Machine Learning based Prediction and Evaluation of COVID-19 Patient's Symptoms Data from Rawalpindi and AJK, Pakistan Region.

By

Hajira Wahid

276947

Supervisor

Dr. Rabia Irfan

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MS CS)

In

School of Electrical Engineering and Computer Science,

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(JUNE 2022)

Approval

It is certified that the contents and form of the thesis entitled "**Machine Learning based Prediction and Evaluation of COVID-19 Patient's Symptoms Data from Rawalpindi and AJK, Pakistan Region.**" submitted by **Hajira Wahid** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Rabia Irfan**

Signature: _____

Date: _____

Committee Member 1: **Dr. RAFIA MUMTAZ**

Signature: _____

Date: _____

Committee Member 2: **Dr. YASIR FAHEEM**

Signature: _____

Date: _____

Committee Member 3:

Signature: _____

Date: _____

Abstract

COVID-19 was discovered to be an infectious and potentially fatal viral disease, and its quick and extensive spread has turned it into one of the world's most critical problems. People across the globe were facing an alarming threat due to limited resources, especially in developing countries. Prediction models incorporating multivariate regression to assess the risk of infection have been designed. Some other models incorporate symptoms-based predictions but with limited and incomplete sets of clinical symptoms. In this thesis work, we proposed a machine learning approach in which we will be able to predict COVID-19 and the severity of its patient. Our model is trained on 6000 clinical records from Holy Family Hospital Rawalpindi and AJK Health Department Pakistan, in which 3000 patients were tested positive. 1365 of the 3000 patients were in serious condition. The proposed model utilized ten features including cough, fever, sore throat, shortness of breath, headache, flu, body ache, loss of taste&smell, and diarrhea. To measure the performance of the model, predictive analysis employs the AUC curve and average precision (AP). The Shapley additive explanations (SHAP) have been utilized for descriptive analysis to investigate the most sensitive features. Machine Learning model random forest outperformed with AUC: (AP=0.98) among other models like Support Vector, KNN, and Logistic Regression. Our approach demonstrates significant prediction accuracy and

can be implemented as a COVID-19 screening tool as well as a technique to identify the severity of this disease. The proposed methodology can be utilized to prioritize testing and evaluation purposes for future investigations and insights.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Hajira Wahid**

Signature: _____

Acknowledgment

First and foremost, I am exceptionally grateful to Allah Almighty Without whose help this thesis work couldn't have been completed. All the help, guidance, and support from my parents and teachers were all because of Allah's will.

I am grateful to my supervisor, Assistant Professor, Dr. RABIA IRFAN, for the tremendous supervision, motivation, and guidance she has conveyed all through my time as his understudy. I have been exceptionally honored to have a supervisor who thought such a great amount about my work, and who acknowledged my requests and questions immediately.

I am especially appreciative to my parents who supported me throughout my life, and who were steadfast with me in all circumstances I had to face during this thesis work. I am grateful to them for providing every possible type of support. My heartfelt dedication to my late grandmother, who didn't survive during this deadly pandemic.

Thank You

Hajira Wahid

.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

The novel coronavirus originated in the Chinese city of Wuhan in December 2019 [1] and was reported to the World Health Organization (W.H.O) on December 31, 2019. COVID-19 was the name given to the virus that caused a global danger W.H.O on February 11th, 2020. The World Health Organization (W.H.O) proclaimed this flare-up a general wellbeing crisis and stated that the virus spreads through the respiratory system when a healthy person comes into contact with a contaminated individual [2]. This pandemic continues to pose several challenges to medical systems around the world [3], including increased demand for hospital beds and critical shortages of medical equipment, as well as the infection of many healthcare personnel. As a result, the ability to make quick clinical choices and make efficient use of healthcare resources is critical [4].

1.1 Introduction

Coronavirus epidemics have raised global concern over the last two decades, including one in 2003 with the Severe Acute Respiratory Syndrome (SARS) and another in 2012 with the Middle East Respiratory Syndrome (MERS)

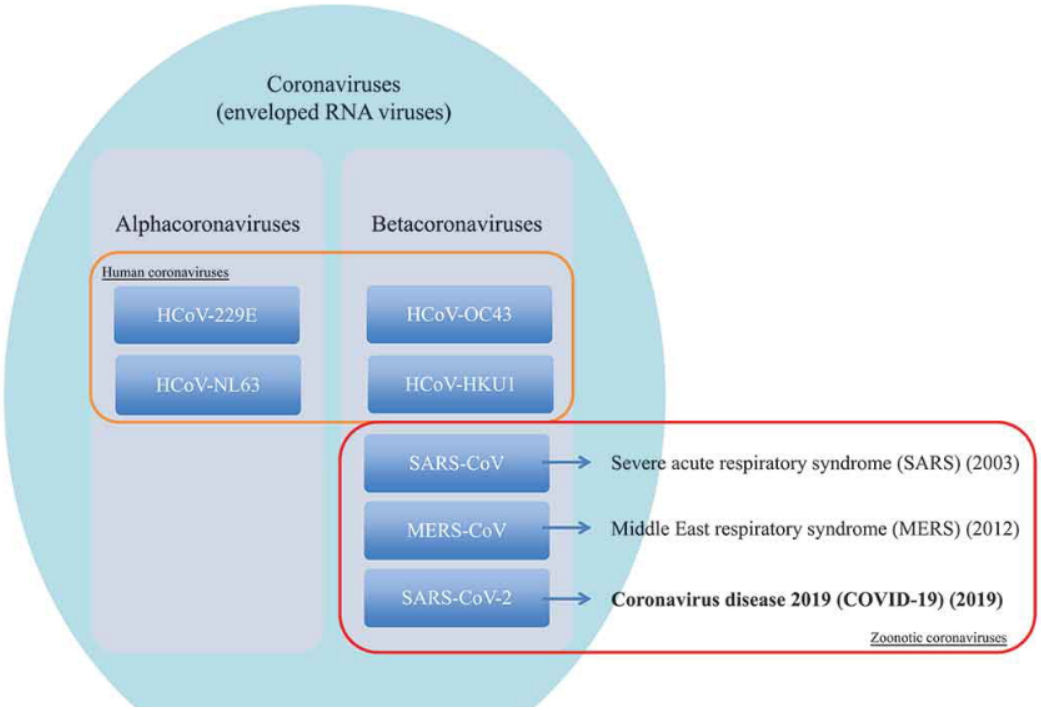


Figure 1.1: Coronavirus Disease–2019 [6]

and now SARS-CoV-2 in 2019, as illustrated in Figure 1.1. The coronavirus disease (COVID-19) has impacted negatively people's lives and killed a great number of individuals around the world. According to World Health Organization, the disease has spread to practically every country, killing over 580,000 individuals among nearly 13,379,000 confirmed cases estimated as of the middle of July 2020 [5]. To reduce the effects of the COVID-19 pandemic, many governments have proposed intervention strategies. Science and technology have made substantial contributions during this unique and turbulent period. For example, in hospitals to bring food and medicine to coronavirus patients robots are employed. Many research scientists are scrambling to examine treatments and therapies to treat sick patients, while some are working on vaccines to prevent the illness [7].

So far, data suggests that the virus spreads from person to person through tiny respiratory droplets. These droplets can also land on neighboring surfaces when a person coughs or sneezes. There is also evidence that the COVID-19 virus can survive for up to three days on surfaces, particularly plastic or metal [8]. That's why the key to avoiding contracting COVID-19 is to wash your hands with soap, use alcohol-based antibacterial soaps, and stay away from persons who seem to be sick. COVID-19 is more infectious than SARS or MERS-CoV, and more importantly, it can propagate unnoticed. Such seems to be as many patients with COVID-19 are asymptomatic and have very minor symptoms, and they may not have been separating themselves properly and spreading the infection [9].

1.2 Identifications and Symptoms

COVID-19 can induce flu-like symptoms such as fever and a dry cough, as well as tiredness, aches and pains, and nasal congestion. Other symptoms, such as a loss of ability to smell or taste, have emerged as the pandemic spreads over the world [10]. Extreme symptoms can result in severe lung infections, including pneumonia. The elderly and persons with underlying medical concerns, such as heart disease or diabetes, are the most vulnerable. Despite the fact that the majority of deaths continue to occur in older people, it is obvious that many young persons infected with the virus can still acquire serious infections that necessitate hospitalization [9]. Fever, dry cough, weariness, aches and pains, sore throat, diarrhea, conjunctivitis, headache, loss of taste or smell, skin rashes, or discoloration of fingers or toes are all symptoms of coronavirus. Some of the most dangerous symptoms are: trouble breathing, difficulty breathing, chest pain or pressure, and loss of communication or mobility [11]. Various algorithms are evaluated, as well as

their accuracy, to get a sense of their potential efficacy towards clarifying the public's physical status [12]. An increasing amount of information suggests that COVID-19 survivors may have chronic symptoms affecting many organ systems following the acute period of illness (also referred to as long-COVID) [13]. Many earlier conceptual studies on long-term COVID investigated the occurrence of short and intermediate severe diseases following COVID-19 exposure.[14] [15]. According to a meta-analysis [15] of 39 trials having seven or eight months of follow-up, the most commonly observed symptoms were weakness, weariness, focus loss, and dyspnea.

1.3 Diagnostic Methods in COVID-19

Severe acute respiratory disease is RNA virus, it can theoretically be detected using any available RNA detection format. PCR is the primary diagnostic method is DNA amplification that has been previously used. Some of the others molecular technologies have also been applied [16]. Blood tests and PCR is the most common methods used so far as mentioned in Figure 1.2. At around 10 days or more following the onset of symptoms, immunoglobulin testing can play a primarily complementary function in the diagnosis of COVID-19 [17]. On the other hand, researchers in computer science have already been able to determine contagious patients early by employing systems that can analyze and understand medical imaging data, such as X-ray pictures and Computed Tomography (CT) scans. Based on the obtained data [18], ML methods are used to acquire necessary information about disease symptoms. When compared with pharmaceutical kits or CT scan procedures, this uses a convenient and speedy way of coronavirus detection [19]. The suggested solution unifies these applications into a single framework.

In several separate applications, smartphones' sensors have been effi-

(a) Method-1

(b) Method-2

Figure 1.2: Diagnostic Methods(most common)[20]

ciently used. For example, data from the thermometer can be utilized to anticipate fever levels. Data acquired by integrated inertial sensors as well as images and videos captured by cellphone cameras can be utilized to diagnose human fatigue [21]. Coronavirus can be determined rapidly and proficiently to have viable screening and facilitate the burden on the medical care system. An approach, the use of a phone-based online poll to capture people's fundamental travel history and common manifestations is recommended in [22]. To support medical staff globally in tracking patients, prediction models that incorporate different factors to evaluate the possibility and prediction of infection have been developed across the globe[23]. These models incorporate variables such as Computer Tomography (CT) scans, clinical symptoms, laboratory testing, and their integration [24]. Other innovative methods can also be seen in Figure 1.3. Considering the virus' significant rate of infec-

tion, it is critical to obtain precise diagnostic data as quickly as possible, as erroneous rejections have been shown to have a particularly harmful demographic influence [25]. With so many asymptomatic carriers, minimizing the incidence of false-negative diagnoses is crucial for considering isolation measures and view of the prevailing for patients [26]. LDTs, especially laboratory formulated test runs, may have the same or similar features as IVD tests. They must, however, be developed and then used in the same facility [27].

The virus could spread between people through additional, as yet unknown routes. Common causes in cases include dry cough, fatigue, and fever reported as clinical symptoms. According to the World Health Organization, dyspnea (shortness of breath), fever, and tiredness are indications and symptoms of severe cases. People who have other illnesses, such as asthma, diabetes, or heart disease, are more susceptible to the virus and may become seriously ill as a result [25]. Several COVID-19 breakout statistical and machine learning models are being used by researchers across the world to make accurate assessments and enforce needed control measures. Among the typical models for COVID-19 global pandemic prediction, state-of-the-art AI, data science and statistical models have acquired increased attention from authorities and public acceptance. [29].

1.4 COVID-19 Testing in Low Resource Settings

The presence of huge testing is critical for tracking the outbreak's course or decline and determining the lockdown exit strategy. Most laboratories in limited settings may lack the expensive platforms [30] required to undertake high-performance commercial tests. Furthermore, in rural hospitals, labo-

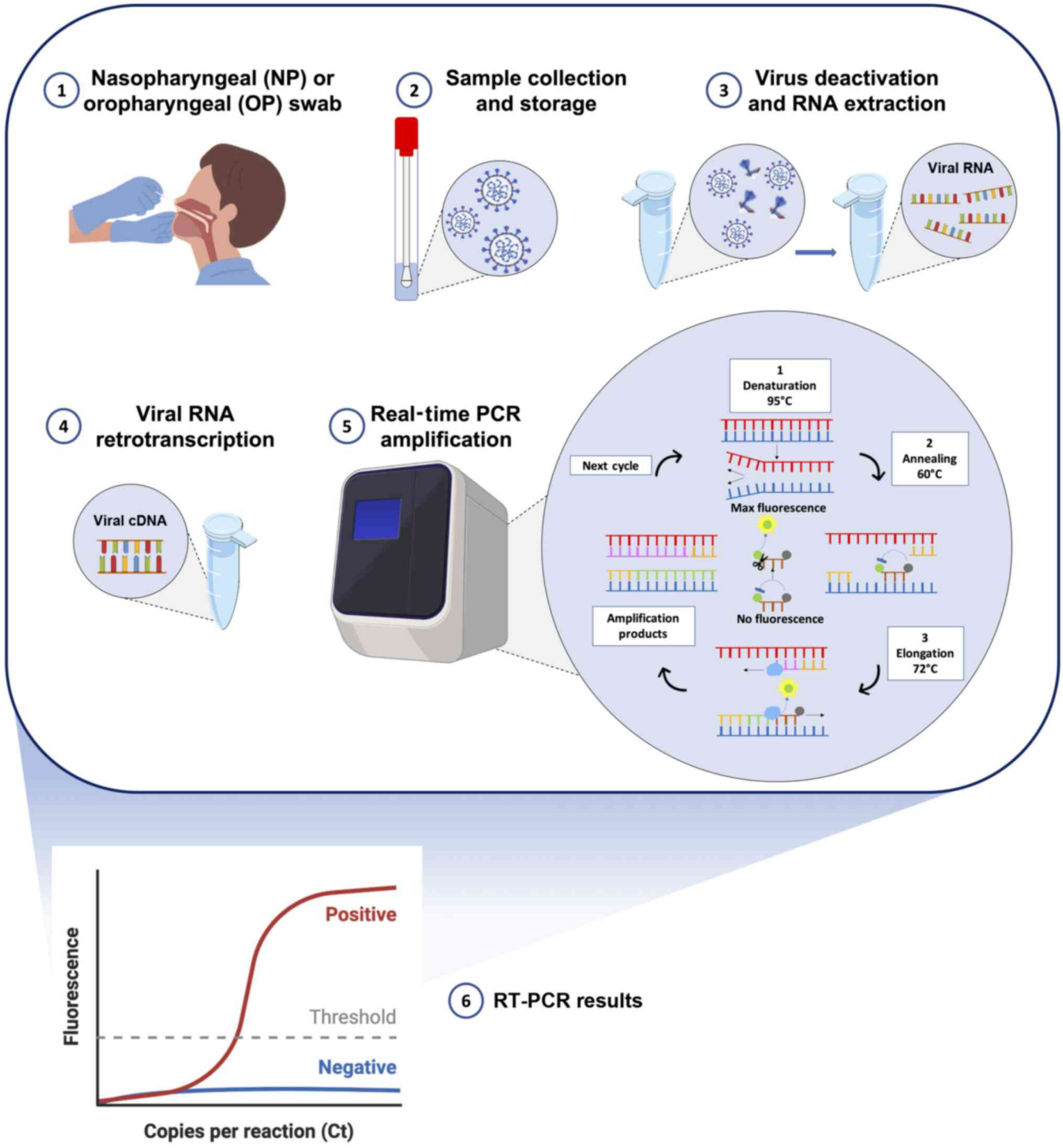


Figure 1.3: Innovative Methods for Diagnosis of Covid-19 [28]

ratories or in politically volatile areas may encounter highly distinct local issues [31]. It should be adhere to the standards for reporting of diagnostic accuracy studies (STARD) standards [32]. The operational properties (sensitivity, specificity, accuracy, and so on) of a COVID-19 diagnostic test also be actively observed and recorded in research. The standards for environmental stability should also be considered contextual. COVID-19 scan tools were created soon before and during the disease's first global outbreak [33]. For upcoming COVID-19 reformations, current technologies can be employed instantly but mostly statistically, allowing for the quick identification of new infected people, quarantine, and deployment of restrictive mechanisms.

