



Paper Type: Original Article

# Prediction of COVID-19 Patients using Machine Learning Algorithms

Ahmed M. Ali <sup>1,\*</sup>  and Shimaa A. Esmail <sup>2</sup> 

<sup>1</sup> Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt; [aabdelmounem@zu.edu.eg](mailto:aabdelmounem@zu.edu.eg).

<sup>2</sup> Faculty of Information Systems and Computer Science, October 6th University, Cairo, Egypt; [shimaaanwar029@gmail.com](mailto:shimaaanwar029@gmail.com).

Received: 05 Mar 2024

Revised: 27 May 2024

Accepted: 26 Jun 2024

Published: 28 Jun 2024

## Abstract

Coronavirus disease (COVID-19), also known as severe acute respiratory syndrome (SARS-COV2), has caused widespread concern for public health worldwide. Based on its rapid spread among those exposed to the wet animal market in Wuhan, China, the city was identified as its origin. The symptoms, reactions, and rates of recovery observed in coronavirus patients around the world have varied. The number of sufferers continues to grow at an exponential rate, and some countries are currently dealing with the third wave. Since the most effective treatment for this disease has yet to be established, early discovery of probable COVID-19 patients can help isolate them socially, slowing the spread and flattening the curve. This study examines current research on coronavirus disease and its impact across age groups. We evaluate the effectiveness of Decision Tree (DT), and Logistic Regression (LR) in detecting COVID-19 in patients based on symptoms. A dataset from a public repository was pre-processed before applying Machine Learning (ML) techniques to it. The results show that all ML algorithms are effective in identifying COVID-19 in potential patients. DT classifiers have the highest accuracy of 98.70%, while SVM, KNN, and LR algorithms achieve 93.60%, 93.50%, and 92.80%, respectively.

**Keywords:** Coronavirus, Machine Learning, Dataset, Decision Tree, Disease Analysis.

## 1 | Introduction

COVID-19 is a terrible and damaging condition that affects the entire world. The first instances were discovered in Wuhan, China, in the last week of December 2019, and then spread rapidly over the world. Most persons infected with this disease receive mild treatment to alleviate respiratory symptoms and improve without requiring a specialized approach. The COVID-19 virus is mostly derived from saliva droplets and nasal discharge, and it travels from person to person [1]. Even after more than a year of generating this condition, the most effective remedy has yet to be successfully introduced. However, several ongoing clinical trials are evaluating the best feasible curative approaches, and vaccines are being produced. Since the vaccinations have not yet matured, other strategies, such as early disease detection, may be more successful in stopping the disease's spread. Machine Learning (ML) techniques [2] and deep learning algorithms [3] are examples of artificial intelligence (AI) methodologies that are crucial for disease prediction, such as outbreak prediction. [4, 5], identification of high-risk patients [6], analysis of COVID-19 with clinical features [7], study



Corresponding Author: [aabdelmounem@zu.edu.eg](mailto:aabdelmounem@zu.edu.eg)



<https://doi.org/10.61356/j.scin.2024.1314>



Licensee SciNexuses. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

of the spread likelihood [8], strategies for fighting this illness [9], and patient mortality risk. The pandemic condition can be improved, as evidenced by the recent advances in AI and data mining approaches to medical problems [10]. A method that demonstrated great accuracy in estimating a patient's chance of survival based on physiological parameters, symptoms, and demographics was created by Mohamad et al. [11]. By creating an ML model, Allae et al. [12] estimated the threshold of COVID-19 cases in a region of Ardabili and associates. [13] proposed that combining soft computing and ML models can aid in the prediction of epidemics. Elderly individuals, particularly those with coexisting conditions such as cancer, diabetes, cardiovascular disease, and persistent lung infections, are observed to be the main victims of this serious illness. People of all ages should take extra precautions because there are indications that the characteristics and demographics of patients dying in China, or another region of the world could not be the same as those in other regions.

ML is playing an important role in several domains that were previously assumed to be solely human-centric. They are used to combine diverse biomedical data sources to develop predictive models based on symptom data collected from clinical test results. This study seeks to examine how reliably a COVID-19-positive patient may be recognized based on their symptoms. Our research aims to evaluate the performance of supervised ML algorithms (DT, KNN, NB, LR, and SVM) by analyzing classification reports to identify the top-performing algorithm.

## 1.1 | Symptoms of Coronavirus

COVID-19 symptoms typically appear after around 5.2 days of incubation [14]. Whereas it has been found that the disease lasts between 6 and 41 days, with an average of roughly two weeks. The disease's containment span has been reported to be proportionate to the patient's age and immune strength. Figure 1 depicts the most typical symptoms reported in coronavirus patients. COVID-19 patients exhibit unique symptoms during the incubation period or shortly thereafter. The most common symptoms recorded of the condition are as follows [15].

