forest (RF)	0.9919	0.6732	0.6341
Naïve Bayes (NB)	0.5317	0.5706	0.5357
Multilayer Perceptron (MLP) Neural	0.0	1.0	0.0
K-Nearest Neighbor(KNN)	0.7327	0.7651	0.7454

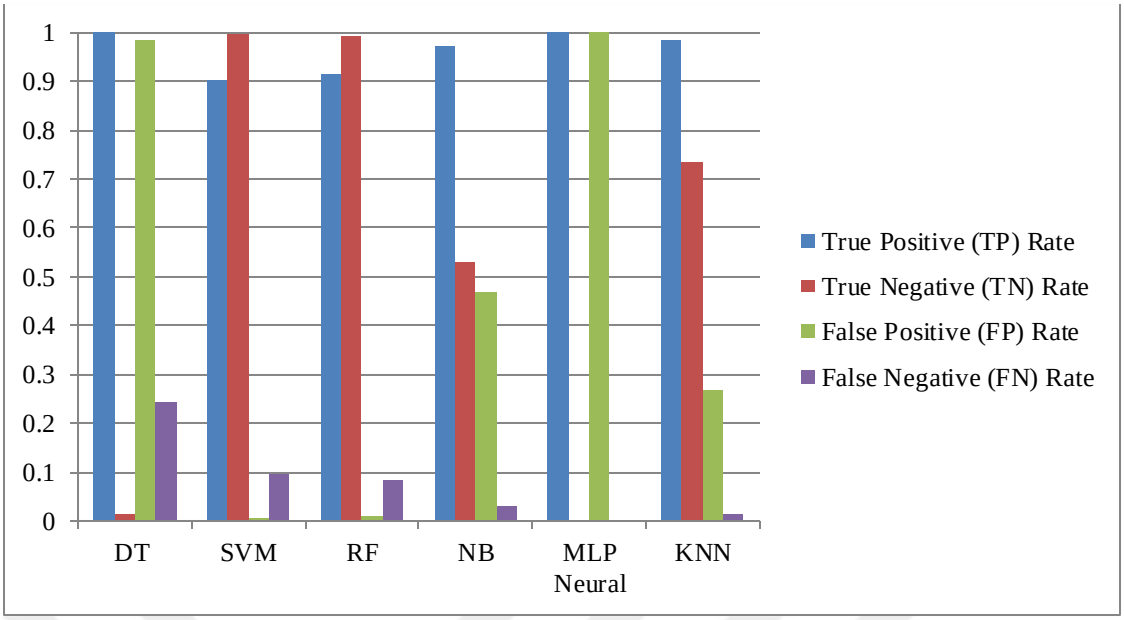


Figure 4.19. shows the confusion matrix of the used dataset of 80 training and 20 of testing.

4.4. Discussions of the Results

The choice between different ratios of training and testing data in machine learning depends on several factors, and there is no one-size-fits-all answer. The decision is often influenced by the size of the used dataset, the complexity of the used model, and the specific goals.

- **70/30 Split (Training/Test):**
    - Advantages: More data for training, which can be beneficial if the used dataset, is relatively small. Faster training times. It is the best-split ratio depending on the specifics of COVID data and problem. Multilayer Perceptron (MLP) Neural accuracy is 99.5300 %, and the required time to build mode is 585496 ms of the first COVID dataset. With 70% of the data allocated to training, the model has a sufficiently large dataset to learn the underlying patterns and relationships within the data. Allocating 30% of the data for testing ensures that there is a separate set of data the model has not seen during training. This allows for a reliable evaluation of the model's performance on unseen data.
    - Considerations: With less data for testing, the evaluation might be less reliable, and the model may not generalize well to new, unseen data.

- **80/20 Split (Training/Test):**
  - Advantages: A slightly larger portion for testing may provide a more reliable evaluation. Still, a good amount of data for training.
  - Considerations: More training data can be beneficial, especially for complex models.
  - The used system results showed that the second dataset, the K-Nearest Neighbor (KNN), is 96.3525 %, and the required time to build mode is 194 ms of 80 Training and 20 of Testing.
- **60/40 Split (Training/Test):**
  - Advantages: More data for testing, which may lead to a more reliable evaluation of your model's performance
  - Considerations: Fewer data for training, which might be a concern if the model is complex or if you have a limited dataset.