1.5 Potential Risks and Challenges

The Coronavirus pandemic is having a significant influence on people's lives in general. Everyone on the planet is likely to be impacted by the disease's devastating effects[34]. Several governments have proclaimed an extraordinary state of emergency and lockdowns. Governments close schools, colleges, universities, bars, marketplaces, supermarkets, and commercial centers among other things. It has fostered a panic, worry, and stress situation in both developed and developing cultures [35] [36]. However, because of its high isolation and lockdown tactics, this sickness causes several other concerns, such as social anxiety, panic situations due to speculation, downturns, as well as intense psychological stress [37].

The number of infections has been expanding globally, with over 1.8 million cases worldwide on April 12, 2020 [38]. The hospital's dermatological activities have to be organized, with a particular concentration on medical emergencies and sick people for primary tumor surgery. Acute facilities, therapeutic consultations, and staffing levels are all being cut [39]. Early

published data show that 25.9% of Severe acute respiratory pneumonia patients required ICU hospitalization, and 20.1% experienced abrupt breathing difficulties [40]. Another challenge is the need for sufficient donor selection with high neutralizing antibody titers and the scarcity of high-quality research [41]. Surveillance is another critical aspect: quick and continuous detection efforts aiming at early detection, isolation, and treatment of persons infected with the virus. Epidemiological viral typing is critical for adequate surveillance and tracing on a regional and worldwide scale [42].

The coronavirus illness (COVID-19) epidemic has posed a significant mental health challenge around the world. In both the general public and healthcare professionals, inconceivable mental afflictions such as anxiety, worry, dread, tension, stress, trauma, mood disorder, and so on have been reported [43] [44]. COVID-19 fear has been linked to suicide attempts in Pakistan's nearby countries such as Bangladesh and India [45]. Economic downturns were defined as contracts within an economic cycle that lower economic productivity, which is often measured by GDP and the rate of unemployment [46]. The COVID-19 catastrophe risks overwhelming undeveloped and emerging countries like Pakistan as a disastrous financial and social crisis in the months and years ahead[47].

The world has been left with a scarcity of established and trustworthy methods to confirm the diagnosis because of the unexpected outbreak of the COVID-19 pandemic [48]. Unfortunately, without rigorous diagnostic testing, determining the real rates at which these events occur is difficult. Whether abnormalities arise what so ever, and how this should change the treatment procedure for future diagnostic standards of care [49] [50]. The performance of current detection kits is likely to have been degraded due to the haste with which these analyses would have to be designed. Kits could

not have been fully tested for medical or scientific accuracy before being made accessible for usage where it was required. In the absence of adequately strict testing criteria, the criteria for an affirmative or negative diagnosis, and also the source and integrity of the original accuracy assessment data, were arbitrary and implausible[51]. Furthermore, the numerous countries that created and implemented the first tests have different medical test development regulations, confounding the capacity to evaluate the reliability and specificity of detection kits [52].

Although machine learning (ML) has long been established as a standard method for modeling and its use in epidemic modeling is still in its infancy, especially in Pakistan due to a lack of clinical data exploration [53]. In this study, we proposed a machine learning model that predicts a positive SARS-CoV-2 and the status of the patient through clinical symptoms. Officials around the world are using several COVID-19 outbreak prediction models to make accurate assessments and enforce needed control measures. Simple epidemiological and statistical models have gained more attention from authorities and are popular with the public among the standard models for COVID-19 global pandemic prediction. Standard models have proven limited accuracy for long term prediction due to a high amount of uncertainty. Random forests, neural networks, bayesian networks, naive bayes, genetic programming, and classification and regression trees are the only ML approaches available (CART)[54] [55]. A lot of challenges have been discussed above in detail. The major challenge was early detection and prediction of the virus by using clinical symptoms specifically for Pakistan, which we have addressed in our proposed methodology. Machine learning methods can be utilized to achieve high accuracy detection of coronavirus using clinical data in Pakistan.

1.6 Problem Statement and Contribution

In developing nations, the most widely used COVID-19 diagnosing test, reverse transcriptase-polymerase chain reaction (RT-PCR), has long been in short supply. This leads to an increase in infection rates and the postponement of crucial preventive actions. The shortage of diagnostic equipment/kits has also made it difficult to effectively identify and manage affected people to prevent future infections. As a result, healthcare staff will continue to require highly sensitive diagnostic techniques to identify suspected instances of COVID-19 more quickly [54]. AI and Machine Learning have been used to screen populations of people for infection risk. China, for example, used AI-powered temperature screening in public places during the COVID-19 pandemic [56].

Pakistan has also faced a tremendous amount of challenges during this epidemic, including a shortage of test kits, early detection, and a slow testing process due to the lack of resources. One of the major problems faced because of the limited resources for remote and under-developing countries like Pakistan to purchase laboratory kits for test purposes was quite an expensive and time taking detection process in far-off and remote areas. The early prediction of the virus and its severity among patients based on clinical symptoms is still a challenge due to the unavailability of patient symptomatic data. However, most systems rely on just constant human supervision [57]. We have further enhanced and speed up this process with some computer-aided digital design in the domain of machine learning.

1.6.1 Justification of Research

COVID-19 can be diagnosed quickly and efficiently with effective screening, easing the burden on healthcare systems. To support medical staff globally in triaging patients, prediction models that include various variables to evaluate the probability of infection have been developed. These models incorporate variables such as computer tomography scans, clinical symptoms, laboratory testing, and their integration [24]. This pandemic continues to present several challenges to medical systems around the world, including increased demand for hospital beds and critical shortages of medical equipment, as well as the infection of many healthcare personnel. As a result, the ability to make quick clinical choices and make efficient use of healthcare resources is critical [58].

The shortage of diagnostic equipment/kits has also made it difficult to effectively identify and manage affected people to prevent future infections. As a result, healthcare staff will continue to require highly sensitive diagnostic techniques to identify suspected instances of COVID-19 more quickly. Machine Learning based prediction and evaluation of COVID-19 detection and prediction of patient severity status will help the research community and medical experts.

1.6.2 Research Questions

- How can machine learning detect and predict COVID-19?

  This above question is divided into sub-questions:

- Which machine learning algorithms to use?
- How will COVID-19 detection accuracy be increased in real time?
- How much dataset to be used for good training results?

- How to gather dataset to apply the proposed approach?

1.6.3 Research Objectives

The foremost objectives of this research are to determine:

- Early prediction of COVID-19 and status of patients on symptomatic data.
- Availability of symptomatic dataset for Pakistan.
- Data analytics on most important and sensitive features.

1.6.4 Research Contribution

As we observed in [59] the Ministry of Israel already has some analysis and predictions on COVID-19 symptoms data. They have reported eight binary features for COVID-19 detection. One of the major drawbacks we have seen in their study was the incomplete set of symptoms. We have gathered a complete set of symptoms from hospitals and extended our work to predict patients' severity as well. The early prediction of the virus based on clinical symptoms was still a big challenge for COVID-19 due to the unavailability of patients' symptoms data, especially in Pakistan. Its detection using clinical symptoms data has extreme importance for the lives of doctors and patients. Early warnings will help to reduce the disease burden and to control it. For Pakistan, it will be a great contribution to health care society for better analysis of this pandemic. The scope and key features of our proposed method:

- As we observed in previous studies they have reported incomplete set of clinical symptoms. We have covered this gap by collecting a complete

set of symptoms from hospitals and extend our work to predict the patient's severity as well.

- It has considered the prediction of Covid 19 using patient symptom data in Pakistan. It would therefore, can be helpful to carry out a more immediate symptomatic analysis to identify suspicious and extremely serious cases. Early warnings will help to reduce the burden of the disease and control it, especially in remote and rural areas.

1.6.5 Organization of Thesis

Chapter-1: Introduction

- This chapter deals with introduction and motivation, diagnostic methods, COVID-19 testing in low resource settings, potential risks and challenges, problem statement and contribution, justification of research, objectives, and research contribution.

Chapter-2: Literature Review

- This chapter includes related work, predictive analysis, descriptive analysis, and discussion.

Chapter-3: Data Acquisition

- This chapter deals with data collection procedure including IRB approval, data collection techniques, data preprocessing, features and study settings.

Chapter-4: Methods

- This chapter includes basic flow: predictive analysis and supervised machine learning algorithms, logistic regression, support vector machine , k-nearest neighbor, random forest and descriptive analysis.

Chapter-5: Results and Discussion

- Parameters for supervised algorithms, classification reports for supervised algorithms, quantitative results for supervised machine learning algorithms, models performance and descriptive statistics.

Chapter-6: Conclusion and Future Work

- This chapter summarize the following: novelty and contribution, limitation and challenges and future work.

1.7 Summary

The different aspects related to COVID-19 and why COVID-19 detection, prediction, and analysis are important are discussed in this chapter. Why the prediction model is important has been discussed also in the situation of Pakistan in terms of COVID-19 analysis. Justification for this research topic is discussed. Moreover, the research objectives are also stated while at the end. Next chapter deals with the literature review.

Chapter 2

Literature Review

Countries all across the world are working to stop the spread of COVID-19 while also developing methods for early detection and treatment. Machine Learning based prediction models were carried out for the analysis and prediction of COVID-19. Healthcare professionals, researchers, and scientists have examined both current and new technologies to forecast early warnings and cautions, track and predict; diagnose and prognosis, and treat and maintain social control. Since 2020, a significant amount of work has been done on COVID-19 prediction and analysis, and to develop an insight into machine learning in literature, we have mostly covered two types of analysis:

1. Predictive Analysis.
2. Descriptive Analysis.

2.1 Predictive Analysis

Machine learning is a novel approach with numerous applications in prediction. This learning technique generates external interference from unlabeled input datasets, which may then be analyzed [60]. The branch of AI is ML, in which specialized algorithms are being created to enable machines to learn

autonomously. The algorithms improved over time, which means that the more data sets they evaluated, the better they got. As a result, in the health care industry ML is a fantastic tool for predictive analytics [61]. Based on the Big data available. Machine learning algorithms gather expertise faster and provide better estimates.

2.1.1 Machine Learning-based Prediction

Machine learning technologies allow for the analysis of massive databases of viral genomes, which helped to improve our understanding of COVID-19 [62]. Machine learning concepts can be used to predict whether or not commercially licensed treatments can be utilized to treat disease [63]. This method makes it easier to repurposed existing medicines on time. Machine learning is a novel approach with numerous applications in prediction [64].

Khanday et al.[65] The study evaluated the 6995 patients. An ICU mortality prediction score, a white platelet count, and the time from the beginning of side effects to confirmation were among the particular qualities the specific characteristics. When it came to predicting patient risk for crucial COVID-19, random forest, classification, and regression decision tree outperformed. Machine-learning models surpassed all other measures by predicting severe COVID-19 with 88.0% sensitivity, 92.7% specificity, and 92.0% accuracy. Evaluation measures were used accuracy, recall, selectivity, and precision. Decision Tree (CRT) with variance threshold outperforms i.e. CRT gives 92% accuracy. A learning model is only as effective as the data that it is fed. As a result, more dependable criteria must be investigated to enhance the models' predictive capacity. There are some drawbacks to this study. For instance, its posterior single-center technique limits its external validity. The minimal number of patients (only 212) with severe illness in this study

calls the statistical power into question.

Assaf et al.[58] Another strategy was applied, in which data mining models were constructed to predict patient recovery using an epidemiological data set from South Korea. The model projected how long it would take COVID-19 patients to recover from the virus. The dataset includes (3,254) patients with (8) attribute including patient ID, gender, age, nation, region, city, and infection. The model estimated the initial and final number of days required for COVID-19 patients to recover from infection, and also the average age of patients with a high risk of not recovering from the COVID-19 pandemic, those who have been able to recover, and those that may recover fast. The current study's findings indicate that the technique developed with the decision tree data mining algorithm is much more effective in predicting the chance of recovering people with a virus, with an overall accuracy of 99.85%. Using data mining or machine learning algorithms, the procedures determine the model's quality and efficacy. Specificity, sensitivity, and accuracy are the three basic approaches for evaluating the data mining model's performance however, in this study, only correctness is used to analyze the produced models. The generated models would be extremely useful in the fight against COVID-19 in healthcare. Only 1500 patients were evaluated so, the limited availability of clinical data was the main issue.

Muhammad et al.[66] categorized textual clinical reports into four categories by using a machine learning approach. Clinical notes are made up of text, whereas attribute findings are made up of the label of the related text. The clinical reports are labeled with the classes (COVID, ARDS, SARS) to which they belong. The following terms were used: identity, gender, age, observation, mortality rate, moved ICU, needed supplemental, fever, directory, filename, etc. In feature extraction, term frequency/inverse (TF/IDF), word

embedding (bag-of-words), and summary length were utilized. These characteristics were fed into standard and ensemble machine learning classifiers. The length of approximately 212 reports was computed. Logistic regression and multinomial naive bayes performed better than other ML algorithms, with 96.2% testing accuracy. Future research may be increasing the amount of data available to models that can improve their efficiency. Furthermore, the sickness may be categorized based on gender, so it can learn whether males or females are more impacted. For improved outcomes, more feature engineering is required, and a deep learning approach may be applied in the future. Symptoms based approach was missing for further analysis.

Villavicencio et al.[67] another method employing Kaggle's COVID-19 symptoms and presence dataset and many supervised machine learning algorithms with comparative analysis using the WEKA tool. The scientists developed a model to explore and predict the presence of viruses. The researchers used a dataset called COVID-19 symptoms and presence from Kaggle to collect data. This dataset contains 20 variables that are potential risk factors for contracting the virus, as well as 1 class attribute that determines the existence of COVID-19. This dataset has global spatial coverage and a date range of 17 April 2020 to 29 August 2020. According to the results, "Support Vector Machine" with "Pearson VII universal kernel" outperforms other algorithms with 98.81% accuracy. This work has the potential to be employed as a decision support system for medical practitioners, with the created model assisting in recognizing the prevalence of COVID-19 in an individual depending on the claimed indications.

Yadav et al.[68] designed a model to outbreak prediction by using the machine learning and soft computing models. Data on total cases lasting more than 30 days were collected from Italians, Germans, Persia, the United

States, and Chinese. The next stage was to identify the optimal model for estimating time-series data. To create the desired model, logistic, linear, logarithmic, quadratic, cubic, compound, power, and exponential equations have been used. One of the objectives of this research was to model time-series data using the logistic microbial growth model. Multi-Layer perceptron and fuzzy logic were two models among a vast number of machine learning models investigated that produced promising results. Future research should compare studies on multiple ML models for single nations to build improved long-term prediction models. Due to fundamental differences in outbreaks between countries, the evolution of generic models with predictive accuracy would be unfeasible. As observed and reported in several papers, an isolated outbreak is unlikely to be duplicated elsewhere.

