

**SARS-COVID 19 PREDICTION: AN ANALYTICAL METHOD USING  
MACHINE LEARNING TECHNIQUES**

**BY**

**NAZMUL HOSSAIN ANANTO**

**ID: 181-15-10642**

**ISHRAT BINTE MAHFUJA**

**ID: 181-15-10517**

**SONIA AKHTER**

**ID: 181-15-10763**

**AND**

**MD. RONY HOWLADER**

**ID: 181-15-10684**

This Report Presented in Partial Fulfillment of the Requirements  
for the Degree of Bachelor of Science in Computer Science and  
Engineering

Supervised By

**Moushumi Zaman Bonny**

Assistant Professor

Department of Computer Science & Engineering  
Daffodil International University

Co-Supervised By

**Md. Sadekur Rahman**

Assistant Professor

Department of Computer Science & Engineering  
Daffodil International University



**DAFFODIL INTERNATIONAL  
UNIVERSITY DHAKA, BANGLADESH  
December 2021**

## APPROVAL

This Project titled “SARS-COVID 19 PREDICTION: AN ANALYTICAL METHOD USING MACHINE LEARNING TECHNIQUES”, submitted by Nazmul Hossain Ananto, Ishrat Binte Mahfuja, Sonia Akhter and Md. Rony Howlader to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 4<sup>th</sup> December, 2021.

### BOARD OF EXAMINERS



Chairman

---

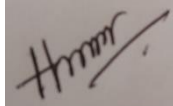
Dr. Touhid Bhuiyan  
Professor and Head  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



Internal Examiner

---

Moushumi Zaman Bonny  
Assistant Professor  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



Internal Examiner

---

Md. Mahfujur Rahman  
Senior Lecturer  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



External Examiner

---

Dr. Md Arshad Ali  
Associate Professor  
Department of Computer Science and Engineering  
Hajee Mohammad Danesh Science and Technology  
University

## DECLARATION

We hereby declare that this project has been done by us under the supervision of **Moushumi Zaman Bonny**, Assistant Professor, Department of Computer Science & Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**



---

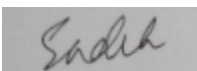
**Moushumi Zaman Bonny**

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

**Co-Supervised by:**



---

**Md. Sadekur Rahman**

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

**Submitted by:**

*Nazmul Hossain Ananto*

---

**Nazmul Hossain Ananto**

ID: 181-15-10642

Department of Computer Science and Engineering

Daffodil International University

*Ishrat Binte Mahfuja*

---

**Ishrat Binte Mahfuja**

ID: 181-15-10517

Department of Computer Science and Engineering

Daffodil International University

Sonia Akhter

---

**Sonia Akhter**

ID: 181-15-10763

Department of Computer Science and Engineering  
Daffodil International University

Md. Rony Howlader

---

**Md. Rony Howlader**

ID: 181-15-10684

Department of Computer Science and Engineering  
Daffodil International University

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Moushumi Zaman Bonny**, Assistant Professor, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Moushumi Zaman Bonny, Md. Sadekur Rahman and Touhid Bhuiyan, Head of Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

## **ABSTRACT**

COVID-19 infections have become prevalent, prompting worldwide efforts to control and treat the virus. Unfortunately, even after the invention of the vaccine so far, no vaccine invention organization has claimed that their vaccine can completely prevent Coronavirus and therefore the virus isn't going to be completely prevented. Since this life-threatening virus has no specific and special treatment and it spreads very easily and very quickly in human habitation. So, in an overpopulated and developing country like Bangladesh in south Asia, it's very difficult to identify every infected person and give them proper treatment for government and health workers. In recent years, artificial intelligence and machine learning have achieved appeal as a part of enhancing healthcare and research in general, especially in the field of the medical sector. To predict "COVID-19", a wide range of machine learning approaches, applications, and algorithms are developed. A machine-learning model is developed through which a potentially infected individual can know how susceptible he/she is to become infected with COVID-19 and their conditions. It may be very helpful for people to detect their problem and get primary treatment from home until they reach the stage of going to the hospital. This may make it possible to reduce the burden of health workers and the government. To acquire the best potential result in this system, more advanced and dynamic algorithms are required, such as K-nearest Neighbor, Decision Tree, Random Forest, AdaBoost, XGBoost, Stochastic Gradient Descent, Linear SVC, Perceptron, Naive Bayes, Support Vector Machines, Logistic Regression, Discriminant Analysis, and other.

## Table of Contents

<b>APPROVAL .....</b>	<b>i</b>
<b>DECLARATION.....</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT.....</b>	<b>iii</b>
<b>ABSTRACT.....</b>	<b>iv</b>
<b>LIST OF FIGURES .....</b>	<b>viii</b>
<b>LIST OF TABLES .....</b>	<b>ix</b>
<b>CHAPTER 1</b>	
<b>INTRODUCTION.....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 Motivation .....	2
1.3 Rationale of the Study .....	3
1.4 Research Questions .....	3
1.5 Expected Outcome .....	5
1.6 Report Layout.....	5
<b>CHAPTER 2</b>	
<b>LITERATURE REVIEW .....</b>	<b>7</b>
2.1 Introduction .....	7
2.2 Related Works .....	7
2.3 Comparative Analysis .....	12
2.4 Scope of the Problem .....	13
2.5 Challenges .....	14
<b>CHAPTER 3</b>	
<b>RESEARCH METHODOLOGY &amp; SYSTEM ARCHITECTURE.....</b>	<b>15</b>
3.1 Introduction .....	15
3.2 Research Subject .....	15
3.3 Machine Learning Techniques .....	15
3.3.1 Supervised Learning .....	16
3.4 Classification Techniques .....	16
3.4.1 Learning .....	17
3.4.2 Classification .....	17
3.5 Algorithmic Details .....	17
3.5.1 Decision Tree.....	17
3.5.2 Support Vector Machine.....	18
3.5.3 K-Nearest Neighbors .....	18

3.5.4 Logistic Regression .....	18
3.5.5 Linear Discriminant Analysis .....	19
3.5.6 eXtreme Gradient Boosting (XGBoost) .....	20
3.5.7 Random Forest.....	21
3.5.8 Gaussian Naïve Bayes .....	21
3.5.9 AdaBoost .....	21
3.5.10 Stochastic Gradient Descent.....	22
3.5.11 Linear SVC.....	22
3.5.12 Perceptron.....	23
3.6 Proposed System .....	23
3.6.1 Data Collection .....	24
3.6.2 Dataset .....	25
3.6.3 Data Pre-processing .....	25
3.6.4 Data Normalization.....	25
3.6.5 Data Splitting .....	25
3.6.6 Algorithm implicate.....	25
3.6.7 Model Analysis.....	26
3.6.8 Extract Appropriate Algorithm.....	26
3.6.9 Creating Model for Web Interface.....	26
3.6.10 Building a Web Interface.....	27
3.6.11 Execute Model.....	27
3.6.12 Input Values.....	27
3.6.13 Predictive Result.....	28
3.7 System Architecture .....	28
3.7.1 User Segment.....	29
3.7.2 Web Insider.....	29
3.7.3 Machine Learning Model .....	30
<b>CHAPTER 4</b>	
<b>EXPERIMENTAL RESULTS &amp; DISCUSSION.....</b>	<b>31</b>
4.1 Introduction .....	31
4.2 Experimental Results.....	31
4.2.1 Data Acquisition .....	31
4.2.2 Data Utilization.....	33
4.2.3 Feature Importance .....	35
4.3 Result & Discussion .....	36

4.3.1 Confusion Matrix.....	36
4.3.2 Classification Report .....	40
4.4 Result Analysis.....	42
4.4.1 Accuracy .....	42
4.4.2 Jaccard Score .....	43
4.4.3 Cross Validated Score .....	44
4.4.6 Standard Deviation .....	45
4.4.7 Misclassification & Error .....	45
4.5 Web Implementation .....	47
4.5.1 Web Interface .....	47
4.5.2 Web Output Analysis.....	48
<b>CHAPTER 5</b>	
<b>IMPACT ON SOCIETY &amp; SUSTAINABILITY .....</b>	<b>50</b>
5.1 Introduction .....	50
5.2 Impact on Society.....	50
5.3 Ethical Aspects.....	50
5.4 Sustainability.....	51
<b>CHAPTER 6</b>	
<b>FUTURE SCOPE &amp; CONCLUSION .....</b>	<b>52</b>
6.1 Introduction .....	52
6.2 Future Scope of this Study .....	52
6.3 Recommendations .....	52
6.4 Conclusion.....	53
<b>REFERENCES.....</b>	<b>54</b>
<b>PLAGIARISM REPORT .....</b>	<b>57</b>



## LIST OF FIGURES

<b>Figures</b>	<b>Page No.</b>
Figure 3.1: Proposed Method to Predict COVID-19	24
Figure 3.2: System Architecture	28
Figure 3.3: System Architecture of Web Interface	30
Figure 4.1: Accuracy Chart	43
Figure 4.2: Jaccard Score Chart	43
Figure 4.3: Cross Validated Score	44
Figure 4.4: Misclassification and Error	46
Figure 4.5: Web Interface	47
Figure 4.6: Web Interface with Negative Output	48
Figure 4.7: Web Interface with Normal Output	48
Figure 4.8: Web Interface with Emergency Output	49

## LIST OF TABLES

<b>Tables</b>	<b>Page No.</b>
Table 4.1: COVID-19 Result frequency of the patients	32
Table 4.2: Data Acquisition & Null Percentage	32
Table 4.3: Dataset Description	34
Table 4.4: Feature Importance	35
Table 4.5: Confusion Matrix	37
Table 4.6: Confusion Matrix for Algorithms	37
Table 4.7: Classification Report	41
Table 4.8: Accuracy, Jaccard and Cross Validated	44
Table 4.9: Standard Deviation	45
Table 4.10: Misclassifications & Errors	46

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

COVID-19 is an infectious disease caused by a novel extreme Acute Respiratory Syndrome Coronavirus 2 virus (SARS-CoV-2). SARS-CoV-2 is the source of Coronavirus Disease 2019 (COVID-19) [1].

Though Scientists first identified a human coronavirus in 1965, but regarding by WHO on 31 December in 2019 the current outbreak of coronavirus was first reported in Wuhan, China, and very quickly spread to almost all countries worldwide. Thousands of people die before the immune system can create the inhibitory antibody as a result of these quick and untraceable viral changes. As a result, in a very short time, almost 258 million people got infected and more than 5.16 million people died all over the world.

Because this virus spreads more rapidly inside and in crowded settings, it has killed a large number of individuals in a short amount of time. Inhaling the virus or touching a contaminated surface and then contacting their eyes, nose, or mouth can infect persons who are in close proximity to someone who has COVID-19. The virus spreads in minute liquid particles from an infected person's lips or nose when they cough, sneeze, speak, sing, or breathe.

The majority of COVID-19 patients will experience mild to moderate symptoms and will recover without the need for treatment. Some, on the other hand, will be severely harmed and will require immediate medical attention.

Respiratory symptoms such as fever, cough, and shortness of breath are signs and symptoms of COVID-19. As the infection mutates, it infects the lungs, producing inflammation in the alveoli, or lung sacs, which then fill with fluid and pus, resulting in pneumonia. Between the eighth and fifteenth days after the onset of symptoms, the inflamed lungs make it difficult for the person to breathe, resulting in Death. Comorbid sufferers have a high death rate, with hypertension being the most frequent, followed by diabetes and coronary heart disease. Since it is highly contagious, A large number of people die as a result of it infecting more people. So, it is possible to reduce the

infected / mortality rate by providing the necessary health care to the infected person at an early stage, and by providing emergency medical treatment to the emergency patients.

Machine Learning Algorithms are divided into numerous categories. Supervised Machine Learning, Unsupervised Machine Learning, Semi-supervised Machine Learning, and Reinforcement Machine Learning are examples of machine learning techniques. We develop models with algorithms from Supervised Machine Learning Algorithms such as K-nearest Neighbour, Decision Tree, Random Forest, AdaBoost, Stochastic Gradient Descent, Linear SVC, Perceptron, Naive Bayes, Support Vector Machines, Logistic Regression, Discriminant Analysis, and eXtreme Gradient Boosting (XGBoost) to predict COVID-19 in this research work. In this system, users can predict COVID-19 by providing their symptoms.

Furthermore, in a pandemic, a web implementation procedure can be the most effective solution for pathology departments and patients who wish to know the actual chances of becoming infected by COVID-19 while staying at home. The purpose of the study is to develop a model using a relevant dataset, establish the model used to predict COVID-19, and determine the patient's condition so that they can find out what kind of treatment they require immediately. After extracting the best algorithm from the most relevant Machine Learning algorithms, the information is transferred to a web application, where patients, doctors, and pathologists can evaluate the patient's condition and input these values to know the possibilities of COVID-19. The modern period will become more modern and hi-tech in the near future. This research will help people better understand the risks of being infected with COVID-19 and take the required steps to prevent the infection by revealing the various outcomes of Web implementation.

## **1.2 Motivation**

Despite having high-efficiency medical instruments, the coronavirus has devastated the healthcare system of many developed countries including the United States, the United Kingdom, Italy, Spain, etc. So, it was very difficult to deal with COVID-19 for overpopulated and developing countries like Bangladesh/ India in south Asia. Even so, Bangladesh's healthcare system is unreliable, responsive, and empathic, and it has been

repeatedly shown to be incapable of providing basic health care to the vast majority of the population. In the COVID-19 pandemic, the Bangladesh government and health ministry struggled enormously to provide Coronavirus detection services and other essential health care to this massive population because, in both government and commercial hospitals, medical facilities such as beds, intensive care units, and ventilators are far insufficient. There is a method based on the application of machine learning which can provide a primary guideline for treatment and this will reduce the burden of hospital workers and the government. As a result, a large section of the population will not be deprived of healthcare, and also the number of infected will decrease.

### **1.3 Rationale of the Study**

People are advised by the government to stay at home because coronavirus is a contagious sickness. If COVID-19 can be predicted and treated at home through a web application, a nation will be benefited in three ways. It will reduce the spread of SARS-CoV-2 (the virus that causes COVID-19). This will alleviate the burden on hospital employees and the government. The risk of death will be reduced if the necessary treatment is given by early detection. As a result, a model that can predict COVID-19 has been proposed and trained using a relevant dataset. The main reason for working on this study is to keep people safe, save time, and make their lives easier. A web interface has been created through which patients can provide data and predict COVID-19 at home. It will be beneficial to doctors and possible COVID-19 patients.

### **1.4 Research Questions**

There are numerous questions that can be highlighted in terms of this study. To make this study more compact, a set of questions was extracted from various individuals.

- **Why is COVID-19 Prediction the target of this study?**

The SARS-CoV-2 virus is the reason for COVID-19 being the biggest problem throughout the world. The virus causes death for millions of people. Almost every single country in the world is suffering from this pandemic situation and

their healthcare facilities are getting stressed. Infected people from this deadly virus can be cured with proper treatment and maintaining many rules and regulations. But there is not enough medical care to serve this huge amount of suffering people. So, to control this huge pressure of infected people COVID-19 prediction from home became the target of this study.

