American International University-Bangladesh (AIUB)

Study on Covid-19 Important Features Selection Using Machine Learning Techniques on Hospital Dataset

Submitted By

SN	Student Name	Student ID
1	**TAMMOY GHOSH**	**16-31993-2**
2	**MD SANOWAR HOSSAIN**	**18-36896-1**
3	**FUAD HASAN**	**18-36870-1**
4	**SHAMMI AKTER**	**18-36445-1**

Department of Computer Science

Faculty of Science & IT

American International University Bangladesh

February, 2022

Declaration

We declare that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from The published and unpublished work of others has been acknowledged in the text and a list of references is given.

Tanmoy Ghosh

TAMMOY GHOSH

16-31993-1

Department of Computer Science

Sanowar

SANOWAR HOSSAIN

18-36896-1

Department of Computer Science

Fuad Hasan

FUAD HASAN

18-36870-1

Department of Computer Science

Shammi Akter

SHAMMI AKTER

18-36445-1

Department of Computer Science

Approval

The thesis titled " " has been submitted to the following respected members of the board of examiners of the Department of Computer Science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science on February, 23, 2022 and has been accepted as satisfactory.

....

DR. S. M. HASAN MAHMUD
Assistant Professor & Supervisor
Department of Computer Science
American International University-Bangladesh

....

TOHEDUL ISLAM
Assistant Professor & Supervisor
Department of Computer Science
American International University-Bangladesh

....

DR. DIP NANDI
Asst. Professor & Head
Department of Computer Science
American International University-Bangladesh

....

Professor Dr. Tafazzal Hossain
Dean
Faculty of Science & Information Technology
American International University-Bangladesh

DR. CARMEN Z. LAMAGNA
Vice Chancellor
American International University-Bangladesh

Acknowledgement

At first, all praise to almighty Allah for giving us the strength and courage to continue working throughout the last four years and especially the last two semesters. With His blessing, we have finally completed the thesis even through a rough pandemic and economic hardships. We also want to take this opportunity to thank our parents for their relentless support that saved us from outside hardships.

We especially like to thank our supervisor, Dr. S. M. Hasan Mahmud Sir for his relentless support in times when we had no direction to go. We are grateful to Hasan sir for always keeping time for us even in busier schedules and communicate with us to show us our flaws and aspects of us that we can improve. We honor the support Sir has provided us and dedicate this work to him.

We also like to thank google for freely providing a useful tool such as Google Colab without which this thesis would not be possible. The computation power google can provide through cloud is game changing and inspired us to be more so that one day we can influence many lives as google influenced ours.

Abstract

Coronavirus pandemic circumstance, precise expectations could extraordinarily help in the wellbeing asset the executives for future waves. However, as a replacement entity, COVID-19’s disease dynamics seemed difficult to predict. We developed a model that employed supervised machine learning algorithms to identify the presentation features predicting COVID-19 disease diagnoses with high accuracy. Features examined included details of the individuals concerned, e.g., age, gender, observation of fever, history of travel, and clinical details just like cough and other infection. We utilized Logistic Regression, Support Vector Machine, Random Forest, XGBoost Classifier, and Decision Tree for our similar investigation. We carried out and applied a few AI calculations to our gathered information and observed that the XGBoost calculation performed with the most noteworthy precision (>90%) to anticipate and choose highlights that accurately show COVID-19 status for all age gatherings. Furthermore, to cut back the size of our dataset, we used SHAP to avoid wasting computation time. Our mission is to attentively examine various data and classify them in step with each algorithm's efficacy in terms of accuracy, precision, recall, and F1 Score. Our greatest accuracy without dimensionality reduction was 92 percent, which was quite good. XGBoost was used to discover it. However, when it involves dimensionality reduction, the foremost important factor is accuracy. The XGBoost Regression technique accomplished a consequence of 95 percent, and the calculation time was under a moment diminished also.

Keywords

Covid-19, Classification, XGBoost, LR, SVM, DT, RF, Feature selection, Importance symptom

Table of Content

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

COVID_19 = Coronavirus disease 2019

SHAP = Shapley Additive exPlanations

RF = Random Forest

DT = Decision Tree

SVM = Support Vector Machine

LR = Logistic Regression

XGBoost = eXtreame Gradient Boosting

ML = Machine Learning

AI = Artificial Intelligence

HIV = Human Immunodeficiency Virus

AUC = Area Under the Curve

ROC = Receiver Operating Characteristic Curve

etc. *etc.*

Symbols	
$\sum$	Summation
ln	Natural Logarithm
ϕ	Flux
etc.	*etc.*

Chapter 1

Introduction

1.1 Covid-19

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2, has become an unprecedented public health crisis and urgent threat to global health [1, 2]. In Wuhan, province of Hubei(china) December 2019 several local health facilities reported cases of pneumonia of unknown origin, which have been identified as the first human cases of COVID-19[3,4,5].But the first human coronaviruses, 229E and OC43 were identified during the 1960s for human nasal secretions[5]. The Covid-19 (SARS-CoV-2) virus pandemic has caused more than 5,903,460 deaths and a total of over 424,352,949 confirmed cases, globally [7].

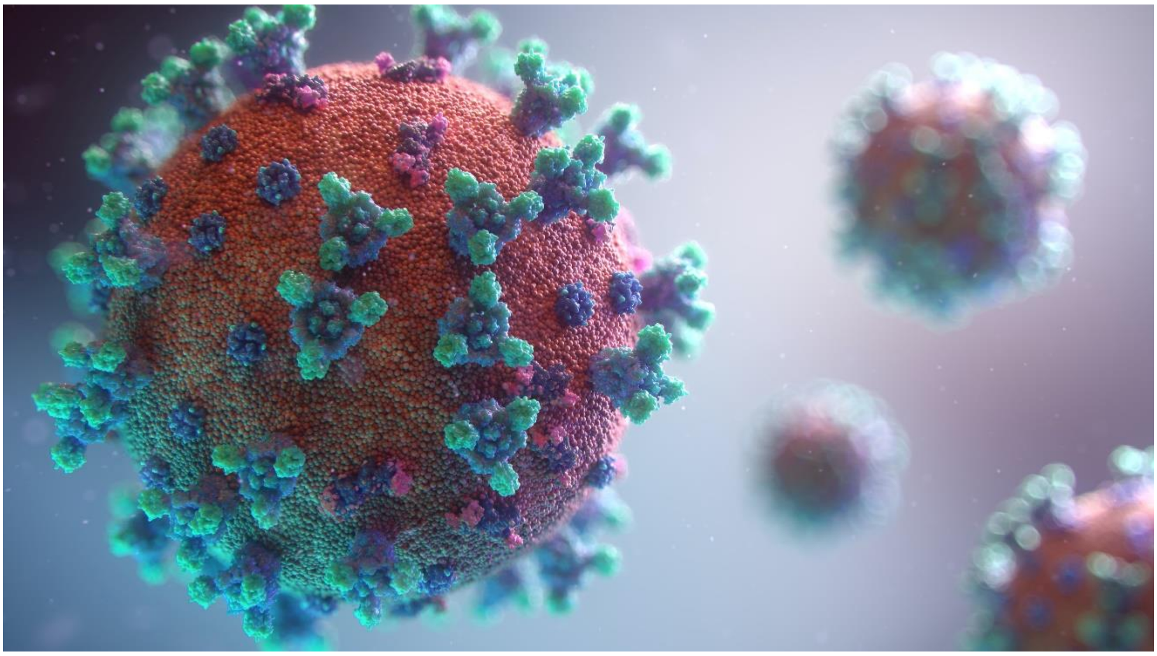

Figure 1.1: Covid-19 Virus [73]

Most patients have mild, self-limiting respiratory infections, with symptoms such as fever, intubation, rate Po2, cough, fatigue and diabetes but some may rapidly develop fatal complications, including acute respiratory distress syndrome (ARDS) or respiratory failure, multiple organ dysfunction, and septic shock that imposes hospitalization and could lead to the death of the patient [3,4,8].

This pandemic has burdened all global health systems and is a huge opportunity to highlight the value of laboratory medicine and focus on new ways to support and accelerate the identify patients at high risk of serious illness. Accurate prediction of COVID-19 deaths and identification of factors associated with the severity of the disease will enable targeted strategies for patients at high risk of dying or developing a serious illness. Thus reducing the burden of unnecessary hospital admissions and putting a strain on the health system. [9]. A better and good understanding of predictive factors for COVID-19 is crucial for the development of clinical decision support systems that can accurately and rapidly detect the patients with increased risk of worsening conditions [1, 10].

Scientists and clinicians have made enormous efforts to this international public health crisis to generate new knowledge and to develop technological tools that may help in combating this infectious disease and mitigate its effects. Some of these efforts include the development of drugs and vaccines [1,2, 11,12], the construction of epidemiological models to forecast the dynamics of disease spreading in the population [13–16], develop mobile applications for monitoring infected patients and new cases [17-19] and develop strategies and use new technologies to manage hospital resources and capacity [20-22].

An emergency non-pharmaceutical prevention measure adopted in many countries has been the reduction or suspension of non-essential activities so on reduce both the speed of recent infections[23] and therefore the risk of exceeding hospital capacities. Undoubtedly, the flexibility to rapidly identify high-risk patients and correctly assign health care priorities is critical, within the first case so on improve hospital capacity planning and within the second case for providing timely treatment for patients [24]. In this sense, the artificial intelligence method is recognized as a powerful and promising technology. This will not only help to identify the serious risk of a particular patient requiring medical care [25, 26], but also for the diagnosis process [27–30], prediction of disease spreading dynamics [31–35], and tracking of infected patients as well as likely future patients [36, 37].

1.2 Motivation

Since the World Health Organization proclaimed the COVID-19 epidemic to be a pandemic, various research have been undertaken employing Artificial Intelligence approaches to enhance these procedures in clinical settings in terms of quality, accuracy, and most crucially time. Automatic medical diagnosis has become extremely crucial, especially when it comes to making timely judgments for deadly infectious diseases like COVID-19 sickness. Corona patients must be diagnosed remotely since their exposure to others increases the number of sufferers on a regular basis. As a result, direct contact with corona patients may endanger the lives of physicians and nursing personnel. To tackle this widespread and hazardous threat, it is critical to examine patient data and then precisely detect them with the least amount of time penalty. In this paper, we has been introduced some AI model in healthcare system to provide more accurate and fast diagnostic results.

To attain this goal AI is incredibly effective technique because it is a wide used, non-invasive and additionally nice for recognizing the symptoms. Recent analysis shows nice when deep learning techniques are used in Covid-19 data set for fast and correct detection. This introduces new opportunities to any develop the understanding of deep learning whereas introducing a lot of fine-tuned algorithms to make sure early and correct detection.

1.3 Covid-19 data

Now a days using data processing and artificial intelligence (AI) to fast detect and analyze covid-19. Can help to improve the precision of Covid-19 diagnosis. Medical data plays an important role in determining the cause of illness, analyzing the fact, evaluating treatment options and predict the symptoms such as fever, headache, dry cough, fatigue, and muscle pain. Medical data refers to the fundamental procedures that will be used to study the information regarding age, gender, habitual residence and examine the condition. Each type of invention provides specific information about the area of the body being treated or researched, the ailment, or the capabilities.

1.4 Background

1.4.1 Different Algorithm in Machine learning Overview

Coronavirus is not a normal respiratory infection brought about by the SARS-CoV-2 infection. Mainstream researchers has zeroed in on this sickness with practically extraordinary force. However, the majority of primary studies published on COVID-19 suffered from small sample sizes [38]. While some primary research studies have reported tens or hundreds of cases, many other studies have reported fewer than 20 patients [39].Therefore, there is an urgent need to collect all available published data on the clinical features of COVID-19 from various studies in order to create a comprehensive dataset for understanding the pathogenesis and clinical features of COVID-19. In this study, we aim to perform a large-scale meta-analysis to synthesize all published studies with clinical data of patients with COVID-19, in order to discover new correlations between clinical variables in patients with COVID-19. We will then, at that point, apply AI to reanalyze the information and fabricate a PC model to foresee assuming somebody has COVID-19 dependent exclusively upon their clinical data.

We believe that the ability to predict patients with COVID-19 based on clinical variables and the use of an easily accessible computational model would be extremely useful in overcoming the widespread shortage of COVID-19 testing capabilities worldwide. Because many countries and hospitals are unable to provide adequate diagnostic resources, health systems lack one of their most effective tools to combat outbreaks: identifying hotspots and targeting affected areas and individuals [39]. The scale of the lack of testing requires the use of COVID-19 diagnostic methods, which currently use the resources of local health facilities. We propose the development of a disease prognosis model based on medical variables and standard medical laboratory tests.

1.4.2 Deep Learning Overview

Deep learning is a kind of AI that elevates preparing the Computer to comprehend individuals' fundamental detects. One of the main type of AI is ML. We use Machine Learning in different reason in medical care, like diagnosing patient credits, anticipating therapy systems, and accuracy medication. AI likewise remember for profound learning and neural organization models are utilized in clinical information examination, expectation and identification of specific illnesses. AI is a technique for information examination that blesses PC calculations with the ability to "learn" from a known informational collection (which incorporates characterizing qualities and a result for every perception), to deliver an expectation about the result given a particular selection of attributes. Obviously, the quality and size of the preparation informational index are essential in deciding the subsequent execution of the calculation. In what follows we exhibit the utilization of neural organizations prepared with the informational collection portrayed in Section II. Our neural organizations are then used to anticipate whether a given patient (excluded from the informational index utilized for preparing) has a place with one of two classes: class 1, which addresses those patients who are bound to get by than to pass on, and class 2 which addresses, on the other hand, those patients who are bound to kick the bucket than to get by. [37]

1.5 Thesis Objective

The main goal of this research is to deconstruct multiple algorithms in order to achieve better results in predicting Covid-19. We sought to create a report that may help people reduce the death rate of Covid-19 patients by saving better clinical treatment and early awareness with a precise, speedy expectation. Algorithms will be used to assess the affectability, precision, time complexity and accuracy of explicit data sets. As it is now we need to read it more for a better and more solid outcome as it is a very important problem in our time.