The results of the first dataset Covid Data 1 is better than the second dataset Covid Data 2 due to the following reasons:

- Better Representativity: A larger dataset is more likely to be representative of the underlying population or distribution. This helps in capturing a more comprehensive view of the patterns and variability present in the data.
- Improved Generalization: Models trained on larger datasets often generalize better to unseen data. They are less prone to overfitting since they have been exposed to a more diverse set of examples, which helps in learning the underlying patterns rather than memorizing the training data.
- More Informative Features: With more data, models can extract more informative features, leading to better performance. Smaller datasets might not provide enough examples to reveal subtle relationships within the data.
- Reduced Variability: Larger datasets help in reducing the variability in performance metrics. In smaller datasets, the choice of a specific subset for training or testing can have a more pronounced effect on results.
- Increased Confidence in Results: Results obtained from a larger dataset are often more statistically significant. This increased confidence is crucial, especially in scientific or critical applications.

- Stability in Model Training: Training machine learning models on larger datasets often leads to more stable convergence. Models are less likely to get stuck in local optima and are more likely to find a global optimum.

Table 4.33 presents a comparison of the system with other relevant works. The case studies demonstrate that the used system yields superior accuracy, particularly in the context of MLP for the first COVID dataset and DT for the second COVID dataset.

Table 4.33. The results of COVID-19 Dataset analysis with the compared systems.

Ref. Auth.	Year	AI Technique	Accuracy
Alakus et al.	2020	CNN-LSTM	92.30 %
Díaz-Pernas et al.	2021	Logistic Regression with CR classifier	84.21 %
Liu et al.	2021	Logistic Regression	94 %
		XGBoost	92 %
Ong et al.	2022	Neural Network	97.32%
		Random Forest	92 %
Moulaei et al.	2022	Random Forest	95.03 %
Used-system First COVID dataset		**Multilayer Perceptron (MLP) Neural**	**99.5300 %**
Used-system Second COVID dataset		**K-Nearest Neighbor(KNN)**	**96.3525 %**

The coronavirus disease (COVID‑19), hospitalized patients are always at risk of death. Machine learning (ML) algorithms can be used as a potential solution for predicting mortality in COVID‑19 hospitalized patients. The used prediction COVID-19 system used machine learning to create a fast detection system like a real-time detection warning system to identify suspected people. We used pre-trained algorithms to classify the dataset contents. We cleaned the dataset with pre-processing methods by removing duplicates and normalized the attributes to increase the model's accuracy. Besides pre-processing the data, the dataset was loaded from the folder as input to the system model. The data has been tested on COVID-19 patients utilizing ten independent variables. The classification scheme utilized in the present study involves two distinct categories: class 0, which denotes the absence of COVID-19 in patients, and class 1, which signifies the

presence of COVID-19 in patients. Next, we were able to improve the model's accuracy by pre-processing the data set with the used Data Mining Pre-processing Methods (Normalization, Attribute-feature-selection, The Missing-values, Nominal to Binary, Nominal to Numeric, and Numeric to Nominal). JAVA library machine learning is used to predict COVID-19.

5. CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

In light of the outbreak of the Coronavirus, which caused the suffering governments worldwide to control this disease and limit its spread, The World Health Organization emphasizes the criticality of safeguarding oneself against COVID-19 infection. Given the preventive measures implemented to mitigate the transmission of the Coronavirus, it is noteworthy that COVID-19 is an infectious ailment that has significantly strained various facets of society, including the economy, due to the emergence of multiple variants. Timely detection of the virus is critical in mitigating its transmission, given its global threat to human life.