### 1.1.1 | Most Common Symptoms

Common symptoms of coronavirus disease 2019 (COVID-19) typically include fever, dry cough, and fatigue, with severe cases often involving difficulty breathing (dyspnea). Many individuals, especially children and young adults, may have no symptoms (asymptomatic) when infected, while older individuals and those with underlying health conditions face a greater risk of experiencing severe illness, respiratory failure, or even death. The incubation period averages around 5 days, with severe symptoms typically manifesting approximately 8 days after the onset of symptoms, and critical illness or death occurring around 16 days post-infection [3-4]. Acute respiratory distress syndrome (ARDS) and intensive care unit (ICU) admission are more likely among COVID-19 patients aged over 60 years and those with significant pre-existing health conditions. Some cases of COVID-19 have also resulted in multiple organ failure. The susceptibility to SARS-CoV-2 infection spans across all age groups, with the median age of infection being approximately 50 years. However, the clinical outcomes vary significantly by age. Generally, older men aged over 60 years with underlying health conditions are at higher risk of developing severe respiratory illness requiring hospitalization or resulting in death. Conversely, most young people and children experience mild symptoms (non-pneumonia or mild pneumonia) or may even be asymptomatic [9].

### 1.1.2 | Other Less Common Symptoms

In some cases, the potential patients showed signs such as production of Sputum, Headache, Hemoptysis, Diarrhea, Dyspnea, and Lymphopenia.

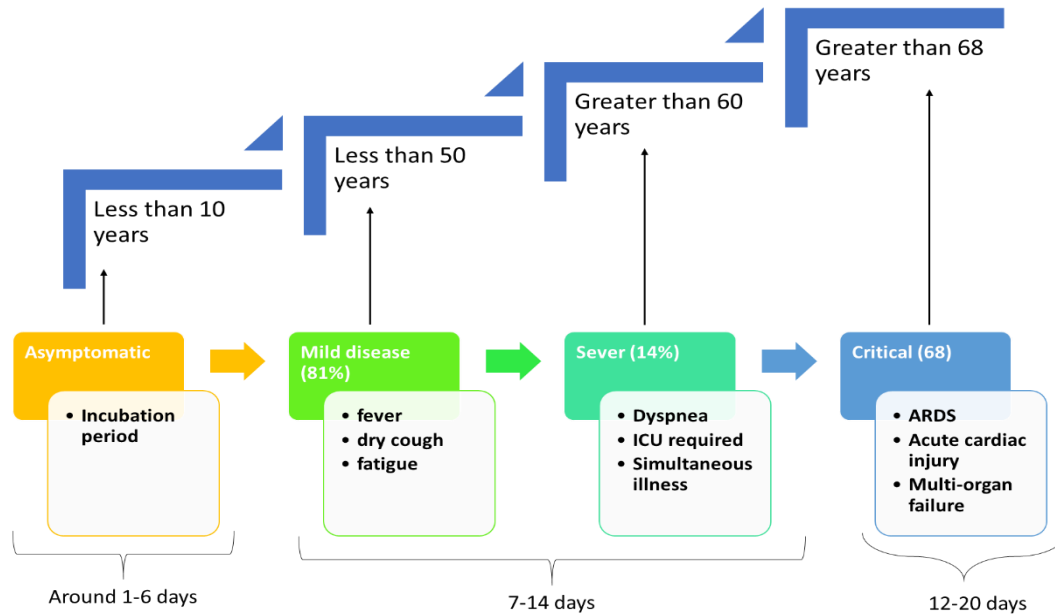


Figure 1. Symptoms of COVID-19.

## 2 | Literature Review

There have been numerous and ongoing investigations, inspections, and attempts to combat this serious virus. One typical tactic for controlling coronavirus toxicity is drug delivery [16]. Using standards from the World Health Organization (WHO), a total of 80 cases that tested positive for COVID-19 were examined and characterized at grade 3 hospitals in Jiangsu. Reverse transcription-polymerase chain reaction (RT-PCR) research on respiratory patients revealed the problem. With an average age of 46.1 years, 41 of the 80 patients were female [17]. 77 individuals were determined to be moderately ill, with three being critically ill. A total of 38 patients had a history of chronic illnesses. Fever symptoms were recorded by 63 patients, while 51 subjects coughed. The patient's lung pictures revealed abnormal shadows on 55 scans, while 25 showed no shade. There were no deaths recorded in the scenario, and 21 patients were discharged within eight days. Jiangsu had milder cases of liver disease and unusual lung activity than Wuhan [18]. According to clinical data from January 13th to February 19th, 2020, 28 COVID-19 patients were infected in Wuhan city 60.7% of infected patients are male, with an average age of 65 years [19]. Seven studies were reviewed for meta-analysis. The results showed that fever (91.3%), fatigue (51.0%), and dyspnea (30.4%) were the most common clinical symptoms [20]. The Chinese center released a substantial publication that included a report of 72,314 cases. Zunyouet al. summarized the key elements of this publication. Of them, 16,186 were suspected cases, 10,567 were diagnosed based on symptoms, 44,672 were confirmed cases, and 889 were asymptomatic. The age range of 87% of the patients was 30 to 79, indicating that older individuals are more susceptible to the coronavirus than younger ones [21].