The major goal of Asteris et al.[69] was to complete the five separate tasks, such as I) predicting the spread of the coronavirus throughout different regions. II) International comparison of growth rates and mitigating actions. III) Making predictions on how the epidemic will end. IV) Investigating the virus's pace of spread. V) Establishing a relationship between the coronavirus and weather conditions. It would be useful to evaluate and assess how quickly or slowly the virus propagated among locations and affected patients to reduce, how effectively the countermeasures were working, how many cases were prevented by these countermeasures, and a rough estimate of the number of patients who will recover from the virus with old treatment, and an understanding of how long it took for this pandemic to finish. It has been able to estimate how rapidly or slowly the virus was moving throughout regions based on a detailed understanding of the relationship between both the transmission and meteorological conditions and affected patients to decrease the spread. Instead of a simple regression line, they used supported

vectors in their work to improve classification accuracy. It would be useful to evaluate and assess how quickly or slowly the virus propagated among locations and affected patients. Future studies should be focused on each country, the differences between climate change and other health factors are also very important. This should be a generalized comparison.

Cheng, Agbehadji et al.[70][71] presented a review paper focusing on big data, artificial intelligence, and nature-inspired computer models that can be utilized in the current pandemic. One method for detecting COVID-19 infection in patients is to check "chest X-ray images" however, due to the enormous number of patients in hospitals, examining a large number of X-ray images is time-consuming and difficult. On the other hand, big data applications have been used to track down contacts. Big data is normally developed from the necessity to evaluate massive amounts of complex data created each minute from a wide variety of sources. In the past different analytics tools, for the most part, are not designed to analyze such complex data for the purpose to extract insights. Further big data is distinguished by its speed, a large volume of data, and a wide range of data sources. AI models can quickly detect the progress of the infection, allowing for the appropriate intervention.

Another approach used by Wang et al.[72] in which they used ensemble methods by integrating the DNN with two other ML algorithms for prediction. They looked into 39 different disorders and diseases in which 86 attributes were collected. Gradient boosting and light gradient boosting were further ensembled with DNN using tensor-flow. Total 5145 samples were collected and then an investigation was made based on clinical importance related features. 81% f1 score and accuracy 91% achieved by ensemble methods. They used confusion matrix and shapely analysis for features im-

portance. Their new model reported high accuracy that can be utilized for the prediction.

In [12] tweets mentioning self-report of COVID-19 symptoms were investigated in this research. The World Health Organization contributed a collection of disease symptom keywords that were used to label the dataset. Followed closely by BERT and Convolution Neural Network the studies revealed that the random forest algorithm produced the greatest results. This research could potentially help in the development of algorithms for recognizing sickness symptoms in social media content.

Zoabi et al.[59] to forecast the COVID-19, a machine learning technique based on symptoms was developed. They have designed a machine learning system that was trained using data from 51,831 participants who had been tested (of whom 4769 were confirmed to have COVID-19). When testing resources are restricted, they might use their methods to prioritize COVID-19 testing. Overall, they developed a model that predicts COVID-19 instances using simple characteristics accessible via basic queries and depending on countrywide data made public by the "Israeli Ministry of Health". When screening resources are insufficient, their technique beyond other considerations, can be utilized for COVID-19 testing. "A gradient-boosting machine model" design with "decision-tree base learners" was utilized to create predictions. SHAP values were obtained to determine the primary factors driving model prediction. SHAP values estimate the contribution of each feature to overall model predictions by averaging across samples.

2.1.2 Deep Learning-based Prediction

Deep learning is another branch of AI that deals with artificial neural networks. Artificial neural networks outperform the human intellect and can be

used to produce accurate predictions with modern multi-layered processing capabilities. They are designed to be biological neural networks similar to those seen in the human brain. Through the study of chest X-ray images, deep convolutional neural networks (CNN), and a frequently utilized deep learning framework, we're able to distinguish between COVID-19 and other causes of pneumonia [73]. AI can assist limit workplace exposure to the virus in addition to improving diagnostic accuracy and efficiency. AI provides a wide range of characteristics and applications that can be used to help us respond to COVID-19. To explore, diagnose, and cure COVID-19, researchers used both machine learning and deep learning models [62]. CT scans and X-rays, to handle rising problems by improving detection accuracy and dependability AI has enabled traditional imaging technologies [74]. The AI-powered visual analysis tool can follow a patient's disease progression [75].

In this study [76] they demonstrated how deep learning models with transfer learning can be better utilized for COVID predictions. They have used the three most commonly used medical image data like X-ray, Ultrasound, and CT scan. The main focus of this study was to conduct a comparative analysis with available deep learning algorithms. CNN is one of the best models which is a high accuracy rate for medical image diagnosis. They also used VGG19 as an optimal approach to check how these models can be utilized for high scarce and challenging datasets. They also highlighted the main challenge like data size and quality and how these adversely impact the model training. The new approach suggested a solution regarding data quality and availability. The main focus was to estimate the specific features and remove noise. The results showed that ultrasound data has superior detection accuracy as compared to X-ray and CT scans.

Another approach [77] in which a review-based study focused on three

main countries China, Korea, and Canada. The medical images dataset considered for this review-based study like X-ray and CT scans. They have used three use cases from these countries and applied deep learning applications for medical image processing. Implementations for COVID-19 medical image processing, finally they also highlighted many issues to drive future research in an outbreak and crisis prevention, resulting in smart healthy cities. They presented an overview of deep learning and its applications in healthcare discovered in the last decade.

In [78] stated that after being exposed to the COVID, VIRUS-infected cases revealed that these types of victims were mostly unwell with internal organ infection. For the rapid transmission of COVID patients, this research is aimed at developing low-cost deep learning models, trained with upper body X-ray images. They have executed almost twenty-five distinct forms of augmentations on the first images to increase the dataset size. For coaching and analyzing categorization models, they tend to use the transfer learning strategy. Furthermore, their method was far more cost-effective than previously disclosed methods.

In [79] authors proposed a system that automatically detects ground-glass opacity (GGO) and pulmonary infiltrates (PIs) in COVID-19 patients. During the patient's follow-up analysis and management, the goal is to assess disease progression. They defined each superpixel cluster using position, grey luminance, and temporal features, which were then classified using a tree random forest (TRF). This can aid in the diagnosis of (RT-PCR). In [80] Corona Virus is detected utilizing chest X-ray radiographs and three major versions of deep convolutional neural network. Three state-of-the-art models were used namely Inception- ResNetV2, InceptionV3, and ResNet50. The ResNet50 model provides the best performance and classification accuracy

among existing systems.

In this paper[81] using multi-task deep learning (DL) approaches, they offered a rapid and efficient method for identifying COVID-19 patients. The proposed technique was evaluated using both X-ray and CT scan pictures. The proposed inception residual network with transfer learning to detect the COVID-19. The result showed 84% accuracy in CT scan images and 94% accuracy in X-ray images. Another approach [82] in which they evaluated 18,479 CXR images. This was the largest dataset for X-rays and lung masks used for COVID-19 detection. A comparison between the proposed U-Net model and the standard model for lung segmentation. For lung segmentation, the innovative U-Net model demonstrated accuracy, (IoU) and Dice coefficients of 98.63%, 94.3%, and 96.94%, respectively.

In this paper [83] to calculate the distance of objects in video streams in real-time, this article described the innovative techniques to detect the COVID-19. The system was made up of research and development to execute a camera stream, including a Livestream, a range and object recognition model, a data packet stream, continuous data collecting, and dissemination of information to support the decision. YOLOV3 is trained using a distance and object identification model and written in Python, while OpenCV captures frames from the video stream. In [84] technique suggested that when compared to raw (unprocessed) X-ray lung images algorithms for preparing normal, COVID-19, and pneumonia X-ray lung images, can improve classification accuracy. In the segmentation of lungs from X-ray images, image quality can be increased. The authors developed an effective preprocessing and classification strategy for detecting respiratory diseases. Intersect over unions scores were used for evaluation purposes to compare the algorithms. For all three classes (normal/COVID-19/pneumonia) preprocessed

X-ray images outperform untreated raw images in classification. To classify the respiratory disorders VGGNet, AlexNet, Resnet, and the suggested deep neural network were used.

The goal of this study [85] was to develop efficient transfer learning strategies by assessing the layer depths and degree of fine-tuning on transfer learning. The data used in the study were gathered from publicly accessible archives and categorized into three categories: COVID-19, pneumonia, and a healthy lifestyle. CNN's VGG-16 and VGG-19 were utilized as backbone networks to study the impact of physical properties in the same CNN design. The experimental findings revealed that the greatest AUC value for the COVID-19 fine-tuned model was 0.95. A suitable degree of fine-tuning can aid in the development of an effective trained model were utilize for pre-trained CNN architecture.

In this approach [86] AI-based technique used for detecting COVID-19 from chest X-ray images. The characteristics collected from X-ray images using the HOG and CNN were fused to construct the classification model through CNN training. To identify the major fracture zone in the raw X-ray pictures, a watershed segmentation approach was applied. The model's performance was evaluated using generalized data throughout the testing step. A 5-fold method could successfully mitigate the overfitting problem, according to cross-validation analysis. K-fold cross-validation revealed that the suggested feature fusion strategy (98.36%) outperformed the individual feature extraction methods in terms of accuracy.

2.2 Descriptive Analysis

A. Aktar et al.[87] in this study used a combination of statistical comparison and correlation methodologies, as well as machine learning algorithms, to

investigate clinical data sets of COVID-19 patients with known outcomes. Their findings showed that many clinical indicators quantifiable in blood samples may distinguish between healthy people and COVID-19-positive patients and demonstrated the utility of these measures. These measurements were found to be useful in predicting the intensity of COVID-19 symptoms in both healthy adults and infected patients. Data from routine hospital laboratory investigations of patient blood might be used in this technique to identify COVID-19 individuals who have been at significant mortality risk, allowing hospital facilities to be optimized for COVID-19 therapy. It has been established several analytical approaches for illness severity prediction that had accuracy and precision scores of more than 90%.

Falesia et al.[88] another study based on using stable psychological traits and machine learning models, could predict experienced stress about the Covid-19 epidemic. They have used descriptive analysis to examine the impact of the pandemic on 2053 Italian people. A set of 18 emotional variables, extended multiple regression analysis, and sophisticated machine learning algorithms were applied. In comparison to Italian normative standards, analysis has discovered increased levels of felt stress in the research group. Women those with lower salaries, and those living with others reported higher degrees of pain. A social media invitation was used to engage participants online because of the lockdown situation, which prevented them from gathering data in the laboratory, they adopted this online recruitment technique. The JASP software [89] was used to analyze the data. This is especially important for those who are more susceptible to "stress", and those individuals must have prompt access to high-quality psychological therapy to minimize the development of chronic repercussions. In the case of the PSS-10 score, a single sample t-test (t, two-sided) was used to determine if the true mean of the

sample differed significantly from the given population.

L. Flesi et al.[88] used a similar approach as in the previous one to use a list of 20 hashtags and scientists analyzed 4 million twitter tweets related to the COVID-19 outbreak for example, coronavirus, COVID-19, and quarantine. They employed (LDA), a machine learning approach, to identify prominent unigrams and bigrams, significant topics and themes and feelings in the obtained data tweets. Furthermore, the research identified four new COVID-19-related discussion topics: (1) the necessity of vaccination to control the spread, (2) isolation and shelter-in-place measures, (3) anti-lockdown objections, and (4) the "COVID-19" epidemic in the US (United States). Future research could gain insight into public trust and confidence in existing measures and policies, both of which are critical. Future studies should look into misinformation and how it spreads through social media. Finally, when people tweet about proven instances and deaths, trust is no longer dominant. Future studies should look at how trust evolves.

B. Milliner et al.[90] authors proposed using a classification-oriented machine learning method and performed a traditional data science process to perform noise cleaning and data processing to perform descriptive statistical analysis in such a way that the most important variables or factors are identified through unsupervised learning. They have the potential to induce major kidney problems. In addition, their strategy has used quality metrics to assess. Finally, this effort paves the way for future research, well to undertake comprehensive examinations of every variable associated with Covid-19 revealing important information that can help increase overall condition. SHAP value was used to analyze the most important features with 90% AUC with Sensitivity. According to the findings of a study conducted by the OMS in collaboration with China, the majority of the 55,924 confirmed COVID-

19 cases examined by the laboratory exhibit clinical characteristics such as fatigue, tossing, and fatigue. Finally, this work opens the door to future research to conduct centralized investigations on each variable associated with Covid-19 to identify useful information that can support an improvement in the existing condition.

J. Xue et al.[91] the main focus of this study was "How are you feeling in the heat of the COVID19 pandemic?". A survey employing psychological and linguistic self-report measures, as well as machine learning, to analyze mental health, individual experience, temperament, and behavior among university students during the COVID19 epidemic. This internet survey investigated the mental well-being, emotional states, and attitude of high school students in Egypt and Germany immediately upon the initial shutdown in May 2020. Data analysis included descriptive cognitive analysis, correlation evaluation of psyche and emotions, with "220" students in which "107" were female, "112" Male and 1 other from "Egypt" and Germany. Consistent public persona traits, consciousness, and state-like cognitive factors were all assessed about (a) mental illnesses (chronic anxiety, fear), and (b) threats (hardly perceptible when it comes to defining, understanding, and showing feelings) were all assessed. For Experimental prospective Machine learning (ML) was utilized to integrate the several normative assessed cognitive factors to see if a machine - based learning algorithms can anticipate and categorize unique personal qualities. From the initial pandemic lockdown in the countries, the majority of students who participated in the study reported observed alterations in all cognitive variables, especially nervousness, mood disorders, and sentiment interpretation.

Another approach has been used in[92] in which Machine learning (ML) was used purely for exploratory purposes, with the ML algorithms chosen to

combine the various psychological variables that were descriptively analyzed to investigate the University Students' Learning during this pandemic. This new mechanism was put to the test by evaluating data from a survey of university students conducted throughout the second wave. The primary goals of this research were as follows: (1) examine student learning throughout the online classes; (2) observed challenges and future concerns for the coming months, and (3) suggest countermeasures and ways to improve this situation that could help for education. Out of 134 pupils, 95 reports that the COVID-19 pandemic will have an influence on their families' level of living, and 77 reports that a decline in the family economic standard just because of this pandemic will affect their education. Furthermore, 34 students answer that the epidemic may cause them to drop out of university. Future work in this regard intends to: (1) increase the participation in this online survey, (2) comparative analysis of other countries as well by different aspects, and (3) complete a comprehensive look and evaluation of distance learning.

S. Sonthali et al.[93] A Machine Learning-based method used for "COVID-19" prediction in "Indonesian health workers" was proposed. The self-reported survey questions included behavioral tendencies (protective, social, and travel), COVID-19 vaccination status, working conditions, and symptoms. The number of participants from Jakarta was 3477, and 502 from Semarang. Descriptive analysis showed approximately 80% of positive cases had symptoms. Cough (48.52%), headache (44.43%), runny nose (39.34%), chills (38.36%), and fever (37.54%) were the top five symptoms reported by those with COVID-19 positive test findings. Ninety percent of those polled said they had no co-morbidities. Lung disease (3.11%), hypertension (2.46%), and pregnancy (2.30%) were the most often reported co-morbidities among COVID-19 positive people. The significance of behavior patterns is demon-

strated by their prestigious position in the SHAP summary graphs for both models, where behavioral tendencies were among the most relevant features for the models.