In this study, a prediction COVID-19 system uses machine learning to create a fast detection system to identify suspected people, like a real-time detection warning system. This research used pre-trained algorithms to classify the dataset contents. We cleaned the dataset with pre-processing methods by removing duplicates and normalized the attributes to increase the model's accuracy. Besides pre-processing, the dataset was loaded from the folder as input to the system model. The data has been tested on COVID-19 patients utilizing ten independent variables. The classification scheme utilized in the present study involves two distinct categories: class 0, which denotes the absence of COVID-19 in patients, and class 1, which signifies the presence of COVID-19 in patients. Next, this study was able to improve the model's accuracy by pre-processing the data set with the used Data Mining Pre-processing Methods (Normalization, B- Attribute-feature-selection, Missing-values, Nominal to Binary, Nominal to Numeric, and Numeric to Nominal). In this project, JAVA library machine learning is used to predict COVID-19.

This research found that the model worked well when applied to the used dataset. This study is based on two steps: first, we uploaded a dataset to train the model. Hence, the study tests the model on those cases to work directly after making a trained classifier so it can directly discover with automatic COVID-19 prediction state of a patient suspected or not.

The used system results showed the high accuracy of Multilayer Perceptron (MLP) Neural as 99.5300 % of the first COVID-19 dataset, and and time to build the model is 1639469 ms. Besides, the better results of the second COVID-19 dataset are K-Nearest Neighbor(KNN) algorithm accuracy is 96.3525 %, better time to build the

model, and early predict the state of patients is 194 ms.

5.2. Recommendations

Several factors can be taken into account for the future expansion of current research by utilizing the following propositions:

- This study aims to detect COVID-19 from chest X-ray images by applying transfer learning techniques using ResNet50, ResNet101, DenseNet121, DenseNet169, and InceptionV3 models. The models underwent training and validation using the most extensive publicly accessible database of COVID-19 CXR images.
- This study aims to explore the detection of severe and mildly infected patients with COVID-19 through the lens of the Resource-Based View of Laboratory Management (RBV) and age data collected at the time of hospital admission. The objective is to
- provide a concise motivation for this approach.

6. REFERENCES

Akhtar, A., Akhtar, S., Bakhtawar, B., Kashif, A. A., Aziz, N., & Javeid, M. S. (2021). COVID-19 detection from CBC using machine learning techniques. International Journal of Technology, Innovation and Management (IJTIM), 1(2), 65-78.

Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-Qays, Z. T., Zaidan, A. A., Zaidan, B. B., ... & Madhloom, H. T. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of medical systems, 44, 1-11.

Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals, 140, 110120.

Ali, R., Al-Jumaili, S., Duru, A. D., Uçan, O. N., Boyaci, A., & Duru, D. G. (2022, October). Classification of Brain Tumors using MRI images based on Convolutional Neural Network and Supervised Machine Learning Algorithms. In 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 822-827). IEEE.

Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., & Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. Multimedia Tools and Applications, 80, 11943-11957.

Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., & Alhyari, S. (2020). COVID-19 prediction and detection using deep learning. International Journal of Computer Information Systems and Industrial Management Applications, 12(June), 168-181.

Abdulkareem, A. B., Sani, N. S., Sahran, S., Alyessari, Z. A. A., Adam, A., Rahman, A. H. A., & Abdulkarem, A. B. (2021). Predicting COVID-19 Based on Environmental FactorsWith Machine Learning. Intelligent Automation & Soft Computing, 28(2).

Altini, N., Brunetti, A., Mazzoleni, S., Moncelli, F., Zagaria, I., Prencipe, B., ... & Bevilacqua, V. (2021). Predictive machine learning models and survival analysis for COVID-19 prognosis based on hematochemical parameters. Sensors, 21(24), 8503.

Arista, A. (2022). Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19. Sinkron: jurnal dan penelitian teknik informatika, 7(1), 59-65.

Arshed, M. A., Qureshi, W., Khan, M. U. G., & Jabbar, M. A. (2021, November). Symptoms based COVID-19 disease diagnosis using machine learning approach. In 2021 International Conference on Innovative Computing (ICIC) (pp. 1-7). IEEE.

Alves, M. A., Castro, G. Z., Oliveira, B. A. S., Ferreira, L. A., Ramírez, J. A., Silva, R., & Guimarães, F. G. (2021). Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Computers in Biology and Medicine, 132, 104335.