- **Why the machine learning approach? Is it reliable?**

Machine learning is a widely used technology for making predictions. Using a massive amount of data, a model may train itself and predict any outcome. COVID-19 can be predicted using a machine learning approach on a clinical dataset. The world is currently undergoing a period of modernization. If we go back about ten years, when Artificial Intelligence and Machine Learning were still in their infancy, these prime points were just a name with some mathematical logic. However, Artificial Intelligence now powers half of the world's technology. As a result, proper practice and more precision in this field can improve its reliability, though it is already adequate.

- **What is the main purpose of using a web interface?**

A user can utilize a web interface to enter values into a machine. It is a generous approach to the general public. This project can be used from anywhere in the world using a web interface and an internet connection. A web interface can also produce results more quickly. One can get a basic idea of the state of his or her COVID-19 condition at home using this web interface. As a result, it is a very practical method of predicting COVID-19.

- **What are the reasons to use 12 separate algorithms?**

One acceptable algorithm that suits the COVID-19 dataset was chosen from 11 distinct algorithms. After comparing and analyzing 12 algorithms, the optimal algorithm with the lowest mistake rate and maximum accuracy rate was discovered. If only one method was chosen, finding the best algorithm would be impossible because no one knows which algorithm will best suit the dataset.

## **1.5 Expected Outcome**

Several times during this research period, the major theme or desired conclusion was modified. It helps to clarify the exact outcome of this study. This research has the potential to ensure the basic treatment to the huge population to prevent the death of COVID-19 also reduce the spread of SARS-CoV-2. It may be determined by the system at what age individuals become more infected by COVID-19. A complete internal evaluation of COVID-19 could be revealed for medical study using exact calculations and algorithms. While this research is being conducted, a team may issue a warning to riverine people and regions regarding COVID-19. The final strategy employed in this study was the creation of a web-based system that identifies any potential patients who should be contacted by a physician.

## **1.6 Report Layout**

In this report, to make the analysis report more compact and efficient for any readers or researchers, six specific chapters are discussed.

Chapter 1 represents a vital introduction to this research. This is about COVID-19 and its details inside a brief. This chapter explains the reason for the research, the rationale for the study, pertinent research questions, expected outcomes, and overall management information.

Chapter 2 contains a full summary on the study's background. Based on this research study, such as machine learning systems, classifications information, and associated work. This chapter also discusses comparative analysis and the extent of this problem statement, as well as perceived difficulties.

Chapter 3 describes the methodology, suggested system, and system architecture for this research project of this report. Each implemented algorithm's algorithmic specifics are detailed, from mathematical conception to present condition.

Chapter 4 presented the entire outcome analysis for each stage. The report concludes with the best algorithm, jaccard score, cross validation score, confusion matrix, and

classification report. The last part of this chapter discusses standard deviation, misclassification, mean absolute error, and mean squared error.

Chapter 5 discussed ethical Aspects, which are essential for any impactful research work, to highlight the research's influence on society. The chapter concludes with a discussion of the research work's Sustainability.

Chapter 6 described the future scope of this research work briefly as the expansion of this research study. This chapter completes the whole study report with a useful conclusion that briefly summarizes the research's main results.



## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Introduction

COVID-19 has been a serious issue in this chapter up to this point. Therefore, the background of this disease is a lot of converting history and a lot of loss. As a result, much work has been done to detect Coronavirus to save lives. This chapter describes the history and background of this disease. Some related work on this issue has been discussed in this chapter. Finally, a comparative analysis was presented to show how much the current work was affected.

#### 2.2 Related Works

A machine learning approach for predicting SARSCoV2 infection is being developed. Prakash, K. B. et al. a survey was done to determine the age groups most affected by COVID-19. The clinical symptoms of COVID-19 include respiratory illness, malaise, dry cough, and malaise, but 80% of patients heal without care. SVM, K-nearest neighbors, + NCA, DT Classifier, GNB Classifier, MR, LR, Random Forest Classifier, and XGboost classifier are just a few of the techniques used in their machine learning models. Machine learning techniques are used to understand COVID-19, its confirmation and the prediction of recovery that affects humans. 20–50 years old are more likely to be infected with COVID-19 [1].

Zoabi, Y. et al. developed a machine-learning approach for predicting SARS-CoV-2 infection in humans, and patients over 60 years of age can get an RT-PCR test to see if they have eight basic symptoms. The model was built using data from all Israelis who were tested for SARSCoV2 during the early months of the COVID-19 outbreak. Predictions were then made using gradient boosting machine models developed with decision-tree base-learners. Machine that boosts gradients uses many algorithms in learning and is widely considered as an industry for predicting tabular data. Furthermore, the methods used in this study could aid the medical system in responding to future outbreaks of this virus, as well as other respiratory viruses in general, which is the cause of this sickness [2].

Amira, F. et al. research et al.'s article Corona Tracker is a community-led research project that tries to use predictive modeling to anticipate the COVID-19 case, mortality, and recovery. This paper is in charge of the website's visualization and does real-time data queries. They use SEIR modeling to predict the occurrence of COVID-19 every day in China and abroad. Their models can be used to understand public opinion trends in the transmission of relevant health information and to estimate the political and economic consequences of the virus spread [3].

This study by Meni, C. et al. focuses on a smartphone-based program called COVID Symptom Tracker, which was COVID Symptom Tracker was previously known as COVID Symptom Tracker. Joe Global collaborated with Massachusetts General Hospital and King's College London to develop the app. This will allow us to collect self-reported information regarding COVID-19. Their algorithm, which combines data from all app users who reported symptoms to predict infection, was applied to all the data from all app users who reported symptoms and indicated that participants were likely to get sick. COVID-19 [4].

Callejon-Leblic, M. A. et al. and colleagues created a complete framework for forecasting the value of odor and taste using machine learning (ML) abnormalities, among other COVID-19 infection symptoms. They are developing a comprehensive LR and two other ML algorithms, RF and SVM, which are part of a Machine Learning (ML) modeling framework (SVM), to evaluate predictive values of odor and dysgeusia in COVID-19 disease. These models can be useful tools for optimizing screening for suspicious individuals. In the context of fresh COVID-19 outbreaks, there are some suspicious instances. They claim that loss of smell and taste can reliably predict COVID-19 infection [5].

Muhammad, L. J. et al. machine learning was examined approaches and produced a COVID-19 transmission prediction model. Only two populations were studied in this study, which included age and gender as well as eight clinical variables such as pneumonia, diabetes, asthma, hypertension, cardiovascular disease (CVD), obesity, chronic kidney disease (CKD), and high risk. The results of the RTPCR test for COVID-19 were considered in a cigarette and patient records included. These were constructed utilizing logistic regression, decision trees, support vector machines, nave

bays, and artificial neural networks, as well as labeled epidemiologically datasets on the favorable and bad features of COVID-19 in Mexico [6].

Aktar, S. et al. to investigate clinical data sets of COVID-19 patients with known outcomes, researchers used statistical comparison and correlation methodologies with machine learning algorithms. They employ statistical methods, such as student tests, chi-square tests, and Pearson correlations, to identify the most critical and collaborative blood indicators that can produce a significant difference between COVID-19 patients and healthy people. They used a range of standard values as reference values for each parameter to compare the blood parameter values of COVID-19 patients with healthy patients. They also create algorithms such as DT, RF, Variants of Gradient Boosting, SVM, K-Nearest Neighbor (KNN), and Deep Learning (DP) [7].

Kwekha-Rashid, A. S. et al. clarified the role of machine learning applications and algorithms in various purposes related to COVID-19. With the COVID-19, machine learning, supervised learning, and unsupervised learning are some of the terms used in this paper. A review of studies published in 2020 on this topic was undertaken at Science Direct, Springer, Hindawi, and MDPI. Their findings suggest that machine learning can help researchers investigate, predict, and identify COVID-19. This review paper evaluates the outcomes of various AI algorithms employed in research to determine the most accurate technique to provide the most progress in COVID-19 diagnosis [8].

Alimadadi, A. et al. COVID-19 was predicted using artificial intelligence and machine learning in this study. Advanced machine learning algorithms were used to collect clinical data from COVID-19 patients in order to better understand viral spread patterns, improve diagnostic speed and accuracy, develop novel effective therapeutic approaches, and identify the most at-risk individuals based on physiological features and possibly personalized basis. They've also designed and created substances like fancy medications against SARSCoV-2 using an in-depth education-based drug discovery process [9].

In this study by Sumika, F. et al., many supervised machine learning algorithms were applied to datasets that included information from patients over the age of 60 who had been tested positive for COVID-19, resulting in computer models that could predict

disease development. There are two primary aspects to their experimental design. The first uses the training dataset to execute a series of iterative tests, while the second uses both the training and validation datasets to run the final test [10].

The work of Alballa, N. et al. provides an overview of current findings on the machine learning techniques employed in COVID-19. They concentrated on two applications that have gotten a great deal of publicity. Utilizing freely available clinical and experimental data, COVID-19 diagnosis and severity, as well as mortality risk, are all factors to consider. Describes the types of algorithms, training datasets, and aspects related to feature selection. Several key points were brought up in this study. For starters, the majority of the in these two applications, the machine learning techniques used are supervised learning algorithms, which are straightforward and easy to understand. Much of the effort was exploratory, and the model produced some interesting results but was never used in a real-world application [11].

Batista, A. F. M. et al. published a research paper. Its goal is to help poor countries make better COVID-19 test priority decisions by estimating the chance of using a positive diagnosis from just routine data gathered during an emergency. To measure predictive performance, they calculated the AUC, sensitivity, specificity, F1-score, Brier score, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) (NPV). Machine learning techniques, they say, can be used to prioritize getting RT-PCR testing in the event of a scarcity, as well as aid While the RT-PCR is being performed, crucial care decisions must be made. The results are being analyzed. Future studies will focus on analyzing the performance of the new fast tests and machine learning techniques when used together [12].

In this study report, Asaph, D. et al. use machine learning methods to anticipate the risk of critical COVID-19 based on clinical baseline features. To forecast degradation, they used three distinct machine-learning algorithms that utilized three distinct techniques: Classification and Regression Decision Tree (CRDT), Random Forest (RF), and Neural Network (NN) (CRT). Each model's samples for testing are given their mean, sensitivity, specificity, positive predictive value, and accuracy, all terms that are used in statistics. The average for each ROC plot, the area under the curve (AUC) was calculated was calculated using receiver operator characteristic (ROC) curves. Their research demonstrated the superiority of ML models in terms of better

accuracy, high NPV, and forecasting when compared to the existing state-of-affairs models. There are some flaws in this research. First, COVID-19's single-center retrospective approach reduces the outward validity due to significant COVID-19 heterogeneity between countries and their people. Second, due to a small proportion of patients suffering from serious illness, the statistical power of the study is limited. This is a common misunderstanding. Third, because there were no multivariate/risk-score predictors, the best univariate/risk-score predictors alternative tools for establishing a baseline had previously been revealed, the diagnostic performance of the AI models was excellent. As a result, prospective validation is required before our findings may be used to clinical practice [13].

Mazurek, J. et al. their research proposes a new measure of model prediction precision called a divergence exponent, which is based on the Lyapunov exponent and addresses the aforementioned shortcomings. In the context of chaotic processes, the suggested approach allows for the measurement and comparison of model prediction precision for time series of unequal length and a specified goal date. Two examples highlight the use of the divergence exponent in evaluating model correctness, and then a set of COVID-19 spread forecasts from other studies is reviewed to indicate its potential [14].

Parro, V. C. et al. this research presents a practical method for measuring the health-care utilization of COVID-19 cases. A dynamic model called Susceptible, Infected, Removed, and Dead was employed to develop the new methodology (SIRD). The model was updated to focus on the dynamics of the healthcare system rather than portraying all cases of the disease. It was fine-tuned using data from each Brazilian state and was updated regularly. A figure of merit was used to optimize the free parameters, which assesses the quality of the model's fit to the data. Using data from the 26 Brazilian states, the parameters of an epidemiological model for the entire country of Brazil, which comprised of a linear combination of the models for each state, were estimated [15].

## 2.3 Comparative Analysis

The prior study advocated focusing on COVID-19 test duration vs. discontinuation rather than appropriately predicting COVID-19. The Neural Network, Random Forest, XGBoost, and Logistic Reg. algorithms were applied as data mining tools. The Neural Network method has an accuracy of 83.12 percent, while the Random Forest algorithm has an accuracy of 85.23 percent, XGBoost offers 85.35 % and Logistic Regression gives 80.43 % at the same time. The authors applied 12 algorithms in this research. Four of the algorithms had a 95 % accuracy rate (XGBoost Classifier with 98.19 %, Decision Tree 97.70 %, KNN 96.22 % and Random Forest with 98.03 % accuracy) which are better than the previous studies [16].

A method for detecting COVID-19 utilizing chest X-ray pictures was developed in a prior study. To predict the number of COVID-19 confirmations, recoveries, and deaths in Jordan and Australia, they used three well-known forecasting algorithms in their system: the prophet algorithm (PA), the autoregressive integrated moving average (ARIMA) model, and the long short-term memory neural network (LSTM). Going to the hospital for an X-ray might be challenging at times. It's also pricey. On the other hand, the author of this research devised a way in which patients can obtain their COVID-19 prediction result by entering their symptoms on a website. As a result, it is less difficult than the preceding one [17].

The study's purpose was to apply machine learning to predict the chance of a positive COVID-19 diagnosis based only on results from emergency room admission tests. There are a total of five machine learning algorithms (neural networks, random forests, gradient boosting trees, logistic regression and support vector machines). The best prediction results were obtained using the support vector machines approach (AUC: 0.85; Sensitivity: 0.68; Specificity: 0.85; Brier Score: 0.16). In this work, a machine-learning model is constructed so that a possibly infected individual can determine how sensitive he or she is to becoming infected with COVID-19 and their conditions. The authors implemented 12 algorithms in this research. Among them XGBoost Classifier produced the best predicting results with 98.19 percent accuracy. It has a Jaccard Score of 96.45 percent and a Cross Validated Score of 99.54 percent [12].

In previous research, an SEIR (Susceptible-Exposed-Infectious-Removed) model was presented to assess the epidemic trend in Wuhan, and an AI model was employed to investigate the epidemic trend in non-Wuhan regions. They discovered that if the closure was relaxed, the outbreak would double in magnitude in non-Wuhan areas of mainland China. The COVID-19 pandemic peaks and magnitude were accurately predicted using the SEIR and AI model. In this study a machine-learning model is developed through which a potentially infected individual can know how susceptible he/she is to become infected with COVID-19 and their conditions. It may be very helpful for people to detect their problem and get primary treatment from home until they reach the stage of going to the hospital [18].