1.6 Thesis Outline

Machine Learning Techniques on Medical dataset for Covid-19 importance feature prediction is summarized in this work. The backdrop to the research is presented in Section 2. The 16

Material and Methods are listed in Section 3. The output of the algorithms is defined in Section 4. This paper's Discussion and Conclusion are included in Section 5 and 6.

Chapter 2

Literature Review

Covid-19 pandemic situation, careful gauges could uncommonly help in the prosperity resource the board for future waves. In any case, as another component, COVID-19's disease components gave off an impression of being difficult to expect.

AI calculations have been acquiring energy in the course of the last a very long time for clinical applications, for example, PC supported conclusion to help doctors for an early finding, which can prompt better-customized treatments and improvement of the clinical consideration proposed to patients.

2.1 Related Works

In this section, previous research efforts on the classification of COVID-19 patients will be reviewed.

Souza et al. [40] reported a research that used supervised machine learning techniques such logistic regression, linear discriminant analysis, naive Bayes, k-nearest neighbors, decision trees, XGBOOST, and support vector machine to identify patients who could develop severe COVID-19 symptoms early. Individual fundamental information such as gender and age range, symptoms, comorbidities, and recent travel history were used to train the machine learning algorithms, which were taught using an available to the public database referring to Brazil. The authors claim that a ROC area under curve (AUC) of 0.92, a sensitivity of 0.88, and a specificity of 0.82 can predict sickness outcome.

Mohammad A. Alzubaidi, Mwaffaq Otoom, Nesreen Otoum, Yousef Etoom, Rudaina Banihani [41] examine the five diverse component choice calculations that gave various rankings to the main top-five indications. They even chosen various side effects for consideration inside the main five. This is on the grounds that every one of the five calculations positions the side effects dependent on various information qualities. Be that as it may, when this large number of five rankings were accumulated (utilizing two diverse amassing strategies) they created two indistinguishable rankings of the five most significant COVID-19 indications. Beginning from the most essential to least significant, they were: Fever/Cough, Fatigue, Sore Throat, and Shortness of Breath. (Fever and hack were positioned similarly in the two collections.)

Moreover, the proposed VBFW technique accomplished an exactness of 92.1 % when used to assemble a one-class SVM model, and a NDCG@5 of 100 %.[41]

Niu et al. [42] incorporated a partner of 150 patients determined to have COVID-19 from Huang gang Central Hospital in the period 23 January–5 March 2020. By taking advantage of univariate and multivariate strategic relapse, the creators investigated which were the most pertinent danger factors related with in-clinic demise. This investigation permitted reasoning that diabetes, a high worth of lactate dehydrogenase on affirmation, and higher successive organ disappointment appraisal score expanded the chances of in-medical clinic passing.

Khaled Mohamad Almustafa[43] It is worth focusing on that pregnancy, diabetes, asthma, inmsupr, hypertension, other_ sicknesses, and tobacco have insignificant or no impact on the characterization of this particular dataset utilizing the J48 classifier.

Results show that J48 classifier gives the best grouping exactness with 94.41% and RMSE = 0.2028 and ROC = 0.919, contrasted with different classifiers of precision of 93.64%, 93.50%, and 92.71% for SGD, RF, and K-NN (K = 1), separately. When utilizing the component determination strategy, J48 classifier can foresee an enduring Covid19MPD case with 94.88% exactness and by utilizing just 10 out of the all out 19 accessible elements, which can be a valuable truth for medical services suppliers in recognizing conceivable tainted Covid19MPD cases. Results for the grouping utilizing highlight determination, in light of the element significance strategy, show that this technique beat the arrangement results for the full-included information with a precision of 94.65%, MAE = 0.0778, and RMSE = 0.214.[43]

Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G.[44] Using an improvement associate (n=3,841) and an orderly AI system, we recognized a COVID-19 mortality indicator that showed high exactness (AUC=0·91) when applied to test sets of review (n= 961) and planned (n=249) patients.

They tracked down that the XGBoost(44.1 delivered the best-performing indicators in the entirety of their examinations. XGBoost is a modern expectation calculation that forms an outfit of choice trees by iteratively zeroing in on harder to foresee subsets of the preparation information.

Aljouie, Abdulrhman Fahad et al.[45] Briefly depict the four classifiers utilized in this review: straight Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), and eXtreme Gradient Boosting (XGB) to demonstrate COVID-19 sickness outcome.

They surveyed the prescient capacity of consolidated CBC, age, sex, comorbidity, and CXRs seriousness explained information. That information was gathered at the hour of COVID-19 finding in our medical clinic. The proposed models can be utilized to anticipate mechanical ventilation (MV), mechanical and harmless ventilation, and mortality right off the bat to focus on patients and deal with the designation of clinic assets. We have observed that adjusting the train set, classifiers reliably yielded preferred execution over utilizing the first slanted information (results not shown). Arbitrary downsampling for making a decent train set with strategic relapse and irregular timberland delivered a superior model contrasted with SMOTE or ADASYN for mechanical ventilation and MV+NIV and mortality expectation. In any case, SMOTE with straight SVM yielded the better model for the three-class characterization issue (MV, NIV, no ventilation). They have thought about the exhibition of arbitrary timberland and XGBoost, which are fit for catching nonlinear connections among highlights and the objective, with strategic relapse and direct SVM since these are broadly utilized classifiers with great speculation. In the entirety of the forecast, assignments endeavored in the current review, consolidated clinical, research center and radiological information with Relief F include choice reliably performed better compared to every informational collection separately. The model LR with RUS and top 20 chose highlights for the consolidated information sources accomplished 0.82 AUC in foreseeing mechanical ventilation necessity in the test set. The top accomplishing model in the test set for mortality forecast yielded an AUC score of 0.83 and a reasonable precision of 0.80, utilizing all highlights from joined information and adjusted irregular forest.[45]

Nachtigall et al. [46] reflectively examined 1904 patients conceded to a public organization of emergency clinics in Germany. The authors analyzed segmented information, comorbidities, and clinical outcomes and found that the most important risk factors for death were older age, baseline lung disease, and male gender.

Banoei et al. [47] played out a multivariate prescient investigation on a subset of 108 out of 250 elements, enveloping comorbidities, blood markers, and clinical highlights. The highlights considered were those caught at the confirmation time from a partner of 250 hospitalized patients with COVID-19. The strongest predictors of mortality were diabetes, coronary heart disease, mental disorders, dementia and being over 65 years of age. Among the biochemical markers, the most important were CRP, lactate and prothrombin.

Sîrbu A, Barbieri G, Faita F, et al [48] highlight choice methodology, applied on information in the wake of editing ("Recursive component disposal") and utilize five standard arrangement models, for example (DT = .935, LR = 0.93, SVM = 0.968, NB = 0.93 and RF =0.93) to accomplish this goal, and they utilize the clinical factors chose at the initial step. They found svm model precision .968 and it is best.

Kuno T, Sahashi Y, Kawahito S, Takahashi M, Iwagami M, Egorova NN.[49] Through LASSO and SHAP, we chose six significant factors; age, hypertension, oxygen immersion, blood urea nitrogen, emergency unit, and endotracheal intubation. AUCs utilizing preparing and testing datasets got from the information before February seventeenth, 2021 were 0.871/0.911. Furthermore, the light GBM model has high consistency for the most recent information (AUC: 0.881). [77]

Tether technique showed 17 factors as significant highlights to anticipate in-hospital mortality; age, race, hypertension, coronary conduit illness, pulse, respiratory rate, systolic circulatory strain, diastolic pulse, oxygen immersion, C-reactive protein, d-dimer, white platelet count, hemoglobin, blood urea nitrogen, eGFR, ICU confirmation and endotracheal intubation. Then, at that point, SHAP showed six significant factors; age, hypertension, oxygen immersion, blood urea nitrogen, ICU confirmation and endotracheal intubation. We made the last model with six factors. AUCs utilizing preparing and testing datasets got from the information before February seventeenth, 2021 were 0.871/0.911 with light GBM, and 0.952/0.918 with the calculated relapse model. Light GBM showed high AUC to foresee in-hospital mortality, which was tantamount to the calculated relapse model.[49]

Zuccaro et al. [50] thought about an accomplice of 426 back to back hospitalized patients from a clinic in Lombardy, Italy, in the period 12 February–30 March 2020. They inferred that male sex, more established age, clinic affirmation after 4 March, and the quantity of comorbidities were autonomous danger factors connected with in-clinic mortality.

Altini N, Brunetti A, Mazzoleni S, et al [51] Starting from the information assortment of 303 patients and 347 separated elements, considering five highlights for every each of the 69 hematochemical boundaries, notwithstanding age and sex data, through factual element determination strategies, the subset of indicators was diminished to just six elements for both objective results.
The best prescient model was the choice tree for the mortality expectation task, with ROC-AUC of 89.66%, and the SVM for the ICU affirmation forecast, with ROC-AUC of 95.07% affirming the chance of using these models for both result forecasts.

Zhou et al. [52] reflectively investigated 116 patients confessed to Chongqing Public Health Medical Center, China, in the period 24 January–7 February 2020, with a determination of gentle or direct COVID-19. As indicated by the creators, three elements were viewed as autonomous indicators of movement to serious infection, during about fourteen days after affirmation: high worth of creatine kinase, low worth of CD4+ T-cell count, and age higher than 65 years.

Altini N, Brunetti A [3] Different machine learning models were also examined in order to develop a reliable classifier that relied on a small number of highly significant parameters to predict the likelihood of death or ICU admission. The most significant laboratory parameter for both outcomes was C-reactive protein min; HR=17.963 (95 percent CI 6.548-49.277, p = 0.001) for mortality, and HR=1.789 (95 percent CI 1.000-3.200, p = 0.050) for admission to ICU, according to the survival analysis. Erythrocytes max was the second most relevant characteristic; HR=1.765 (95 percent CI 1.141-2.729, p = 0.05) for mortality and HR=1.481 (95 percent CI 0.895-2.452, p = 0.127) for ICU admission. [3]
The decision tree was the best method for explaining the probability of death, with a ROC-AUC of 89.66 percent, while the support vector machine was the best method to estimate ICU admission, with a ROC-AUC of 95.07 percent. The hematochemical predictors discovered in this study can be used to describe the degree of disease in COVID-19 patients as a better prediction signature. [3]

Yoshida et al. [53] found sex incongruities in clinical and natural boundaries of serious results in 776 grown-ups with COVID-19, hospitalized in a U.S. medical services framework. The information from the accomplice were obtained in New Orleans, LA, between 27 February and 15 July 2020.

Dan Assaf et al. [54] use a database from a tertiary medical facility to identify individuals who are at risk of worsening throughout their hospital stay. The authors use historical and clinical characteristics such APACHE II score, white blood cell count, time from symptoms to admission, oxygen saturation, and blood lymphocytes count to train three distinct machine-learning approaches (neural networks, random forest and classification, and regression tree). The findings reveal that the sensitivity is 88.0 percent, the specificity is 92.7 percent, and the accuracy is 92.0 percent.

Pourhomayoun and Shakibi [55] present machine learning algorithms, including support vector machines, neural networks, random forest, decision tree, logistic regression, and k-nearest neighbors, to predict the mortality rate of COVID-19 patients. To prepare the calculations, the creators use research facility affirmed cases having a place with 76 nations all over the planet. The dataset utilized contains segment information, travel history, general clinical data like comorbidities, and side effects. Their results show that the neural organization calculation accomplishes the best exhibition with a precision of 93.75%.

Bezzan and Rocco [56] use research facility information, gathered from Sirio Libanes Hospital in Brazil, to distinguish patients requiring unique consideration at the clinic and to foresee lengths of stay at the particular consideration units. The creators test a few ML calculations to choose the best presentation. The last determination is the XGBOOST calculation for the two targets, which accomplishes 0.94 ROC AUC for the primary objective and 0.77 for the subsequent objective.

Li Yan et al. [57] propose a choice rule based absolutely at the managed XGBoost classifier to are anticipating victims at the most noteworthy possibility. The prescient variant is right off the bat taught with three qualities: lactic dehydrogenase (LDH), lymphocytes, and unnecessary awareness C-receptive protein (hs-CRP). The results show that the rendition can precisely distinguish the result of patients with extra than 90% exactness. a few distinct endeavors mindfulness on sorting out victims requiring specific consideration, especially hospitalization and additionally particular consideration units [58-59], or patients at a superior casualty danger.

Wu C, Chen X, Cai Y, et al.[60] The most commonly self-reported symptoms at onset of illness were fever (n = 188 [93.5%]), cough (n = 163 [81.1%]), productive cough (n = 83 [41.3%]), dyspnea (n = 80 [39.8%]), and fatigue or myalgia (n = 65 [32.3%]). The majority (n = 154 [76.6%]) of patients had fever with cough; 74 (36.8%) had fever with dyspnea; 66 (32.8%) had fever with fatigue, myalgia, or headache; and only 13 (6.5%) presented with fever alone. Results of 201 patients, the median age was 51 years (interquartile range, 43-60 years), and 128 (63.7%) patients were men. 84 (41.8%) patients developed ARDS and of these 84 patients, 44 (52.4%) died.