Heshui Shi et al. conducted a study in the same league to describe the CT scans of 81 patients and found that the early diagnosis of COVID-19 disease can be aided by a combination analysis of imaging features and clinical data [22]. Only a few candidates have shown promise in vitro studies, and few have progressed to a randomized creature therefore, the use of human preliminary measures to combat COVID-19 contamination may be limited [23]. Ying et al.'s analysis [24]. The reproduction number ( $R_0$ ) indicates how easily the illness can spread from one person to another. The findings demonstrated that  $R_0$ 's value is significantly greater than the WHO's predicted value. A multicenter analysis of 68 deaths and 82 discharge reports of patients diagnosed with COVID-19 was carried out by Qiorong Ruan et al. They concluded that cases involving elderly individuals who had elevated blood markers for inflammation and subsequent infections had a catastrophic outcome. They added that cytokine storm syndrome patients have an increased chance of dying [25]. Neurological side effects are classified into three types: focal sensory system, which are the side effects of the

central nervous system (CNS), ailments (migraine, discombobulation, impaired awareness, ataxia, severe cerebrovascular infection, and epilepsy), and fringe sensory system, which are the side effects of the peripheral nervous system (PNS) and skeletal muscle injury [26]. It is determined that RT-PCR testing is the most practical and efficient diagnosis test for COVID-19 determination, but Chest CT scan analysis is also a competent diagnosis method. To compare the diagnostic value and consistency, Tao et al. [27] conducted a study and the results showed that Chest CT scans are highly sensitive to COVID-19 detection and can be used as the main tool. Analysis of symptoms and patients' history shows that COVID-19 spreads through direct contact [28]. Although the cases reported in Jiangsu are comparatively moderate as compared with Wuhan [29]. Detection of COVID-19 carriers is vital because, in the absence of a possible vaccine, the only effective way to combat the spread of this disease is to identify possibly infected persons and isolate them from healthy people [30]. Mizner et al. [31] offered an assessment of existing detection methods, demonstrating that each method has limitations, necessitating the development of a more efficient system.

Prabira et al. analyzed 11 different Convolutional Neural Network (CNN) systems and proposed an accurate support vector framework that incorporates SVM with classification models such as Residual Neural Network (ResNet50) to classify patients showing signs of coronavirus from X-ray scans of potential sufferers. The ResNet50 plus model performed better on data collected from numerous online data repositories [32]. Table 1 shows the state-of-the-art research conducted to identify the symptoms, gender, age, number of days the patient remained infected, and additional diseases the patient was undergoing corresponding to the country, and city where the patient was monitored. Mohammad et al. proposed a framework to detect coronavirus-based thermal screening of patients using an Internet of Things (IoT) based smart helmet. They reviewed 6 case studies that analyzed COVID-19-infected patients, using qRT-PCR, and CT scan analysis methods. The study concludes that the accuracy of qRT-PCR (81.3%) was lower than the CT scan (89.8%), therefore the CT scan method is more substantial [33]. It projected an open-source CNN that uses resizing and recurring learning rate discovery techniques and an altered version of the ResNet50 system which results in 96.23% accuracy on a public dataset (COVID) with an epoch count of 41 [34]. Ioannis et al. gauged the performance of the CNN framework generated through a transfer learning approach to detect several irregularities in small medical image data repositories. The result of analysis on two different datasets showed 96% accuracy, 98.66% sensitivity, and 96.46% specificity [35].

**Table 1.** Patient details and symptoms.

| Year/ Month | Author              | Number of Cases                     | Additional Disease                                 | Death/Recovery Rate | Days Infected       | Country and City | Gender and Age Group                      | Symptoms                               |
|-------------|---------------------|-------------------------------------|--|---------------------|---------------------|------------------|---|--|
| Feb/2020    | Jian et al. [25]    | 80                                  | 38 patients with a history of chronic disease      | 23% recovered       | 8 days, 21 patients | Jiangsu, China   | 41 females, average age 46.1 years        | Fever and cough in 63 and 51 cases     |
| March/2020  | Zhang et al. [15]   | 18-28                               | Cancer   | 28.6% recovered     | 14 days             | Wuhan, China     | 17 (60.7%) male patients, 65 years        | Fever, dry cough, fatigue, and dyspnea |
| May/2020    | J. Yang et al. [16] | 77658 include 576 infected patients | Hypertension, diabetes, and cardiovascular disease | 2663 deaths         | 17 days             | China            | Female (890:686), Male (890), Age (45-57) | Fever, cough, fatigue, and dyspnea     |
| April/2020  | Z. Wu et al. [17]   | 72314                               | Asymptomatic cases with lung diseases              | 5 deaths            | 30 days             | China            | 10-80 years                               | Fever, dry cough, and fatigue          |
| April/2020  | H. Shiet al. [18]   | 81                                  | Chest CT imaging abnormalities                     | 4 deaths            | 10.5 days           | China            | 42 men, 39 women, and 50 years age        | Fever, dry cough, and fatigue          |