## **2.4 Scope of the Problem**

Bangladesh is the world's eighth most populous country. For every 10,000 inhabitants in Bangladesh, there are just 3.05 physicians and 1.07 nurses [19]. Because of the limitations and unavailability of healthcare in Bangladesh, 79 % of COVID-19 affected patients are unable to test. Bangladesh was rated the second-worst country in South Asia in a study of 1000 individuals. According to the most recent WHO figures, COVID-19 claimed the lives of 27,970 persons between 2020 and 2021. Many may not even have tested and hence are not included on the list. Testing at hospitals is particularly risky due to the high number of persons involved and the lack of adequate security to prevent the spread of COVID-19. As a result, uninfected persons get infected and transmit the virus to their relatives and friends. COVID-19 prevention necessitates extreme caution and meticulous preparation. Self-preservation is essential in this country's current position. Preliminary testing and, once the result is known, following the essential guidelines can save the life of a COVID-19 infected individual. Otherwise, both the infection and fatality rates would climb, making it difficult to halt the spread of this contagious virus. A modern application for forecasting COVID-19 might alleviate hospital overcrowding and lower the fatality rate.

## 2.5 Challenges

To forecast the expected, Machine Learning requires a large amount of data. One of the most difficult tasks was gathering massive amounts of data. Clinical data was utilized to predict COVID-19 in this investigation. Hospitals are well-known for their aversion to sharing patient information. The data required for this study is COVID-19 suspects data, which is relatively uncommon. COVID-19 is mostly tested at government facilities, and acquiring data from a government like Bangabandhu Sheikh Mujib Medical University (BSMMU) was much more difficult. There are always a lot of missing values in clinical datasets. Hospitals frequently fail to fill out required data due to time constraints. Handling these null data is a difficult task. This might be a lengthy procedure. There were a lot of factors in the dataset, and not all of them were required to predict COVID-19. Because data visualization was required to choose the useful characteristics from the raw dataset, feature selection became more difficult. It was also difficult to build an algorithm that was appropriate for the scenario. The dataset was trained using a variety of algorithms, and the researchers picked the methods that predicted COVID-19 with the highest accuracy. One of the most difficult jobs for the team was to create an intuitive user interface for this technology, which is user-friendly and allows anybody to input data and forecast COVID-19 to know their status and receive essential advice prior to actually going to the hospital.



## CHAPTER 3

### RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE

#### 3.1 Introduction

The research approach should be considered for getting started with research. First, a limitation should be identified, and then techniques must be developed to solve that problem. The methodology is explained in this chapter. The algorithms that were developed to solve the problem were then explained, detailed description of the approach with a schematic diagram for easier understanding. For better visualization, a system architecture was also shown.

#### 3.2 Research Subject

The main goal of the research is to identify a challenge to study and then develop a workable solution to solve this problem. The SARS-CoV-2-caused new coronavirus disease (COVID-19) pandemic caused a major and exigent circumstance to world health. It has killed millions of people worldwide. Unless the corona is not detected at an early stage, it destroys the human immune system and spreads rapidly to the lungs of the infected person, increasing the risk of death. So, if we can identify the symptoms of coronavirus, detect the COVID-19 patients, and take the required precautions for them, it will reduce the rate of infection and death at the same time. So, it's important to study Coronavirus and develop a model that can predict at different ages and genders. It is also possible to know how long it will take for them to recover according to age. A web interface that will be user-friendly and people will be able to get the necessary guidelines based on their condition at home.

#### 3.3 Machine Learning Techniques

Classification and supervised machine learning can be used together to classify objects. Self-contained and able to continually integrate data, a machine learning system is described for decision-making purposes as one. This can be done by relying on previous

experience, conducting analytical observations, and employing other methods. There is a vast range of machine learning methods available.

### **3.3.1 Supervised Learning**

Supervised machine learning algorithms are frequently used in this research. Supervised machine learning approaches implement labeled samples of previous events to form predictions. The learning technique develops an inferred function from a study of a well-trained dataset to provide predictions about the output values. After that, there are no boundaries to how much training the system can handle. When the learning algorithm compares its output to what was intended, it may detect mistakes and make necessary adjustments to the model. This thesis uses numerous supervised learning techniques to classify and predict COVID-19 in its early stages.

### **3.4 Classification Techniques**

Data classification is a type of data analysis in which models defining relevant data classes are generated. It's the most frequently used and effective machine learning technology. Under supervision learning, these models, known as classifiers, may predict categorical target class. The predictions are unordered and discrete. In supervised machine learning, there are various methods of classification, and multiclass classification is one of them. In some situations, the number of class labels can be quite enormous. There are more than two classes in a multi-classification system. Each sample is allocated to only one target label in multi-class categorization. For example, classify a group of photos of leaves that could be aspen, maple, or lupin. A leaf can be either aspen or maple, but not both at the same time.

On labeled data, the classification learning technology can be implemented. In classification learning, there are two types of data. One type is considered to be training data, while the other is referred to as test data. The model is built using training data, and the model is tested using test data. There are two steps to the process of classification.

### **3.4.1 Learning**

A suitable technique and training data are used to develop a classifier during the learning phase, which is subsequently evaluated against the real world. A classifier is essentially a set of rules that may be applied to a range of different events in appropriate environments when a classification algorithm and training data are combined.

### **3.4.2 Classification**

It is possible to determine which class of unknown data will be predicted using the classifier or model developed during the prediction phase as a result of the learning phase. This segment uses the test data to determine whether or not a model's predictions are accurate.

## **3.5 Algorithmic Details**

Twelve of the top Supervised Machine Learning algorithms are applied to conduct this research project. In general, an algorithm can be defined as an organized set of instructions that instructs computer software how to modify a set of input data into usable data. Statistics are facts, and valuable information is any knowledge that is important to individuals, technology, or systems. Machine learning algorithms work in a comparable pattern that uses a process and some mathematics. In general, not all Machine Learning Algorithms have the same mathematical transformation. This research study, on the other hand, contains the most significant machine learning algorithms, as well as relevant algorithmic processes in the complete system architecture.

### **3.5.1 Decision Tree**

Decision trees can be used to approach classification and regression problems in supervised learning. They are, perhaps, most often developed for solving classification problems. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node represents the conclusion in this tree-structured classifier. To do so, you'll need to use the Decision Tree to create a training model that employs standard

decision rules to predict the class or value of input variables based on the training dataset.

$$E(S) = \sum_{i=1}^C -p_i - \log_2 p_i \dots \dots (i)$$

Here, S stands for the current state, and  $p_i$  indicates the probability of an event, I in state S, or the percentage of class I in the node state S.

### 3.5.2 Support Vector Machine

As the Support Vector Machine (SVM) is a machine learning approach, its performance must be continuously reviewed. It can be used to perform classification and regression problems, respectively. Even though it is used for various applications, it is most frequently implemented for categorization. It was discovered that the method's creators attempted but failed to plot data items in n-dimensional space using a graphing calculator. This study looked into data charting in two dimensions, with the researchers determining to implement the COVID-19 Positive or Negative and condition of patient techniques for their analysis.

### 3.5.3 K-Nearest Neighbors

By implementing a modest supervised machine learning methodology, the k-nearest neighbors' method (KNN) can be used to perform both classification and regression issues. KNN is easy to comprehend and put into practice. The fundamental theorem of KNN is the Distance measure. Because the dataset is divided into two classes, KNN is used in this classification. The exact formula for the K-Nearest Neighbor algorithm is generated from equation (ii).

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \dots \dots (ii)$$

### 3.5.4 Logistic Regression

Logistic regression is a supervised classification model. It's for datasets with quantitative model parameters and a category-specific value with discrete choices or categories. The positive category or result is assigned to 1 while the negative category

or result is assigned to 0. The matching model estimates how likely an occurrence is to correspond to category 1. Typically, logistic regression is not an ideal approach for multiclass classification. Rather, it has to be tweaked to accommodate multi-class categorization tasks. Splitting the multi-class classification issue into separate binary classification tasks and fitting a conventional logistic regression model per each sub-problem is a typical strategy for applying logistic regression to multi-class classifiers. One-vs-rest and one-vs-one wrapper concepts are examples of this sort of method. Multinomial Logistic Regression is a customized implementation of logistic regression that forecasts a multinomial likelihood (i.e. more than two categories) for every given case.

Multinomial logistic regression is indeed an extended logistic system that includes a collection of predictors  $X$  to calculate the likelihood for both the classes of a subjective endogenous class  $Y$ .

$$Pr(Y_{ik}) = Pr(Y_i = k | x_i; \beta_1, \beta_1, \dots, \beta_m) = \frac{\exp(\beta_{0k} + x_i \beta'_k)}{\sum_{i=1}^m \exp(\beta_{0k} + x_i \beta'_k)}$$

*with  $k = 1, 2, 3, \dots, m$*

Here  $\beta_k$  is indeed the  $k$ th classification of  $Y$ 's set of vectors of  $X$ 's predicted values.

However, the probability for every  $\beta_{0k} + q, q R$  and  $\beta_{0k} + q, q R_c$  are equal; this regression model is unidentifiable. As a result, a suitable normalization approach must be used.

### 3.5.5 Linear Discriminant Analysis

To estimate possibilities, the Linear Discriminant Analysis model employs Bayes' Theorem. They generate forecasts based on the likelihood that a current input dataset will fall into each of the classes. The resulting category is the one with the greatest chance, and the Linear Discriminant Analysis makes a prediction based on it. The probability is calculated using Bayes' Theorem, which predicts the likelihood of the output variable provided the entry. They also utilize the likelihood of every category and the possibility of each category's data:

$$P(Y = x | X = x) = \frac{[(P_{lk} \times f_k(x))]}{[(\text{sum}(P_{li} \times f_i(x)))]}$$

Where  $x$  denotes the entry.

$k$  denotes the output category.

$P_{lk} = \frac{N_k}{n}$ , which is the basic probability of each class in the training phase.

In Bayes' Theorem, it's also known as a probability value.

$f_k(x)$  is the possibility of  $x$  relating to class  $k$  as assessed by  $f_k(x)$ .

The  $f(x)$  is visualized by using the Gaussian Distribution function before being fed through into the formula above, yielding the following formulae:

$$D_k(x) = x * \left( \frac{mean}{\sum 2} \right) - \left( \frac{mean^2}{(2 * \sum 2)} \right) + \ln(P_{lk})$$

The classifier equation  $D_k(x)$  is known as the dependent variable for class  $k$  given entry  $x$ ,  $mean$ ,  $2$ , and  $P_{lk}$  are all determined from either the data and the class are determined as having an enormous number, which will be regarded in the final classifying.

### 3.5.6 eXtreme Gradient Boosting (XGBoost)

XGBoost is a decision-making Machine Learning approach that uses a gradient boosting architecture. It comprises a sequence of categorized and reverted (also known as CART) trees as ensemble learning, then improves tree performances by constructing a collection of trees that reduces a normalized goal function to the smallest possible value. Concepts such as split-wisdom discovery in each tree, memory compatible approximate techniques for identifying splits, and appropriate computation of gradient boosting techniques were used to create an approach with a significant computational performance and strong forecasting ability.

The XGBoost model may be written as follows for the set of data

$D = \{(x_i, y_i)\} (x_i \in R^m, y_i \in R, i = 1, 2, \dots, n)$  with  $n$   $m$  aspects:

$$y_i = \sum_{k=1}^k f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \dots, n) \dots \dots \dots (iii)$$

Here  $F = \{f(x) = w_{q(x)}\} (q: R^m \rightarrow \{1, 2, \dots, T\}, w \in R^T)$  exemplifies the collection of CART decision dendrogram,  $q$  denotes the tree structure of the sample map to the intermediate node,  $T$  refers to the collection of the intermediate node, and  $w$  indicates the real score of the intermediate nodes in the CART decision tree structure collection.

### 3.5.7 Random Forest

Random forests, also known as random decision-making forests, are a supervised learning approach that uses numerous decision-making trees to categorize, retrograde, and perform other tasks. For classification problems, the random forest output is the category picked by the preponderance of trees. For regression problems, the mean or average prediction of the different trees is presented.

$$\sigma = \sqrt{\frac{\sum_b^B (f_b(x') - f)^2}{B - 1}} \dots \dots \dots (iv)$$

The experimental value, B, is an uncharged variable in formula (vi). Hundreds to thousands of trees are commonly used, based on the scale and nature of the workshop. The average validation loss for each sample of training X', and just the trees with no X' in their random subset, can help determine an ideal number of B trees by validation set or analysis of the out-of-bag error. The training and validation mistakes start to degrade after a few trees are fitted.

### 3.5.8 Gaussian Naïve Bayes

Gaussian Naive Bayes is a variant of Naive Bayes that allows for Gaussian normal distribution and continuous data. Naive Bayes is a set of related supervised learning algorithms for classification algorithms based on the Bayes theorem. The categorizing method is simple but effective. When working with continuous data, it's acceptable to assume that the values for each class follow a normal (or Gaussian) distribution. The complete formulation approach for the

Gaussian Naive Bayes algorithm is derived in equation (v).

$$P(x_i|y) == \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\pi\sigma_y^2}\right) \dots \dots \dots (v)$$

### 3.5.9 AdaBoost

Adaptive classification boosting, usually known as AdaBoost classification, is a group learning method. It constructs a strong classifier from a collection of weak classifiers. In this strategy, a weak classifier learns from its prior classifier errors. Take a look at the n sample dataset. Initially, each sample is given a weight of 1/n. With this dataset, a weak classifier is created.

The total  $\epsilon$  error is calculated by this classifier. And the influence of such a classification system is evaluated by the overall inaccuracy in data sample categorization.

$\alpha = \frac{1}{2} \ln \left( \frac{1-\epsilon}{\epsilon} \right) \dots \dots \dots (vi)$  This is used in equation (ix) to modify the weight of the dataset samples, resulting in the creation of a new dataset.

### 3.5.10 Stochastic Gradient Descent

The Stochastic Gradient Descent (SGD) approach is a straightforward yet effective way to learn linear and regressive classifiers using convex loss functions, such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around for a long time in the field of machine learning, it has only recently gained popularity in the context of large-scale learning. A basic stochastic gradient descent learning process is provided by the class SGD Classifier, which supports a range of classification loss functions and penalties.

The hinge loss, which is identical to a linear SVM, is used to train an SGDClassifier. The SGD approach's mathematical fundamentals are detailed here. Here are several instructive instances, such as  $(x_1, y_1) \dots \dots (x_n, y_n)$  where  $x_i \in R^m$  and  $y_i \in R$  ( $y_i \in -1, 1$  for classification). This system must learn a linear scoring function  $f(x) = w^T x + b$  with model parameter  $w \in R^m$  and intercept  $b \in R$ . The system must consider the sign-off while producing binary classification predictions  $f(x)$ . The normalized training error given by the equation is used to discover model parameters (vi).

$$E(x, b) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \alpha R(w) \dots \dots \dots (vii)$$

where L is a loss function that examines model fit and R is a normalization term that prevents the model from becoming complicated;  $\alpha > 0$  is a non-negative linear combination that determines the normalization strength.

### 3.5.11 Linear SVC

A Linear SVC (Support Vector Classifier) is a machine learning technique that is meant to match the data presented, culminating in the "best fit" hyperplane that separates or classifies the data as per its categorization. It may be necessary to submit certain



attributes to the classifier after the hyperplane has been created to obtain the "projected" classifier performance. This enables this method uniquely suited to a specific goal; nonetheless, it can be applied to a variety of situations. When SVC is given training examples, it receives two arrays of shape as input: an array X of shape holding the training data and an array y of shape class labels, both of which are groupings of form.