As the accurate identification of patients with COVID-19 becomes more important with each passing day, new detection methods have begun to emerge. There is much research in the literature to detect COVID-19 using CT images, X-rays and audio signals. Also most of the available works in the literature are considered demographic data, comorbidities, and blood markers. In this work, our purpose was to realize a predictive model based on algorithm in our covid-19 dataset. In previous works, as Banoei et al. [47], which considered blood markers at admission time, we included time series data for hematochemical factors, allowing the construction of a more reliable predictive model. Niu et al. [42] considered the evolution of parameters over time, but based their conclusions on a cohort smaller than ours, being composed of only 150 patients. As predictive models, they mainly considered univariate and multivariate logistic regression, whereas we compared a wide variety of methods: Decision tree (DT), random forest (RF), support vector machines (SVM), Logistic regression (LR), and XGBoost (xg) boosting. Finally,Mohammad A. Alzubaidi, Mwaffaq Otoom, Nesreen Otoum, Yousef Etoom, Rudaina Banihani [41], Beginning from the most essential to least significant, they were: Fever/Cough, Fatigue, Sore Throat, and Shortness of Breath. (Fever and hack were positioned similarly in the two collections.) Moreover, they proposed VBFW technique accomplished an exactness of 92.1 % when used to assemble a one-class SVM model, and a NDCG@5 of 100 %.[2]. Therefore, our paper can be considered a contribution over the existing literature, especially because we performed, in a SHAP features model for find importance features which are effective role in covid-19 of data, and systematic comparison of predictive models our dataset.

Chapter 3

Methods

3.1 Data Descriptions

There are 27 attributes or column in our dataset. They behold different information of several data. The attributes are discussing below:

TABLE 3.1 Available features and their descriptions

DATA	Descriptions
Age	Amount of time someone or something lived or existed
Intubation	The process in which a health worker inserts a tube through a person's mouth or nose and then down into the windpipe (windpipe/breathing tube)
Rate Po2	Measure the oxygen pressure in the arterial blood. The extent to which oxygen can move from the lungs to the blood
Cancer	Reflects the various diseases that can begin in almost any organ or tissue of the body when abnormal cells grow uncontrollably
Liver disease	Any of several conditions that can affect and damage the liver
Diabetes	A chronic disease that occurs when the pancreas is unable to produce insulin
Hematologic disease	Disorders that primarily affect the blood and blood-forming organs
HIV/AIDS	A potentially life-threatening chronic condition due to HIV
Immune Deficiency(Acquired or Congenita)	T cell deficiency, often causing disorders Secondary forms such as Acquired Immunodeficiency Syndrome (AIDS)
Peragnancy	A term used to describe the period during which a fetus develops in the woman's womb or uterus

Heart disease	Diseases affecting the heart or blood vessels
Renal disease	Cases in which the kidneys stop working and are unable to Help remove waste and extra water from the body blood or maintain a balance of chemicals in the body
Asthma	Conditions in which your airways narrow and may swell and may produce extra mucus
COPD	COPD, or COPD, refers to a group of diseases that cause airflow obstruction and breathing problems
Chronic neurological disorders	Defined as gangrene a substance that affects the brain and nerves located throughout the human body Spinal cord
Other chronic diseases	Long term z chronic illness lasting 3 months or more
Condition when entering the hospital	Record the patient's condition/how severe the patient is
Hospital duration	Average length of stay refers to the average number of days spent spent in hospital
Cough	Cough is your body's way of responding when something is bothering your throat or airway
Myalgia or Fatigue	Cause Fever or chills if caused by infection and can also cause symptoms such as joint pain or extreme weakness (fatigue)
Shortness of breath	A feeling of pressure over the chest when you cannot take a deep breath
Loss of consciousness	A condition where a person lacks normal awareness of themselves and the environment
Fever	Temporary elevated temperature, often due to of disease
Sample for test	Checks patient tests

Cantact Coronadisease	Symptoms or has been in contact with someone diagnosed with covid-19
Gender	Male(1) / Female(0)
Death	Indicates whether the patient has been approved or has recovered from covid19

3.2 SHAP

SHAP (SHapley Additive Explanations) by Lundberg and Lee (2017) 69 is a way of explaining individual predictions. The authors of SHAP proposed Kernel SHAP, an alternative core-based estimation approach to Shapley values inspired by local surrogate models. And he proposed Tree SHAP, an effective approach to estimating tree patterns. SHAP comes with many world-class interpretation methods based on Shapley's combination of values.

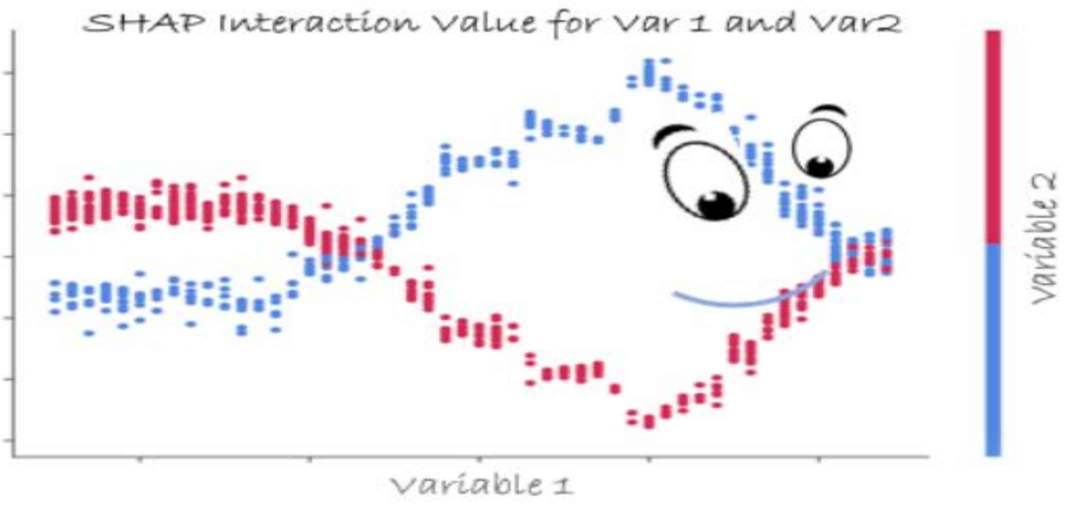

Figure 3.2: SHAP Variable [74]

The purpose of SHAP is to predict x examples by calculating the contribution of each feature in the forecast. SHAP assigns each function a meaningful value for this expectation. Its new components include: (1) identification of a new class of importance targets of additive properties, and (2) theoretical results that show that there is a unique solution in this class with a set of desirable properties. SHAP specifies the explanation as:

g(z′)=ϕ0+M∑j=1ϕjz′jg(z′)=ϕ0+∑j=1Mϕjzj′

where g is the explanation model, z′∈{0,1}Mz′∈{0,1}M is the coalition vector, M is the maximum coalition size and ϕj∈Rϕj∈R is the feature attribution for a feature j, the Shapley values. The formula simplifies to:

g(x′)=ϕ0+M∑j=1ϕjg(x′)=ϕ0+∑j=1Mϕj

Shapley values are the main arrangement that fulfills properties of Efficiency, Symmetry, Dummy and Additivity. SHAP additionally fulfills these, since it registers Shapley values.

3.2.1 SHAP Feature Importance

The idea of the importance of the SHAP function is simple: functions with large absolute Shapley values are important. Since we want global significance, we average the absolute Shapley values per function on the data:

1nIj n∑i 1 | (i) j | 1nIj i 1n | j(i) |

Next, we sort the functions by reducing their meanings and plotting them. The following image shows the importance of the SHAP function for the pre-trained XGBoost in predicting the Covid-19 dataset.

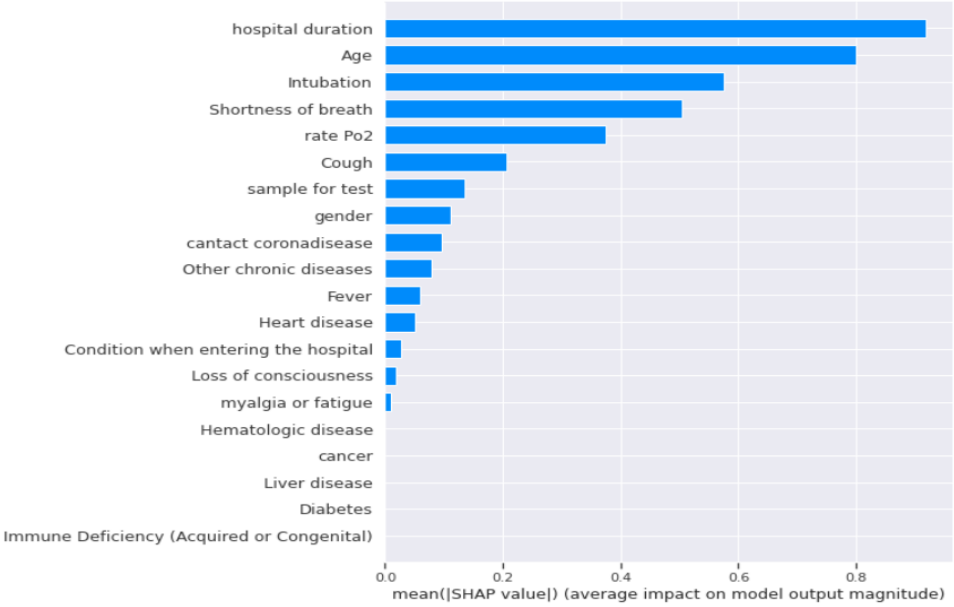

Figure 3.2.1: SHAP Feature Importance

The importance of the SHAP function is an alternative to the importance of the permutation function. There is one important difference between the two levels of importance. The significance of the permuted features is based on the reduction in model performance. Depends on the attribute size of the SHAP attribute.

3.2.2 SHAP Summary Plot

The summary graph combines functional significance with functional effects. Each point on the summary plot is a Shapley value for properties and instances. The position of the y-axis is determined by the function and Shapley's value on the x-axis. The color represents the value of the function from bottom to top. The overlapping points are distorted in the y-axis direction, so we get an idea of the distribution of the Shapley values for each function. Functions are arranged according to their values. The SHAP value plot can also show the positive and negative relationships of the predictors with the target variable.

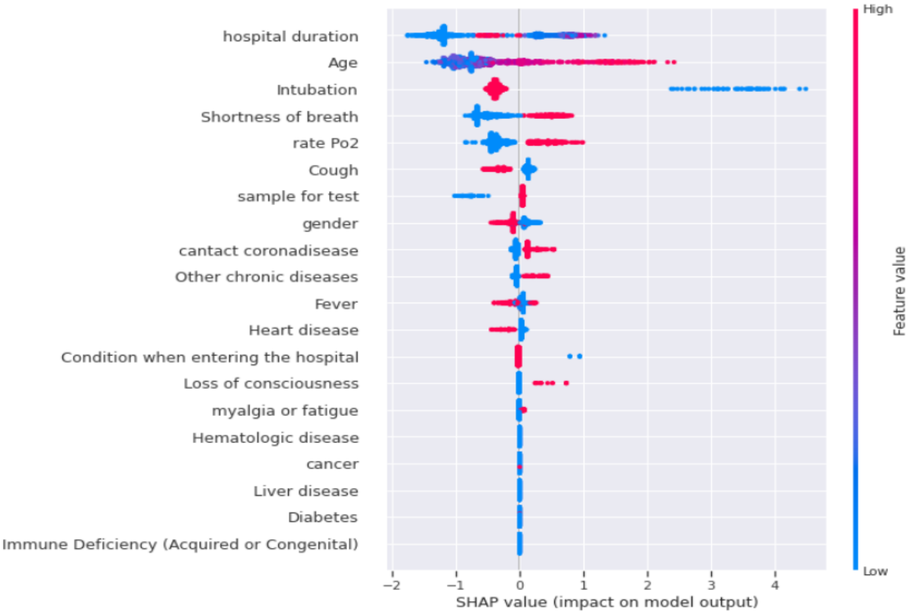

Figure 3.2.2: SHAP Summary Plot

3.2.3 Train-Test Split

Supervised learning algorithms frequently use training and verify knowledge. A dataset is divided into two parts: a coaching set and a check set. We have all forms of knowledge in the globe, such as money data or client data. This is applicable to supervised learning algorithms in Machine Learning. Because the coaching set comprises a well-known output, the model learns from it in order to generalize to different data.

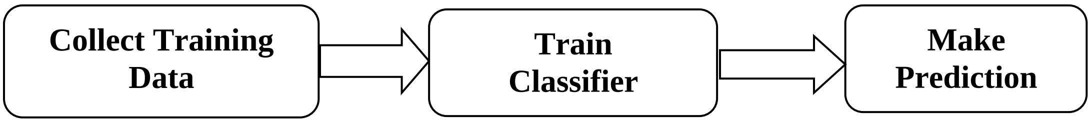

Figure 3.2.3: Class Distribution

XGBoost (eXtreme Gradient Boosting Algorithm), a joint machine learning method widely known for its superior performance over other machine learning methods, was selected for this study [12]. We first split our data into 80% training dataset and 20% testing dataset. The expected improvement acquisition function was used. We also performed XGBoost classification on subgroups of COVID-19 patients, stratifying them by gender, age, and other importance attribute.

Table 3.2.4: Test Set Result.