|            |                             |                             |   |                               |          |  |  |  |
|------------|-----------------------------|-----------------------------|---|-------------------------------|----------|--|--|--|
| April/2020 | Q. Ruan, K. Yanget al. [21] | 150                         | Mild flu, myocardial damage, and circulatory failure                    | 68 deaths                     | 14 days  | China                                  | 30-85 years                              | Fever, dry cough, and fatigue                  |
| Feb/2020   | L. Mao et al. [22]          | 75569                       | Lung CT abnormalities (hypogeusia, hyposmia, hypoplasia, and neuralgia) | 2239 deaths                   | 10 days  | China, Europe, North America, and Asia | 52.7 ±15.5 years and 127 (59.3%) females | Headache, acute cerebrovascular, and dizziness |
| Feb 2020   | T. Aiet al. [23]            | 77658                       | RT-PCR essays   | 2663 recovered, and 33 deaths | 7 days   | 33 Countries                           | Age, 51 ± 15 years, 46% male             | Fever, and dry cough                           |
| April 2020 | L. Van Cu et al. [24]       | 21 infected                 | Mild chest pain, and blood pressure                                     | Some recovered                | 5 days   | Wuhan                                  | 25 years old woman                       | Cough, flu, fever, and chest pain              |
| Feb/2020   | J. Wu et al. [25]           | 66577, 80 patients infected | A syndrome is known as an (ARDS) metabolic acidosis                     | 21 cases recovered            | 8 days   | China, and other countries             | 41 females, aged 46 years                | Chest pain or other                            |
| May/2020   | Z. Hu et al. [26]           | 51857, 24 infected cases    | CT images of a glass chest, a shadow in the lungs                       | 1121 deaths                   | 9.5 days | Nanjing, Jiangsu Province, China       | Males, ages ranging from 5 to 95 years   | Fever, cough, and fatigue                      |

Biraja et al. proposed a framework that trains a Bayesian deep learning classifier using a transfer method to find out vulnerability in the X-ray scans from an open COVID-19 dataset. The outcome determines that susceptibility results in higher reliability in the estimate as it alarms radiologists about incorrect forecasts [36]. Charmaine et al. summarized that radiographic patterns of observation in CT chest scans and RT-PCR are significant methods for the recognition of coronavirus. Their research is the comparison of 2D, and 3D deep neural networks which resulted in 0.966% AUC, 98.2% sensitivity, and 92.2% specificity [37]. Bin et al. [38] demonstrated an experiment of Lopinavir–Ritonavir on elderly people hospitalized with severe coronavirus impact which caused breathing hindrance. Ying et al. illustrated that the period was secured from 1 January to 2 February 2020. During that time, they recognized 12 investigations. assessed the essential conceptive number for coronavirus cases from China or abroad. The assessments ranged from 1.4 to 6.49 where the mean calculated was 3.28, the middle was 2.79, and the inter-quartile. was 1.16 [39].

Ganyani et al. [40,41] determined that an essential key irresistible sickness constraint of this disease is quintessential to demonstrating and managing the intercession techniques. T. Thiruvalluan et al. stated that the coronavirus arising in Wuhan city in China is spreading throughout the world with the ACE II receptor as a binding site via human transmission and is called SARS CoV-2. There is currently no officially approved cure for COVID-19 that has been controlled by symptomatic relief and some antiviral medication, so avoidance plays an important role in suppressing the spread [42–44]. Akib Mohi et al. state that over 100 countries were affected by COVID-19 in no time. It is important to develop a control system that will detect coronavirus. Disease diagnosis may be one of the remedies for handling the current havoc with the help of various AI resources [45]. Shi Zhao et al. stated that since December 2019, the extreme acute respiratory disease coronavirus (SARS-CoV-2), has exhibited a large spread (of COVID-19) in other parts of the world starting from Wuhan, China. As of 15 February, there were 56 COVID-19 confirmed cases in Hong Kong after the onset of the first symptom on 23, 2020 January [46]. Table 2 shows the previous research conducted to analyze the various methodologies applied to detect coronavirus. The table states the input features utilized in the commonly applied detection methods, the source of data used, classifiers used, and the result obtained. The literature review implies the following points: First, fever and dry cough are frequent symptoms reported by coronavirus patients worldwide. Second, death rates were higher in cases when patients had significant chronic conditions, such as cancer or advanced age. Third, the detection approaches that performed best in COVID-19 situations were ML, deep learning, and CNN, which demonstrated higher levels of accuracy. These frameworks outperform the findings achieved using molecular biology approaches.