Albahri et al.[94] focused on a complete systematic review based on COVID-19 screening and diagnostics using automated AI solutions, mining methods and machine learning technologies between 2010 and 2020. They evaluated three search query sequences utilising five databases: IEEE Xplore, Web of Science, PubMed, Science-Direct, and Scopus. According to the test findings, the two most relevant criteria for the prediction model were age and symptoms. This study gathered the descriptions of the primary features, the performance evaluation methodology used, and the correct level of each article in the literature. To address the severe public health concern for CoV, brief motivation, difficulties, limitations, and recommendations were taken from the examined studies. According to the results, the KNN classifier performed well for binary classification, for multi-label problems, the decision tree, and naive bayes were the best classifier. The decision tree classifier outperforms the other models in terms of prediction proficiency. According to the experimental results, age and symptoms were the two most important criteria for the prediction model. As a result, this review explained why research on the CoV problem is being done to save lives.

This paper[95] carried out a descriptive investigation of COVID-19 with a focus on India. Age, gender, travel history, communication medium, and present status are all considered. This paper's significant research findings are as follows: (1) Data analysis is carried out from the standpoint of India. (2) For analysis and visualization, many statistical programs including SPSS, r-studio, and MS-Excel are employed. (3) According to the findings of this study, age is not an important determinant in a person's sensitivity to this

condition. Maharashtra possesses a high rate of intensive care.

In this Paper [96] statistical analysis of regions that have been significantly affected by COVID-19 in comparison to communities that have been relatively influenced by COVID-19, including a focus on the social factors within them. Regions with intermediate COVID-19 percentages and their contrast with exceptionally high COVID-19 rates. From March 1st to April 17th, 2020 they choose six regional areas in Queens and New York which shows the percentage of affirmative COVID-19 instances by zip code. The study's findings revealed that COVID-19 instances were 30% higher in areas with exceptionally high infections than in those having moderate cases. Reduced educational attainment, limited access to services, and an increase in chronic illness were the major issues with extreme high cases.

Another study found [97] [98] in which a descriptive analysis has majorly focused on 15 European countries and the United States based on age groups and gender. In these nations based on age and gender, they compared to the matching previous all-cause death per year. People over the age of 40 have faced all-cause mortality during this analysis. Males face a higher risk than females but in the second half of the year, male relative risks for COVID-19 death were lower.

In this paper [99] The World Health Organization provided COVID-19 statistics for the United States and Germany between 20/01/2020 and 18/09/2020. For 35 weeks, the dataset is composed of weekly cases reported and weekly aggregate reported cases. Moreover, machine learning was suggested to derive the infection curve and predict the outbreak proclivity. Machine learning approaches such as linear regression, support vector machines (SVM), random forest, and multi-layer perceptron were applied. It was discovered that SVM had the best trend. The global pandemic, according to

predictions, will climax around the end of January 2021.

In this study [100] based on 14 clinical variables, this work creates six prediction models for COVID-19 diagnosis using six distinct classifiers (BayesNet, Logistic, IBk, CR, PART, and J48). In China's Zhejiang province study examined 114 cases. The results revealed that the CR meta-classifier, with an accuracy of 84.21%, has been the most reliable classifier for predicting positive and negative COVID-19 cases. It could also benefit developing countries particularly when RT-PCR kits are insufficient for testing the infection. The findings could aid in early detection.

An online questionnaire was created as a data collection method in this paper[101]. This data is being used for several statistical model-based prediction models. Based on their indications and symptoms, these models were used to predict prospective COVID-19 patients. There are thirteen attributes in the obtained dataset having symptoms and signs. Because of its capacity to record extremely complex information in the hidden layers and the use of exponential activation functions, the MLP has the highest results in comparison to the other approaches. Most typically, machine learning models trade generalization ability in exchange for predictive power.

F. Fernandes et al.[102] developed multipurpose algorithms to predict the probability of acquiring critical illnesses in COVID-19 patients in Brazil. It has been provided descriptive statistics regarding the patients' demographic features. It has been provided descriptive statistics regarding the patients' demographic features. The study's sample (1040 COVID-19 patients) was largely made up of men (53.3%), with an average age of 51.7 years, and the majority of patients (63.8%) were white. The complete descriptive statistics for all variables are shown. Finally, they have used the 95% confidence interval of the AUROC to compare the performance of the two methodologies

(individual versus aggregated models) and (AUROC). The AUROC value was used to choose the best model. The model estimated the shapley values of each variable to better understand their separate contributions to the predictive models. The model estimated the shapley values of each variable to better understand their separate contributions to the predictive models.

2.3 Discussion

Since descriptive and predictive analysis is critical for generating insights into data and the most crucial elements of any model. So far, the gap has been caused by a restricted dataset and a country-by-country study based on symptoms. Some countries reported self collected data which is so far not validated by any medical expert. Pakistan, on the other hand, is also deficient in the analysis and prediction of COVID-19 based on patient symptom data. A similar approach[59] is adopted by the Ministry of Israel, however, they have not disclosed some significant symptoms, therefore, a complete collection of symptoms was not provided for a good grasp of the most critical criteria to forecast the COVID-19. Every country has a different development rate and severity in terms of features, thus it is necessary to notice country-specific symptom-based predictions and analysis.

Table 2.1: Literature Summary for COVID-19 ML based Prediction

Study	**Dataset**	**Approach**
Khanday et al.[65]	ICU COVID-19 patients	Machine learning techniques
Assaf et al.[58]	South Korea dataset	Data mining models
Muhammad et al.[66]	Clinical reports	Ensemble ML classifiers
Villavicencio et al.[67]	Kaggle's COVID-19 symptoms	Supervised ML algorithms
Yadav et al.[68]	Total cases of five countries	Optimal model
Asteris et al.[69]	Recovered patients dataset	ML and Data science models
Zoabi et al.[59]	Symptoms data	Predictive analysis
Killeen et al.[103]	US county level data	Epidemiological analysis
Lachmann et al.[104]	Reported case	ML for epidemiological analysis
Malki et al.[1]	Weather data and COVID-19	Predicting mortality rate
kokudo2020call et al.[2]	Self collected survey	Tricks to tackle pandemic
Pham et al.[23]	Survay data	(AI) models
Cheng et al.[70]	In-hospital patients data	ML based detection

2.4 Summary

Related work is discussed in the chapter; it tends to be seen that by utilizing AI and ML quite significant amount of work is done as some of the studies mentioned in literature summary in Table 2.1. Since the gaps associated with symptom-based studies have been specifically identified in Pakistan, so the proposed methodology is primarily based on a comprehensive collection of clinical symptoms that may assist us in predicting the COVID-19 and the status of deceased and recovered patients.

The next chapter deals with the data acquisition and preparation.

Chapter 3

Dataset Acquisition and Methodology

This chapter deals with the procedure followed for dataset collection and methods. This thesis work used two datasets, one from Holy Family Hospital Rawalpindi and the second from AJK Health Department.

3.1 Data Acquistion

The entire process of collecting and preparing data is divided into the following main steps:

3.1.1 IRB Approval

The first step in doing any clinical research or collecting clinical data is to obtain approval from an Institutional Review Board. As per government guidelines and institutional arrangements, IRB has the position to endorse, object to, screen, and constrain changes in all examination tasks that fall under its influence. This study was reviewed and approved by RMU (Rawalpindi Medical University) Ethical Review Board under the study protocol. After the

approval of the ethical board, the data collection procedures were applied.

3.1.2 Data Collection Techniques

This section describes the complete procedures and data acquisition techniques used for this analysis.

1. Primary Data
2. Secondary Data

1. Primary Data

The term primary data refers to information gathered by the researcher directly. The data set previously used in [59] have some limitations in term of incomplete set of symptoms. We gathered data from the COVID-19 unit at Holy Family Hospital on our own with a complete set of symptoms. The patients' files were gathered and analyzed to extract the relevant features. The features extracted are all possible symptoms of COVID-19 through clinical files. Sample size: 3500 clinical files were analyzed from RMU.

2. Secondary Data

Secondary data is information that was previously obtained by the department or by someone else. With the Director General's approval, we obtained an electronic CSV file containing simply the symptoms of 2500 patients from the AJK Health Department. The data from AJK Health Department is exempted from IRB Approval.

3.1.3 Data Preprocessing

This section deals with data preprocessing. The first step is to manually extract characteristics from clinical patient records and save the entire symptom

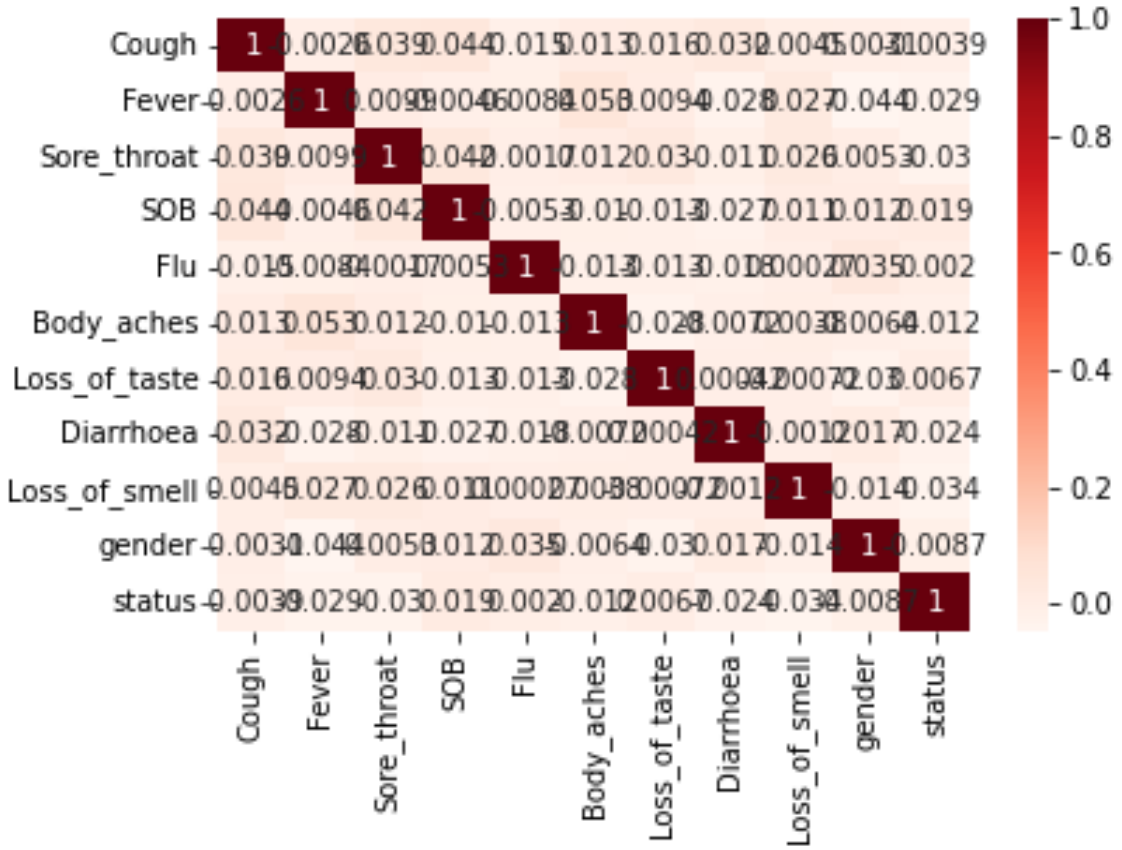


Figure 3.1: Features Correlation

in a CSV file. The detail of entire features we have utilized is mentioned in Table 3.1, then preprocess the data by using Python tools to keep the data within the same range. The Python libraries, Pandas, and NumPy were used to convert the data in categorical order and create data frames for prediction. Binary values and additional weights such as normal, mild, and severe are replaced with categorical values. These weights aid in predicting the state of deceased and recovering patients. Correlation matrices are used to summarize the data, as inputs for advanced analysis, and as diagnostics for

advanced analysis. In our case, the observable pattern is that all variables are highly correlated with each other. The correlation Matrix can be seen in Figure 3.1.

Table 3.1: Features used in Dataset

Features	State
Bio Data	
Gender	Male
	Female
Symtoms Data	
Fever	Yes
	No
Cough	Normal
	Mild
	Severe
Flu	Yes
	No
Fatigue	Normal
	Mild
	Severe
Sore-Throat	Yes
	No
SOB	Normal
	Mild
	Severe
diarrhoea	Yes
	No
Lost of Taste and Smell	Yes
	No
Headache	Yes
	No

3.1.4 Features and Study Settings

In this section features and study settings are discussed in detail. Clinical data were collected between August and November 2021 to analyze and evaluate all previous clinical records and extract relevant features. The data was

	S.no	Fever	Cough	Sore_throat	SOB	Flu	Body_aches	Loss_of_taste	Diarrhoea	Loss_of_smell	gender	status
0	1	1	1	1	0	1	1	0	1	0	1	0
1	2	1	1	0	2	1	0	0	1	0	0	0
2	3	0	0	1	0	1	2	1	0	1	0	0
3	4	1	1	2	1	1	2	0	0	1	1	1
4	5	1	0	1	1	1	1	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
2995	2996	1	0	2	2	1	2	1	0	0	0	0
2996	2997	1	0	1	1	1	2	1	1	0	1	0
2997	2998	1	0	2	0	0	1	0	1	0	1	0
2998	2999	1	1	0	0	0	0	1	1	1	1	1
2999	3000	1	0	0	0	0	0	0	1	1	1	0

Figure 3.2: CSV Data File for Status Prediction

acquired from hard files, which were subsequently divided into first, second, and third waves based on test date and outcome. The collection provides daily initial records for all individuals who have been tested for COVID-19. Aside from the test date and outcome, other information is provided, such as clinical symptoms. We established a model based on this data that predicts COVID-19 test outcomes and status of admitted patients using ten features: gender, age, and eight initial clinical symptoms. The training-validation set included data from 6000 tested people, (3000 of which have been confirmed to have COVID-19). The training-validation set is subdivided into training and validation sets at a 4:1 ratio. The overall dataset for status prediction is mentioned in Figure 3.2. In this chapter, the entire journey of data collection and preparation are discussed, as well as the process of clinical feature extraction is explained.

The next section deals with the proposed methodology and complete predictive and descriptive analysis.

Figure 3.3: Workflow for Supervised Machine Learning

3.2 Methodology

This section deals with the methodology used for machine learning algorithms on both datasets described in the previous chapter. Data used is directly exported from the CSV file created earlier. After dataset collection and pre-processing data is directly imported from the CSV file created previously and after normalization machine learning algorithms are applied. A typical supervised machine learning workflow is shown in Figure 3.3. A detailed overview of the proposed methodology is shown in Figure 3.4. In the machine learning phase for training and testing, datasets are used. This next sections mainly covers the following analysis of proposed methods.