	Precision	**Recall**	**F1-score**	**Support**

0	0.96	0.99	0.97	211
1	0.82	0.50	0.62	18
Accuracy			0.95	229
Macro AVG	0.89	0.75	0.80	229
Weighted AVG	0.95	0.95	0.94	229

The supervised study approach incorporates data algorithms that analyze the data and make predictions for the future. Classification and regression are two distinct categories in this method. In contrast to regression, classification is a strategy for determining the label of data and using it for discrete answers. In the calibration phase, the initial step is to read the given data. Various classification methods are commonly used in machine learning applications. For classification and accuracy results, this study used logistic regression, support vectors, random forests, decision tree, and XGBoost algorithms. Each method was run on three independent datasets, each with its own set of features. All independent attributes were included in the first dataset, substantial correlations were found in the second dataset, and low correlations were found in the third dataset. Three different data sets were employed for each machine learning method, and precise results were obtained for comparison. XGBoost is one of the most often used techniques for solving classification problems. It examines the interaction between a categorically dependent variable and independent variables using a sigmoid function.

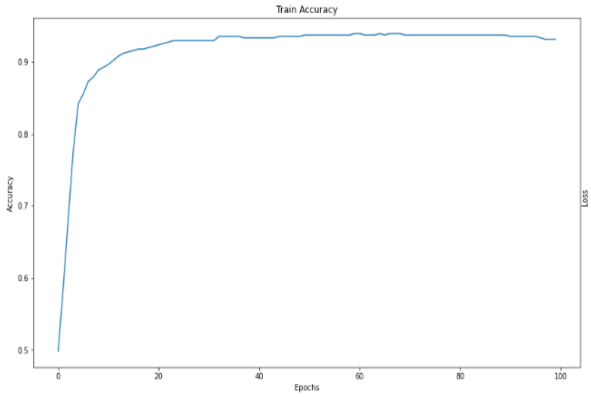

Figure 3.2.5: Train Accuracy graph

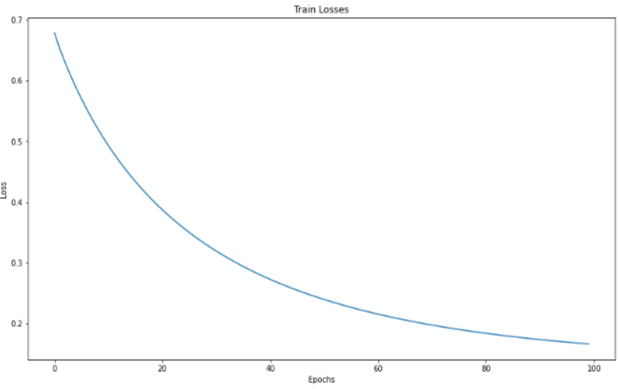

Figure 3.2.6: Train Loss graph.

3.3 Algorithm

3.3.1 Support Vector Machine

Support Vector Machine (SVM) is one of the important tools in machine learning. The working principle of SVM is as follows: Some secret datasets are trained by algorithms to achieve a set of classification models. Which can help to predict new categories of data [61, 62]. Its scope of application is widely used in various fields, such as disease or medical imaging diagnosis [63–65], financial crisis prediction [66], biomedical engineering, and bioinformatics classification [67, 68]. Although SVM is an efficient machine learning method, its classification accuracy needs to be further improved for multivariate classification of spaces and data set for feature interaction variables [69].

Recently, Support Vector Machine (SVM) has excellent classification and prediction performance and is widely used for disease diagnosis or medical assistance.

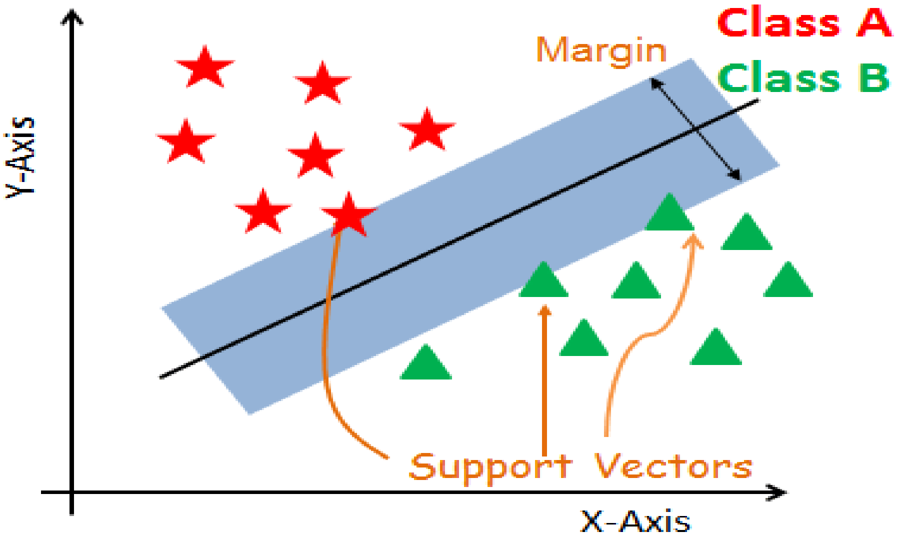

Figure 3.3.1: Graph of SVM [75]

3.3.2 Decision Tree

The estimation of a Decision Tree has a place with the class of coordinated learning computations. Dissimilar to other controlled learning computations, the Decision tree estimation can likewise be utilized to manage backslide and plan issues. The objective of utilizing a Decision Tree is to make a preliminary model that can be utilized to anticipate the class or worth of the genuine variable by including straightforward Decision rules got from past information (preparing data). In Decision Trees, we start at the lower part of the tree to foresee a class mark for a record. We think about the advantages of the root trademark comparable to the record's quality. We follow the branch relating to that worth and leap to the following center point dependent on our examination.

3.3.2.1 Sorts of Decision Trees:

Decision tree sorts are determined by the type of target variable we have. It usually falls into one of two categories:

Categorical Variable Decision Tree: A Categorical Variable Decision Tree is a Decision Tree with an all-out target variable. Continuous Variable Decision Tree: When the objective variable in a decision tree is constant, it is referred to as a Continuous Variable Decision Tree.

The decision to make crucial components with vigor has an impact on the precision of a tree. For grouping and relapse trees, the decision rules are varied.

Decision trees use a variety of algorithms to determine whether or not to split a hub into at least two sub-hubs. The generation of sub-hubs increases the homogeneity of the sub-hubs that result. As a result, we may argue that the hub’s virtue increases in terms of the objective variable. The Decision Tree separates the hubs based on each and every available variable, then chooses the split that results in the most homogeneous sub-hubs.

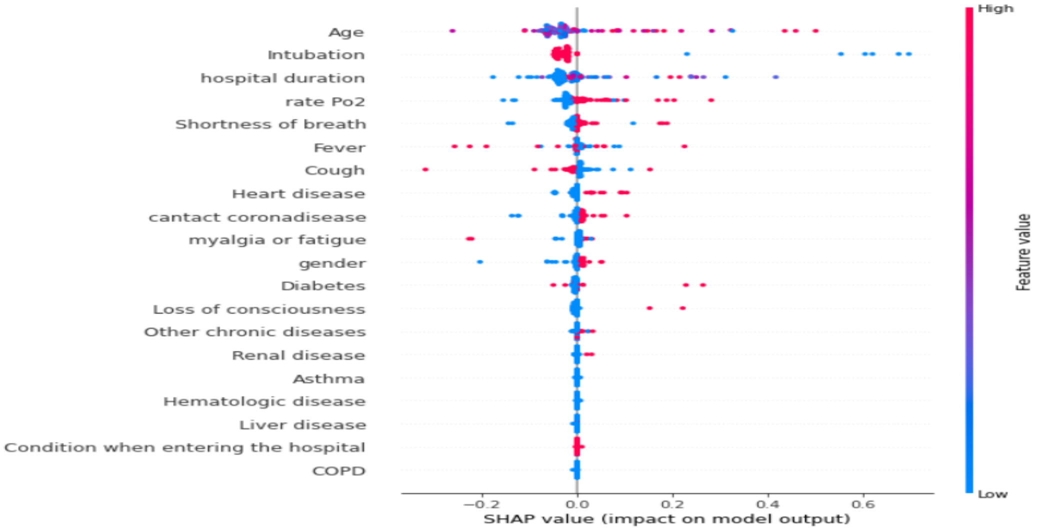

Figure 3.3.2.1: Features Importance using DT

3.3.3 Random Forest

Random Forest is a supervised learning algorithm. This creates a forest and makes it a bit random. The "forest" you are building is a set of decision trees that have been trained most of the time with the "bagging" methods. The general idea of the bagging method is that the combination of learning models increases the overall result. [70]

Random forest has almost the same hyper parameters as a decision tree or bag classifier. With Random Forest, you can also solve regression problems using algorithmic regression coefficient. Random forest adds more randomness to the model as trees grow. Instead of looking for the most important feature when splitting a node, it looks for the best feature from a random subset of features. This results in a great variety which generally leads to a better model. [71]

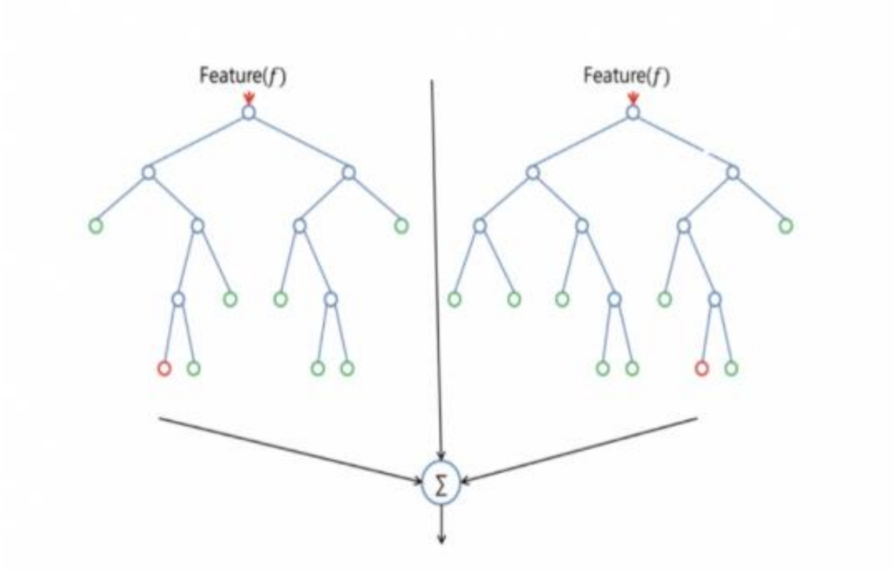

Figure 3.3.3(a): Random Forest [76]

Another good thing about random forests is that it is very easy to measure the relative importance of each feature. Sklearn measures the importance of a resource by looking at how many tree nodes using that resource on average reduce impurities (across all trees in the forest). It automatically calculates this score for each post-workout function and scales the results so that the sum of all meanings is equal to 1.

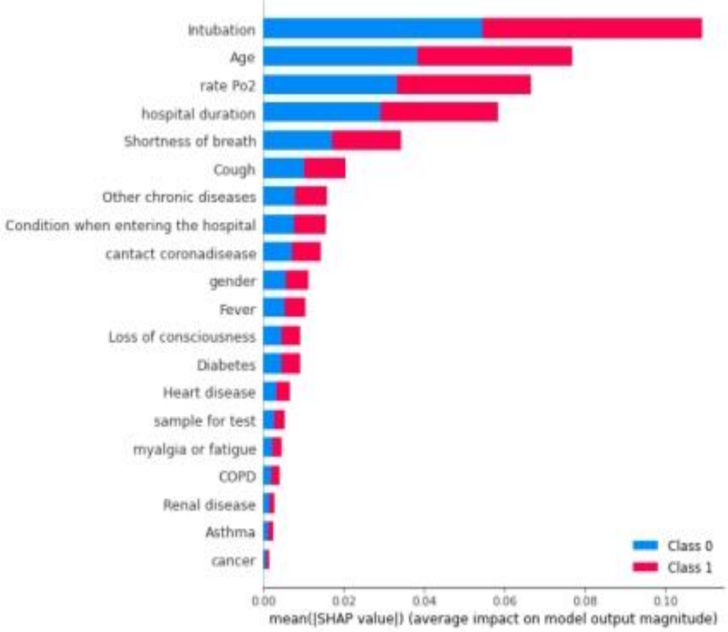

Figure 3.3.3(b): Features Importance using RF

3.3.4 Logistic Regression

Logistic regression is one of the most popular machine learning algorithms, and it comes under the supervised learning technique. It is used to predict category-dependent variables using a given set of independent variables.

Logistic regression predicts the output of a categorically dependent variable. Thus, the result must be a categorical or discrete value. It can be Yes or No, 0 or 1, true or false, etc. but instead of giving the exact value as 0 and 1, give the probabilistic values between 0 and 1.