### 3.5.12 Perceptron

The perceptron algorithm is a type of neural network model that is applied as a method in setting a binary classification system. It is made from a single cluster or neuron that takes data from a row and forecasts how the data will be classified into classifications. In this approach of creating forecasts, a forecast function and a feature vector, both of which are linear classification techniques, are used to make predictions. The training of a threshold function that converts an input x to an output f(x) in the form of a binary value is explained in this section. The input x's binary value is mapped to the output f's binary value (x). The entire functionalized technique for the Perceptron algorithm is obtained in equation (v).

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0, \\ 0 & \text{otherwise} \end{cases} \dots \dots \dots (viii)$$

### 3.6 Proposed System

After considering all of the aforementioned algorithms, the desired system can be proposed. A system diagram, as illustrated in Figure 3.1, is better appropriate for understanding the system's detailed procedure

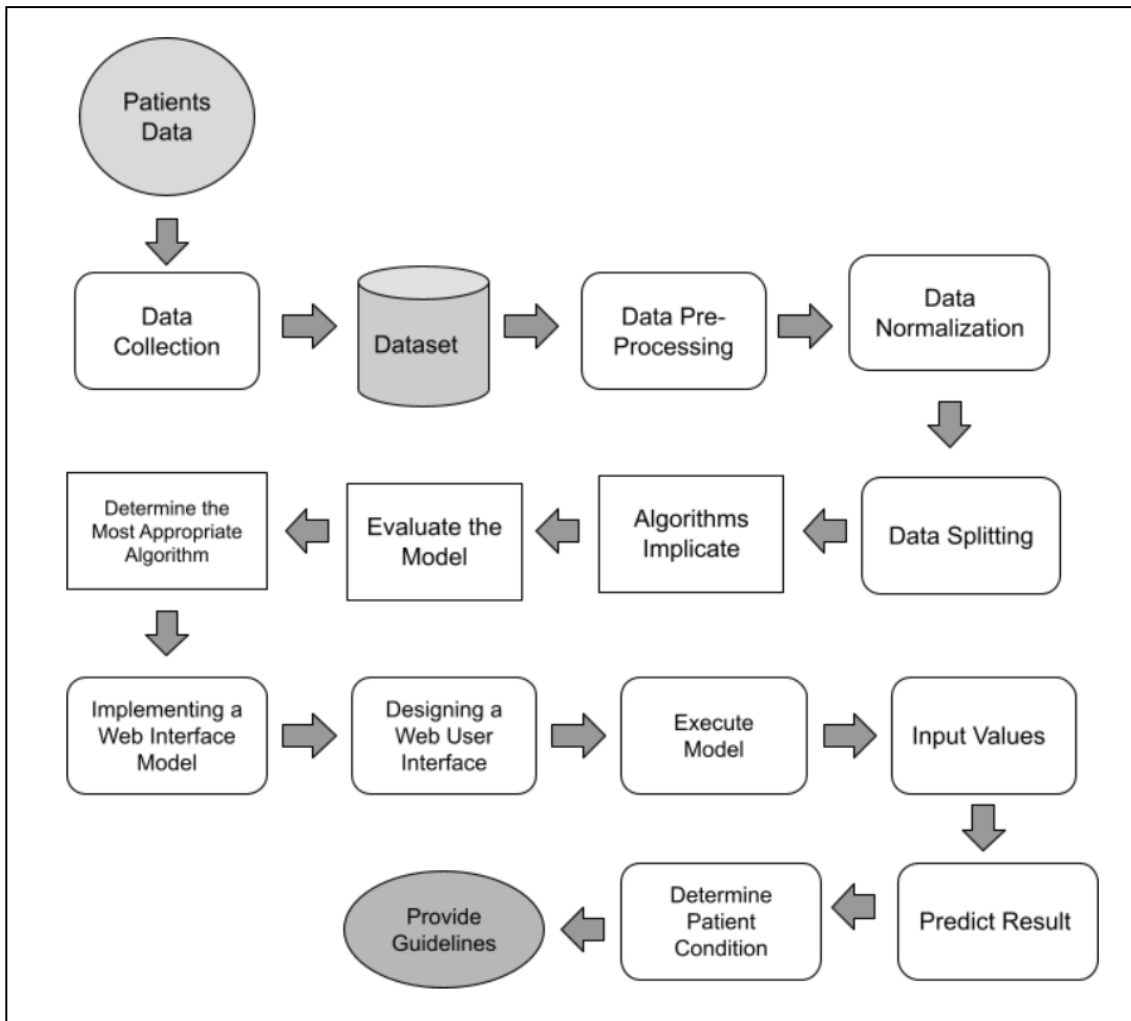


Figure 3.1: Proposed Method to Predict COVID-19

### 3.6.1 Data Collection

To predict COVID-19 and provide immediate guidelines to COVID-19 patients, the system required real-life data. Since very few hospitals in Bangladesh have tested and provided treatment to COVID-19 patients, so, in the context of this study, we collected clinical data from Bangabandhu Sheikh Mujib Medical University (BSMMU), which is the most modern COVID-19 hospital in the country. Also, we collected data through google form and social media from COVID-19 infected patients. collecting all of the required data from various sources is the initial step in this study. Following that, all of the data was combined into a single CSV file for analysis and interpretation.

### **3.6.2 Dataset**

The frequency of rows remained ideal for integrating various Machine Learning Algorithms after combining all of the combined data into a digital Comma Separated Value file. To forecast anything, machine learning techniques require a large amount of data. There are some discrepancies in the original data, which has 3039 rows and 23 columns.

### **3.6.3 Data Pre-processing**

The dataset had to be changed because it had a lot of missing entries and nominal data. First and foremost, the nominal data were transformed into numerical data. Then there's the issue of filling in the blanks. The mean value was used to fill in all of the empty variables. The dependent variables were X and Y, while the predictor variables were X and Y.

### **3.6.4 Data Normalization**

Normalization is the process of transforming quantitative values to a scale based while keeping the value variations intact. Min-max scaler was used as most of the features are in binary form. For improved accuracy, the predictor variables (X) were then standardized.

### **3.6.5 Data Splitting**

The set of data must be separated into two parts: training and testing, even before the machine learning method can be used. The model has been tested using 20% of the data, then trained with the rest 80%. The particular model may be developed to predict anything using the Training half of the dataset, and the Testing component can then be used to assess how precisely the data is being forecasted.

### **3.6.6 Algorithm implicate**

To determine the best accuracy and choose the optimal method, twelve different algorithms have been used. K-nearest Neighbour, Decision Tree, Random Forest, AdaBoost, XGBoost, Stochastic Gradient Descent, Linear SVC, Perceptron, Naive

Bayes, Support Vector Machines, Logistic Regression, Discriminant Analysis are the names of the twelve methods. Different analytical outcomes were discovered using all of these approaches.

### **3.6.7 Model Analysis**

The data were transformed into tables after evaluating the Confusion Matrix, Accuracy Score, Jaccard Score, Cross Validated Score, AUC Score, Mean Absolute Error, and Mean Squared Error for all of the algorithms. The confusion matrix summarizes the accuracy with which the data is predicted. The percentage of accuracy of predicted data is provided by the Accuracy score, Jaccard Score, Cross Validated Score, AUC Score. The algorithms' error rate is calculated using Mean Absolute Error and Mean Squared Error.

### **3.6.8 Extract Appropriate Algorithm**

By evaluating and analyzing all of the essential information from the tables, the optimum method was discovered. In that given dataset, the extracted method has the highest accuracy rate and lowest error rate. First and foremost, an appropriate algorithm must be designed to make effective use of the dataset. In this instance, it's better to use a variety of algorithms as models and then select the best one. In this study, many analytical criteria were used to find the most effective method, including accuracy score, Jaccard score, cross-validation score, AUC score, and so on. The best strategy that has been demonstrated to be acceptable for the Clinical Dataset of COVID-19 in this study has been the Decision tree Classifier. It achieved the highest scores in each of the above-mentioned criteria. The procedure then moves on to the next stage after choosing the appropriate algorithm.

### **3.6.9 Creating Model for Web Interface**

The team develops a web-based interface after selecting the optimal model. "joblib" was used to link the interface to the proposed method. It's a component of the SciPy ecosystem and includes Python task pipeline capabilities. It includes tools for quickly storing and loading Python structures that use NumPy data structures. Some machine

learning methods that need a large number of features or the storage of the complete dataset may benefit from it. Joblib is a Python library that stores the machine learning model called finalized model.sav and separates every NumPy array inside the model into its file. For developing a model, an object is needed that corresponds to that algorithm. The trained dataset is then utilized to train the model using the .fit() method. The model will be able to use it after being trained with the proper algorithm.

### **3.6.10 Building a Web Interface**

The developers used the Python "Django" framework to create the web interface. To design a user-friendly interface, basic HTML, CSS, Bootstrap, and JavaScript have also been used. The joblib file was linked to the website inside the website's backbone.

### **3.6.11 Execute Model**

The model must be stored in the website's root folder when it has been created. This is where joblib comes into play. Joblib is a SciPy ecosystem component that adds Python job pipeline features. As a result, Python objects are transformed to byte streams, which may have been stored in data files, utilized to maintain the sequence of operations between sessions, and even sent over the internet. The model's name is specified in the views.py file, and then the file is loaded and the model's method is utilized to forecast the output. This is how the model is developed and saved to a folder.

### **3.6.12 Input Values**

After building the model and integrating it with the Django framework, it can be used to quickly create an interface for predicting the COVID-19. Django is a Python development framework that encourages rapid growth and practical design. It was created by professional developers to address the issue of web development hassles so that developers may focus on developing apps rather than reinventing the wheel. It is freely available. The input data is collected using a form created with the assistance of jQuery. It's a simpler method to develop a web interface because Django does the majority of the work for you.

### 3.6.13 Predictive Result

The interface helps any user to provide necessary data (user's symptoms) into the website to predict COVID-19. If a patient had a COVID-19, the website will show a positive result and determine his/her condition. If he is in the early stages of the infection, the system will give him primary guidelines, and if he is in an emergency condition, it will suggest going to Corona service provider hospital. And if the website will display a negative result and tell him that everything is alright.

### 3.7 System Architecture

To enhance the practical formation in machine learning technique and web implementation, a system architecture formation is needed to assume the complete project system. A basic system architecture that is a wider representation of the proposed system is shown in Figure 3.2

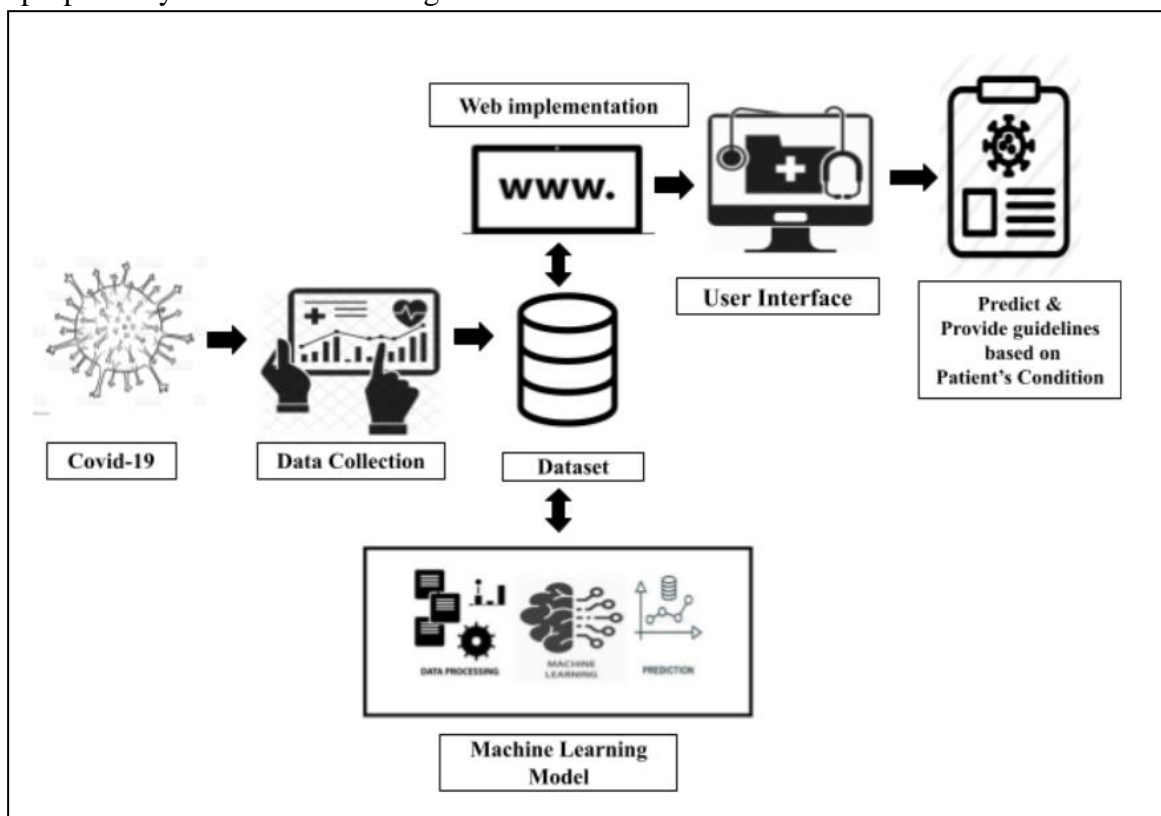


Figure 3.2: System Architecture

### **3.7.1 User Segment**

The user section depicts how a user uses a machine to determine whether or not she is infected with the Coronavirus and to examine his or her health condition. A user will use a device such as a laptop to enter basic health information, symptoms, and other information. The system will then predict whether or not the person is COVID-19 infected. It also gives her some important guidelines depending on her condition. The user section has been kept as simple as possible. The user segment's interface is quite user-friendly; therefore, anyone can use this system. All that is required is to enter the value and click the submit button, and the computer will predict the patient's present state.

### **3.7.2 Web Insider**

The developers used the Python "Django" platform to develop the web interface. To design a user-friendly UI, basic HTML, CSS, Bootstrap, and JavaScript were also required. With the aid of python, the joblib file was linked to the website in the website's backend. The user interface makes it easy for anybody to enter the information needed to forecast COVID-19 onto the website. The website will indicate a positive with the normal condition or a positive with the emergency condition if a patient has been infected with COVID-19. The appropriate recommendations will be displayed based on the data entered by the user and the forecasted condition. And if the patient is expecting a good outcome, the website will show him a negative test and assure him that everything is well. In Figure 3.3, system architecture of web interface is shown.