Agrawal, T., & Choudhary, P. (2021). FocusCovid: automated COVID-19 detection using deep learning with chest X-ray images. Evolving Systems, 1-15.

Andreu-Perez, J., Perez-Espinosa, H., Timonet, E., Kiani, M., Girón-Pérez, M. I.,

Benitez-Trinidad, A. B., ... & Rivas, F. (2021). A generic deep learning based cough analysis system from clinically validated samples for point-of-need covid-19 test and severity levels. IEEE Transactions on Services Computing, 15(3), 1220-1232.

Bhattacharya, S., Maddikunta, P. K. R., Pham, Q. V., Gadekallu, T. R., Chowdhary, C. L., Alazab, M., & Piran, M. J. (2021). Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey. Sustainable cities and society, 65, 102589.

Bai, L., Yang, D., Wang, X., Tong, L., Zhu, X., Zhong, N., ... & Tan, F. (2020). Chinese experts' consensus on the Internet of Things-aided diagnosis and treatment of coronavirus disease 2019 (COVID-19). Clinical eHealth, 3, 7-15.

Babukarthik, R. G., Adiga, V. A. K., Sambasivam, G., Chandramohan, D., & Amudhavel, J. J. I. A. (2020). Prediction of COVID-19 using genetic deep learning convolutional neural network (GDCNN). Ieee Access, 8, 177647-177666.

Booth, A. L., Abels, E., & McCaffrey, P. (2021). Development of a prognostic model for mortality in COVID-19 infection using machine learning. Modern Pathology, 34(3), 522-531.

Chuenyindee, T., Ong, A. K. S., Prasetyo, Y. T., Persada, S. F., Nadlifatin, R., & Sittiwatethanasiri, T. (2022). Factors affecting the perceived usability of the COVID-19 contact-tracing application "Thai chana" during the early COVID-19 omicron period. International Journal of Environmental Research and Public Health, 19(7), 4383.

Crespo, F., Crespo, A., Sierra-Martínez, L. M., Peluffo-Ordóñez, D. H., & Morocho-Cayamcela, M. E. (2022). A Computer Vision Model to Identify the Incorrect Use of Face Masks for COVID-19 Awareness. Applied Sciences, 12(14), 6924.

Chiroma, H., Ezugwu, A. E., Jauro, F., Al-Garadi, M. A., Abdullahi, I. N., & Shuib, L. (2020). Early survey with bibliometric analysis on machine learning approaches in controlling COVID-19 outbreaks. PeerJ Computer Science, 6, e313.

Díaz-Pernas, F. J., Martínez-Zarzuela, M., Antón-Rodríguez, M., & González-Ortega, D. (2021, February). A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. In Healthcare (Vol. 9, No. 2, p. 153). MDPI.

Dellière, S., Dudoignon, E., Voicu, S., Collet, M., Fodil, S., Plaud, B., ... & Alanio, A. (2022). Combination of mycological criteria: a better surrogate to identify COVID-19-associated pulmonary aspergillosis patients and evaluate prognosis?. Journal of Clinical Microbiology, 60(3), e02169-21.

Ezat, W. A., Dessouky, M. M., & Ismail, N. A. (2020). Multi-class image classification using deep learning algorithm. In Journal of Physics: Conference Series (Vol. 1447, No. 1, p. 012021). IOP Publishing.

Estiri, H., Strasser, Z. H., & Murphy, S. N. (2021). Individualized prediction of COVID-19 adverse outcomes with MLHO. Scientific reports, 11(1), 5322.

Frater, J. L., Zini, G., d'Onofrio, G., & Rogers, H. J. (2020). COVID-19 and the clinical hematology laboratory. International journal of laboratory hematology, 42, 11-18.

Farid, A. A., Selim, G. I., & Khater, H. A. A. (2020). A novel approach of CT images feature analysis and prediction to screen for corona virus disease (COVID-19).

Fuhrman, J. D., Gorre, N., Hu, Q., Li, H., El Naqa, I., & Giger, M. L. (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. Medical Physics, 49(1), 1-14.