Logistic regression is similar to linear regression, except for how they are used. Linear regression is used to solve regression problems, while it is used to solve classification problems. Logistic Regression is an important machine learning algorithm because it can provide probabilities and classify new data using continuous and discrete datasets. Now show our features important using SHAP by this algorithm:

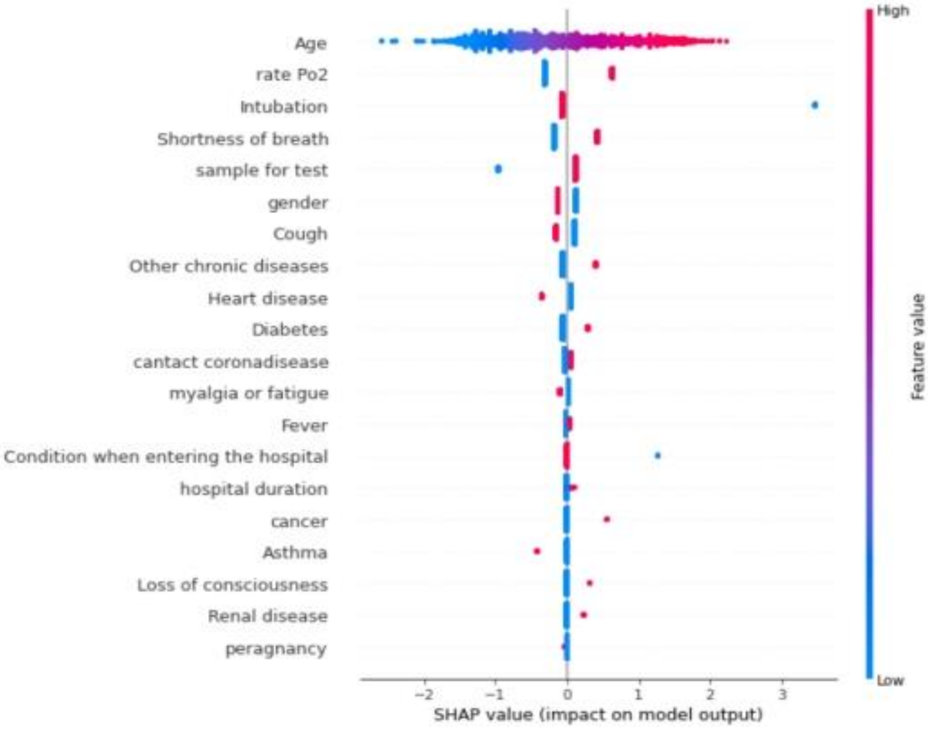

Figure 3.3.4: Features Importance using LR

3.3.5 XGBoost

XGBoost is a gradient-based decision tree implementation designed for speed and performance where competitive machine learning dominates. The XGBoost classification model is called XGBClassifier. We can create and customize this for our training dataset. The model fits using the scikit-learn API and the model.fit() function.

Predictions created by XGBoost are probabilistic by default. Since this is a binary classification problem, each prediction is the probability that the input model belongs to the first class. We can easily convert them to binary class values by rounding them to 0 or 1. [72]

One of the advantages of using gradient boost is that once reinforced trees have been built, it is relatively easy to obtain a significance point for each attribute.

Overall, significance gives a score that indicates how useful or valuable each function is in constructing the model's complementary decision trees. The more the attribute is used to make important decisions with decision trees, the higher its relative importance.

This significance is explicitly calculated for each attribute in the dataset. This makes it possible to rank attributes and compare them.

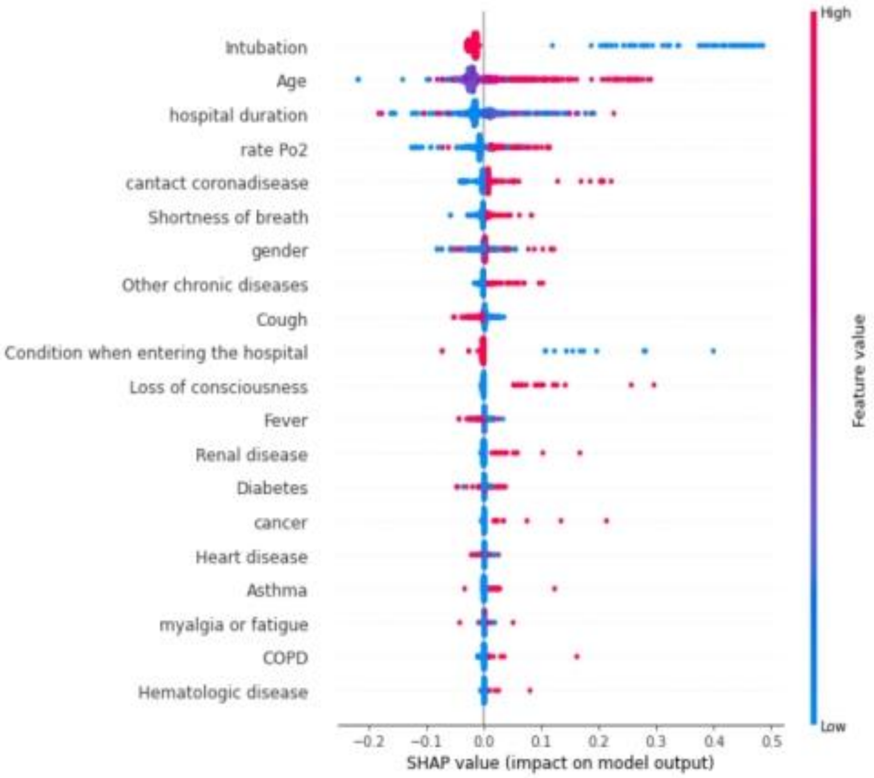

Figure 3.3.5: Features Importance using XGBoost

Chapter 4

Results

In this section, we will review the results obtained from the various mentioned methods and use of classification and compare the performance of classifiers by method, then the best approach to classification. Comparison between the selection method and prediction of Covid-19. Statistical parameters of simulation results to compare classifier performance, such as confusion matrix, model accuracy, precision, recall, and area under the curve (AUC), were used to characterize how well the classifier performed in modulation in terms of identifying a particular data point.

4.1 Performance Metrics

This study focuses on comparing different classification issues and then focusing on classification using the classified performance matrix. The tagged variable 1(recovery) signifies it is a positive occurrence and clearly relates to the patient having in good health. On the other hand, 0(dead) denotes a negative occurrence, indicating that the patient died.

4.1.1 Confusion Matrix

Confusion matrix is universally acknowledged as a simple to comprehend matrix, and it is likely the most popular matrix used to determine the precision and correctness of a prototype. A confusion matrix is a summary of classification problem prediction outcomes. The format of the confusion matrix allows users to visualize the effectiveness of the matrix. In this case, instances are actual classes that each row of the matrix in figure 4-1 represents. A single column, on the other hand, attaches the exemplification to a pre-defined class or in the opposite direction. For a better understanding of a Confusion Matrix, the following words and their explanations are provided.

True Positives (TP):

In this case, both the expected and actual classes are correct (true) i.e., a classifier predicts that the patient will develop a health condition, but the patient actually has Covid-19. In this case, the classifier predicts the correct decision.

False Positives (FP):

This circumstance occurs when a classifier predicts that a patient has a complication of covid-19 but the patient does not have corona disease, i.e., when a classifier predicts that a patient has a complication of disease but the patient does not have corona. As a result, the classifier is unable to forecast the proper decision in that situation.

True Negatives (TN):

True Negatives say that the anticipate class and the actual class are both false (0), i.e., when a classifier predicts that a patient would have no corona virus complications, the patient actually does not have corona. As a result, this classifier correctly predicts the outcome.

False Negatives (FN):

Essentially, in this instance, the Classifier's expected class is false (0), while the actual class is correct (true) (1), i.e., the Classifier predicts that the patient has no corona virus, but the patient actually has corona effect [4]. As a result, the classifier, like False Positives, is unable to anticipate right decisions.

As a result, when more TP and TN are detected inside the confusion matrix, the classifier's accuracy improves. Similarly, a classifier's accuracy is important. In a Confusion Matrix, as the amount of FP and FN increases, decreases. As a result, the ideal circumstance is for none of the FP and FN to be detected inside the model. If this occurs, the model will be able to provide us with 100 percent accuracy [13].

Table 4.1.1: Confusion Matrix.

	Predictive Negative	**Predictive Positive**
Actual Negative	**True Negative (TN)**	**False Positive (FP)**
Actual Positive	**False Negative (FN)**	**True Positive (TP)**

4.1.2 Accuracy

The proportion of correct expectation assembled by the classification data model over the total number of anticipations assembled by the classifier is known as accuracy. We can expect good accuracy if the target variable classes in a dataset are approximately balanced.

$$\textbf{Accuracy} = \frac{(\mathbf{TP} + \mathbf{TN})}{(\mathbf{TP}+\mathbf{FP}+\mathbf{TN}+\mathbf{FN})} \qquad \mathbf{1}$$

4.1.3 Precision

Precision is called that which generate ratio of True Positives to the summation of True Positives and False Positives. Simply high precision means that an algorithm generated mostly appropriate results than inappropriate.

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad 2$$

4.1.4 Recall

Recall is a measure of the proportion of patients that were predicted to have the complications among those patients that actually have the virus. A high recall show that an algorithm generated maximum appropriate results. The formula of recall is given below-

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad 3$$

4.1.5 F1 Score

F1 Score is the Harmonic Mean between precision and recall. Additionally, weighted average of precision and recall is known as F1 score. The span of F1 score is from 0 to 1 [8]. F1 score represents how accurate the classifier is and also shows how durable that is at the same time.

$$\textbf{F1 Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad 4$$

4.2 Model Performances

4.2.1 Logistic Regression (LR)

Before using SHAP, the accuracy of the Logistic Regression model was 91 percent, but after using SHAP, the accuracy jumped to 92 percent. Along with Accuracy, additional performance measures Precision, Recall, and F1 score all improve after SHAP is used, as seen in Figure (4.2.1.a) and Figure (4.2.1.b) which depicts the confusion matrix of the model before SHAP.

Table 4.2.1.a: LR Classification

	Precision	Recall	F1-score	Support
0	0.93	0.98	0.95	204
1	0.69	0.36	0.47	25
Accuracy			0.91	229
Macro AVG	0.81	0.67	0.71	229
Weighted AVG	0.91	0.91	0.90	229

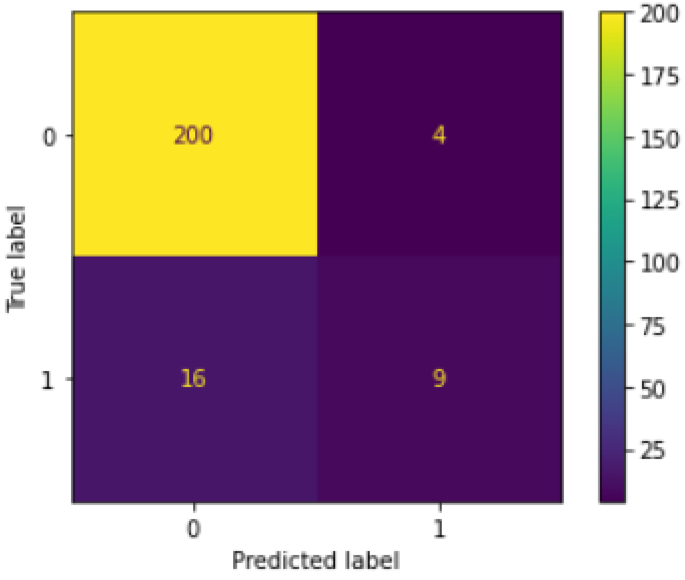

Figure 4.2.1.b: LR Confusion Matrix

4.2.2 XGBoost

Before using SHAP, the accuracy of the XGBoost model was 92 percent, but after using SHAP, the accuracy increased to 95 percent. Other performance measures Precision, Recall, and F1 score increase following the inclusion of SHAP, as shown by Figure (4.2.2.a – 4.2.2.b), and Figure (4.2.2.c) which represents the confusion matrix of the model. Clear understanding show the below figure :

Table 4.2.2.a: XGBoost Classification without SHAP

	Precision	Recall	F1-score	Support
0	0.93	0.98	0.96	204
1	0.73	0.44	0.55	25
Accuracy			0.92	229
Macro AVG	0.83	0.71	0.75	229
Weighted AVG	0.91	0.92	0.91	229

Table 4.2.2.b: XGBoost Classification with SHAP

	Precision	Recall	F1-score	Support
0	0.96	0.99	0.97	211
1	0.82	0.50	0.62	18
Accuracy			0.95	229
Macro AVG	0.89	0.75	0.80	229
Weighted AVG	0.95	0.95	0.94	229

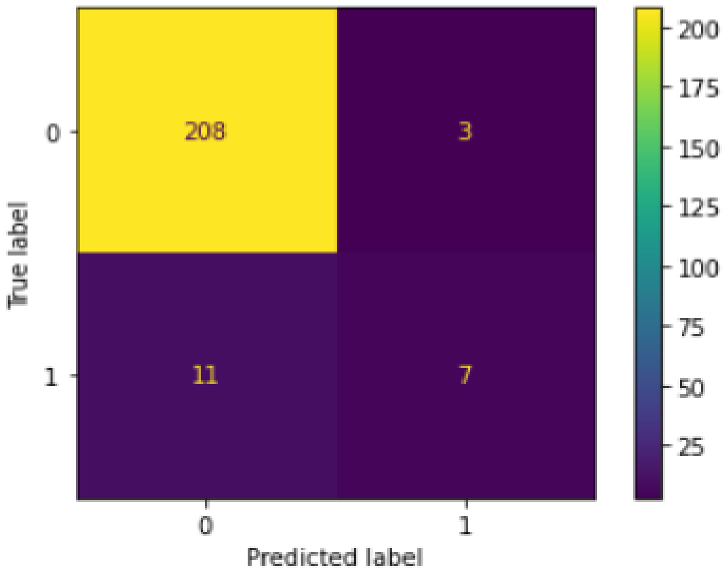

Figure 4.2.2.c: XGBoost Confusion Matrix

4.2.3 Support Vector Machine(SVM)

Before using SHAP, the accuracy of the SVM model was 89 percent, but after using SHAP, the accuracy increased 91 percent. Other performance measures Precision, Recall, and F1 score increase after the inclusion of SHAP, as shown in Figure (4.2.3.a and 4.2.3.b). The confusion matrix of the model is Figure (4.2.3.c). For greater understanding and clarity, show the performance of the Support Vector Machine using SHAP.