1. Performing Predictive Analysis

2. Performing Descriptive Analysis

3.3 Performing Predictive Analysis

To predict COVID-19 diagnosis and status of died and recovered patients, we have trained several machine learning models by using clinical symptoms. For this purpose, we trained four ML models including support vector, KNN, logistic regression, and random forest. All these models are implemented

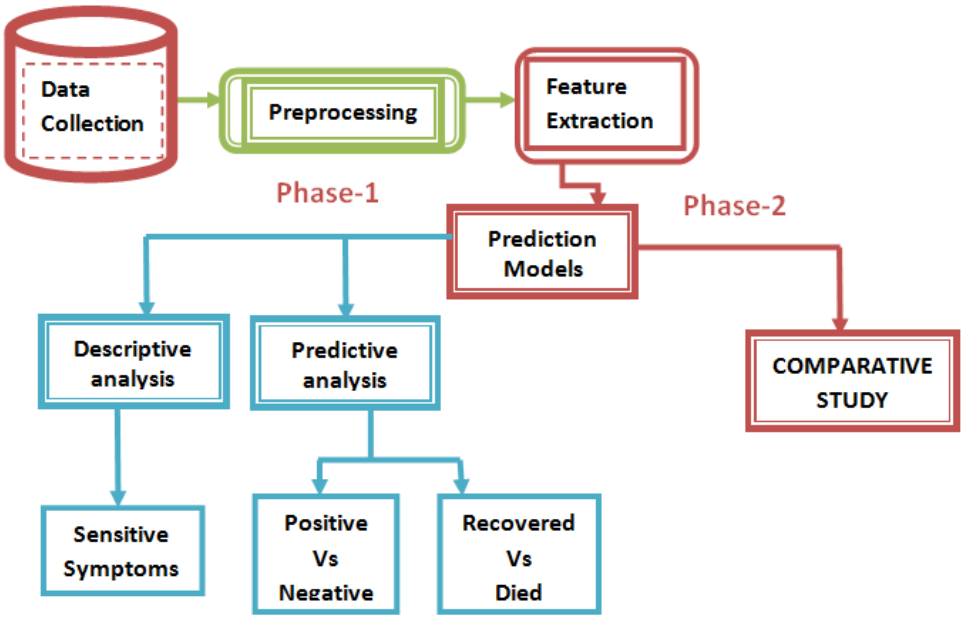


Figure 3.4: Methodology Workflow

using the Scikit-learn Python library [105]. The results are quite promising in our dataset. Random forest [106] outperformed among other three Models. Preprocessed features are used as inputs for each model, yielding an output prediction risk score ranging from 0 to 1. A thresholding function is then used to transform the output into a class label. In sci-kit-learn, the sequential feature selection method is used to implement feature selection. The area under the receiving-operating characteristic curve (AUC) and area under the precision-recall (PR) curve, as well as average precision (AP), are used to examine and interpret model performance.

3.3.1 Techniques/Approaches

Moving towards the pipeline followed for machine learning algorithms, the initial step is to load dataset; after that the dataset is normalized because some features have binary values, some features have continuous values, whereas some features have discrete values. To overcome this problem, all data is normalized using min-max scaling. The normalized dataset is then fed to the following four machine learning algorithms for training.

1. Logistic Regression
2. Support Vector Machine (SVM)
3. K- Nearest Neighbors (KNN)
4. Random Forest

1. Logistic Regression

Logistic Regression (LR) is utilized to find the connection between all out dependent factors and independent factors. LR is utilized when the reliant variable has two qualities, like 0 and 1, yes and negative, or valid and invalid, and is thus known as binary logistic regression. Multinomial logistic regression Whenever the reliant variable has multiple qualities [107] is utilised. The LR transformation is denoted formally as:

$$i = LogisticRegression(p) = 1n(P/P - 1) \tag{3.1}$$

based on its link with the label, this method guesses the class of numerical variable. The ten features with values chosen in feature engineering are expressed in the form of a table and delivered as an input.

2. Support Vector Machine SVM

SVM classification tasks comprise testing and training data that include some occurrences of the data. Each observation in the preparation dataset has at least one objective qualities; accordingly, the essential motivation behind svm is to develop a model that will anticipate target worth or values [108]. SVM is utilized for relapse by presenting different loss functions, which might be linear or nonlinear.

$$y(x) = sign\left[\sum_{k=1}^{n} \alpha kyk\psi(x, xk) + b\right] \tag{3.2}$$

It requires 'n' features for the specific text with the provided label. Because the dataset is small, we chose 10 features that are of the unigram and bigram variety. The training set's data points are listed here. Here is the data points $(\mathrm{yk,xk})_1^n, where is the number of features used.$

3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a directed machine learning classifier that is utilized for supervised ML classifiers. The k-nearest classification model inside the higher dimensional space is used as the input variable in both tasks. KNN depends on labeled input information to get familiar with a capacity that will deliver reasonable result when unlabeled information is inputted [104]. The result of KNN order is a class membership where information cases are arranged by a majority vote of their neighbors, with the information occurrence being designated to the class that is generally pervasive among its k-closest neighbors.

4. Random Forest

This approach employs a modified tree learning algorithm that picks and divides each learning process by a random subset of features [109]. The 10 features selected in feature engineering with values are expressed in the form of a table and delivered as an input. The approach constructs a forest from a subset of randomly selected data and sums the votes for the decision trees to determine the final class of the object using several decision trees.

3.3.2 Evaluation Metrices

In our methodology we use these parameters (recall, precision, F1 score, accuracy, support) to estimate the best one, and descriptions of these parameters are given below. The following evaluation measures are used for evaluation.

1. Accuracy
2. Confusion Matrix

1. Accuracy

Different evaluation measures are used to evaluate results produced by supervised and unsupervised algorithms. Accuracy is one of the most important evaluation measure used almost by everyone to check reliability of their results. In data science accuracy means how well the data points are predicted correctly.

$$\text{Accuracy} = \text{No. of correct predictions made} \div \text{Total prediction made} \quad (3.3)$$

2. Confusion Matrix

A technique for measuring performance in machine learning classification is called confusion matrix. In order to determine the true values it is a type of table that allows you to determine the performance of the classification model on a collection of test data. We can determine the model's various parameters, such as precision, Recall, F1 score and so on, using the confusion matrix.

Precision

Precision measures the number of positive class forecasts that actually belong in that category.

$$\text{Precision} = tp/tp + fp \tag{3.4}$$

Recall

The number of positive class expectations generated from all positive examples in the data set is measured by a recall.

$$\text{Recall} = tp/tp + fn \tag{3.5}$$

F1 score

F-Measure gives a solitary score that adjusts both the worries of precision and recall in one number.

$$\text{F1 score} = 2 * precision * recall(\text{precision+recall}) \tag{3.6}$$

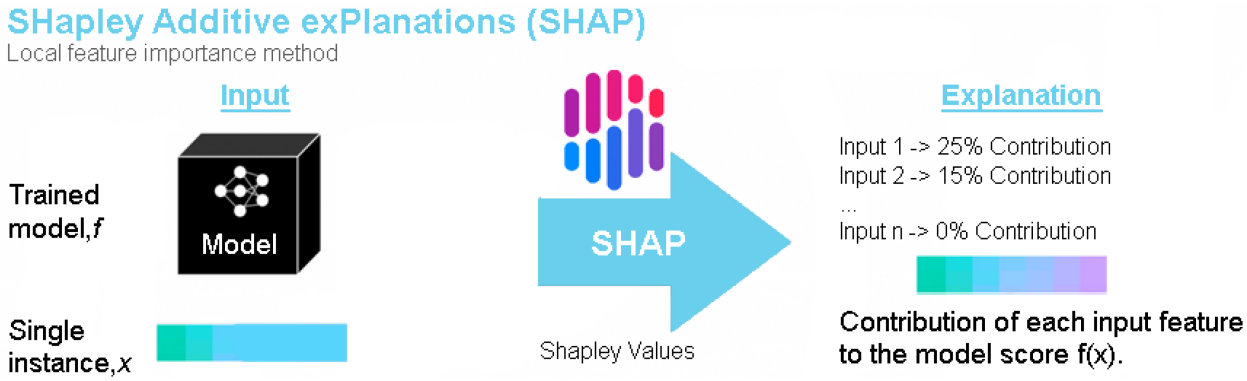


Figure 3.5: SHapley Additive Values

3.4 Performing Descriptive Analysis

This section deals with the complete descriptive analysis. Descriptive Statistical analysis is a crucial aspect of machine learning since it helps you analyze your data. This is simply because machine learning is all about creating predictions. Statistics, on the other hand, are all about deriving conclusions from data, which is essential to having an abstract understanding of the problem. Shapley or SHAP additive explanations is a visualization tool that may be used to visually explain the output of a machine learning model. It can be used to explain any model's prediction by estimating the influence of each characteristic on the prediction [110]. The Shapley value represents all potential combinations and marginal contributions as shown in Figure 3.5. In data analytics, descriptive analysis is one of the concepts that are related to gaining a new perspective on the data and its various existing patterns. In our approach to see more insights into data, we used SHAP values that are obtained to determine the primary factors in driving the model prediction[111].

In the proposed methodology, we utilized SHAP values that divide each sample's prediction result into the contribution of each constituent feature value. SHAP values estimate the significance of each parameter to aggregate predicted results by averaging across samples. The concept of force plot is normally considered to show the influence of each parameter. In our case, we have seen the most sensitive symptoms in predictions. Shapley Additive Explanations (SHAP) beeswarm plot for predicting COVID-19 diagnosis, displaying SHAP values for the model's most relevant features. The summary plots' (y-axis) features are arranged by their mean absolute SHAP values. Each point in the study corresponds to a specific person in the study. The position of each point on the x-axis indicates the impact of that feature on the classifier's prediction for a certain individual. Colors show the values of those qualities. The detailed analysis and results of SHAP values will be discussed in the next chapter.

This chapter outlines the proposed method, details of the algorithm used for predictive analytics, and the SHAPLEY value used for descriptive analytics. The next chapter deals with experiments and results.

Chapter 4

Experiments and Results

This chapter describes experimentation and a complete analysis of all results. Details of the parameters used for algorithms, classification reports, model performance, and descriptive analysis.

4.1 Tuned Parameters of Supervised Machine Learning Algorithms

Parameter tuning is an important step for obtaining accurate and best results. Parameter tuning is a technique in which different parameter values are used and the values which give the best results are kept. Sometimes setting the values to default also gives convenient results. The tuned parameter values we have set are used for supervised machine learning algorithms, which can be seen in Table-4.1.

Metrics such as accuracy, sensitivity (also known as recall), specificity, positive predictive value (PPV) (also known as precision), negative predictive value (NPV), and f1 score are used to evaluate the models' performance. The AUROC value used to choose the best model. We estimate the shapley

Table 4.1: Supervised Algorithms Tuned Parameters Values

Algorithms		Parameters values after tuning
Random Forest	Nestimators=500	MaxDepth=10 RandomState=0
SVM		Gamma='auto' Probability=True
KNN		No. of neighbors: 200
Logistic Regression		Default values

values of each variable to better understand their separate contributions to predictive models. The Python programming language and the scikit-learn module were used to execute all of the analyses.

4.2 Classification Reports for COVID-19 Prediction

A classification report shows precision, recall, f1-score and accuracy. Precision provides information about how many of the predictions are correct; its value ranges between 0-1; worst results give 0 precision, whereas best results give a 1.0 precision value.

Table 4.2: Report: RF for COVID-19 Prediction

	0	1	Accuracy	Macro avg	Weighted avg
Precision	0.7061	0.8667	0.8506	0.7867	0.8361
Recall	0.5761	0.9612	0.8506	0.8563	0.8506
F1-score	0.5831	0.9125	0.8506	0.8301	0.831
Support	142.000	601.000	0.8506	743.000	743.000

Table 4.3: Report: SVM for COVID-19 Prediction

	0	1	Accuracy	Macro avg	Weighted avg
Precision	0.6861	0.8607	0.8438	0.7767	0.8261
Recall	0.5761	0.9632	0.8438	0.6563	0.8406
F1-score	0.5831	0.9025	0.8438	0.6301	0.8210
Support	142.000	601.000	0.8438	743.000	743.000

Similarly, recall tells how much of the positive class predictions are matched, this also ranges from 0-1 and f1-score tells how many correct positive predictions are done (value range from 0-1)[103]. Best results by random forest and SVM as very high f1-score, recall, and precision, can be seen for Covid-19 Positive class. Classification reports of Random Forest and SVM can be seen in Table-4.2 and Table-4.3, respectively. To predict the severity of patients, Random Forest again outperformed as illustrated in Table-4.5. SVM has very low accuracy during the status prediction of hospitalized patients. We can claim that Random Forest is the best Prediction model as compared to the other three. The quantitative results can also be seen in the Table-4.4 and Table-4.5 for supervised machine learning algorithms.

Table 4.4: Classification Report COVID-19 Prediction: Cross Validation -1

Algorithms	Precision	Recall	F1 Score	Accuracy
Random Forest	**0.86**	**0.96**	**0.91**	**0.85**
SVM	**0.86**	**0.96**	**0.90**	**0.84**
KNN	0.85	0.96	0.90	0.83
Logistic Regression	0.85	0.96	0.90	0.83

Table 4.5: Classification Report Status Prediction: Cross Validation -2

Algorithms	Precision	Recall	F1 Score	Accuracy
Random Forest	**0.89**	**0.88**	**0.88**	**0.90**
KNN	**0.83**	**0.84**	**0.83**	**0.85**
SVM	0.51	0.52	0.52	0.51
Logistic Regression	0.51	0.52	0.52	0.50

4.3 Graphical Results of Supervised Machine Learning Algorithms

This section covers all the graphical results. The model's performance is also shown by utilizing ROC curves. One of the important classification

evaluation techniques is the confusion matrix.

4.3.1 Confusion Matrix

A confusion matrixis used over test data to check the number of false and correct predictions done by the model. A Confusion matrix calculates the following:

1. True Positive (TP)
2. True Negative (TN)
3. False Positive (FP)
4. False Negative (FN)

1. True Positive (TP)

A true positive shows the correct number of predictions for the positive class. For example, in this case, if a recovered item is classified as recovered, it will be considered as true positive. In Figure 4.1, the high true positive numbers for status prediction can be seen for random forest. Similarly, for Test indications, the number of True positives can be seen in Figure 4.2 for both KNN and SVM.

2. True Negative (TN)

A true negative shows the correct number of predictions of the negative class. For example, in this case, if a status decease sample is classified as decease, it will be considered as true negative. The high true negative numbers for prediction can be seen in Figure 4.1 and Figure 4.2.

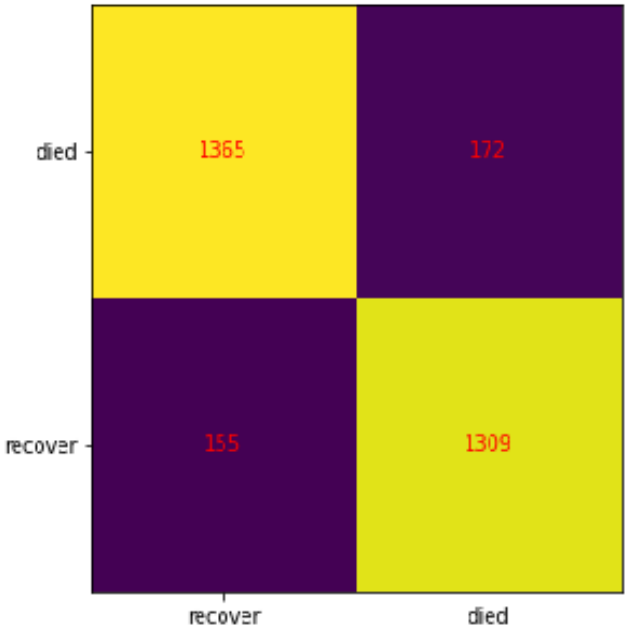


Figure 4.1: Confusion Matrix: Random forest

3. False Positive (FP)

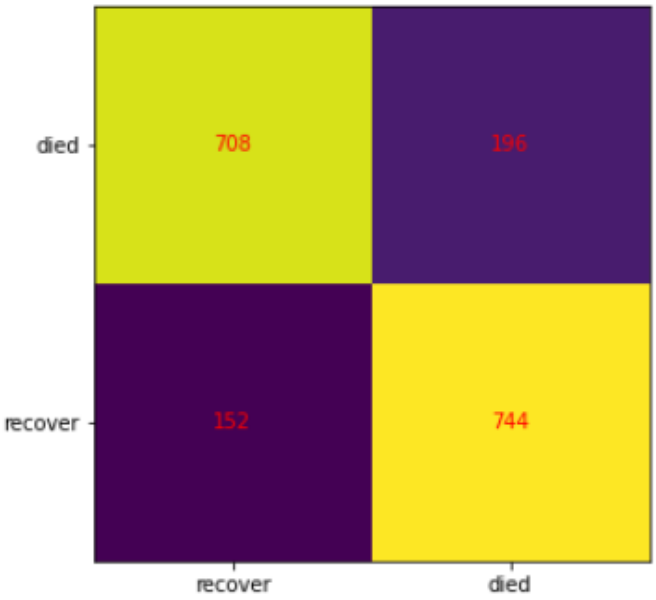
False-positive shows a negative sample classified as positive. For example, in this case, if a critical patient status is classified as died, it will be considered as false positive.