Fayyoumi, E., Idwan, S., & AboShindi, H. (2020). Machine learning and statistical modelling for prediction of novel COVID-19 patients case study: Jordan. International Journal of Advanced Computer Science and Applications, 11(5).

Guo, T., Wang, J., Liang, X., Zheng, Y., Zhu, Y., Zhou, K., ... & Shen, B. (2022). Pulmonary and Renal Long COVID at Two-year Revisit.

Hemdan, E. E. D., Shouman, M. A., & Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055.

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., & Xie, P. (2020). Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. medrxiv, 2020-04.

Islam, M. M., Karray, F., Alhajj, R., & Zeng, J. (2021). A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19). Ieee Access, 9, 30551-30572.

Jamshidi, M. B., Lalbakhsh, A., Talla, J., Peroutka, Z., Roshani, S., Matousek, V., ... & Lotfi, S. (2021). Deep learning techniques and covid-19 drug discovery: Fundamentals, state-of-the-art and future directions. Emerging Technologies During the Era of COVID-19 Pandemic, 9-31.

Jewell, N. P., Lewnard, J. A., & Jewell, B. L. (2020). Predictive mathematical models of the COVID-19 pandemic: underlying principles and value of projections. Jama, 323(19), 1893-1894.

Jamshidi, E., Asgary, A., Tavakoli, N., Zali, A., Setareh, S., Esmaily, H., ... & Mansouri, N. (2022). Using machine learning to predict mortality for COVID-19 patients on day 0 in the ICU. Frontiers in Digital Health, 3, 210.

Kumar, A., Alsadoon, A., Prasad, P. W. C., Abdullah, S., Rashid, T. A., Pham, D. T. H., & Nguyen, T. Q. V. (2022). Generative adversarial network (GAN) and enhanced root mean square error (ERMSE): deep learning for stock price movement prediction. Multimedia Tools and Applications, 1-19.

Khan, I. U., Aslam, N., Aljabri, M., Aljameel, S. S., Kamaleldin, M. M. A., Alshamrani, F. M., & Chrouf, S. M. B. (2021). Computational intelligence-based model for mortality rate prediction in COVID-19 patients. International journal of environmental research and public health, 18(12), 6429.

Kumari, A., & Sood, M. (2021). Implementation of SimpleRNN and LSTMs based prediction model for coronavirus disease (Covid-19). In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012015). IOP Publishing.

Kang, J., Chen, T., Luo, H., Luo, Y., Du, G., & Jiming-Yang, M. (2021). Machine learning predictive model for severe COVID-19. Infection, Genetics and Evolution, 90, 104737.

Karthikeyan, A., Garg, A., Vinod, P. K., & Priyakumar, U. D. (2021). Machine learning based clinical decision support system for early COVID-19 mortality prediction. Frontiers in public health, 9, 626697.

Liu, J. E., & An, F. P. (2020). Image classification algorithm based on deep learning-kernel function. Scientific programming, 2020, 1-14.

Lunagaria, M., Katkar, V., & Vaghela, K. (2022). Covid-19 and Pneumonia Infection Detection from Chest X-Ray Images using U-Net, EfficientNetB1, XGBoost and

Recursive Feature Elimination. International Journal of Advanced Computer Science and Applications, 13(9).

Le, D. N., Parvathy, V. S., Gupta, D., Khanna, A., Rodrigues, J. J., & Shankar, K. (2021). IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification. International journal of machine learning and cybernetics, 1-14.

Levin, A. T., Owusu-Boaitey, N., Pugh, S., Fosdick, B. K., Zwi, A. B., Malani, A., ... & Meyerowitz-Katz, G. (2022). Assessing the burden of COVID-19 in developing countries: systematic review, meta-analysis and public policy implications. BMJ Global Health, 7(5), e008477.

MacGowan, S. A., & Barton, G. J. (2020). Missense variants in ACE2 are predicted to encourage and inhibit interaction with SARS-CoV-2 Spike and contribute to genetic risk in COVID-19. BioRxiv, 2020-05.

Ma, J., Deng, Y., Zhang, M., & Yu, J. (2022). The role of multi-omics in the diagnosis of COVID-19 and the prediction of new therapeutic targets. Virulence, 13(1), 1101-1110.