Table 4.2.3.a: SVM Classification without SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.89	1.00	0.94	204
1	0.00	0.00	0.00	25
Accuracy			0.89	229
Macro AVG	0.45	0.50	0.47	229
Weighted AVG	0.79	0.89	0.84	229

Table 4.2.3.b: SVM Classification with SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.93	0.97	0.95	204
1	0.62	0.40	0.49	25
Accuracy			0.91	229
Macro AVG	0.78	0.69	0.72	229
Weighted AVG	0.90	0.91	0.90	229

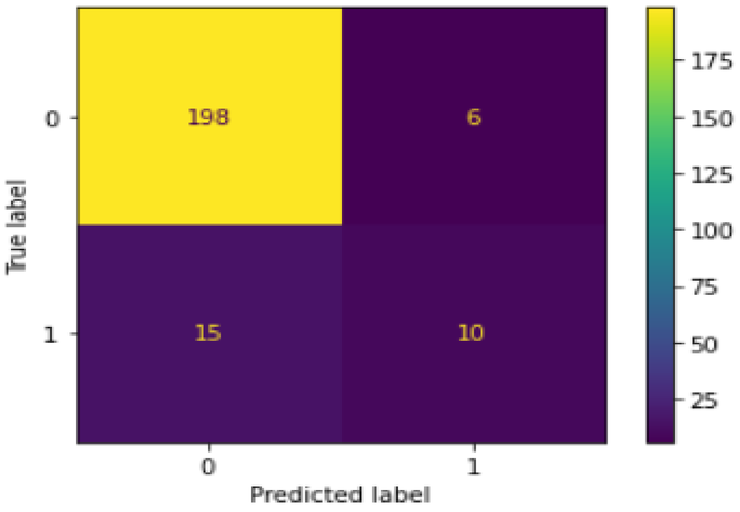

Figure 4.2.3.c: SVM Confusion Matrix

4.2.4 Decision Tree(DT)

Before using SHAP, the accuracy of the Decision Tree model was 86 percent, however after using SHAP, the accuracy increased to 91 percent. Other performance indicators, in addition to accuracy, Precision, recall, and F1 score all after SHAP is implemented, as seen in Figure (4.2.4.a and 4.2.4.b). Figure 4.2.4.c which depicts the confusion. For better understanding and clarity, show the performance of the Decision Tree before and after employing SHAP.

Table 4.2.4.a: DT Classification without SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.93	0.91	0.92	204
1	0.40	0.48	0.44	25
Accuracy			0.86	229
Macro AVG	0.67	0.70	0.68	229
Weighted AVG	0.88	0.86	0.87	229

Table 4.2.4.b: DT Classification with SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.96	0.94	0.95	211
1	0.43	0.56	0.49	18
Accuracy			0.91	229
Macro AVG	0.70	0.75	0.72	229
Weighted AVG	0.92	0.91	0.91	229

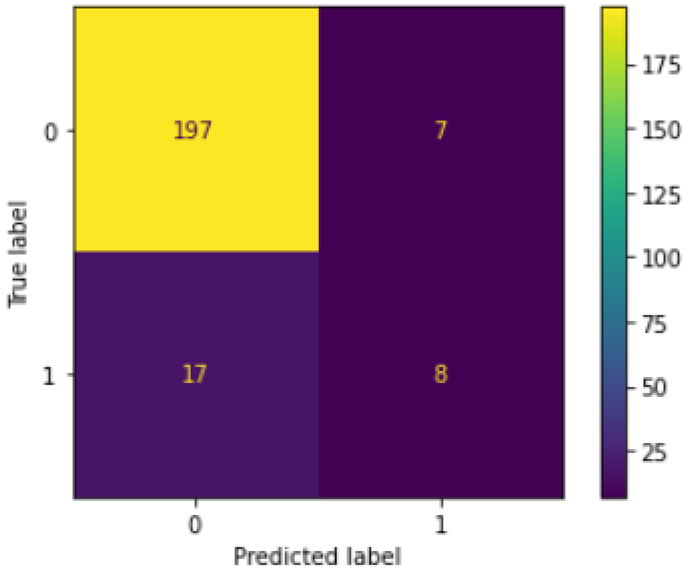

Figure 4.2.4.c: DT Confusion Matrix

4.2.5 Random Forest(RF)

Before using SHAP, the accuracy of the Random Forest model was 87 percent, but after using SHAP, the accuracy to 93 percent. Other performance indicators, in addition to accuracy, Precision, recall, and F1 score all for better understanding and clarity Figure (4.2.5.a to 4.2.5.c) show the performance of Random Forest before and after employing SHAP.

Table 4.2.5.a: RF Classification without SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.93	0.93	0.93	204
1	0.42	0.44	0.43	25
Accuracy			0.87	229
Macro AVG	0.68	0.68	0.68	229
Weighted AVG	0.88	0.87	0.87	229

Table 4.2.5.b: RF Classification with SHAP

	Precision	**Recall**	**F1-score**	**Support**
0	0.95	0.98	0.97	211
1	0.64	0.39	0.48	18

Accuracy			0.93	229
Macro AVG	0.79	0.68	0.72	229
Weighted AVG	0.92	0.93	0.93	229

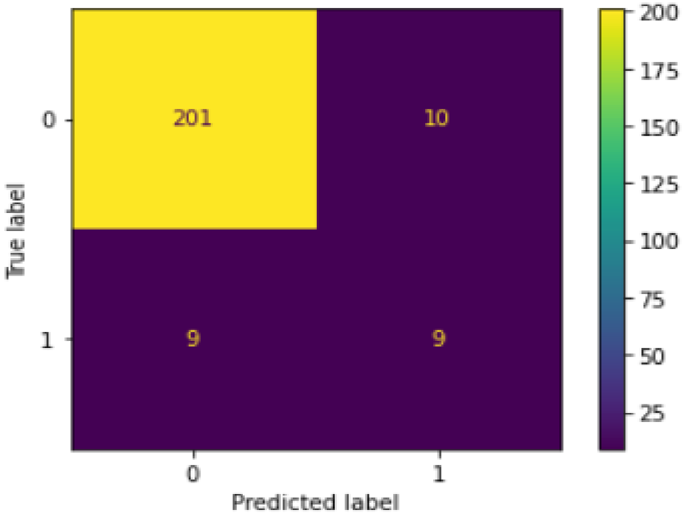

Figure 4.2.5.c: RF Confusion Matrix

4.3 Graphical Comparison Among the Algorithms

Basically here we used five type of algorithms to get best accuracy. From these five algorithm's we got the best accuracy from XGBoost algorithm about 20.7%. Using Logistic Regression and Random Forest (RF) algorithm we got 20.2% accuracy. Accordingly, from support vector machine and Decision tree algorithm we got 19.5% accuracy. From the above discussion with seeing the graph we can easily declared that XGBoost algorithm is the best algorithm to find the best accuracy.

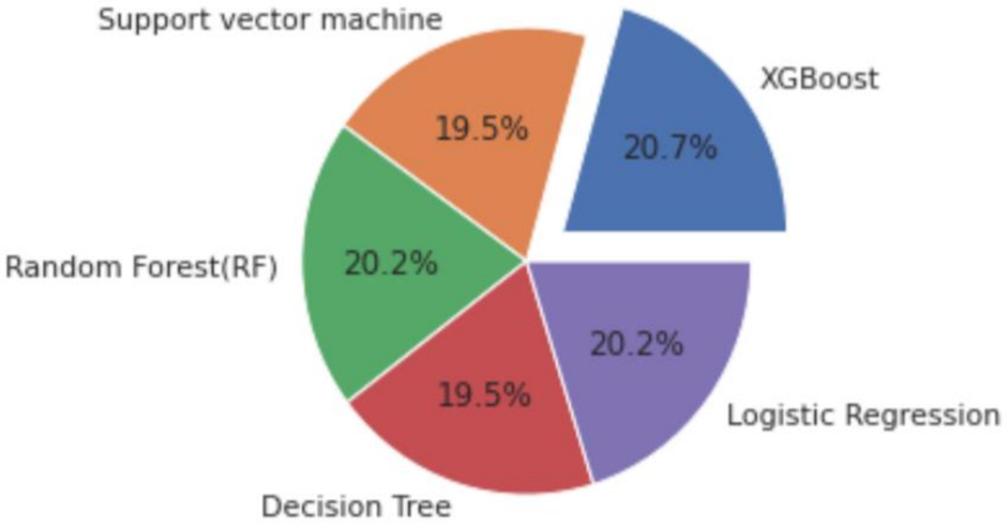

Figure 4.3: Percentage Chart of All Method

4.4 Comparison and Analysis Between the Algorithms

In this table,we can see five algorithm models. These are: XGBoost (xg), Random Forest Classifier, Decision Tree Classifier, Support Vector Machine, Logistic Regression. We can see here some accuracy value. We find and map all this accuracy without using SHAP, and also using SHAP. When we use the SHAP we get better accuracy than the without using SHAP. In the XGBoost algorithm we get 95% accuracy using SHAP and without using SHAP get 92% accuracy. Accordingly using the SHAP maximum value of the algorithm greater than to without using SHAP

Table 4.4(a): All Method Score Comparison Table

	Model	**Accuracy without SHAP%**	**Accuracy with SHAP%**
1	**XGBoost (xg)**	.92	.95
2	**Random Forest Classifier (RF)**	.87	.93
3	**Decision Tree Classifier (DT)**	.86	.91
4	**Support Vector Machine (SVM)**	.89	.91
5	**Logistic Regression(LR)**	.91	.92

Table 4.4(b): Matrix value of Classifier

Classifier for 5-class	Accuracy	Precision		Recall		F1 Score	
		Death (1)	Recover(0)	Death (1)	Recover(0)	Death (1)	Recover (0)
RF	.93	.42	.93	.44	.93	.43	.93
SVM	.91	.62	.93	.40	.97	.49	.95
DT	.91	.43	.96	.56	.94	.49	.95
LR	.91	.69	.93	.36	.98	.47	.95
XGBoost	.95	.82	.96	.50	.99	.62	.97

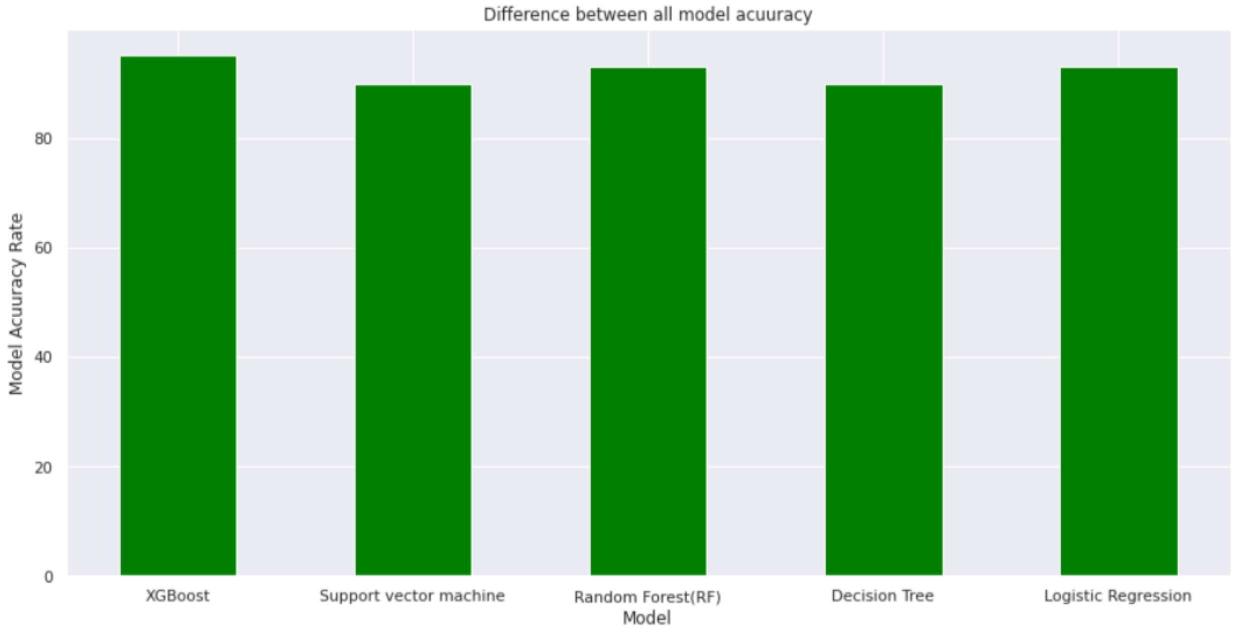

Figure 4.4(c): All Method Comparison graph

4.5 AUC & ROC Curves

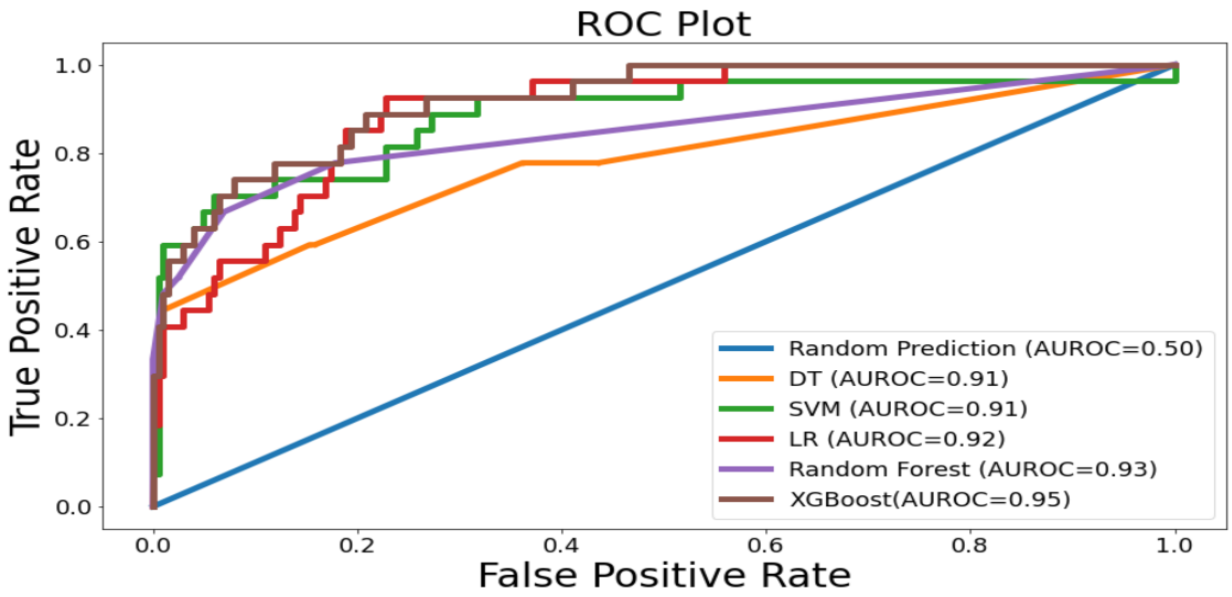

Figure 4.5: AUC & ROC Curve (XGB, SVM, LR, RF and DT)