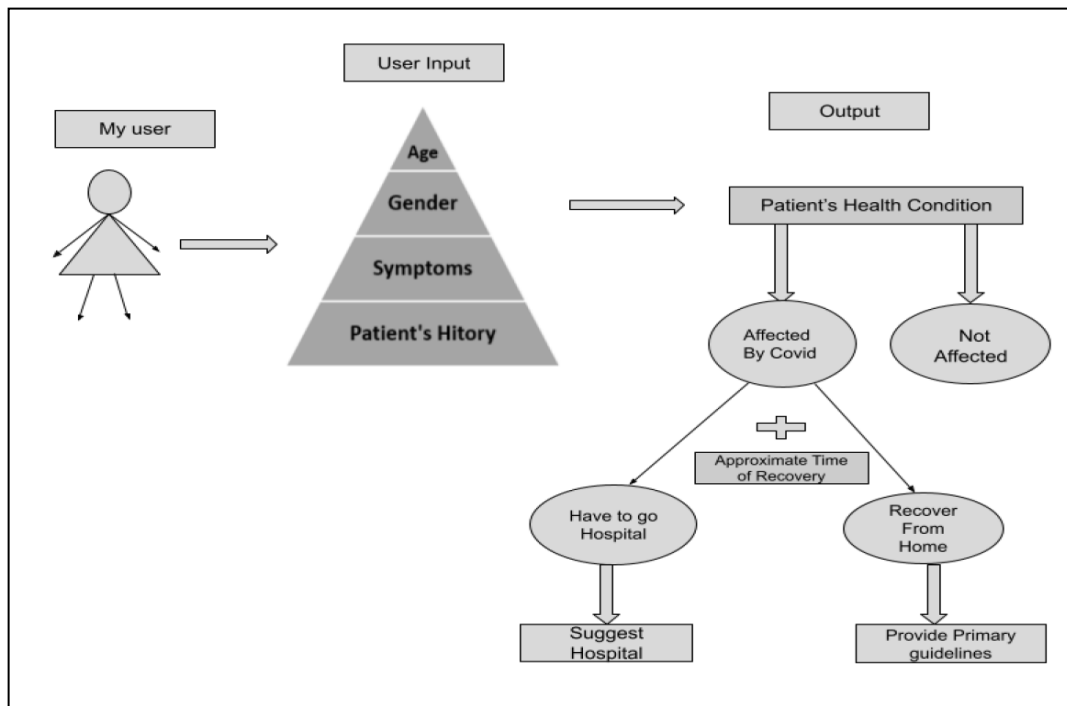


Figure 3.3: System Architecture of Web Interface

### 3.7.3 Machine Learning Model

By analyzing and monitoring all of the essential findings from the tables, the optimum method was discovered. Within this given dataset, the retrieved method has the highest accuracy rate and lowest error rate. It was important to create a web-based application after identifying the optimal algorithm. "joblib" was used to link the interface to the best algorithm. Joblib is a Python package that saves the "finished model.sav" machine learning model. In this scenario, the primary filename has been used. In this work, the `joblib.load()` module is used to load the .sav file in Python.



## CHAPTER 4

### EXPERIMENTAL RESULTS & DISCUSSION

#### 4.1 Introduction

The result is essential for every type of research and project. So a project also contains the outcome. All of the results in this chapter are presented in the table. This chapter explains the dataset for COVID-19 in-depth, including data collecting, data usage, and feature significance. Then, to demonstrate the results of various algorithms, a confusion matrix table was created. Precision, recall, and F1-Score were also presented in a table with the assistance of a classification report. Different graphs presented the accuracy, Jaccard score, cross-validation score, AUC score, and ROC curve. For the convenience of explanation, the details have been organized in a table. The standard deviation is sometimes shown as a table. Finally, the problem was shown in a tabular format. The output display was supplied as a screenshot in this chapter because this project requires a web implementation.

#### 4.2 Experimental Results

Following the successful execution of Machine Learning Model Development, each algorithm demonstrated its unique accuracy and scores, leading the best method to predict COVID-19 to be identified. As a consequence, the experimental outcomes section is an analytical part where all potential scores for all algorithmic applications and procedures may be examined.

##### 4.2.1 Data Acquisition

The Dataset was used to train the algorithm collected from Bangladesh's most advanced COVID-19 hospital, Bangabandhu Sheikh Mujib Medical University (BSMMU). The hospital records the daily patient's information who were tested for COVID-19 from all over the country. Also, through a google form, social media, and other sources to collect data. The model was trained using around 3038 tested individual data samples. Each class has three numerical values: COVID-19 Negative, If the result is positive then it shows normal or emergency according to the condition of the patient. Condition

of 55.57% data of those who resulted positive from the COVID-19 test is used in this dataset shown in Table 4.1.

Table 4.1: COVID-19 Result frequency of the patients

<b>Not Affected</b>	<b>Affected</b>	
44.42%	55.57%	
	<b>Cabin</b>	<b>Emergency</b>
	33.92%	21.65%

To predict COVID-19 and get the result, various information is available in the dataset, including basic information like age, gender, etc., and clinical symptoms, and some basic questions about the patient's history. There are 22 significant components or features in each sample in the COVID-19 dataset, of these, 2 of Numerical (Age & Oxygen Saturation (%)) plus a target variable (class), and the remaining 20 are nominal variables. Each class has three numerical values: COVID-19 Negative, If the result is positive then it shows normal or emergency according to the condition of the patient. There are a few missing values in the dataset. A brief description of the dataset may be found in Table 4.2

Table 4.2: Data Acquisition & Null Percentage

<b>Feature</b>	<b>Scale</b>	<b>Data Type</b>	<b>Missing Values (%)</b>
Age	Age in year	Numerical	0.69
Gender	(Male, Female)	Nominal	0.43
Front Liner	(Yes, No)	Nominal	0.39
Travel History	(Yes, No)	Nominal	0.33
Went Crowd	(Yes, No)	Nominal	0.56
Comorbid	(Yes, No)	Nominal	0.3
Contact Infected COVID-19	(Yes, No)	Nominal	0.43
Fever	(Yes, No)	Nominal	0.13

Dry Cough	(Yes, No)	Nominal	0.56
Sore Throat	(Yes, No)	Nominal	0.16
Diarrhea	(Yes, No)	Nominal	0.26
Tiredness	(Yes, No)	Nominal	0.3
Headache	(Yes, No)	Nominal	0.26
Muscle Pain	(Yes, No)	Nominal	0.36
Rash on Skin	(Yes, No)	Nominal	0.33
Chest Pain	(Yes, No)	Nominal	0.3
Shortness of Breath	(Yes, No)	Nominal	0.33
Lost Taste or Smell	(Yes, No)	Nominal	0.33
Loss Speech or Movement	(Yes, No)	Nominal	0.26
Oxygen Saturation (Less than 92%)	In percentage	Numerical	0.53
Unable to Complete Short Sentence	(Yes, No)	Nominal	0.43
Any Symptoms Manifested From	(Yes, No)	Nominal	0.39
Condition	(Neg, Norm, Emgr)	Numerical	0.1

#### 4.2.2 Data Utilization

The data in a computer system might be more easily handled if each category (nominal) variable was coded independently. The yes/no responses were classified as 1 and 0, with 1 representing yes and 0 representing no. To determine the value of gender, male gender was given a value of 0, and female gender was given a value of 1. A factorization approach was used to transform all of the categorical variables. To differentiate the samples, a random number between 1 and 3038 was assigned to each. There were a lot of blank spaces in this dataset. For several different reasons, patients may ignore critical measures before a diagnosis is determined. As a result, if the diagnostic categories of the samples are unknown, missing values will appear in the data, requiring the use of an appropriate imputation approach. After the categorical variables had been encoded,

the missing values of the main COVID-19 dataset were processed and filled. For a better understanding of the dataset, the data descriptions for the 22 attributes were extracted. Table 4.3 shows the Count, Mean, Standard Deviation, Min, 25%, 50%, 75%, and Max values derived from the dataset.

Table 4.3: Dataset Description

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Age	3038	48.66809	19.6265	0.8	34	51	64	95
Gender	3038	0.653209	0.475003	0	0	1	1	1
FL	3038	0.040316	0.196346	0	0	0	0	1
TH	3038	0.359645	0.479184	0	0	0	1	1
WC	3038	0.477999	0.498198	0	0	0	1	1
Comorbid	3038	0.335109	0.471401	0	0	0	1	1
CIC	3038	0.332222	0.470082	0	0	0	1	1
Fever	3038	0.633483	0.481614	0	0	1	1	1
DC	3038	0.526995	0.497955	0	0	1	1	1
SR	3038	0.425329	0.494066	0	0	0	1	1
Diarrhea	3038	0.356116	0.478295	0	0	0	1	1
Tiredness	3038	0.361172	0.479708	0	0	0	1	1
Headache	3038	0.384806	0.485991	0	0	0	1	1
MP	3038	0.396116	0.48828	0	0	0	1	1
RS	3038	0.179987	0.383607	0	0	0	0	1
CP	3038	0.284569	0.450621	0	0	0	1	1
SOB	3038	0.292923	0.548944	0	0	0	1	9
LTS	3038	0.36659	0.481156	0	0	0	1	1
LSM	3038	0.165688	0.371361	0	0	0	0	1
Oxygen	3038	0.251481	0.432796	0	0	0	0.8125	1

UCSS	3038	0.141481	0.347835	0	0	0	0	1
ASMF	3038	0.296445	0.455856	0	0	0	1	1
Condition	3038	0.76761	0.782996	0	0	1	1	2

### 4.2.3 Feature Importance

Methods that assign a value to input qualities based on their predictive capacity for a target variable are referred to as "feature importance." The phrase "feature importance" refers to a collection of methods for assigning scores to the input features of a predictive model, representing the relative importance of each information when making a prediction. The feature significance score may be used to get insight into the dataset and model, as well as to develop a prediction model. The feature relevance of different qualities for different algorithms is represented in Table 4.4.

Table 4.4: Feature Importance

Attribute	Algorithms					
	Decision Tree	Random Forest	XGBoost Classifier	Logistic Regression	AdaBoost Classifier	Discriminate Analysis
Age	0.22782	0.19024	0.02133	-2.80550	0.49000	-2.96580
Gender	0.02682	0.03448	0.01445	-0.10162	0.01000	-0.10949
Front Liner	0.01403	0.01518	0.03911	-1.94800	0.01000	-1.05872
Travel History	0.03052	0.02663	0.02597	-0.44580	0.03000	-0.36033
Went Crowd	0.03250	0.03716	0.02832	-0.59514	0.02000	-0.70985
Comorbid	0.03805	0.03917	0.03473	0.14728	0.03000	0.10714
Contact Infected COVID-19	0.05373	0.03710	0.02958	-0.31989	0.03000	-0.36391
Fever	0.01388	0.03368	0.02010	-0.50575	0.05000	-0.29111
Dry Cough	0.02786	0.03568	0.02029	-0.67819	0.02000	-0.56960
Sore Throat	0.03102	0.03444	0.02042	-0.54766	0.02000	-0.40234
Diarrhea	0.02809	0.03733	0.02512	-0.58325	0.01000	-0.47626
Tiredness	0.02185	0.03392	0.02095	-0.52628	0.03000	-0.44147
Headache	0.02239	0.03753	0.02319	-0.56108	0.01000	-0.39805
Muscle Pain	0.04096	0.03703	0.02436	-0.46665	0.01000	-0.38199
Rash on Skin	0.02203	0.02173	0.02926	0.67964	0.02000	0.38718
Chest Pain	0.00800	0.03014	0.01720	-0.29442	0.03000	-0.27213
Shortness of Breath	0.02448	0.04881	0.03486	0.39784	0.03000	-0.31985

Lost Taste or Smell	0.02809	0.04132	0.02419	-0.55875	0.01000	-0.38982
Loss Speech or Movement	0.03235	0.04812	0.06063	-2.39234	0.04000	-0.69895
Oxygen Saturation (Less than 92%)	0.21574	0.10855	0.42662	-3.17999	0.03000	-1.93800
Unable to Complete Short Sentence	0.01957	0.03358	0.02889	-0.40743	0.03000	-0.30628
Any Symptoms Manifested From	0.04023	0.03817	0.03040	-0.78118	0.04000	-0.61801

### 4.3 Result & Discussion

In this study, the value of Normal and Emergency (patient's condition) was made positive, whereas the value of Covid Negative was made negative. The confusion matrix has been used to demonstrate desired outcomes and evaluate the efficacy of machine learning algorithms. The template for the confusion matrix for various algorithm types is shown in Table-4.5.

#### 4.3.1 Confusion Matrix

It is necessary to develop a confusion matrix to validate the results of the implementation phase. A Confusion matrix is a  $N \times N$  matrix used with  $N$  target classes for evaluating the performance of a classification model. The matrix compares the actual target values to the predicted ones to determine the machine learning model's accuracy. This indicates how efficiently the algorithmic model is performing and what errors it is making. With some mathematical equations, it will be possible to determine Precision, Recall, and Accuracy using binary classification. Furthermore, for multi-class classification, these average values must be analyzed using a micro or macro average. Before we get into this, it's vital to understand the four-building pieces that go into calculating various measurement parameters. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the four types of false positives and false negatives (FN). The actual development of the Confusion Matrix,

which includes True Positive, False Positive, False Negative, and True Negative, is shown in Table 4.5. The confusion matrix for each algorithm is shown in Table 4.6.

Table 4.5: Confusion Matrix

Confusion Matrix			
	Actual Class		
Predicted Class	True Positive (TP)	False Negative (FN)	False Negative (FN)
	False Positive (FP)	True Negative (TN)	True Negative (TN)
	False Positive (FP)	True Negative (TN)	True Negative (TN)

Table 4.6: Confusion Matrix for Algorithms

Algorithm	Confusion Matrix				Confusion Matrix Percentage			
		Negative	Normal	Emergency		Negative	Normal	Emergency
KNN	Negative	275%	3%	0%	Negative	45%	1%	0%
	Normal	11%	184%	2%	Normal	2%	30%	0%
	Emergency	0%	7%	126%	Emergency	0%	1%	21%
Decision Tree	Negative	277%	0%	1%	Negative	46%	0%	0%
	Normal	7%	186%	4%	Normal	1%	31%	0%
	Emergency	0%	2%	131%	Emergency	0%	0%	22%
Random Forest	Negative	278%	0%	0%	Negative	46%	0%	0%
	Normal	6%	187%	4%	Normal	1%	31%	0%
	Emergency	0%	2%	131%	Emergency	0%	0%	22%
AdaBoost	Negative	254%	18%	6%	Negative	42%	3%	1%
	Normal	61%	94%	42%	Normal	10%	15%	7%
	Emergency	3%	12%	118%	Emergency	1%	2%	19%
XGBoost	Negative	278%	0%	0%	Negative	46%	0%	0%

	Normal	6%	188%	3%	Normal	1%	31%	0%
	Emergency	0%	2%	131%	Emergency	0%	0%	22%
Stochastic Gradient Descent	Negative	265%	13%	0%	Negative	44%	2%	0%
	Normal	88%	83%	26%	Normal	15%	14%	4%
	Emergency	14%	33%	86%	Emergency	2%	5%	14%
Linear SVC	Negative	248%	30%	0%	Negative	41%	5%	0%
	Normal	56%	113%	28%	Normal	9%	18%	5%
	Emergency	5%	29%	99%	Emergency	1%	5%	16%
Perceptron	Negative	249%	29%	0%	Negative	41%	5%	0%
	Normal	64%	116%	17%	Normal	10%	19%	3%
	Emergency	5%	58%	70%	Emergency	1%	10%	11%
Naive Bayes	Negative	272%	6%	0%	Negative	45%	1%	0%
	Normal	129%	35%	33%	Normal	21%	6%	5%
	Emergency	9%	18%	106%	Emergency	2%	3%	17%
Support Vector Machines	Negative	259%	19%	0%	Negative	43%	3%	0%
	Normal	31%	158%	8%	Normal	5%	26%	1%
	Emergency	1%	23%	109%	Emergency	0%	4%	18%
Logistic Regression	Negative	243%	35%	0%	Negative	40%	6%	0%
	Normal	57%	116%	24%	Normal	9%	19%	4%
	Emergency	3%	31%	99%	Emergency	1%	5%	16%
Discriminant Analysis	Negative	247%	31%	0%	Negative	40%	5%	0%
	Normal	61%	108%	28%	Normal	10%	18%	4%
	Emergency	6%	25%	102%	Emergency	1%	4%	17%



### **A. True Positive (TP)**

Optimistic tuples are those that have been correctly categorized by the classifier. The letter TP in the acronym indicates that 46% of the values in XGBoost, Random Forest, and Decision Tree were True Positive (TP) values. KNN and Naïve Bayes came in second with 45% and Logistic Regression third with 44%.