4. False Negative (FN)

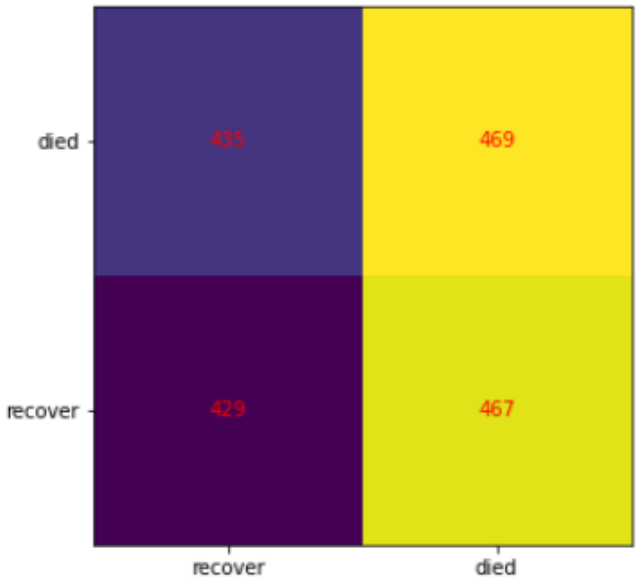
False-negative shows a positive classified as negative. For example in this case, if a recovered sample is classified as died it will be considered as true positive.

4.3.2 Cross Validation for Severity Prediction

In Figure 4.1 large number of samples termed as true positive and slightly better true negative samples can be seen for the random forest; a very small



(a) Confusion Matrix:KNN



(b) Confusion Matrix:SVM

Figure 4.2: Confusion Matrix

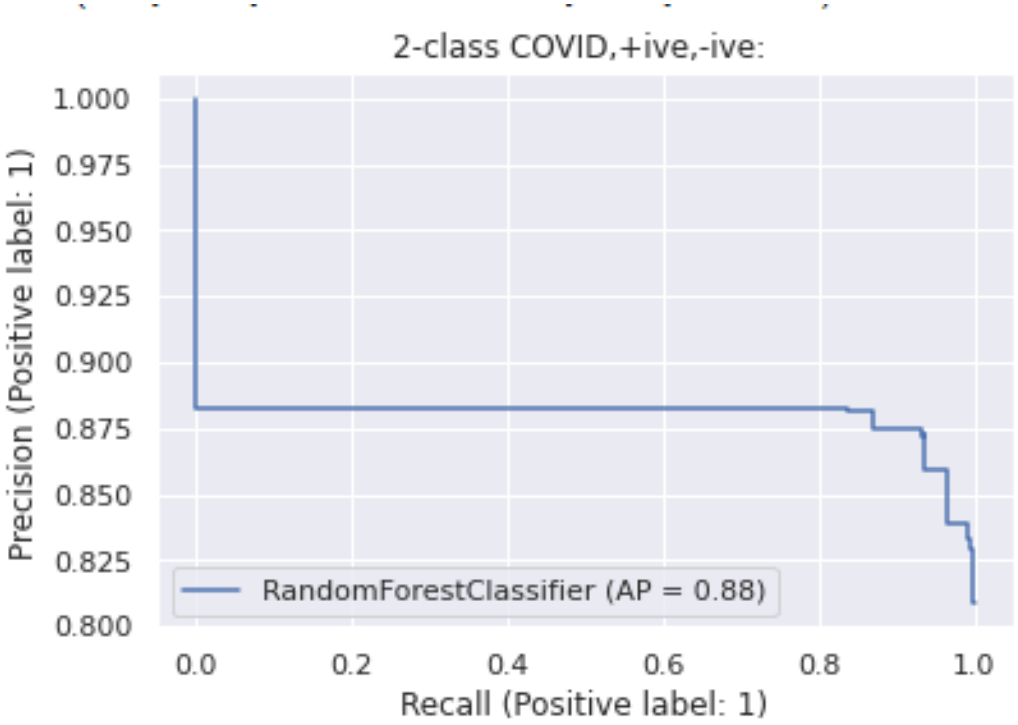
number of samples are classified as true negative. In Figure 4.2 the confusion matrixes for each of the applied machine learning algorithms using a self-collected dataset as a training dataset: Both true negative and true positive results are moderate using k-Nearest Neighbors. In Figure 4.2(b) for SVM Lower number of true positives and true negatives as compared to Random Forest and KNN. Moving towards the result analysis of supervised algorithms, it is observed that Random Forest and KNN gave the most reliable results, using cross-validation. On the other hand, SVM shows unreliable prediction upon the dataset for status prediction.

4.3.3 Models Performance

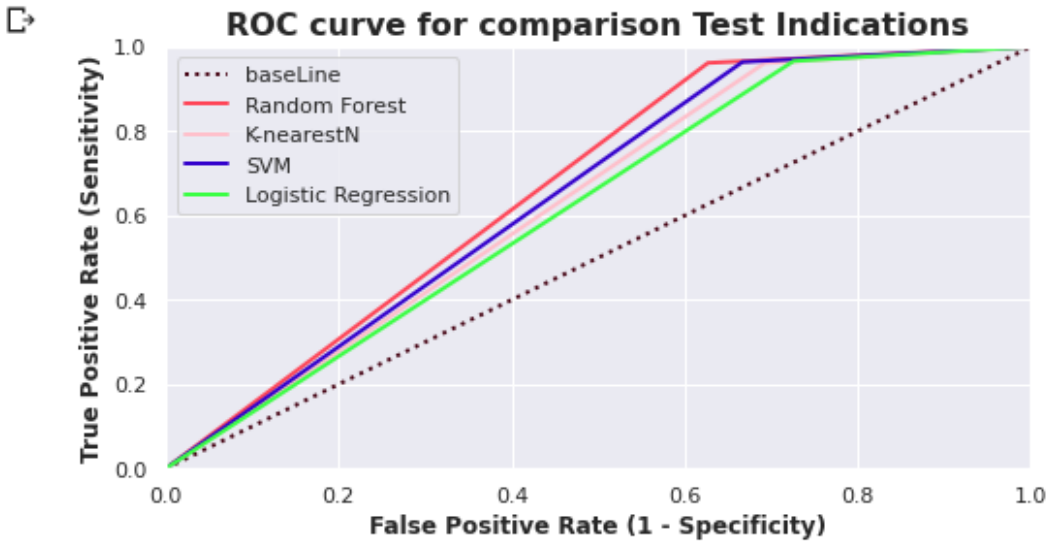
In this section we thoroughly discussed the model performance. First, we evaluated the machine learning algorithms predicted performance for a given individual outcome. It also displays test set ROC and PR curves as well as cross validation calibration curves for all the algorithms applied to the full training set.

We trained our model initially for COVID-19 diagnosis on 6000 clinical reports, of which 3000 were found to be positive. The predictive performance of COVID-19 detection models yielded encouraging results as shown in Figure 4.3. All of the models performed admirably well, with very minor differences. The classification reports for the experiments and reasonable results can be seen for class predictions and test indications. To predict the positive and negative results, all the models performed well, but Random Forest outperformed among others. The machine learning algorithms sensitivity and specificity are also quite high, in most cases exceeding 0.90, with an average sensitivity of 0.93 and specificity of 0.84.

At the second stage, we analyzed the clinical files of the admitted patients

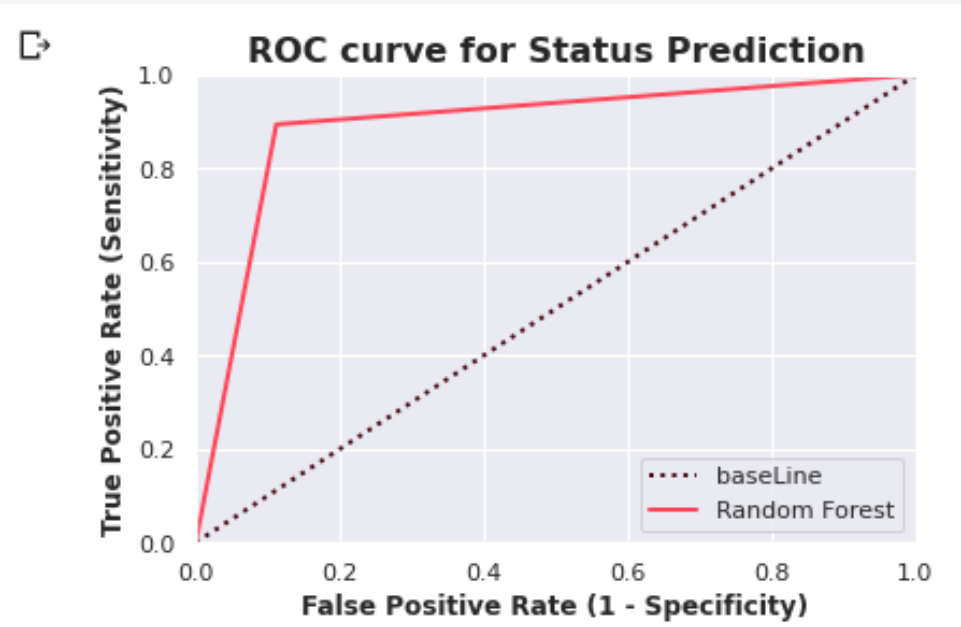


(a) Random Forest

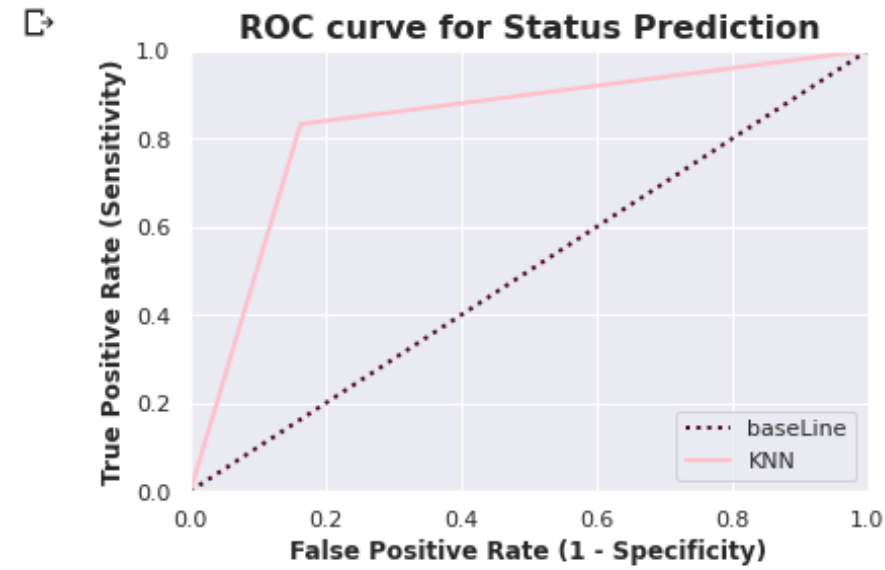


(b) Performance of all Models

Figure 4.3: COVID-19 Symptom based Prediction

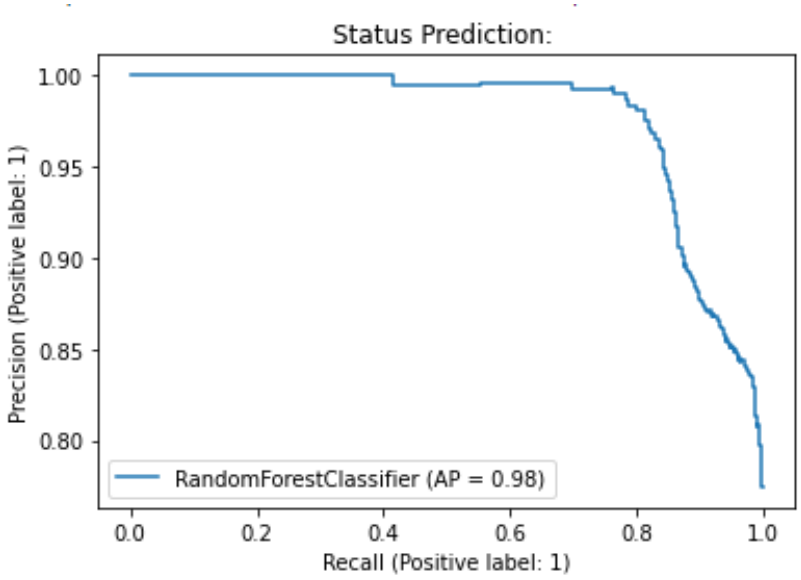


(a) Random Forest

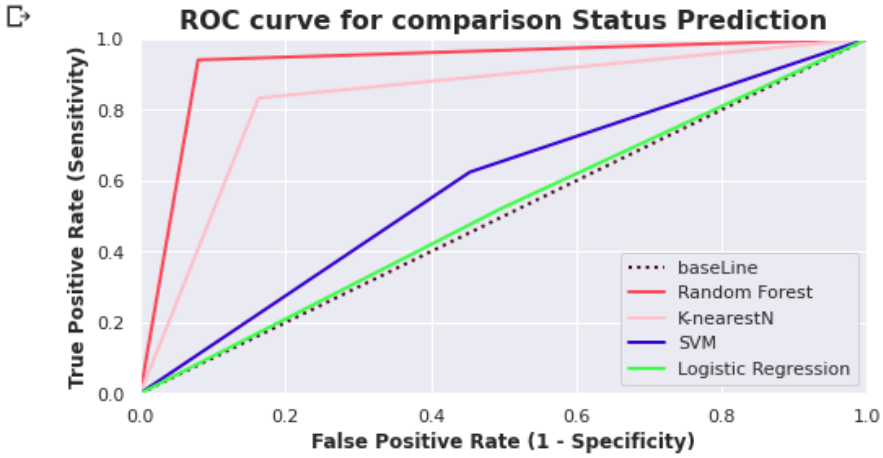


(b) KNN

Figure 4.4: Prediction Models(died vs recovered)



(a) PR Curve



(b) Models Comparison

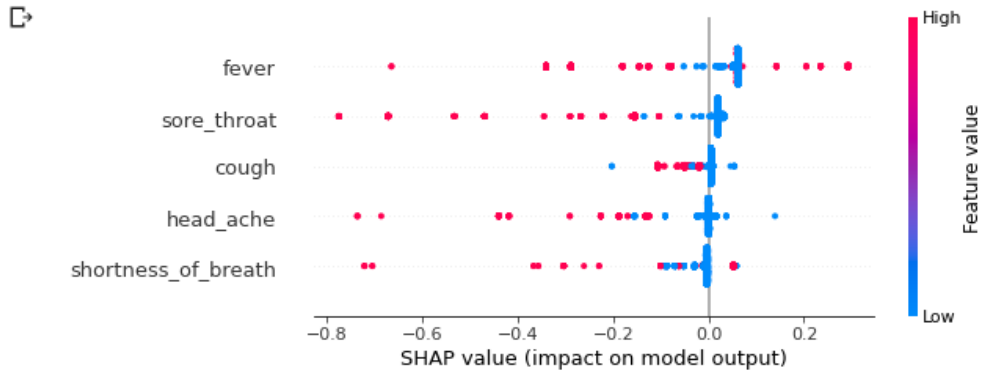
Figure 4.5: Performance Comparison

to predict their status based on the severity of their symptoms. We thoroughly examined clinical data to collect relevant features. Random Forest performed well with the AUC(AP=98) in Figure 4.5(a) and KNN followed with 0.84 accuracies in predicting the status of a patient, whether it was deceased or recovered. SVM and Logistic Regression show the most unreliable results for severity prediction as shown in Figure 4.5(b).