4.6 Important Feature

Table 4.6: Top 5 features for four different age groups

Datasets	**Algorithms**	**Top Feature Age Wise**				
		1st Feature	2nd Feature	3rd Feature	4th Feature	5th Feature
Age (0-34)	**XGBoost**	Shortness of breath	Hospital duration	Contact corona diseases	Age	Intubation
	LR	Shortness of breath	Contact corona disease	Fever	Age	Cough
	RF	Intubation	Condition when entering the hospital	Hospital duration	Age	Shortness of breath
	SVM	Fever	Shortness of breath	Other chronic diseases	Hospital duration	Gender
	DT	Condition when entering the hospital	Age	Hospital duration	Shortness of breath	Intubation
Age (35-47)	**XGBoost**	Rate Po2	Gender	Hospital duration	Age	Cough
	LR	Rate Po2	Gender	Cough	Fever	Shortness of breath
	RF	Intubation	Condition when entering the hospital	Hospital duration	Rate Po2	Age
	SVM	Intubation	Hospital duration	Rate Po2	Heart diseases	Age
	DT	Intubation	Rate Po2	Condition when entering the hospital	Hospital duration	Age

Age (48-64)	**XGBoost**	Contact Corona diseases	Hospital duration	Rate Po2	Gender	Age
	LR	Rate Po2	Intubation	Age	Hospital duration	Cough
	RF	Intubation	Hospital duration	Age	Rate Po2	Gender
	SVC	Intubation	Gender	Contact corona disease	Rate Po2	Cancer
	DT	Intubation	Hospital duration	Age	Other Chronic diseases	Gender
Age (64+)	**XGBoost**	Intubation	Diabetes	Rate Po2	Contact corona disease	Hospital duration
	LR	Intubation	Age	Rate Po2	Heart disease	Cough
	RF	Intubation	Age	Hospital duration	Contact Corona diseases	Rate Po2
	SVC	Intubation	Gender	Contact corona disease	Rate Po2	Cancer
	DT	Intubation	Hospital duration	Age	Gender	Contact corona diseases

4.7 Top Ten Features

Table 4.7: Comparison of top 10 features (XGB, SVM, LR, RF and DT)

XGBoost	Logistic Regression	SVM	Decision Tree	Random Forest
Hospital duration	Age	Gender	Age	Intubation
Age	Rate P02	Contact corona diseases	Intubation	Age
Intubation	Intubation	Intubation	Hospital duration	Rate Po2
Shortness of breath	Shortness of breath	Rate Po2	Rate Po2	Hospital duration
Rate Po2	Gender	Cancer	Shortness of breath	Shortness of breath
Cough	Cough	Liver disease	Fever	Cough
Gender	Contact corona	Diabetes	Cough	Other

	diseases			chronic diseases
Contact corona diseases	Other chronic diseases	Hematologic disease	Heart diseases	Condition when entering the hospital
Other chronic diseases	Fever	HIV/AIDS	Contact corona diseases	Contact corona diseases
Fever	Heart diseases	Immune deficiency(Acquired or Congenital)	Myalgia or Fatigue	Gender

4.8 XGBoost Top Ten Chart

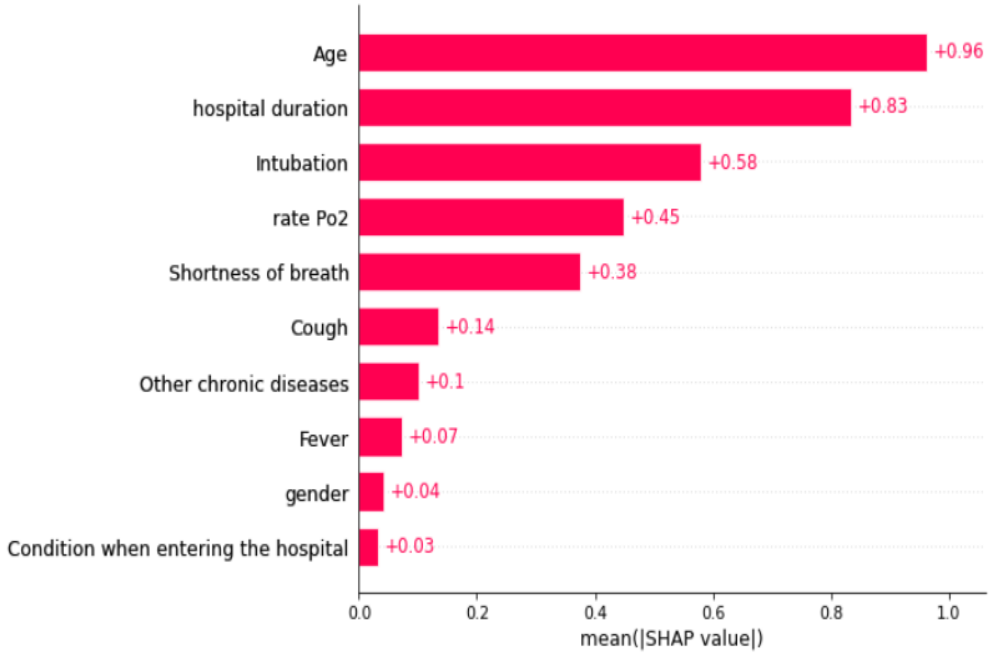

Figure 4.8: Top Ten Bar Chart

4.9 XGBoost Top Ten Summary Plot

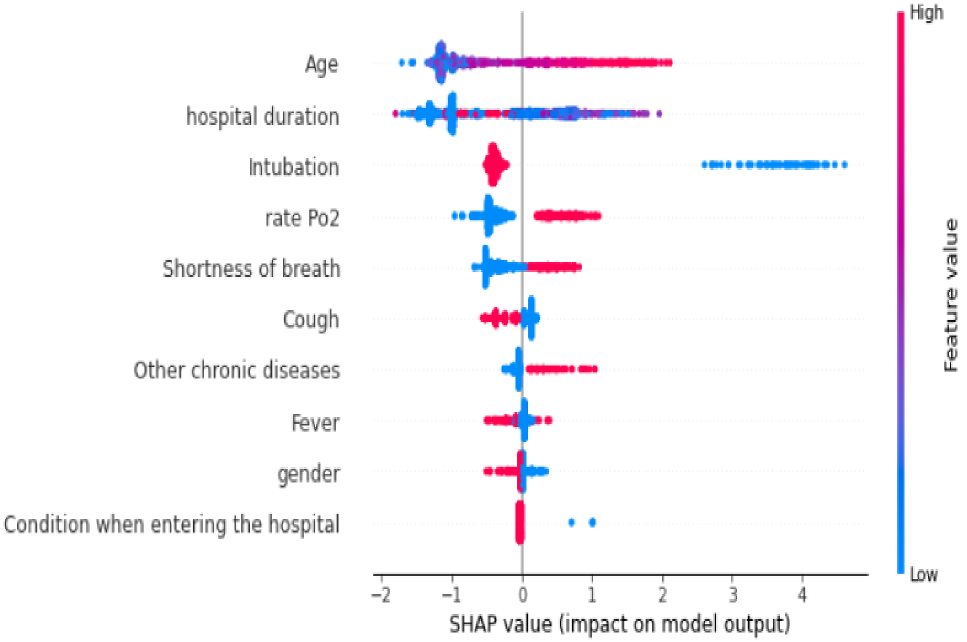

Figure 4.9: Top Ten Summery Plot

Chapter 5

Discussion

The contagious disease COVID-19 has shocked the world and continues to threaten the lives of billions of people. For this reason, early detection of COVID-19 patients is an important process for the treatment and control of the disease. The literature review work indicates that the optimal technique has not yet been determined. Therefore, our challenge is to find a sufficiently fast and accurate detection strategy. In this work, we introduced an accurate and intelligent detection strategy that can provide a smart medical diagnosis. In our detection strategy, the essential part is to find out what the best feature is causing the covid-19 and also find out the accuracy of the feature, recall, precision, etc. The proposed feature selection methodology is called SHapley Additive exPlanations (SHAP), which combines between the benefits of both filter and wrapper selection methods.

The results presented in Table 4.6 indicate that the Ten most important symptoms of COVID-19, starting from the most important to least important:

Using XGBoost Method: Hospital duration, Age, Intubation, Shortness of breath, Rate Po2, Cough, Gender, Contact corona diseases, Other chronic diseases, Fever.

Using LR Method: Age, Rate P02, Intubation, Shortness of breath, Gender, Cough, Contact corona diseases, Other chronic diseases, Fever, Heart diseases.

Using SVM Method: Gender, Contact corona diseases, Intubation, Rate Po2, Cancer, Liver disease, Diabetes, Hematologic disease, HIV/AIDS, Immune deficiency(Acquired or Congenital).

Using DT Method: Age, Intubation, Hospital duration, Rate Po2, Shortness of breath, Fever, Cough, Heart diseases, Contact corona diseases, Myalgia or Fatigue.

Using RF Method: Intubation, Age, Rate Po2, Hospital duration, Shortness of breath, Cough, Other chronic diseases, Condition when entering the hospital, Contact corona diseases, Gender.

In summary, for our COVID-19 dataset, the use of different feature selection measures provided different importance levels for the top-ten ranked symptoms, and even different sets of important symptoms. This is due to the fact that each of the ten measures for feature selection (or importance) ranks the symptoms based on different data characteristics.

Here we have used five different algorithms. Among the five algorithms, we found that XGBoost has the best accuracy. The top 5 features selected by XGBoost are given below:

Age (0-34): Shortness of breath, Hospital duration, Contact corona diseases, Age, Intubation.

Age (35-47): Rate Po2, Gender, Hospital duration, Age, Cough.

Age (48-64): Contact Corona diseases, Hospital duration, Rate Po2, Gender, Age.

Age (64+): Intubation, Diabetes, Rate Po2, Contact corona disease, Hospital duration.

Table 4.4 shows the precision for every technique. Four out of the five models offer scoring precision above 78%, with the best outcomes XGBoost (exactness 0.93) trailed by LR, RF, DT and SVM. Our XGBoost accuracy gives better execution looked at than any remaining accuracy. By utilizing SHAP, the accuracy of our model is more than 89% and XGBoost is 95%.

Chapter 6

Conclusion

ML algorithms, such as XGBoost, LR, RF, DT, and SVM Classifiers were applied on the Covid19 to select the best possible classification algorithm for the selection of the death and recovered cases, then the performance enhancement of the specified classifiers in terms of features selection was performed, such a task can be useful to the healthcare providers in identifying and diagnose covid19 cases in a better efficient way, also, a feature importance algorithm was applied on the mentioned dataset to evaluate all features importance on the available dataset and to understand what was the main contributors to the severe cases of patients.

Results show without feature selection method that XGBoost classifier gives the accuracy with 93.32% and ROC = 0.9318, compared to the other classifiers of accuracy of 80.70%, 78.77%, 91.21% and 78.14% for RF, DT, LR and SVM, respectively. When using the feature selection method (SHAP), XGBoost classifier gives outstanding accuracy Covid-19 case with 95.20% and by using only 10 out of the total 26 available features, which can be a useful fact for healthcare providers in identifying possible infected Covid19 cases.

Taken together, the results suggest that symptoms of hospital duration, cough, intubation, rate Po2 and shortness of breath should be considered as important symptoms in the diagnosis of patients with COVID-19, with particular emphasis on short of breath and intubation symptoms.

Going forward, this work will encourage researchers to investigate and develop a critical assessment of the nature of the coronavirus for different countries/regions. Such an investigation would assess the extent of potential mutations of Covid19 by country or region and, if so, the significance of the multiregional characteristics. May reflect potential virus mutations by country/region. It will help vaccine developers determine the necessary treatment/vaccination. According to the selection of important things, characteristics that contribute to the well-being of Covid-19 patients.

References

[1] C. Liu, Q. Zhou, Y. Li, L. V. Garner, S. P. Watkins, and L. J. Carter, “Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases,” *ACS Central Science*, vol. 6, no. 3, 2020.

[2] X. Chen, W. Han, G. Wang, and X. Zhao, “Application prospect of polysaccharides in the development of anti-novel coronavirus drugs and vaccines,” International Journal of Biological Macromolecules, vol. 164, 2020.

[3] N. Altini et al., “Predictive machine learning models and survival analysis for COVID-19 prognosis based on hematochemical parameters,” Sensors (Basel), vol. 21, no. 24, p. 8503, 2021.