Marques, G., Agarwal, D., & de la Torre Díez, I. (2020). Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. Applied soft computing, 96, 106691

Matsuda, T., Tianlong, W. A. N. G., & Mehmet, D. İ. K. (2023). Decentralized Machine Learning Approach on ICU Admission Prediction for Enhanced Patient Care Using COVID-19 Data. Proceedings of International Mathematical Sciences, 5(2), 91-102.

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. BMC medical informatics and decision making, 22(1), 1-12.

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. BMC medical informatics and decision making, 22(1), 1-12.

Morand, A., Fabre, A., Minodier, P., Boutin, A., Vanel, N., Bosdure, E., & Fournier, P. E. (2020). COVID-19 virus and children: What do we know?. Archives de pediatrie, 27(3), 117-118.

Meo, S. A., Meo, A. S., & Klonoff, D. C. (2023). Omicron new variant BA. 2.86 (Pirola): Epidemiological, biological, and clinical characteristics-a global data-based analysis. European Review for Medical & Pharmacological Sciences, 27(19).

Mayr, M., Gutmann, C., Takov, K., Burnap, S., Singh, B., Theofilatos, K., ... & Shankar-Hari, M. (2020). SARS-CoV-2 RNAemia and proteomic biomarker trajectory inform prognostication in COVID-19 patients admitted to intensive care.

Nagi, A. T., Awan, M. J., Mohammed, M. A., Mahmoud, A., Majumdar, A., & Thinnukool, O. (2022). Performance Analysis for COVID-19 Diagnosis using custom and state-of-the-art deep learning models. Applied Sciences, 12(13), 6364.

Ong, A. K. S., Chuenyindee, T., Prasetyo, Y. T., Nadlifatin, R., Persada, S. F., Gumasing, M. J. J., ... & Sittiwatethanasiri, T. (2022). Utilization of random forest and deep learning neural network for predicting factors affecting perceived usability of a COVID-19 contact tracing mobile application in Thailand "Thaichana". International Journal of Environmental Research and Public Health, 19(10), 6111.

Ong, A. K. S., Chuenyindee, T., Prasetyo, Y. T., Nadlifatin, R., Persada, S. F., Gumasing, M. J. J., ... & Sittiwatethanasiri, T. (2022). Utilization of random

forest and deep learning neural network for predicting factors affecting perceived usability of a COVID-19 contact tracing mobile application in Thailand "Thaichana". International Journal of Environmental Research and Public Health, 19(10), 6111.

Omran, N. F., Abd-el Ghany, S. F., Saleh, H., Ali, A. A., Gumaei, A., & Al-Rakhami, M. (2021). Applying deep learning methods on time-series data for forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. Complexity, 2021, 1-13

Ong, A. K. S., Prasetyo, Y. T., Yuduang, N., Nadlifatin, R., Persada, S. F., Robas, K. P. E., ... & Buaphiban, T. (2022). Utilization of random forest classifier and artificial neural network for predicting factors influencing the perceived usability of COVID-19 contact tracing "Morchana" in Thailand. International Journal of Environmental Research and Public Health, 19(13), 7979.

Ong, A. K. S., Prasetyo, Y. T., Yuduang, N., Nadlifatin, R., Persada, S. F., Robas, K. P. E., and Buaphiban, T. (2022). Utilization of random forest classifier and artificial neural network for predicting factors influencing the perceived usability of COVID-19 contact tracing "Morchana" in Thailand. International Journal of Environmental Research and Public Health, 19(13), 7979.

Podder, P., Bharati, S., Mondal, M. R. H., & Kose, U. (2021). Application of machine learning for the diagnosis of COVID-19. In Data science for COVID-19 (pp. 175-194). Academic Press.

Rahman, M. M., Islam, M. M., Manik, M. M. H., Islam, M. R., & Al-Rakhami, M. S. (2021). Machine learning approaches for tackling novel coronavirus (COVID-19) pandemic. SN Computer Science, 2, 1-10.