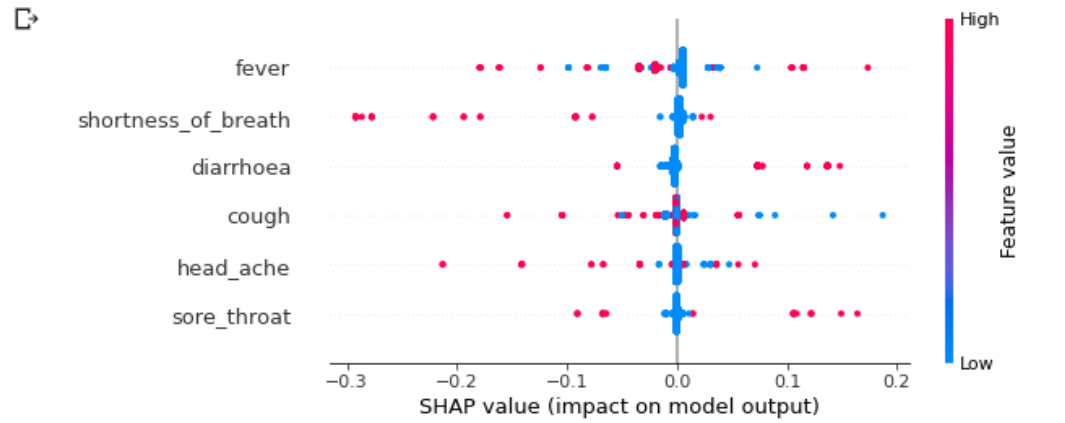
4.4 Interpretation and Comparative Analysis

This section describes a comprehensive comparative analysis and interpretation of models. The ROC curve explains how much better the model distinguishes between both classes for hospitalized patients. AUC means the area under the curve; it shows how much the model is capable of distinguishing between the two classes. AUC having a score of 0.5 means no discrimination ability seen by the model [112]. A score of 0.6-0.7 is termed poor discrimination; the more the value is closer to 1.0; the better the discrimination ability of the model. If the curve is closer to the top left corner, the model has performed very well. Four of the most well-known structured data machine learning models (Logistic Regression, Random Forest, KNN, and SVM) were trained on 70% of the data and tested on the remaining 30%, which represented new unknown data. The test set produced all of the data presented in this study. For both predictions, we compare all four models. All of the models practically offered good results for COVID-19 positive case prediction.

Moving towards the result analysis of supervised algorithms, it is observed that random-forest gives the most reliable results in Figure 4.4(a), using roc random forest is the closest to the top left corner, proving its distinguishing ability between recovered and decease patient. On the other hand, SVM had



(a) SHAPLEY Analysis for Most Sensitive Features of Positive Cases



(b) SHAPLEY Analysis of Most Sensitive Features of Third Wave

Figure 4.6: Descriptive Statistics

the most unreliable prediction upon the dataset of recovered and decease Patients, whereas Logistic Regression didn't perform well. This is because the targeting does not have a linear correlation with the characteristic.

4.5 Descriptive Statistics

In Shapley Additive Explanations (SHAP) COVID-19 diagnosis, for which SHAP values are shown, the most important features of the model. In the

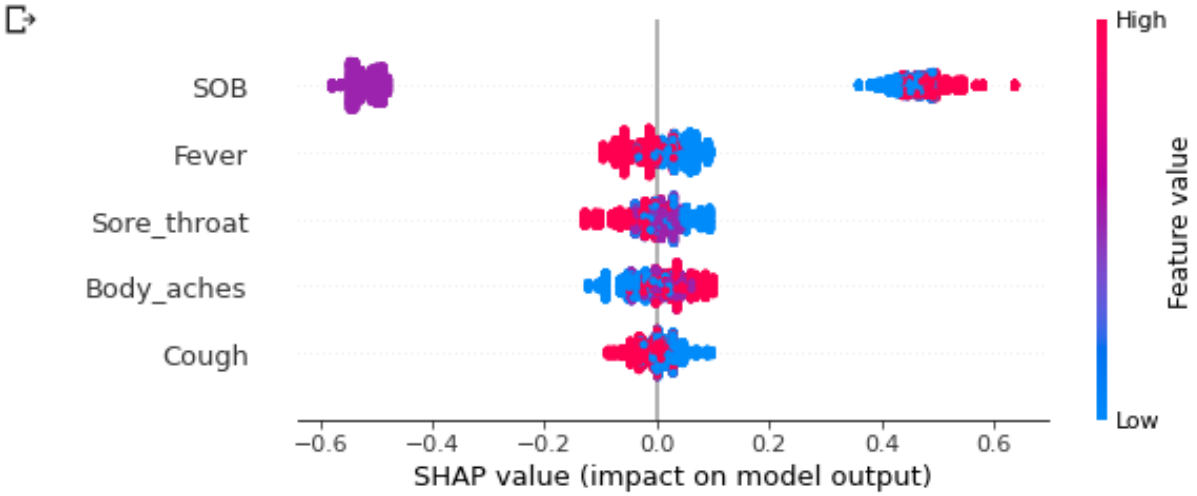


Figure 4.7: SHAPLEY Analysis of Most Sensitive Features severity prediction.

abstract plot (y-axis) the properties are arranged according to their average absolute SHAP values. The point of the study is equal to one person. The position of each point on the X-axis reflects the effect it has on the feature. Rating predictor for an individual the values of these traits (i.e.fever) are reflected in their color. Figure 4.6 shows the descriptive statistics for the demographic characteristics[113] of the patient's symptoms, according to the Shapley values. The variables are ranked according to the contribution of each specific algorithm. SHAP summary plot for the voting classifier model is shown. The SHAP analysis ranked being symptomatic as the most relevant factor, and it was highly associated with a high predicted chance of COVID-19 infection.

Among the COVID-19 symptoms, In Figure 4.6(a) the top five symptoms for the positive cases. It shows the top five symptoms, fever, sore-throat, cough, body ache, and shortness of breath of more sensitive patients that

most contributed to predicting a positive class. The discriminatory results in Figure 4.6(b) for the third wave, like diarrhea and shortness of breath, are discovered to be quite common during patients' hospital stays.

A severe outcome in the Shapley Values in Figure 4.7 where SOB has significant importance for severe cases. Another observation made throughout the analysis at some aspects, such as diarrhea, had a very low impact on the first and second waves of COVID-19. But for the third wave, most individuals reported this symptom. Similarly, SOB for a deceased patient has a high repeat rate. All algorithms yielded promising results in the prediction of COVID-19. At the end of severity prediction, SVM and logistic regression failed to accurately assess the patient's condition. Later, the RF performed better for both predictions due to its versatile nature. There is no denying the fact that RF is the best choice for this dataset and analysis as it makes all possible combinations.

In the next chapter we intend to discuss the conclusion and future work.

Chapter 5

Conclusion and Future work

COVID-19 has now become a global issue, with the World Health Organization declaring it a pandemic. This virus propagated from China to the rest of the world. Reliable procedures for confirming COVID-19 diagnosis are critical to the disease's efficient management and eradication [114].

5.1 Summary

This study examined the ability of clinical data, including patient symptoms and to predict COVID-19 diagnosis. In this thesis work, we reported that machine learning algorithms, notably the random forest, KNN, SVM, and Logistic Regression of these models, are capable of accurately predicting COVID-19 diagnosis and patient status. Random forest performed well during testing and training, so it was chosen as the best classifier of choice for the entire model in terms of both AUC and AP. The whole model gave promising results and may have a stronger potential for generalization to larger data sets. The current study sought to evaluate the ML classifiers for the prediction of COVID-19 and its severity among patients based on clinical symptoms. Our technique differs from previous research [59], which also collects

and evaluates the absence of some important symptoms like diarrhea and lost of taste and smell. This is especially significant in resource-constrained poor countries. Our findings reported good accuracy scores with a complete set of clinical symptoms which were missing in previous studies.

5.2 Novelty and Contribution

Many current investigations are being conducted to investigate the specific etiological mechanisms of SARS-Cov-2 and the associated spectrum of symptoms in different countries. In Pakistan, no study was found based on clinical symptoms. So, we focused on complete analysis and insights of patient's clinical symptoms, which will help the health care professionals and policymakers. As a result, the ability to make quick clinical choices and decisions and make efficient use of healthcare resources in Pakistan. Furthermore, our findings differ from prior studies[59] in terms to built correlation between symptoms and different waves, as well as severity predictions. So, the technique given in this study may help the healthcare system respond to future outbreaks of this disease and other respiratory viruses in general. Based on simple clinical manifestations, the model we developed can be utilized for a COVID-19 test screening. Further developing clinical needs might lighten the current strain on the medical care system by enabling even more efficient use of health services in the event of subsequent waves of the SARS-Cov-2 pandemic.

5.3 Limitations and Challenges

The current study has some limitations as well. We totally rely on regional clinical files for AJK and Rawalpindi. Because the lack of digital record of reported symptoms in our hospitals, only a limited number of clinical files are analysed due to time constraints. The most challenging part is gathering

data from hospitals across the whole country and then formulating broad predictions. The whole process is very time taking and require extra effort to initially focus on data availability. Some other ethical approval and protocols are also require for expanding this work to other hospitals and regions .

5.4 Future Work

Future research should concentrate on a vast number of clinical trials around the country. Medical records should be more consistent and up to date, so that data collection should take less time. Dataset requires to maximize. The minimal number of patients with serious diseases in the trial brings statistical power into question. A future study might create projections based on provincial data and then compare its severity in rural and urban areas of Pakistan for upcoming waves.

Bibliography

[1] Zohair Malki **andothers**. "Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches". **in***Chaos, Solitons & Fractals*: 138 (2020), **page** 110137.

[2] Norihiro Kokudo **and** Haruhito Sugiyama. "Call for international cooperation and collaboration to effectively tackle the COVID-19 pandemic". **in***Global Health & Medicine*: 2.2 (2020), **pages** 60–62.

[3] Herbert W Hethcote. "The mathematics of infectious diseases". **in***SIAM review*: 42.4 (2000), **pages** 599–653.

[4] Thirumalaisamy P Velavan **and** Christian G Meyer. "The COVID-19 epidemic". **in***Tropical medicine & international health*: 25.3 (2020), **page** 278.

[5] Ahmed Mohammed Obaid Al Saidi **andothers**. "Decisive leadership is a necessity in the COVID-19 response". **in***The Lancet*: 396.10247 (2020), **pages** 295–298.

[6] So Yat Wu **andothers**. "The diagnostic methods in the COVID-19 pandemic, today and in the future". **in***Expert review of molecular diagnostics*: 20.9 (2020), **pages** 985–993.

[7] Adam J Kucharski **andothers**. "Early dynamics of transmission and control of COVID-19: a mathematical modelling study". **in***The lancet infectious diseases*: 20.5 (2020), **pages** 553–558.

[8] Biswaranjan Paital. "Nurture to nature via COVID-19, a self-regenerating environmental strategy of environment in global context". **in***Science of the Total Environment*: 729 (2020), **page** 139088.

[9] Alexander E Loeb **andothers**. "Departmental experience and lessons learned with accelerated introduction of telemedicine during the COVID-19 crisis". **in***The Journal of the American Academy of Orthopaedic Surgeons*: (2020).

[10] Athalia Christie **andothers**. "Decreases in COVID-19 cases, emergency department visits, hospital admissions, and deaths among older adults following the introduction of COVID-19 vaccine—United States, September 6, 2020–May 1, 2021". **in***Morbidity and Mortality Weekly Report*: 70.23 (2021), **page** 858.

[11] Nathan Ford, Marco Vitoria **and** Meg Doherty. "World Health Organization Guidance to Support Human Immunodeficiency Virus Care Models During the Coronavirus Disease 2019 Era". **in***Clinical Infectious Diseases*: 74.9 (2022), **pages** 1708–1710.

[12] Clıstenes Fernandes da Silva, Arnaldo Candido Junior **and** Rui Pedro Lopes. "Predictive Analysis of COVID-19 Symptoms in Social Networks through Machine Learning". **in***Electronics*: 11.4 (2022), **page** 580.

[13] Qing Han **andothers**. "Long-Term sequelae of COVID-19: A systematic review and meta-analysis of one-year follow-up studies on post-COVID symptoms". **in***Pathogens*: 11.2 (2022), **page** 269.

[14] César Fernández-de-Las-Peñas **andothers**. "Symptoms Experienced at the Acute Phase of SARS-CoV-2 Infection as Risk Factor of Long-term Post-COVID Symptoms: The LONG-COVID-EXP-CM Multicenter Study". **in***International Journal of Infectious Diseases*: (2022).

[15] Giuseppe Maglietta **andothers**. "Prognostic factors for post-COVID-19 syndrome: a systematic review and meta-analysis". **in***Journal of clinical medicine*: 11.6 (2022), **page** 1541.

[16] Ralph Weissleder **andothers**. "COVID-19 diagnostics in context". **in***Science translational medicine*: 12.546 (2020), eabc1931.

[17] Saverio Caini **andothers**. "Meta-analysis of diagnostic performance of serological tests for SARS-CoV-2 antibodies up to 25 April 2020 and public health implications". **in***Eurosurveillance*: 25.23 (2020), **page** 2000980.

[18] Dasari Naga Vinod **and** SRS Prabaharan. "Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19". **in***Chaos, Solitons & Fractals*: 140 (2020), **page** 110182.

[19] Rajarshi Guha **andothers**. *What is the role of cheminformatics in a pandemic?* 2021.

[20] P Cao **andothers**. "Chapter II: diagnostic methods". **in***European Journal of Vascular and Endovascular Surgery*: 42 (2011), S13–S32.

[21] Erfan Maddah **and** Borhan Beigzadeh. "Use of a smartphone thermometer to monitor thermal conductivity changes in diabetic foot ulcers: a pilot study". **in***Journal of Wound Care*: 29.1 (2020), **pages** 61–66.

[22] Arni SR Srinivasa Rao **and** Jose 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". **in***Infection Control & Hospital Epidemiology*: 41.7 (2020), **pages** 826–830.

[23] Quoc-Viet Pham **andothers**. "Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts". **in***IEEE access*: 8 (2020), **page** 130820.

[24] Zixin Hu **andothers**. "Artificial intelligence forecasting of covid-19 in china". **in***arXiv preprint arXiv:2002.07112*: (2020).

[25] Ying Liu **andothers**. "The reproductive number of COVID-19 is higher compared to SARS coronavirus". **in***Journal of travel medicine*: (2020).

[26] Sung-mok Jung **andothers**. "Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases". **in***Journal of clinical medicine*: 9.2 (2020), **page** 523.

[27] Dennis Normile. "Singapore claims first use of antibody test to track coronavirus infections; Science; AAAS". **in***Science Magazine*: (2020).

[28] Luca Falzone **andothers**. "Current and innovative methods for the diagnosis of COVID-19 infection". **in***International journal of molecular medicine*: 47.6 (2021), **pages** 1–23.

[29] Sina F Ardabili **andothers**. "Covid-19 outbreak prediction with machine learning". **in***Algorithms*: 13.10 (2020), **page** 249.

[30] Jasper Fuk-Woo Chan **andothers**. "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster". **in***The lancet*: 395.10223 (2020), **pages** 514–523.

[31] Mashura Shammi **andothers**. "COVID-19 pandemic, socioeconomic crisis and human stress in resource-limited settings: A case from Bangladesh". **in***Heliyon*: 6.5 (2020), e04063.

[32] Patrick M Bossuyt **andothers**. "STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies". **in***Clinical chemistry*: 61.12 (2015), **pages** 1446–1452.