[4] M. Ciotti et al., “COVID-19 outbreak: An overview,” Chemotherapy, vol. 64, no. 5–6, pp. 215–223, 2019.

[5] X. Li et al., “Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan,” J. Allergy Clin. Immunol., vol. 146, no. 1, pp. 110–118, 2020.

[6] G. Lippi and M. Plebani, “Procalcitonin in patients with severe coronavirus disease 2019 (COVID-19): A meta-analysis,” Clin. Chim. Acta, vol. 505, pp. 190–191, 2020.

[7] “COVID live - Coronavirus statistics - worldometer,” Worldometers.info. [Online]. Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1? [Accessed: 01-Mar-2022].

[8] Q. Liu, K. Xu, X. Wang, and W. Wang, “From SARS to COVID-19: What lessons have we learned?,” J. Infect. Public Health, vol. 13, no. 11, pp. 1611–1618, 2020.

[9] R.-H. Du et al., “Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study,” Eur. Respir. J., vol. 55, no. 5, p. 2000524, 2020.

[10] A. S. Yadaw, Y.-C. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, “Clinical features of COVID-19 mortality: development and validation of a clinical prediction model,” The Lancet Digital Health, vol. 2, no. 10, pp. e516–e525, 2020.

[11] E. Kim, G. Erdos, S. Huang, T. W. Kenniston, S. C. Balmert, and C. D. Carey, “Microneedle array delivered recombinant coronavirus vaccines: Immunogenicity and rapid translational development,” EBioMedicine, vol. 55, 2020.

[12] K. Dhama, K. Sharun, R. Tiwari, M. Dadar, Y. S. Malik, and K. P. Singh, “COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics,” Human vaccines & immunotherapeutics, vol. 16, 2020.

[13] U. De León, Á. G. Pérez, and E. Avila-Vales, “An SEIARD epidemic model for COVID-19 in Mexico: mathematical analysis and state-level forecast,” Chaos, Solitons & Fractals, vol. 140, 2020.

[14] R. H. Mena, J. X. Velasco-Hernandez, N. B. Mantilla-Beniers, G. A. Carranco-Sapiéns, L. Benet, and D. Boyer, “Using posterior predictive distributions to analyse epidemic models: COVID-19 in Mexico City,” Physical biology, vol. 17, no. 6, 2020.

[15] B. Ivorra, M. R. Ferrández, M. Vela-Pérez, and A. Ramos, “Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China,” Communications in nonlinear science and numerical simulation, vol. 88, 2020.

[16] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, and S. Merler, “The effect of

travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," Science, vol. 368, no. 6489, 2020.

[17] M. Zens, A. Brammertz, J. Herpich, N. Südkamp, and M. Hinterseer, "App-based tracking of self-reported COVID-19 symptoms: analysis of questionnaire data," Journal of medical Internet research, vol. 22, no. 9, 2020.

[18] K. Yamamoto *et al.*, "Health observation app for COVID-19 symptom tracking integrated with personal health records: Proof of concept and practical use study," *JMIR MHealth UHealth*, vol. 8, no. 7, p. e19902, 2020.

[19] D. A. Drew, L. H. Nguyen, C. J. Steves, C. Menni, M. Freydin, and T. Varsavsky, "Rapid implementation of mobile technology for real-time epidemiology of COVID-19," Science, vol. 368, no. 6497, 2020.

[20] J. Singh, M. B. Green, S. Lindblom, M. S. Reif, N. P. Thakkar, and A. Papali, "Telecritical Care Clinical and Operational Strategies in Response to COVID-19," Telemedicine and e-Health.

[21] M. P. Mcrae, I. P. Dapkins, I. Sharif, J. Anderman, D. Fenyo, and O. Sinokrot, "Managing COVID-19 With a Clinical Decision Support Tool in a Community Health Network: Algorithm Development and Validation," Journal of medical Internet research, vol. 22, no. 8, 2020.

[22] M. Zawiah, F. Y. Al-Ashwal, R. M. Saeed, M. Kubas, S. Saeed, and A. H. Khan, "Assessment of Healthcare System Capabilities and Preparedness in Yemen to Confront the Novel Coronavirus 2019 (COVID-19) Outbreak: A Perspective of Healthcare Workers," Frontiers in public health, vol. 8, 2020.

[23] M. A. Acuña-Zegarra, M. Santana-Cibrian, and J. X. Velasco-Hernandez, "Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance," Mathematical Biosciences, vol. 325, 2020.

[24] M. Nemati, J. Ansary, and N. Nemati, "Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data," Patterns, vol. 1, no. 5, 2020.

[25] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial Intelligence (AI) applications for COVID-19 pandemic," Diabetes & Metabolic Syndrome, vol. 14, no. 4, 2020.

[26] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," Chaos, Solitons & Fractals, vol. 139, 2020.

[27] M. B. Jamshidi et al., "Artificial Intelligence and COVID-19: Deep Learning approaches for diagnosis and treatment," IEEE Access, vol. 8, pp. 109581–109595, 2020.

[28] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, "New machine learning method for image-based diagnosis of COVID-19," Plos one, vol. 15, no. 6, 2020.

[29] M. T. Vafea, E. Atalla, J. Georgakas, F. Shehadeh, E. K. Mylona, and M. Kalligeros, "Emerging technologies for use in the study, diagnosis, and treatment of patients with COVID-19," Cellular and molecular bioengineering, vol. 13, no. 4, pp. 249–257, 2020.

[30] E. Kana, M. Kana, A. Kana, and R. Kenfack, A web-based diagnostic tool for covid-19 using machine learning on chest radiographs (cxr). medRxiv. 2020.

[31] L. Yan, H. T. Zhang, J. Goncalves, Y. Xiao, M. Wang, and Y. Guo, "An interpretable mortality prediction model for COVID-19 patients," Nature machine intelligence, vol. 2, no. 5, pp. 283–288, 2020.

[32] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, and U. Reuter, "Covid-19 outbreak prediction with machine learning," Algorithms, vol. 13, no. 10, 2020.

[33] A. Ahmad, S. Garhwal, S. K. Ray, G. Kumar, S. J. Malebary, and O. M. Barukab, “The number of confirmed cases of covid-19 by using machine learning: Methods and challenges,” Archives of Computational Methods in Engineering, vol. 1, 2020.

[34] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. .

[35] R. Pal, A. A. Sekh, S. Kar, and D. K. Prasad, “Neural network based country wise risk prediction of COVID-19,” Applied Sciences, vol. 10, no. 18, 2020.

[36] A. Rao and J. A. Vazquez, “Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine,” Infection Control & Hospital Epidemiology, vol. 41, no. 7, pp. 826–830, 2020.

[37] Quiroz-Juárez MA, Torres-Gómez A, Hoyo-Ulloa I, León-Montiel RdJ, U’Ren AB (2021) Identification of high-risk COVID-19 patients using machine learning. PLoS ONE 16(9): e0257234.https://doi.org/10.1371/journal.pone.0257234

[38] M. G. Chang *et al.*, “Time Kinetics of Viral Clearance and Resolution of Symptoms in Novel Coronavirus Infection,” *Am J Respir Crit Care Med*, vol. 201, no. 9, pp. 1150–1152, 2020.

[39] W. T. Li et al., “Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis,” BMC Med. Inform. Decis. Mak., vol. 20, no. 1, 2020.

[40] F. Souza, N. S. Hojo-Souza, E. B. Santos, C. M. Silva, and D. L. Guidoni, Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data. medRxiv. 2020.

[41] M. A. Alzubaidi, M. Otoom, N. Otoum, Y. Etoom, and R. Banihani, “A novel computational method for assigning weights of importance to symptoms of COVID-19 patients,” Artif. Intell. Med., vol. 112, no. 102018, p. 102018, 2021.

[42] Y. Niu et al., “Development of a predictive model for mortality in hospitalized patients with COVID-19,” Disaster Med. Public Health Prep., pp. 1–9, 2021.

[43] K. M. Almustafa, “Covid19-Mexican-patients’ dataset (Covid19MPD) classification and prediction using feature importance,” Concurr. Comput., vol. 34, no. 4, 2022.

[44] A. S. Yadaw, Y. C. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, “Clinical features of COVID-19 mortality: development and validation of a clinical prediction model,” Lancet Digit Health, vol. 2, no. 10, pp. e516–e525, 2020.

[44.1] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794.

[45] Aljouie, Abdulrhman Fahad et al. “Early Prediction of COVID-19 Ventilation Requirement and Mortality from Routinely Collected Baseline Chest Radiographs, Laboratory, and Clinical Data with Machine Learning.” Journal of multidisciplinary healthcare vol. 14 2017-2033. 30 Jul. 2021, doi:10.2147/JMDH.S322431

[46] I. Nachtigall et al., “Clinical course and factors associated with outcomes among 1904 patients hospitalized with COVID-19 in Germany: an observational study,” Clin. Microbiol. Infect., vol. 26, no. 12, pp. 1663–1669, 2020.

[47] M. M. Banoei, R. Dinparastisaleh, A. V. Zadeh, and M. Mirsaeidi, “Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying,” Crit. Care, vol. 25, no. 1, 2021.

[48] Sîrbu A, Barbieri G, Faita F, et al. Early outcome detection for COVID-19 patients. Sci Rep. 2021;11(1):18464. Published 2021 Sep 16. doi:10.1038/s41598-021-97990-1

[49] Kuno T, Sahashi Y, Kawahito S, Takahashi M, Iwagami M, Egorova NN. Prediction of in-hospital

mortality with machine learning for COVID-19 patients treated with steroid and remdesivir [published online ahead of print, 2021 Oct 14]. J Med Virol. 2021;10.1002/jmv.27393. doi:10.1002/jmv.27393

[50] V. Zuccaro et al., “Competing-risk analysis of coronavirus disease 2019 in-hospital mortality in a Northern Italian centre from SMAtteo COvid19 REgistry (SMACORE),” Sci. Rep., vol. 11, no. 1, 2021.

[51] Altini N, Brunetti A, Mazzoleni S, et al. Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters. Sensors (Basel). 2021;21(24):8503. Published 2021 Dec 20. doi:10.3390/s21248503

[52] Y.-H. Zhou et al., “Predictive factors of progression to severe COVID-19,” Open Med. (Warsz.), vol. 15, no. 1, pp. 805–814, 2020.

[53] Y. Yoshida et al., “Clinical characteristics and outcomes in women and men hospitalized for coronavirus disease 2019 in New Orleans,” Biol. Sex Differ., vol. 12, no. 1, 2021.

[54] Assaf D, Gutman Y, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine. 2020;15(8):1435–1443. pmid:32812204

[55] M. Pourhomayoun and M. Shakibi, Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health. 2021.

[56] Bezzan V, Rocco CD. Predicting special care during the COVID-19 pandemic: A machine learning approach. arXiv preprint arXiv:201103143. 2020.

[57] L. Yan, H. T. Zhang, Y. Xiao, M. Wang, C. Sun, and J. Liang, Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan. .

[58] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, and B. Zhu, “An interpretable machine learning framework for accurate severe vs non-severe covid-19 clinical type classification,” vol. 3638427, 2020.

[59] S. Subudhi, A. Verma, A. B. Patel, C. C. Hardin, M. J. Khandekar, and H. Lee, Comparing Machine Learning Algorithms for Predicting ICU Admission and Mortality in COVID19. medRxiv. 2020.

[60] Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med. 2020;180(7):934–943. doi:10.1001/jamainternmed.2020.0994

[61] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods. Cambridge, England: Cambridge University Press, 2013.

[62] J. Luts, F. Ojeda, R. Van De Plas Raf, B. De Moor, S. Van Huffel, and J. A. K. Suykens, “A tutorial on support vector machine-based methods for classification problems in chemometrics,” Analytica Chimica Acta, vol. 665, no. 2, pp. 129–145, 2010.

[63] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Systems with Applications, vol. 36, no. 2, pp. 3240–3247, 2009.

[64] C.-Y. Chang, S.-J. Chen, and M.-F. Tsai, “Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images,” Pattern Recognition, vol. 43, no. 10, pp. 3494–3506, 2010.

[65] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis,” Expert Systems with Applications, vol. 38, no. 7, pp. 9014–9022, 2011.

[66] P. Danenas and G. Garsva, “Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach,” Procedia Computer Science, vol. 9, pp. 1324–1333,

2012.

[67] C. L. Huang, H. C. Liao, and M. C. Chen, “Prediction model building and feature selection with support vector machines in breast cancer diagnosis,” Expert Systems with Applications, vol. 34, no. 1, pp. 578–587, 2008.

[68] H. F. Liau and D. Isa, “Feature selection for support vector machine-based face-iris multimodal biometric system,” Expert Systems with Applications, vol. 38, no. 9, pp. 11105–11111, 2011.

[69] Y. Zhang, Z. Chi, and Y. Sun, “A novel multi-class support vector machine based on fuzzy theories,” in Intelligent Computing: International Conference on Intelligent Computing, Part I (ICIC ’06), vol. 4113, D. S. Huang, K. Li, and G. W. Irwin, Eds. Berlin, Germany: Springer, pp. 42–50.

[70] https://www.kaggle.com/niklasdonges/end-to-end-project-with-python

[71] https://builtin.com/data-science/random-forest-algorithm

[72] https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

[73] https://images.app.goo.gl/wPMeauJGUdCxkWMFA

[74] https://images.app.goo.gl/CYbTh585H7VwbJv89

[75] https://images.app.goo.gl/yzWPXDWiJV8uVBjK7

[76] https://images.app.goo.gl/XuG22KnhPpXxiBR79

[77] https://hazard model.herokuapp.com/covid