**Table 2.** Previous methodologies applied to detect coronavirus.

| Month/<br>Year | Author               | Detection<br>Method  | Input<br>Parameters  | Data Source  | Study Detail  | Classifier<br>Used   | Result<br>Obtained   |
|----------------|----------------------|--|--|--|---|--|--|
| Feb/2020       | Minzhe et al. [31]   | Review of an available nucleic acid method for detecting coronaviruses,    | Segments of the gene for PCR based methods, DNA, RNA for lamp, and Micro Array | GitHub Open-i  | Analysis of various coronavirus detection methods   | PCR, DNA, RNA,   | Each detection method has some drawbacks, thus new methods should be examined                                    |
| March/2020     | Muhammad et al. [34] | CNN system by retuning ResNet- 50 model                                    | Resized input images   | Publicly available COVID-Net dataset on GitHub   | System is proposed by <i>/fine-tuning</i> , ResNet-50 for input size, and learning rate             | CNN, ResNet  | The model showed 96.23% accuracy on the COVID-19 patient's data with only 41 epochs                              |
| March/2020     | Arabia et al. [32]   | Deep Learning, ResNet 50 plus, For deep feature extraction                 | X-ray  | GitHub Open-i  | Selection of train/validation/ test ratio: random 60: 20: 20  | SVM  | FPR = 95.38%<br>F1 = 95.52%<br>MCC = 91.41%<br>Kappa = 90.76%  |
| March/2020     | Mohammad et al. [33] | IoT based thermal helmet for recording temperature of subjects             | Image Processing, IoT, GSM interference, GPS and Mobile application            | GLH to track places visited by infected  | Input source of a thermal and optical camera Microcontroller, Arduino IDE Output source             | Proteus software for the basic model   | Capability to detect the the temperature of the subject from a distant place in a crowded location               |
| March/2020     | Biraja et al. [36]   | Analysis of dropwort based convolutional neural network                    | X-ray Dataset images   | Coronavirus X-ray of chest dataset to identify lung swelling, and enlargement of lymph nodes                 | NN parameters for independent, and identical training samples to analyze                            | Bayesian deep learning MC drop weight to enhance the accuracy of predictions | Results depict a correlation between uncertainty, and accuracy in prediction                                     |
| March/2020     | Binn et al. [38]     | Adults hospitalized a trial with severe coronavirus of Lopinavir Ritonavir | The clinical, and the electronic data set used                                 | Male and non-pregnant female patients, 18 years old diagnostic specimen positive on RT-PCR                   | clinical recorded data in the forms and electronic entered double database validated by trial staff | Lopinavir Ritonavir treatment or 100 to the standard care group              | Lopinavir in 13 patients because of events   |
| March/2020     | Loannis et al. [35]  | Evaluation of convolutional neural network architecture                    | X-ray images from cohen dataset available on GitHub                            | 1428 X-ray dataset confirmed with 24 positive cases, 700 confirmed, common with bacterial and infected cases | CNN model for spotting a variety of abnormalities in the small image data repository                | Untrainable layers at the bottom   | Good performance of the deep learning framework with 96.78% accuracy, 98.66% sensitivity, and 96.46% specificity |

|            |                       |  |   |   |   |   |   |
|------------|-----------------------|--|---|---|---|---|---|
| March/2020 | Tipwa et al. [40]     | Estimate the onset symptom of data with generation interval          | Shopping mall dataset, Singapore, and Tianjin. Data set available in GitHub | Dataset of shopping center 45 cases, Singapore 45 cases   | Previous estimates with no change so it does not distribute the big impact on estimates | Dataset of shopping center and Singapore  | Singapore, the mean age was 5.2 with 75 days<br>China 3-4 days depending on the formula of the brooding period with a mean value of 5.2 in 2.8 days |
| April/2020 | Charmaine et al. [37] | Deep learning and CNN frameworks analysis for detection of COVID-19  | CT scans  | 618 sloping CT scans which included 110/219 COVID-19 patients, 224/399 Influenza-A viral such as H1N1, H7N9 | 2d or 3D deep Learning algorithm done   | ResNet for Feature extraction   | AUC 0.966 for COVID-19 positive and negative, sensitivity 98.2%, and specificity 92.2%  |
| April/2020 | Joel et al. [41]      | Transmission stochastic model used to parametrize to the coronavirus | Dataset available on GitHub   | 24550 cases, 190 infected 490 deaths  | coronavirus outbreak controlling of feasibility   | Stochastic parametrize  | Stochastic transmission 2.5, 20 cases, and 15% of transmission before symptom   |
| May/2020   | Muhammad et al. [42]  | qRT-PCR, and CT analysis   | RT-PCR and CT scan  | Total subjects 1275, male patients 599, and female patients 676   | Review of qtr.-PCR, and CT scan methods   | Higher bilateral lobe (51.4%) than single lobe (21.5%) in coronavirus infected patients | qRT-PCR showed 81.3% positive, abnormal CT scan showed in 89.8% of patient  |

The following part describes the strategy we used to detect coronavirus patients based on their symptoms, utilizing LR, KNN, DT, NB, and SVM. The findings will assist the healthcare sector in making decisions, particularly in nations where the disease is projected to have a significant impact [43].