Rahman, T., Al-Ishaq, F. A., Al-Mohannadi, F. S., Mubarak, R. S., Al-Hitmi, M. H., Islam, K. R., ... & Chowdhury, M. E. (2021). Mortality prediction utilizing blood biomarkers to predict the severity of COVID-19 using machine learning technique. Diagnostics, 11(9), 1582.

Rasheed, J., Hameed, A. A., Djeddi, C., Jamil, A., & Al-Turjman, F. (2021). A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisciplinary Sciences: Computational Life Sciences, 13, 103-117.

Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Rahman, M. A., Wang, Q., ... & Zhao, X. (2020). Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. Journal of X-ray Science and Technology, 28(5), 821-839.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.

Russell, B., Moss, C., George, G., Santaolalla, A., Cope, A., Papa, S., & Van Hemelrijck, M. (2020). Associations between immune-suppressive and stimulating drugs and novel COVID-19—a systematic review of current evidence. ecancermedicalscience, 14.

Sepehrinezhad, A., Shahbazi, A., & Negah, S. S. (2020). COVID-19 virus may have neuroinvasive potential and cause neurological complications: a perspective review. Journal of neurovirology, 26, 324-329.

Swapnarekha, H., Behera, H. S., Nayak, J., & Naik, B. (2020). Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. Chaos, Solitons & Fractals, 138, 109947.

Sarwinda, D., Paradisa, R. H., Bustamam, A., & Anggia, P. (2021). Deep learning in image classification using residual network (ResNet) variants for detection of

colorectal cancer. Procedia Computer Science, 179, 423-431.
Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. SN Computer Science, 1, 1-15.
Solayman, S., Aumi, S. A., Mery, C. S., Mubassir, M., & Khan, R. (2023). Automatic COVID-19 prediction using explainable machine learning techniques. International Journal of Cognitive Computing in Engineering, 4, 36-46.
Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., ... & Shi, Y. (2020). Lung infection quantification of COVID-19 in CT images with deep learning. arXiv preprint arXiv:2003.04655.
Toquero, C. M. (2020). Challenges and opportunities for higher education amid the COVID-19 pandemic: The Philippine context. Pedagogical Research, 5(4).
Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using weka. Algorithms, 14(7), 201.
Wiwoho, J., Firdaus, S. U., Hidayat, M. A., Nanda, F., & Arkananta, S. (2023, December). Precautionary Health Protection in the Conduct of the 2024 Simultaneous Elections to the Spread of New COVID-19 Mutations in Indonesia. In International Conference for Democracy and National Resilience (ICDNR 2023) (pp. 242-258). Atlantis Press
Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters, 141, 61-67.
Xiao, A., Wu, F., Bushman, M., Zhang, J., Imakaev, M., Chai, P. R., ... & Alm, E. J. (2022). Metrics to relate COVID-19 wastewater data to clinical testing dynamics. Water research, 212, 118070.
Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., ... & Wang, G. (2020). Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Frontiers in cell and developmental biology, 683.
Zimmerman, A., & Kalra, D. (2020). Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications. Reviews in Cardiovascular Medicine, 21(3), 345-352.

CURRICULUM VITAE

Student Information	
Name/Surname:	AHMED JADDOA ENAD AL-MAMOORI
Nationality:	IRAQ
Orcid No:	0009-0005-6534-7560

School Information	
Undergraduate Study	
University	BABYLON UNIVERSITY
Faculty	COLLEGE OF SCIENCE
Department	DEPARTMENT OF COMPUTER SCIENCE
Graduation Year	2004/2005
Graduate Study	
University	KIRŞEHİR AHİ EVRAN UNIVERSITY
Institute	INSTITUTE OF NATURAL AND APPLIED SCIENCES
Department	DEPARTMENT OF ADVANCED TECHNOLOGIES
Graduation Year	2024

Articles and Papers Produced from the Thesis
AHMED J. E. AL-MAMOORI, Mustafa AKSU (2024) Machine Learning-Based COVID-19 Diagnosis and Prediction System: Performance Analysis of Various Learning Algorithms and Classification of Related Diseases . *Journal (Bulletin of Electrical Engineering and Informatics BEEI)*