### **B. True Negative (TN)**

Positive tuples that were improperly classified by the classifier are known as negative tuples. The letter TN can be used to represent these situations. Decision Tree, XGBoost and Random Forest have a value of 53%, followed by KNN which has True Negative (TN) scores of 52%. And there's 45% TN of Discriminate Analysis.

### **C. False Positive (FP)**

Today's concentration is on these negatively labeled tuples that were incorrectly identified as positive by the classifier. This sort of connection can be denoted by the symbol FP. Random Forest, Decision Tree, and XGBoost provided fewer False Positive (FP) values in this section of the investigation, with just 1%. KNN contributed 2% of the False Positive (FP) results, whereas SVM generated 5%.

### **D. False Negative (FN)**

These positive tuples were misclassified as negative by the classifier. The letter FN stands for it. The Random Forest, Decision Tree, and XGBoost False Negative (FN) values had a 0% result. Then there's KNN and Naive Bayes, both of which have 1% False Negative (FN) values, and SGD, which has 2% FN values.

### **E. Precision**

Precision may be considered as a statistic for measuring how accurate something is. To look at it another way, it's the proportion of retrieved events that are truly relevant. The mathematical formula for measuring precision is shown in Equation (x).

$$Precision = \frac{TP}{TP + FP} \dots \dots (ix)$$

## F. Recall

In machine learning, it is a measure of completeness the number of optimistic tuples recognized as such. In comparison to the total number of relevant cases, a relevant instance is defined as the fraction of relevant examples that have been found. The mathematical formula for determining Recall is shown in Equation (xi).

$$Recall = Sensitivity = \frac{TP}{TP + FN} \dots \dots \dots (x)$$

## G. F1-Measure

The visual mean, often known as the F measure, is a measurement that evaluates a test's accuracy and recall. The mathematical formula for measuring F1-Measure is shown in Equation (xii).

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots \dots \dots (xi)$$

## H. Accuracy

The proportion of test set tuples correctly recognized by the classifier on that test set is the accuracy of a classifier on that test set. It may be easier to evaluate accuracy using equation (xiii).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \dots \dots \dots (xii)$$

### 4.3.2 Classification Report

In machine learning, a classification report is a statistic that is used to analyze the performance of the system. It's used to evaluate the accuracy, recall, F1 Score, and support of a trained classification model. It's a standard methodology for a classification-based machine learning model. The model's accuracy, recall, F1 score, and support are all represented. It provides a more accurate representation of the overall performance of the trained model. To be evaluated, classification results from machine learning models require knowledge of all metrics given in the research study. The

classification report for each method is shown in Table 4.7, which includes Precision, Recall, F1-Score, and Accuracy percentage value.

Table 4.7: Classification Report

Algorithm	Class	Precision	Recall	F1-Score	Accuracy (%)
KNN	Negative	0.96	0.99	0.98	96.22
	Cabin	0.95	0.93	0.94	
	Emergency	0.98	0.95	0.97	
	Macro Avg.	0.96	0.96	0.96	
	Weighted Avg.	0.96	0.96	0.96	
Decision Tree	Negative	0.98	1.00	0.99	97.70
	Cabin	0.99	0.96	0.98	
	Emergency	0.98	0.98	0.98	
	Macro Avg.	0.98	0.98	0.98	
	Weighted Avg.	0.98	0.98	0.98	
Random Forest	Negative	0.98	1.00	0.99	98.03
	Cabin	0.99	0.95	0.97	
	Emergency	0.97	0.98	0.98	
	Macro Avg.	0.98	0.98	0.98	
	Weighted Avg.	0.98	0.98	0.98	
AdaBoost	Negative	0.80	0.91	0.85	76.64
	Cabin	0.76	0.48	0.59	
	Emergency	0.71	0.89	0.79	
	Macro Avg.	0.76	0.76	0.74	
	Weighted Avg.	0.77	0.77	0.75	
XGBoost	Negative	0.98	1.00	0.99	98.19
	Cabin	0.99	0.95	0.97	
	Emergency	0.98	0.95	0.98	
	Macro Avg.	0.98	0.98	0.98	
	Weighted Avg.	0.98	0.98	0.98	
Stochastic Gradient Descent	Negative	0.86	0.82	0.84	71.38
	Cabin	0.66	0.63	0.65	
	Emergency	0.71	0.83	0.77	
	Macro Avg.	0.75	0.76	0.75	
	Weighted Avg.	0.76	0.76	0.76	
Linear SVC	Negative	0.80	0.89	0.84	75.66
	Cabin	0.66	0.57	0.61	
	Emergency	0.78	0.74	0.76	
	Macro Avg.	0.75	0.74	0.74	
	Weighted Avg.	0.75	0.76	0.75	
Perceptron	Negative	0.78	0.90	0.84	71.55
	Cabin	0.57	0.59	0.58	
	Emergency	0.80	0.53	0.64	
	Macro Avg.	0.72	0.67	0.68	
	Weighted Avg.	0.72	0.72	0.71	

Naive Bayes	Negative	0.66	0.98	0.79	67.93
	Cabin	0.59	0.18	0.27	
	Emergency	0.76	0.80	0.78	
	Macro Avg.	0.67	0.65	0.61	
	Weighted Avg.	0.66	0.68	0.62	
Support Vector Machines	Negative	0.89	0.93	0.91	86.51
	Cabin	0.79	0.80	0.80	
	Emergency	0.93	0.82	0.87	
	Macro Avg.	0.87	0.85	0.86	
	Weighted Avg.	0.87	0.87	0.86	
Logistic Regression	Negative	0.80	0.87	0.84	75.33
	Cabin	0.64	0.59	0.61	
	Emergency	0.80	0.74	0.77	
	Macro Avg.	0.75	0.74	0.74	
	Weighted Avg.	0.75	0.75	0.75	
Discriminate Analysis	Negative	0.79	0.89	0.83	75.16
	Cabin	0.66	0.55	0.60	
	Emergency	0.78	0.77	0.78	
	Macro Avg.	0.74	0.73	0.74	
	Weighted Avg.	0.74	0.75	0.75	

## 4.4 Result Analysis

This is essential to go over the results analysis after calculating all feasible elements such as Precision, Recall, F1-Measure, Accuracy, and so on. It will be determined in this study which algorithm performs the best among all algorithms and which algorithm performs this when compared to others.

### 4.4.1 Accuracy

The accuracy of an algorithm is a measure of its optimum performance. Based on the facts presented to it, how well it operates. The performance of accuracy may be measured using a probabilistic technique. In comparison to the other methods, eXtreme Gradient Boosting is the most accurate, whereas Gaussian Naive Bayes is the least accurate. The eXtreme Gradient Boosting method has demonstrated that it is one of the most powerful and scalable implementations of gradient boosting machines, capable of pushing the limits of processing power for boosted trees algorithms. It was created with the primary objective of improving model performance and boosting the speed with which computers could analyze data. The accuracy chart and percentage of each method used to predict in this model are shown in Figure 4.1 and Table 4.8.

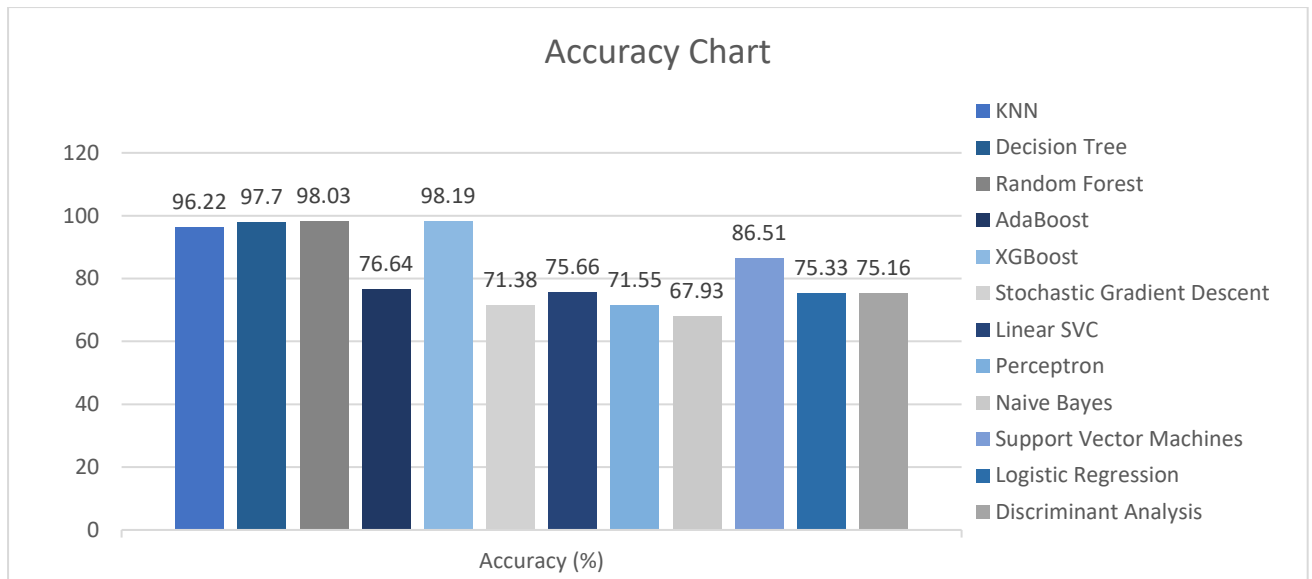


Figure 4.1: Accuracy Chart

### 4.4.2 Jaccard Score

The Jaccard score is a statistic used to determine the similarity and diversity of sample sets. They are comparable in terms of the Intersection to Union ratio. The Jaccard coefficient, which is defined as the intersection size divided by the union size, may be used to compare the similarity of two finite sample sets quantitatively. The accuracy chart and percentage of each method used to forecast in this model are shown in Equation (xiv), Figure 4.2, and Table 4.8.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \dots \dots \dots (xiii)$$

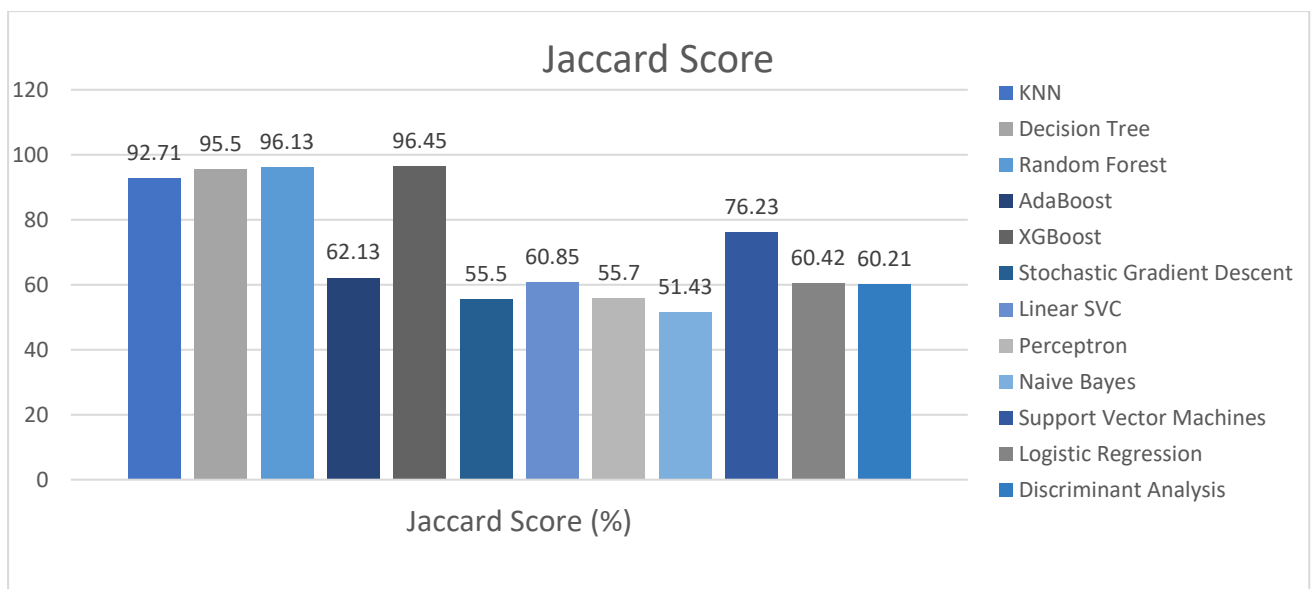


Figure 4.2: Jaccard Score Chart

### 4.4.3 Cross Validated Score

Machine learning models are evaluated using the statistical approach of cross-validation. Cross-Validation begins with the data being shuffled and divided into  $k$  folds. The data is then fitted using  $k$  models and  $\frac{1}{k}$  of the data is assessed. The final score is derived by averaging the results  $\left(\frac{k-1}{k}\right)$  of each evaluation, and the resulting model is then fitted to the whole dataset for implementation. The cross-verified score chart and percentage of each method used to estimate in this model are shown in Figure 4.3 and Table 4.8.