### 3 | Research Methodology

ML classification algorithms take up data to process, classify, or predict. The flow of the process involves pre-processing which includes data cleaning, data transformation, and feature selection, followed by the application of ML algorithms.

#### 3.1 | Data and Pre-Processing

Our study presented in this paper is based on publicly available databases. The dataset is obtained from the Israeli government website and is accessible worldwide, <https://data.gov.il/dataset/covid-19/resource/d337959a-020a-4ed3-84f7-fca182292308>. Such type of datasets is also used by several other research publications [6]. Although the website is constantly updated with the latest data, we used a dataset from 15/12/2020 till 21/01/2021. A total of 1,048,576 entries of patient records are contained in the dataset, containing symptoms and actual results of the potential COVID-19 patients. The columns are as follows: test

date, cough, fever, sore throat, shortness of breath, headache, corona result, age\_60\_and\_above, gender, and test indication.

Firstly, the data is obtained from the source and analyzed. However, the data obtained is noisy and needs to be handled, otherwise, it could be misinterpreted, which could result in erroneous outcomes for the algorithm. The missing values are dropped, and the data types of the required features are transformed. For the analysis, the features contained in the file, and their correlation is examined. The features are shortlisted by peer review to obtain the set of features around which the model will be revolved. The correlation between the selected variables can be seen in Figure 2.

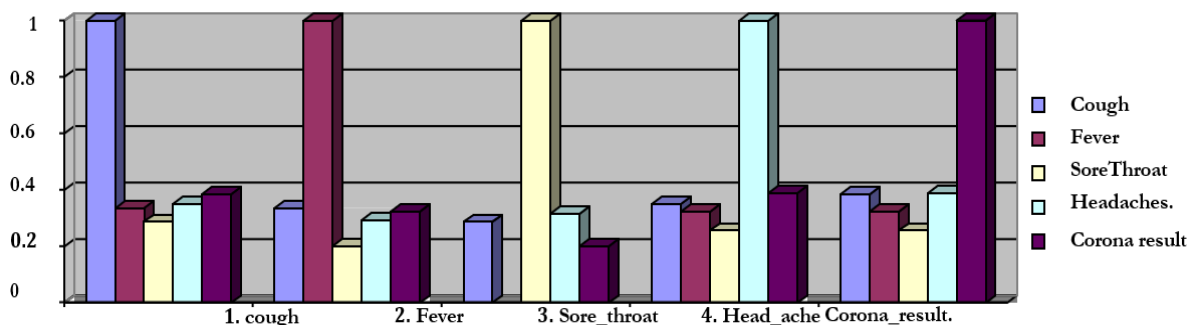


Figure 2. Correlation features of COVID-19.

## 3.2 | Machine Learning Models

The model formed is an intelligent system that is empowered by ML techniques' implementation on the pre-processed dataset. Moreover, several statistical measures are incorporated for the evaluation of the prediction of the suggested model. The algorithms applied are as follows; Logistic Regression, K-Nearest Neighbor, Decision Tree, Naïve Base, and Support Vector Machine.

### 3.2.1 | Logistic Regression

LR is a statistical method used for binary classification that models the probability of a binary outcome based on one or more predictor variables. It uses the logistic function to map predicted values to probabilities, which are then used to classify data points into one of the two categories. The model estimates the coefficients for each predictor variable through maximum likelihood estimation, optimizing them to best separate the classes. Logistic Regression is widely used due to its simplicity, interpretability, and efficiency in handling linear relationships between the predictors and the outcome [11].

### 3.2.2 | K-Nearest Neighbors (K-NN)

KNN is a non-parametric, instance-based learning algorithm used for both classification and regression tasks. It classifies a data point based on the majority class of its 'k' nearest neighbors in the feature space. The distance between points is typically measured using Euclidean distance (or other types of distances), though other distance metrics can be used. K-NN is simple to implement and effective for small datasets with well-defined class boundaries, but it can be computationally expensive and sensitive to the choice of 'k' and the distance metric, especially in high-dimensional spaces [12].

### 3.2.3 | Decision Trees

DT are a type of supervised learning algorithm used for both classification and regression. They work by recursively splitting the data into subsets based on the value of input features, creating a tree-like model of decisions. Each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. Decision Trees are intuitive, easy to visualize, and can handle both numerical and categorical data. However, they are prone to overfitting and can create complex trees that generalize poorly to unseen data if not properly pruned [11].