[33] Sien Ombelet **andothers**. "Clinical bacteriology in low-resource settings: today's solutions". **in***The Lancet Infectious Diseases*: 18.8 (2018), e248–e258.

[34] Jaspreet Singh **and** Jagandeep Singh. "COVID-19 and its impact on society". **in***Electronic Research Journal of Social Sciences and Humanities*: 2 (2020).

[35] S Freeman. "Systemic social issues reflected in coronavirus outbreak". **in***Dossier Covid 19. Impactos socioculturales de la pandemia*: (2020), **pages** 18–20.

[36] DB Lora Jones. "Coronavirus: Eight charts on how it has shaken economies". **in***Retrieved from BBC News*: (2020).

[37] Satyendra Pratap Singh, Mahak Nischal **and** Aditi Saxena. "7 Strategies for Prevention". **in***Health Informatics and Technological Solutions for Coronavirus (COVID-19)*: (2021), **page** 97.

[38] Uwe Wollina. "Challenges of COVID-19 pandemic for dermatology". **in***Dermatologic therapy*: 33.5 (2020), e13430.

[39] G Radi **andothers**. "Global coronavirus pandemic (2019-nCOV): implication for an Italian medium size dermatological clinic of a II level hospital". **in***J Eur Acad Dermatol Venereol*: 34.5 (2020), e213–e214.

[40] Lisheng Wang **andothers**. "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence". **in***International journal of antimicrobial agents*: 55.6 (2020), **page** 105948.

[41] Giuseppe Marano **andothers**. "Convalescent plasma: new evidence for an old therapeutic tool?" **in***Blood Transfusion*: 14.2 (2016), **page** 152.

[42] Feng Wang **andothers**. "The laboratory tests and host immunity of COVID-19 patients with different severity of illness". **in***JCI insight*: 5.10 (2020).

[43] Mohammed A Mamun **and** Mark D Griffiths. "First COVID-19 suicide case in Bangladesh due to fear of COVID-19 and xenophobia: Possible suicide prevention strategies". **in***Asian journal of psychiatry*: 51 (2020), **page** 102073.

[44] Kanika K Ahuja **andothers**. "Fear, xenophobia and collectivism as predictors of well-being during Coronavirus disease 2019: An empirical study from India". **in***International Journal of Social Psychiatry*: 67.1 (2021), **pages** 46–53.

[45] Kapil Goyal **andothers**. "Fear of COVID 2019: First suicidal case in India!" **in**(2020).

[46] Mayowa Oyesanya, Javier Lopez-Morinigo **and** Rina Dutta. "Systematic review of suicide in economic recession". **in***World journal of psychiatry*: 5.2 (2015), **page** 243.

[47] Mohammed A Mamun **and** Mark D Griffiths. "A rare case of Bangladeshi student suicide by gunshot due to unusual multiple causalities". **in***Asian journal of psychiatry*: 49 (2020).

[48] Emily Hu. "COVID-19 testing: challenges, limitations and suggestions for improvement". **in**(2020).

[49] Liu Wenjun Fangyao **andothers**. "Potential false-positive rate among the'asymptomatic infected individuals' in close contact with COVID-19 patients Zhuang Guihua, Shen Mingwang, Zeng Lingxia, Mi Baibing, Chen". **in***Shen Mingwang, Zeng Lingxia, Mi Baibing, Chen*: ().

[50] Qinjian Hao, Hongmei Wu **and** Qiang Wang. "Difficulties in false negative diagnosis of coronavirus disease 2019: a case report". **in**(2020).

[51] Wenjie Yang **and** Fuhua Yan. "Patients with RT-PCR-confirmed COVID-19 and normal chest CT". **in***Radiology*: 295.2 (2020), E3–E3.

[52] Yang Zhou **andothers**. "Clinical reports on early diagnosis of novel coronavirus (2019-nCoV) pneumonia in stealth infected patients". **in**(2020).

[53] Junaid Shuja **andothers**. "COVID-19 open source data sets: a comprehensive survey". **in***Applied Intelligence*: 51.3 (2021), **pages** 1296–1325.

[54] Dong Jin Park **andothers**. "Development of machine learning model for diagnostic disease prediction based on laboratory tests". **in***Scientific reports*: 11.1 (2021), **pages** 1–11.

[55] Lingzhong Meng, Fang Hua **and** Zhuan Bian. "Coronavirus disease 2019 (COVID-19): emerging and future challenges for dental and oral medicine". **in***Journal of dental research*: 99.5 (2020), **pages** 481–487.

[56] Giacomo Grasselli, Antonio Pesenti **and** Maurizio Cecconi. "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response". **in***Jama*: 323.16 (2020), **pages** 1545–1546.

[57] Yichun Cheng **andothers**. "Kidney disease is associated with in-hospital death of patients with COVID-19". **in***Kidney international*: 97.5 (2020), **pages** 829–838.

[58] Dan Assaf **andothers**. "Utilization of machine-learning models to accurately predict the risk for critical COVID-19". **in***Internal and emergency medicine*: 15.8 (2020), **pages** 1435–1443.

[59] Yazeed Zoabi, Shira Deri-Rozov **and** Noam Shomron. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms". **in***npj digital medicine*: 4.1 (2021), **pages** 1–5.

[60] Zaid Abdi Alkareem Alyasseri **andothers**. "Review on COVID-19 diagnosis models based on machine learning and deep learning approaches". **in***Expert systems*: 39.3 (2022), e12759.

[61] Senthilkumar Mohan **andothers**. "An approach to forecast impact of Covid-19 using supervised machine learning model". **in***Software: Practice and Experience*: 52.4 (2022), **pages** 824–840.

[62] Gurjit S Randhawa **andothers**. "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study". **in***Plos one*: 15.4 (2020), e0232391.

[63] Sovesh Mohapatra **andothers**. "Repurposing therapeutics for COVID-19: rapid prediction of commercially available drugs through machine learning and docking". **in***PLoS One*: 15.11 (2020), e0241543.

[64] Shashi Kushwaha **andothers**. "Significant applications of machine learning for COVID-19 pandemic". **in***Journal of Industrial Integration and Management*: 5.04 (2020), **pages** 453–479.

[65] Akib Mohi Ud Din Khanday **andothers**. "Machine learning based approaches for detecting COVID-19 using clinical text data". **in***International Journal of Information Technology*: 12.3 (2020), **pages** 731–739.

[66] LJ Muhammad **andothers**. "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery". **in***SN Computer Science*: 1.4 (2020), **pages** 1–7.

[67] Charlyn Nayve Villavicencio **andothers**. "COVID-19 Prediction applying supervised machine learning algorithms with comparative analysis using WEKA". **in***Algorithms*: 14.7 (2021), **page** 201.

[68] Milind Yadav, Murukessan Perumal **and** M Srinivas. "Analysis on novel coronavirus (COVID-19) using machine learning methods". **in***Chaos, Solitons & Fractals*: 139 (2020), **page** 110050.

[69] Panagiotis G Asteris **andothers**. "A novel heuristic algorithm for the modeling and risk assessment of the COVID-19 pandemic phenomenon". **in***Computer Modeling in Engineering & Sciences*: 125.2 (2020), **pages** 815–828.

[70] Shi Cheng **andothers**. "Swarm intelligence in big data analytics". **in***International conference on intelligent data engineering and automated learning*: Springer. 2013, **pages** 417–426.

[71] Israel Edem Agbehadji **andothers**. "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing". **in***International journal of environmental research and public health*: 17.15 (2020), **page** 5330.

[72] Peipei Wang **andothers**. "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics". **in***Chaos, Solitons & Fractals*: 139 (2020), **page** 110058.

[73] Ali Narin, Ceren Kaya **and** Ziynet Pamuk. "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks". **in***Pattern Analysis and Applications*: 24.3 (2021), **pages** 1207–1220.

[74] Ahmed Hosny **andothers**. "Artificial intelligence in radiology". **in***Nature Reviews Cancer*: 18.8 (2018), **pages** 500–510.

[75] Ophir Gozes **andothers**. "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis". **in***arXiv preprint arXiv:2003.05037*: (2020).

[76] Michael J Horry **andothers**. "COVID-19 detection through transfer learning using multimodal imaging data". **in***Ieee Access*: 8 (2020), **pages** 149808–149824.

[77] Sweta Bhattacharya **andothers**. "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey". **in***Sustainable cities and society*: 65 (2021), **page** 102589.

[78] P Pandiaraja **andothers**. "A Scrutiny on COVID-19 Detection using Convolutional Neural Network and Image Processing". **in***Annals of the Romanian Society for Cell Biology*: (2021), **pages** 3831–3843.

[79] Santiago Tello-Mijares **and** Luisa Woo. "Computed tomography image processing analysis in covid-19 patient follow-up assessment". **in***Journal of Healthcare Engineering*: 2021 (2021).

[80] R Dhaya **andothers**. "Deep net model for detection of covid-19 using radiographs based on roc analysis". **in***Journal of Innovative Image Processing (JIIP)*: 2.03 (2020), **pages** 135–140.

[81] Md Zahangir Alom **andothers**. "COVID_MTNet: COVID-19 detection with multi-task deep learning approaches". **in***arXiv preprint arXiv:2004.03747*: (2020).

[82] Tawsifur Rahman **andothers**. "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images". **in***Computers in biology and medicine*: 132 (2021), **page** 104319.

[83] Sadettin Melenli **and** Aylin Topkaya. "Real-time maintaining of social distance in covid-19 environment using image processing and big data". **in***The International Conference on Artificial Intelligence and Applied Mathematics in Engineering*: Springer. 2020, **pages** 578–589.

[84] Umashankar Subramaniam **andothers**. "An Expert System for COVID-19 Infection Tracking in Lungs Using Image Processing and Deep Learning Techniques". **in***BioMed Research International*: 2021 (2021).

[85] Ki-Sun Lee **andothers**. "Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest X-ray images using explainable deep-learning algorithm". **in***Journal of Personalized Medicine*: 10.4 (2020), **page** 213.

[86] Mominul Ahsan **andothers**. "COVID-19 detection from chest X-ray images using feature fusion and deep learning". **in***Sensors*: 21.4 (2021), **page** 1480.

[87] Sakifa Aktar **andothers**. "Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development". **in***JMIR Medical Informatics*: 9.4 (2021), e25884.

[88] Luca Flesia **andothers**. "Predicting perceived stress related to the Covid-19 outbreak through stable psychological traits and machine learning models". **in***Journal of clinical medicine*: 9.10 (2020), **page** 3350.

[89] Jonathon Love **andothers**. "JASP: Graphical statistical software for common statistical designs". **in***Journal of Statistical Software*: 88.1 (2019), **pages** 1–17.

[90] Brett Milliner. "The effects of combining timed reading, repeated oral reading, and extensive reading". **in***Reading in a Foreign Language*: 33.2 (2021), **pages** 191–211.

[91] Jia Xue **andothers**. "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach". **in***Journal of medical Internet research*: 22.11 (2020), e20550.

[92] Cornelia Herbert, Alia El Bolock **and** Slim Abdennadher. "How do you feel during the COVID-19 pandemic? A survey using psychological and linguistic self-report measures, and machine learning to investigate mental health, subjective experience, personality, and behaviour during the COVID-19 pandemic among university students". **in***BMC psychology*: 9.1 (2021), **pages** 1–23.

[93] Shreyash Sonthalia **andothers**. "COVID-19 Likelihood Meter: a machine learning approach to COVID-19 screening for Indonesian health workers". **in***medRxiv*: (2021).

[94] Ahmed Shihab Albahri **andothers**. "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review". **in***Journal of medical systems*: 44 (2020), **pages** 1–11.

[95] Vaibhav Bhatnagar **andothers**. "Descriptive analysis of COVID-19 patients in the context of India". **in***Journal of Interdisciplinary Mathematics*: 24.3 (2021), **pages** 489–504.

[96] Gunness Harlem. "Descriptive analysis of social determinant factors in urban communities affected by COVID-19". **in***Journal of Public Health*: 42.3 (2020), **pages** 466–469.

[97] Peter Bauer **andothers**. "An international comparison of age and sex dependency of COVID-19 deaths in 2020: a descriptive analysis". **in***Scientific reports*: 11.1 (2021), **pages** 1–11.

[98] Andrew T Levin **andothers**. "Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications". **in***European journal of epidemiology*: 35.12 (2020), **pages** 1123–1138.

[99] Serkan Ballı. "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods". **in***Chaos, Solitons & Fractals*: 142 (2021), **page** 110512.

[100] Ibrahim Arpaci **andothers**. "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms". **in***Multimedia Tools and Applications*: 80.8 (2021), **pages** 11943–11957.

[101] Ebaa Fayyoumi, Sahar Idwan **and** Heba AboShindi. "Machine learning and statistical modelling for prediction of novel COVID-19 patients case study: Jordan". **in***Machine Learning*: 11.5 (2020), **pages** 3–11.

[102] Fernando Timoteo Fernandes **andothers**. "A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil". **in***Scientific reports*: 11.1 (2021), **pages** 1–7.

[103] S Kohli. "Understanding a classification report for your machine learning model". **in***India: Medium. com*: (2019).

[104] H Onel. *Machine learning basics with the K-nearest neighbors algorithm, towards data science. 2018*. 2020.

[105] Lars Buitinck **andothers**. "API design for machine learning software: experiences from the scikit-learn project". **in***arXiv preprint arXiv:1309.0238*: (2013).

[106] Leo Breiman. "Random forests". **in***Machine learning*: 45.1 (2001), **pages** 5–32.

[107] Safial Islam Ayon, Md Milon Islam **and** Md Rahat Hossain. "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques". **in***IETE Journal of Research*: (2020), **pages** 1–20.

[108] Md Milon Islam **andothers**. "Prediction of breast cancer using support vector machine and K-Nearest neighbors". **in***2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*: IEEE. 2017, **pages** 226–229.

[109] Rakesh Katuwal **and** Ponnuthurai N Suganthan. "Enhancing multi-class classification of random forest using random vector functional neural network and oblique decision surfaces". **in***2018 International Joint Conference on Neural Networks (IJCNN)*: IEEE. 2018, **pages** 1–8.

[110] Samanta Knapič **andothers**. "Explainable artificial intelligence for human decision support system in the medical domain". **in***Machine Learning and Knowledge Extraction*: 3.3 (2021), **pages** 740–770.

[111] Scott M Lundberg **and** Su-In Lee. "A unified approach to interpreting model predictions". **in***Advances in neural information processing systems*: 30 (2017).

[112] Karimollah Hajian-Tilaki. "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation". **in***Caspian journal of internal medicine*: 4.2 (2013), **page** 627.

[113] Lehana Thabane **and** Noori Akhtar-Danesh. "Guidelines for reporting descriptive statistics in health research". **in***Nurse researcher*: 15.2 (2008).

[114] AM Whittington **andothers**. "Coronavirus: rolling out community testing for COVID-19 in the NHS". **in***BMJ opinion*: (2020).

[115] Jianfeng Xie **andothers**. "Critical care crisis and some recommendations during the COVID-19 epidemic in China". **in***Intensive care medicine*: 46.5 (2020), **pages** 837–840.

[116] Remigio Ismael Hurtado Ortiz, Juan Carlos Barrera Barrera **and** Katherine Michelle Barrera Barrera. "Analysis model of the most important factors in Covid-19 through data mining, descriptive statistics and random forest". **in***2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*: **volume** 4. IEEE. 2020, **pages** 1–8.

[1] [2] [4] [115] [3] [23] [24] [25] [29] [53] [57] [56] [65] [58] [66] [67] [29] [68] [69] [70] [71] [59] [87] [89] [88] [90] [91] [116] [92] [93] [94] [102] [105] [106] [107] [108] [104] [109] [111] [112] [113] [114]