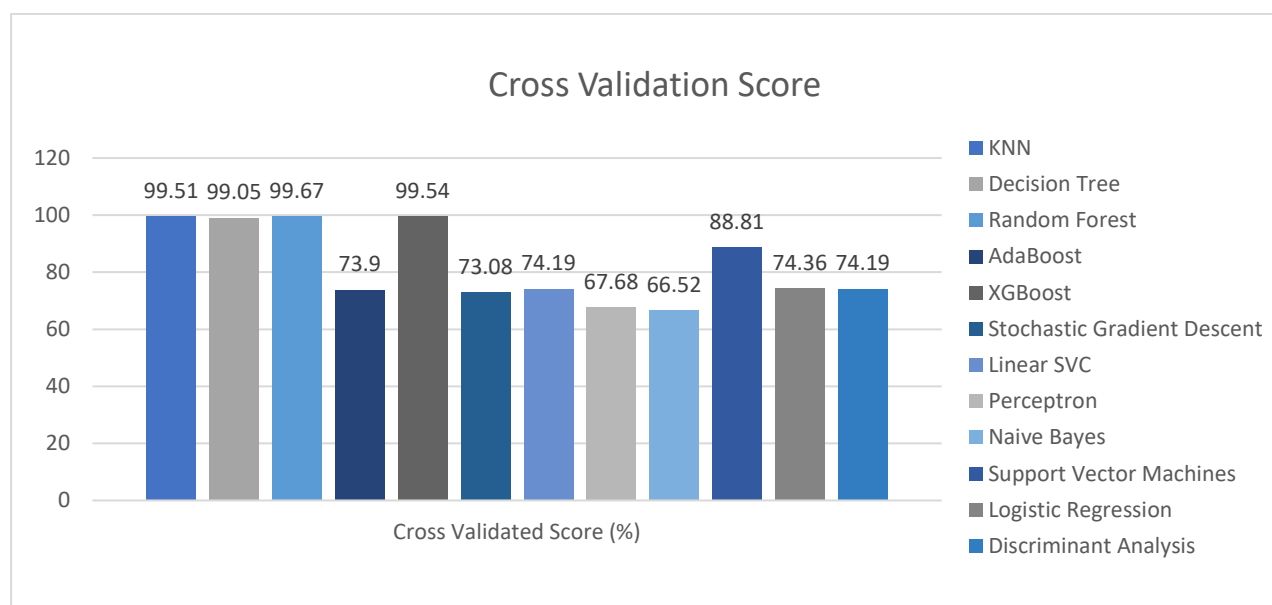


Figure 4.3: Cross Validated Score

Table 4.8: Accuracy, Jaccard and Cross Validated

Algorithm Name	Accuracy Score (%)	Jaccard Score (%)	Cross Validated Score (%)
KNN	96.22	92.71	99.51
Decision Tree	97.70	95.50	99.05
Random Forest	98.03	96.13	99.67
AdaBoost	76.64	62.13	73.90
XGBoost	98.19	96.45	99.54
Stochastic Gradient Descent	71.38	55.50	73.08
Linear SVC	75.66	60.85	74.19
Perceptron	71.55	55.70	67.68
Naive Bayes	67.93	51.43	66.52

Support Vector Machines	86.51	76.23	88.81
Logistic Regression	75.33	60.42	74.36
Discriminate Analysis	75.16	60.21	74.19

#### 4.4.6 Standard Deviation

The standard deviation may also be evaluated using the data from this study. A standard deviation indicates how far a dataset deviates from its mean. The standard deviation is calculated by computing the square root of each data point's variance. As the data points become further apart from the mean, the standard deviation increases. Table 4.9 displays the standard deviation for the top six algorithms.

Table 4.9: Standard Deviation

Algorithm Name	Standard Deviation
Decision Tree	0.06
Random Forest	0.04
AdaBoost	0.1
XGBoost	0.08
Logistic Regression	0.95
Discriminate Analysis	0.67

#### 4.4.7 Misclassification & Error

Any algorithm defect becomes a challenge when determining its efficiency. After misclassification, absolute error and mean square error are included in a machine-learning model's accuracy. When an inappropriate attribute is chosen, misclassification might occur. When all classes, groups, or categories of a variable have the same error rate, it is considered misclassification.

The absolute error refers to the level of inaccuracy in measurement. The average of all absolute mistakes in measurement is considered the Mean Absolute Error (MAE). The formula for Mean Absolute Error is represented by equation (xv).

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x| \dots \dots \dots (xiv)$$

The Mean Squared Error indicates how accurate a regression line is to a set of points (MSE). The formalized computation for Mean Squared Error is shown in Equation (xvi).

$$MSE = \frac{1}{n} \sum_{i=1}^n |y_i - y| \dots \dots \dots (xv)$$

The misclassification, mean absolute error and mean square error in the algorithms are shown in Table 4.10 and Figure 4.5. Misclassification, Mean Absolute Error, and Mean Squared Error all had lower error rates with the XGBoost, with 1.81 %, 1.81 %, and 1.81 %, correspondingly.

Table 4.10: Misclassifications & Errors

<b>Algorithm Name</b>	<b>Misclassification (%)</b>	<b>Mean Absolute Error (%)</b>	<b>Mean Squared Error (%)</b>
KNN	3.78	3.78	3.78
Decision Tree	2.30	2.47	2.80
Random Forest	1.97	1.97	1.97
AdaBoost	23.36	24.84	27.80
XGBoost	1.81	1.81	1.81
Stochastic Gradient Descent	28.62	30.92	35.53
Linear SVC	24.34	25.16	26.81
Perceptron	28.45	29.28	30.92
Naive Bayes	32.07	33.55	36.51
Support Vector Machines	13.49	13.65	13.98
Logistic Regression	24.67	25.16	26.15
Discriminate Analysis	24.84	25.82	27.80



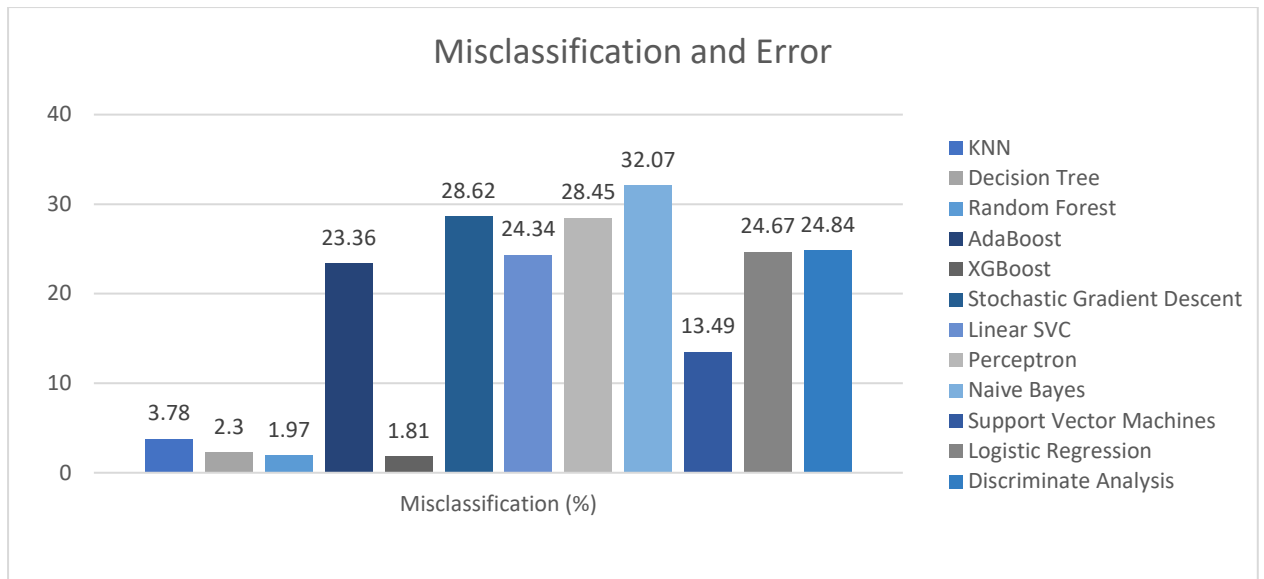


Figure 4.4: Misclassification and Error

## 4.5 Web Implementation

The development of a machine learning model was completed successfully, and the appropriate approach was found and executed. It's now time to demonstrate this model's implementation on the Web Interface, as identified in the following chapter's System Architecture section.

### 4.5.1 Web Interface

COVID-19 will be forecasted on a webpage that uses the "Django framework," as outlined in Chapter 3. As previously mentioned in chapter 3, a web interface based on "Django" will be constructed. A very appropriate and beneficial website has been designed to ensure that this assignment is accomplished effectively. Figure 4.5 depicts the Web Interface in question.

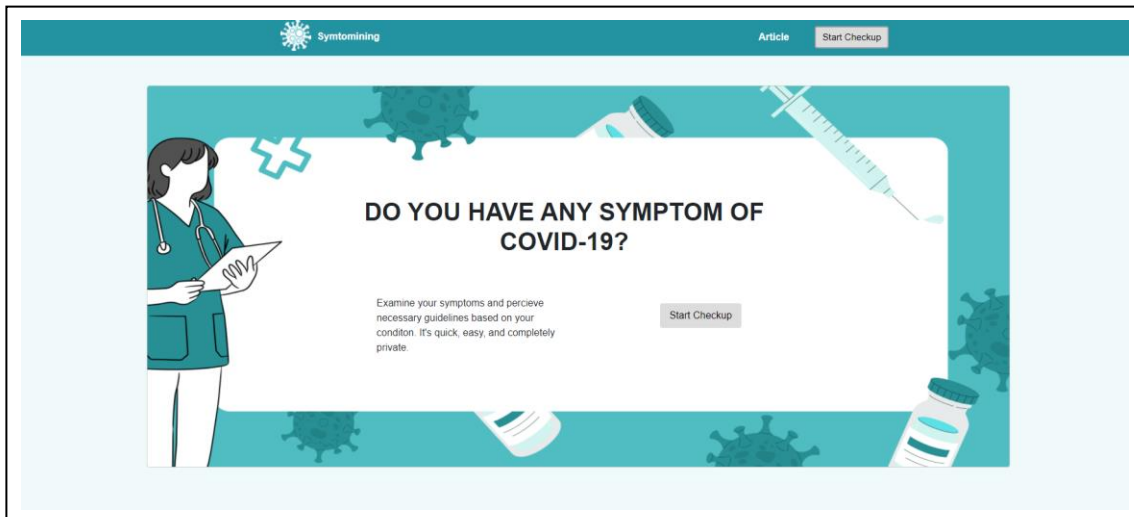


Figure 4.5: Web Interface

### 4.5.2 Web Output Analysis

Exactly three categorization alternatives are available based on this study approach and gathered data: Negative, Cabin, and Emergency, which represent for COVID-19 not infected, infected with the normal condition, and infected with an emergency condition, correspondingly. This part will require three separate result evaluations to be completed in their totality. In Figure 4.6, all data was entered at random and the desired outcome was obtained using a previously trained model to predict COVID-19. In addition, in Figures 4.7 and 4.8, all data was randomly input from the test case and the needed output was obtained as a previously trained model to forecast COVID-19.

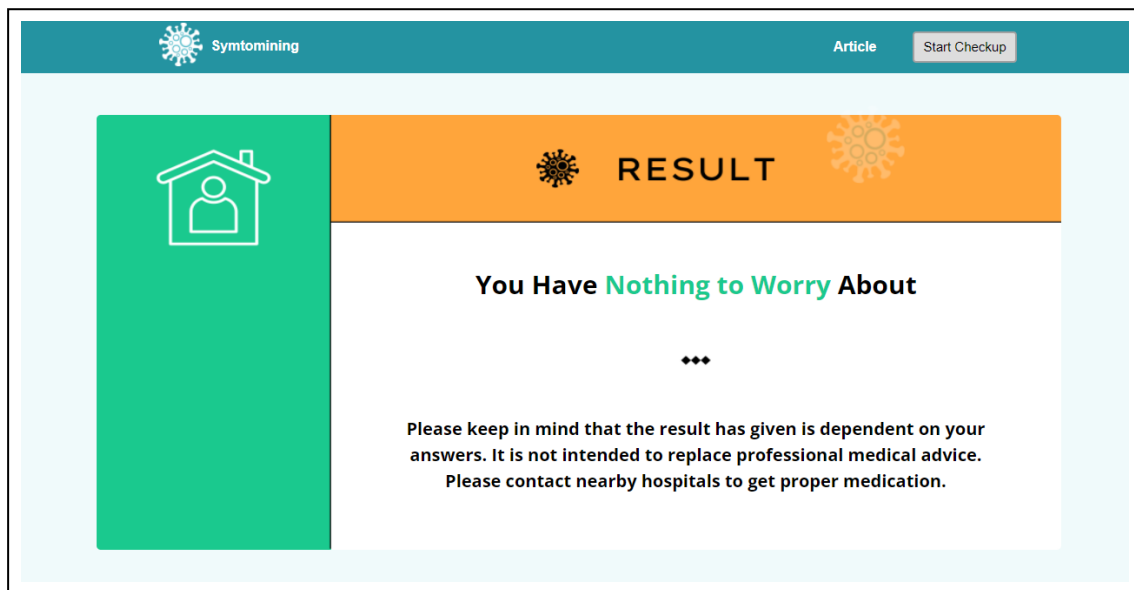


Figure 4.6: Web Interface with Negative Output

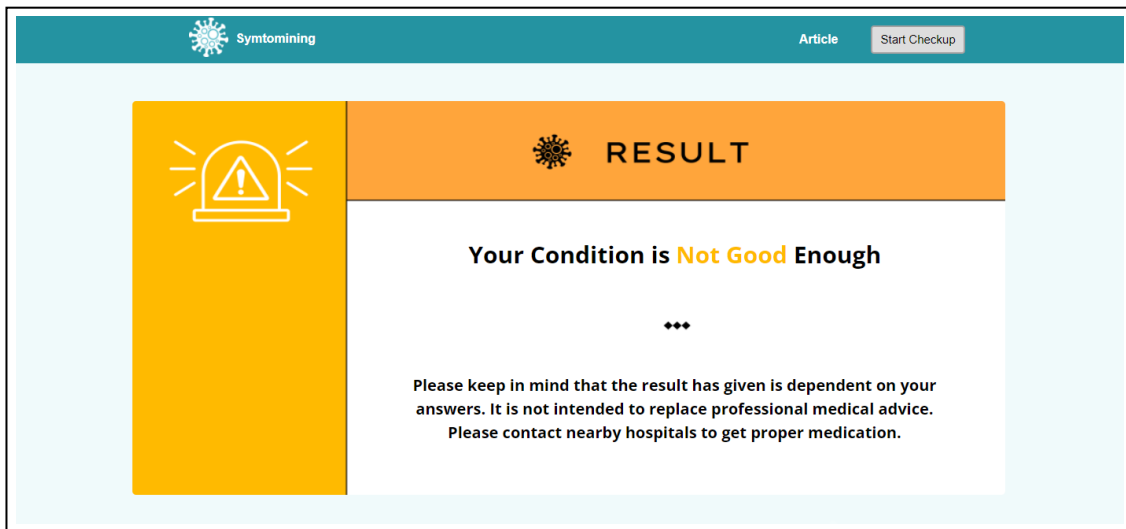


Figure 4.7: Web Interface with Cabin Output

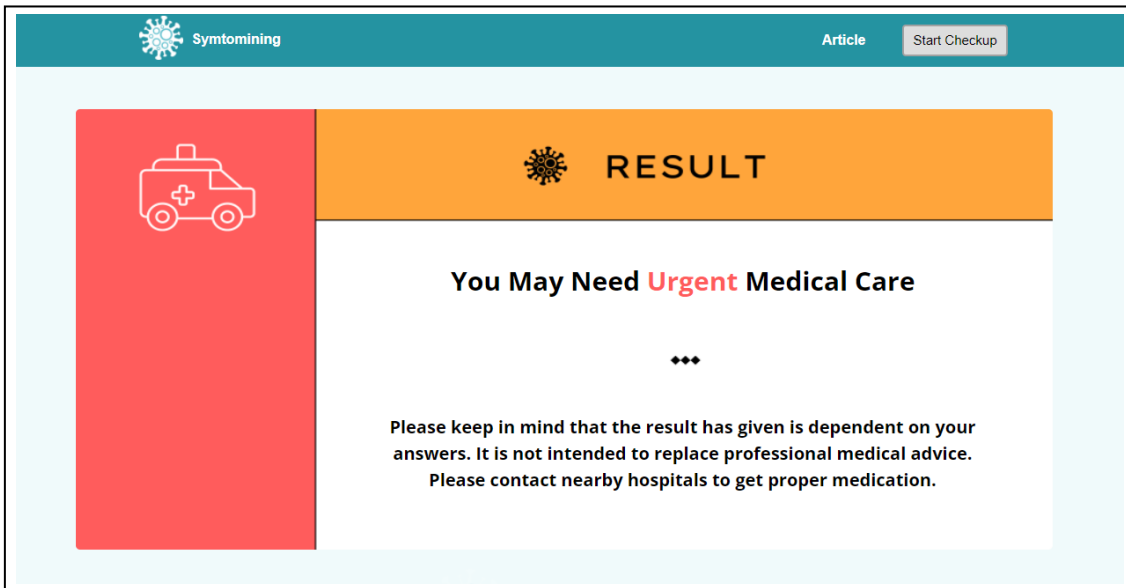


Figure 4.8: Web Interface with Emergency Output

## **CHAPTER 5**

### **IMPACT ON SOCIETY & SUSTAINABILITY**

#### **5.1 Introduction**

If a project has been completed, the project's influence on society must be examined and assessed. The impact of the Prediction of COVID-19 project has been described in three parts in this chapter. The influence on society or the state section discusses how this project will significantly affect mankind. Then there are ethical considerations. The ethical component of the study was thoroughly examined to comprehend how this initiative can benefit patients and the medical sector of a country. Finally, the project's long-term viability was considered. It is addressed how this idea might expand in the future and serve more individuals.