### 3.2.4 | Naïve Bayes

NB is a family of probabilistic classifiers based on applying Bayes' theorem with the assumption of conditional independence between every pair of features given the class label. Despite its 'naïve' assumption, it often performs surprisingly well in many complex real-world situations. NB classifiers are particularly effective for large datasets and are commonly used in text classification problems such as spam detection and sentiment analysis. They are computationally efficient, simple to implement, and work well with both binary and multiclass classification problems [9].

### 3.2.5 | Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find the optimal hyperplane that maximizes the margin between different classes in the feature space. Data points closest to the hyperplane, known as support vectors, are critical in defining the position and orientation of the hyperplane. SVM is effective in high-dimensional spaces and is particularly useful for cases where the number of dimensions exceeds the number of samples. It can use different kernel functions to handle non-linear relationships, making it a versatile and robust classifier. However, SVM can be computationally intensive and less effective on large datasets with many noisy features [10].

## 3.3 | Evaluation Metrics

One technique to assess an ML algorithm's precision in classifying a data point is to look at its accuracy. Specifically, it is determined by dividing the total number of true positives, false positives, true negatives, and false negatives by the number of true positives and true negatives. Eq. (1) shows how accuracy is calculated. Furthermore, the computation of precision and recall is done in conjunction with accuracy. The accuracy is determined by dividing the total number of positive results by the total number of false positive results. On the other hand, recall is calculated by dividing genuine positives by the total of false negatives and true positives. The following equations provide additional proof of the calculation:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{4}$$

$$\text{F1 score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

ML has an important role in measuring the performance of the used classifiers. The accompanying confusion matrices help to visualize the algorithms' performance. A confusion matrix is used to address ML issues, such as statistical classification using supervised learning. The confusion matrix is made up of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) values. TP is the instance where the model correctly identifies the positive class. TN is the outcome where models correctly identify the actual negative class. FP is the outcome at which the model incorrectly predicts the class when it is not present. Lastly, FN is the outcome where the model does not identify even in its presence. Table 3 shows how a confusion matrix is plotted.

**Table 3.** Confusion matrix.

|          | Predicted 0         | Predicted 1         |
|----------|---------------------|---------------------|
| Actual 0 | True Negative (TN)  | False Positive (FP) |
| Actual 1 | False Negative (FN) | True Positive (TP)  |

This study's machine learning classifiers use supervised learning for binary classification. The confusion matrix's TP value indicates how many patients with COVID-19 and how accurately the classifier identifies them. The FP figure is the number of patients that were incorrectly classified as having COVID-19. The TN represents the number of patients in the dataset who do not have COVID-19, and the classifier accurately identifies them. Finally, the FN represents the number of patients who have COVID-19 but are incorrectly identified by the model as not having the illness. The findings of each classifier, including accuracies, confusion matrices, and classification reports, are provided below.

## 4 | Results and Discussion

### 4.1 | First Experiment

This work uses ML classifiers as supervised learning for binary classification. The confusion matrix's TP value indicates how many patients with COVID-19 and how accurately the classifier identifies them. The FP figure is the number of patients that were incorrectly classified as having COVID-19. The TN represents the number of patients in the dataset who do not have COVID-19, and the classifier accurately identifies them. Finally, the FN represents the number of patients who have COVID-19 but are incorrectly identified by the model as not having the illness. The findings of each classifier, including accuracies, confusion matrices, and classification reports, are provided below. Table 4 shows the accuracies of the five classifiers used.

**Table 4.** Classifier accuracies for first dataset.

| Classifier | KNN   | SVM  | LR    | DT   | NB   |
|------------|-------|------|-------|------|------|
| Accuracy   | 93.54 | 93.6 | 92.77 | 93.7 | 93.7 |

From the study, we determine that supervised machine-learning algorithms can be utilized for the prediction of COVID-19 in patients with potential symptoms of this disease. For the evaluation of the models, we have used the metrics of accuracy. Accuracy is the fraction of predictions that the model got correct. Upon comparing the obtained results critically, we determine that logistic regression performs most inexactly with an accuracy of 92.80% whereas, NB and DT show the highest accuracy of 93.70%.

### 4.2 | Second Experiment

Furthermore, another dataset with 1048576 cases, handling missing values was investigated using several models, and the findings were obtained. Table 5 shows the outcomes of the new dataset.

**Table 5.** Classifier accuracies for second dataset.

| Classifier | RF    | LR   | DT   |
|------------|-------|------|------|
| Accuracy   | 94.04 | 93.6 | 93.7 |

## 5 | Conclusion

This study's ML classifiers use supervised learning for binary classification. The confusion matrix's TP value indicates how many patients with COVID-19 and how accurately the classifier identifies them. The FP is the number of patients that were incorrectly classified as having COVID-19. The TN represents the number of patients in the dataset who do not have COVID-19, and the classifier accurately identifies them. Finally, the FN represents the number of patients who have COVID-19 but are incorrectly identified by the model as not having the illness. The findings of each classifier, including accuracies, confusion matrices, and classification reports, were calculated. A diverse set of models and algorithms for data processing and visualization were considered and tested. However, following comparison and peer evaluation, we choose the algorithms for our investigation. The comparison of their results reveals that the NB and DT outperform the LR in terms of accuracy. This research aims to assist future researchers in examining the ML approaches to solve COVID-19 problems and to determine whether the patient is sick with this illness. Moreover, this

research can be extended in the future to address relevant problems of other diseases and act accordingly in case of a potential pandemic. In addition to the research, I improved the results of the decision tree model to 98%, another data set was used, and a different model was applied, such as RF, DT, and LR.

## Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

## Author Contributions

"Conceptualization, A.A., and S.E.; Methodology, A.A.; Software, A.A.; Validation, A.A., and S.E.; formal analysis, A.A.; Investigation, A.A.; resources, A.A.; data maintenance, A.A.; writing-creating the initial design, S.E.; writing-reviewing and editing, A.A.; visualization, S.E.; monitoring, A.A.; project management, A.A.; funding procurement, S.E. All authors have read and agreed to the published version of the manuscript.

## Funding

This research has no funding source.

## Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

## Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- [1] V. K. Deshwal, "COVID-19: A comparative study of Asian, European, American continent," *International Journal of Scientific Research and Engineering Development*, vol. 3, no. 2, pp. 436–440, 2020.
- [2] Y. A. Alrazaq, A. M. Alajlani, D. Alhuwail, J. Schneider, S. A. Kuwari, et al., "Artificial intelligence in the fight against COVID-19: Scoping review," *Journal of Medical Internet Research*, vol. 22, no. 12, pp. 1–18, 2020.
- [3] X. Wang, X. Deng, Q. Fu, Q. Zhou, Feng, et al., "A weakly-supervised framework for covid-19 classification and lesion localization from chest ct," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, Aug. 2020.
- [4] R. Zagrouba, M. A. Khan, M. A. Saleem, M. F. Mushtaq, et al., "Modelling and simulation of COVID-19 outbreak prediction using supervised machine learning," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 2397–2407, 2021.
- [5] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, et al., "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset," *Springer Nature: SN Computer Science*, vol. 2, no. 11, pp. 1–13, 2021.
- [6] M. A. Quiroz-Juarez, A. T. Gomez, I. H. Ulloa, R. D. J. L. Montiel et al., "Identification of high-risk COVID-19 patients using machine learning," *MedRxiv*, vol. 66, no. 3, pp. 1–10, 2021.
- [7] I. Arpaci, S. Huang and M. A. Emran, "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms," *Multimedia Tools & Applications*, vol. 1, no. 1, pp. 1–15, 2021.
- [8] N. S. Punna, S. K. Sonbhadra, and S. Agarwal, "COVID-19 epidemic analysis using machine learning and deep learning algorithms," *Health Informatics*, vol. 1, no. 1, pp. 1–10, 2020.
- [9] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, et al., "Artificial intelligence and machine learning to fight COVID-19," *Physiological Genomics*, vol. 52, no. 4, pp. 200–202, Apr. 2020.

- [10] A. Albahri and R. A. Hamid, "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus: A systematic review," *Journal of Medical Systems*, vol. 44, no. 7, pp. 1–10, 2020.
- [11] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making," *Health Informatics*, vol. 1, no. 1, pp. 1–10, 2020.
- [12] A. Erraissi, M. Azouazi, A. Belangour and M. Banane, "Machine learning model to predict the number of cases contaminated by COVID-19," *International Journal of Computing and Digital Systems*, vol. 9, pp. 1–11, 2020.
- [13] S. F. Ardabili, "COVID-19 outbreak prediction with machine learning," *Specialized Research Networks Journal*, vol. 1, pp. 1–11, 2020.
- [14] Q. Li, "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, Mar. 2020.
- [15] C. Huang, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *Lancet*, vol. 395, no. 10223, pp. 497–506, Feb. 2020.
- [16] Y. He, H. Yu, E. Ong, Y. Wang, L. Huffman, et al., "Cid, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis," *Scientific Data*, vol. 7, no. 1, pp. 1–5, 2020.
- [17] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of COVID-19 epidemic declaration on psychological consequences: A study on active Weibo users," *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, pp. 2032, Mar. 2020.
- [18] J. Wu, "Clinical characteristics of imported cases of coronavirus disease 2019 COVID-19 in Jiangsu province: A multicenter descriptive study," *Clinical Infectious Diseases*, vol. 1, pp. 1–9, Feb. 2020.
- [19] L. Zhang, "Clinical characteristics of COVID-19 infected cancer patients: A retrospective case study in three hospitals within Wuhan, China," *Annals of Oncology*, vol. 31, no. 7, pp. 894–901, Jul. 2020.
- [20] J. Yang, "Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: A systematic review and meta-analysis," *International Journal of Infectious Diseases*, vol. 94, pp. 91–95, May 2020

**Disclaimer/Publisher's Note:** The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.