#### **5.2 Impact on Society**

This research has had a substantial impact on society. Because coronavirus is an infectious disease, the government encourages people to stay at home to avoid being infected by other infected persons. Going to the hospital for COVID-19 pandemic testing is quite dangerous. People can provide their symptoms into an interface that has been developed. When they notice the symptoms of COVID-19 in themselves and get an immediate result with the necessary guidelines, it would be highly beneficial if anyone could see their report at home and receive preliminary guidance depending on their condition. It will also help to prevent the spread of CoronaVirus and reduce the pressure on hospital workers and governments in overpopulated countries with insufficient medical facilities. A survey of patients and healthcare professionals is now being performed to study the efficiency of this research. And it's safe to assume that the survey's outcomes on the impact on society and patients-doctors will be beneficial.

#### **5.3 Ethical Aspects**

People are not required to visit a diagnostic hospital for any type of test. At home, they can obtain a basic understanding of COVID-19. They get the capability to identify their problem on their own. All data will be kept in the database when a web interface is developed, and the model will become stronger day by day. Because this is a machine

learning-based project, it's difficult to be completely dependent on this model right now. Nothing can be accurately predicted by a machine. It will require some time. When the database has millions of records, the model will be stronger, and it is hoped that it will be able to accurately predict over time. It will be unnecessary to visit a diagnostic hospital in the future if Artificial Intelligence and the Internet of Things can be interconnected to a database. People will be able to predict COVID-19 at home before consulting a doctor about their condition. And if a machine that could be portable that could test blood can be developed, people will be able to test COVID-19 more accurately at home.

## **5.4 Sustainability**

Because every single variant of the coronavirus is incredibly dangerous, and this life-killing virus constantly changes its genetic sequence, so this research has long-term potential. Mutations with a frequency of more than 50 and a spike of even more than 30 have the possibility of causing the vaccine to fail and spread rapidly and deadly. As a result, using a website to monitor his or her COVID-19 condition is a very effective approach. The website mainly uses machine learning techniques to predict Coronavirus Disease and the patient's condition. However, there are many possibilities for deep learning, AI, and the Internet of Things to provide more accurate results for this project in the future. As a result, numerous projects may be done using this project in the future if it is sustainable. It's even possible to make a mobile app. Only COVID-19 is predicted and its stage is checked in this research. Additionally, in the future, various diseases and organ failure can be predicted using this machine learning research and the web method.

## CHAPTER 6

### FUTURE SCOPE & CONCLUSION

#### 6.1 Introduction

The future scope was discussed in this chapter. What methods may be employed to produce a better machine in the future, and how can this project help the organization develop more efficiently in the future. This chapter has come to a tidy conclusion, which is available at the end of the chapter. Finally, a list of references is provided at the end of this chapter.

#### 6.2 Future Scope of this Study

Regardless of how challenging the concept is, it may be implemented to develop a website for any hospital that provides COVID-19 treatments, regardless of location. The authors have created a convenient and straightforward interface for this project. A website based on the Internet of Things may be built in the future and made available to the general public over the internet. Entering symptoms and all of the relevant information on the website that is needed to detect Coronavirus is a simple act of predicting COVID-19 at home. After that, the person will be able to determine whether or not they have COVID-19. The website will save the information because it is an Internet of Things-based website and users will submit new data each time. The model will learn from each new piece of data in the following phases, leading it to become more accurate and effective at the time. Deep Learning using Neural Network methods can be highly successful over time, and Artificial Intelligence can also be used.

#### 6.3 Recommendations

Each diagnosis component has a normal level, and this procedure can detect any unexpected stages based on the normal level. Fever, Dry Cough, Sore Throat, muscle pain Diarrhea, Tiredness, Headache, and other symptoms are primary symptoms of Coronavirus Disease in predicting the COVID-19. In extreme situations, when the virus affects the patient's lungs extensively, the infected individual has chest pain and shortness of breath. In severe cases, the affected individual has trouble moving and speaking, and their oxygen saturation drops, leading to a critical condition and death.

According to our data set, 68.87%, 63.05%, 50.80%, 49.39%, 45.73% of infected patients had a fever, dry cough, muscle pain, sore throat, and loss of taste & smell respectively. And 41.56%, 40.56%, 31.12% infected patients had shortness of breathing, chest pain, and difficulties in speech and movement respectively. Most of the symptoms of the coronavirus are similar to the common cold, flu, and normal allergies. So it is very difficult to detect infected and susceptible individuals. In the data set, 61.04% of the infected patients are normal patients and 38.96% are emergency patients. As a result, if appropriate steps are taken, such as diagnosing the corona in the early stages of infection, following proper guidelines, and taking necessary treatments at the appropriate time, the risks of the infected individual being critically infected are reduced. Although the coronavirus cannot be completely prevented, some precautionary measures can be taken to reduce the death rate and thereby minimize the spread of coronary heart disease. Such as following government rules, maintaining social distances, avoiding crowded situations, not going out unnecessarily, modifying food habits, exercising, and obtaining immediate treatment in extreme circumstances.

## **6.4 Conclusion**

If COVID-19 is detected, diagnosed, and treated at an early stage, there is a high probability that it will be controlled properly and efficiently, the patient will probably recover in a short time and the risk of death will be reduced. To determine if they have COVID-19, people must find a hospital where the COVID-19 test kit is available, wait a long time, and get tested. It is quite difficult to provide sufficient kits to test COVID-19 for such a huge population under the current situation. Most suspected patients are unable to diagnose the COVID-19 in time and take immediate action due to an insufficient diagnostic system. A model trained on a relevant dataset, according to the researchers who conducted this study, can predict the development of coronavirus infection at any stage. The authors created an interface for this project that allows users to obtain their COVID-19 report by simply filling out a form with the required information. Twelve machine learning algorithms are trained on the dataset and evaluated based on their accuracy, Jaccard score, Cross Validated score, and AUC scores to determine the best model for the dataset. The accuracy score, the Jaccard score, and the Cross Validated score are the three separate scores computed. When compared to other classifiers, XGBoost exceeds the competition in terms of

performance. that after a website for this research is established, it will be made available to any hospital that provides COVID-19 treatments. COVID-19 reports will be accessible online and available 24 hours a day, seven days a week, so people will not have to leave their homes to get them.

## REFERENCES

- [1] Kolla, B. (2021), “Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms”, Volume 8. No. 5,  
doi: 10.30534/ijeter/2020/117852020
- [2] Zoabi, Y., Deri-Rozov, S. & Shomron, N., “Machine learning-based prediction of COVID-19 diagnosis based on symptoms”, *npj Digit. Med.* 4, 3 (2021).  
<https://doi.org/10.1038/s41746-020-00372-6>
- [3] Hamzah FAB, Lau C, Nazri H, Ligot D V, Lee G, Tan CL. “CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction”, *Bull World Heal Organ.* 2020; 1:32. <https://doi.org/10.2471/BLT.20.255695>
- [4] Menni, C., Valdes, A.M., Freidin, M.B. et al., “Real-time tracking of self-reported symptoms to predict potential COVID-19”, *Nat Med* 26, 1037–1040 (2020).  
<https://doi.org/10.1038/s41591-020-0916-2>
- [5] Callejon-Leblic, M.A.; Moreno-Luna, R.; Del Cuvillo, A.; Reyes-Tejero, I.M.; Garcia-Villaran, M.A.; Santos-Peña, M.; Maza-Solano, J.M.; Martín-Jimenez, D.I.; Palacios-Garcia, J.M.; Fernandez-Velez, C.; Gonzalez-Garcia, J.; Sanchez-Calvo, J.M.; Solanellas-Soler, J.; Sanchez-Gomez, S., “Loss of Smell and Taste Can Accurately Predict COVID-19 Infection: A Machine-Learning Approach”, *J. Clin. Med.* 2021, 10, 570. <https://doi.org/10.3390/jcm10040570>
- [6] Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al., “Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset”, *SN COMPUT. SCI.* 2, 11 (2021). <https://doi.org/10.1007/s42979-020-00394-7>
- [7] Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, Alyami SA, Lin PI, Islam SMS, Quinn JM, Eapen V, Moni MA., “Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data:



Statistical Analysis and Model Development”, *JMIR Med Inform.* 2021 Apr 13;9(4): e25884. doi: 10.2196/25884. PMID: 33779565; PMCID: PMC8045777.

[8] Kwekha-Rashid, A.S., Abduljabbar, H.N. & Alhayani, B, “Coronavirus disease (COVID-19) cases analysis using machine-learning applications”, *Appl Nanosci* (2021). <https://doi.org/10.1007/s13204-021-01868-7>

[9] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, and Xi Cheng, “Artificial intelligence and machine learning to fight COVID-19”, 03 APR 2020. <https://doi.org/10.1152/physiolgenomics.00029.2020>

[10] De Souza FSH, Hojo-Souza NS, Dos Santos EB, Da Silva CM and Guidonia DL (2021), “Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study with Brazilian Data”, *Front. Artif. Intell.* 4:579931. doi: 10.3389/frai.2021.579931

[11] Norah Alballa, Isra Al-Turaiki, “Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review, *Informatics in Medicine Unlocked*”, Volume 24, 2021, 100564, ISSN 2352-9148, <https://doi.org/10.1016/j.imu.2021.100564>.

[12] André Filipe de Moraes Batista, João Luiz Miraglia, Thiago Henrique Rizzi Donato, Alexandre Dias Porto Chiavegatto Filho, “COVID-19 diagnosis prediction in emergency care patients: a machine learning approach”, *medRxiv* 2020.04.04.20052092; doi: <https://doi.org/10.1101/2020.04.04.20052092>

[13] Assaf, D., Gutman, Y., Neuman, Y. et al., “Utilization of machine-learning models to accurately predict the risk for critical COVID-19”, *Intern Emerg Med* 15, 1435–1443 (2020). <https://doi.org/10.1007/s11739-020-02475-0>

[14] Mazurek J (2021), “The evaluation of COVID-19 prediction precision with a Lyapunov-like exponent”, *PLoS ONE* 16(5): e0252394. <https://doi.org/10.1371/journal.pone.0252394>

[15] Parro VC, Lafetá MLM, Pait F, Ipólito FB, Toporcov TN (2021), “Predicting COVID-19 in very large countries: The case of Brazil”, *PLoS ONE* 16(7): e0253146. <https://doi.org/10.1371/journal.pone.0253146>

- [16] Elkin ME, Zhu X (2021), “Understanding and predicting COVID-19 clinical trial completion vs. cessation”, PLoS ONE 16(7): e0253789.  
<https://doi.org/10.1371/journal.pone.0253789>
- [17] Salah Alazab, Moutaz and Awajan, Albara and Mesleh, Abdelwadood and Abraham, Ajith and Jatana, Vansh and Alhyari, Salah. (2020), “COVID-19 Prediction and Detection Using Deep Learning. International Journal of Computer Information Systems and Industrial Management Applications”, 12. pp. 168-181.
- [18] Feng S, Feng Z, Ling C, Chang C, Feng Z (2021), “Prediction of the COVID-19 epidemic trends based on SEIR and AI models”, PLoS ONE 16(1): e0245101.  
<https://doi.org/10.1371/journal.pone.0245101>
- [19] Al-Zaman M. S. (2020), “Healthcare Crisis in Bangladesh during the COVID-19 Pandemic”, The American journal of tropical medicine and hygiene, 103(4), 1357–1359. <https://doi.org/10.4269/ajtmh.20-0826>
- [20] An XS, Li XY, Shang FT, Yang SF, Zhao JY, Yang XZ, Wang HG, “Clinical Characteristics and Blood Test Results in COVID-19 Patients”, Ann Clin Lab Sci. 2020 May;50(3):299-307. Retraction in: Ann Clin Lab Sci. 2020 Jul;50(4):560. PMID: 32581016.

# PLAGIARISM REPORT

## COVID-19 Prediction (White Walkers)

### ORIGINALITY REPORT

<b>11</b> %	<b>7</b> %	<b>7</b> %	<b>5</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<a href="http://v3r.esp.org">v3r.esp.org</a> Internet Source	<b>1</b> %
<b>2</b>	<a href="http://dokumen.pub">dokumen.pub</a> Internet Source	<b>1</b> %
<b>3</b>	V. C. Parro, M. L. M. Lafetá, F. Pait, F. B. Ipólito, T. N. Toporcov. "Predicting COVID-19 in very large countries: The case of Brazil", PLOS ONE, 2021 Publication	<b>1</b> %
<b>4</b>	<a href="http://www.medrxiv.org">www.medrxiv.org</a> Internet Source	<b>&lt;1</b> %
<b>5</b>	David Paper. "Hands-on Scikit-Learn for Machine Learning Applications", Springer Science and Business Media LLC, 2020 Publication	<b>&lt;1</b> %
<b>6</b>	"Software Engineering Perspectives in Intelligent Systems", Springer Science and Business Media LLC, 2020 Publication	<b>&lt;1</b> %
<b>7</b>	Submitted to Savannah State University Student Paper	