University of Essex

Department of Mathematical Sciences

MA981 DISSERTATION

An approachable way to detect Covid-19 using Machine learning

Md Momen Hasan

Registration no: 2201937

Supervisor: **Dr Tao Gao**

August 29, 2025
Colchester

Contents

List of Figures

List of Tables

Abstract

The Covid-19 is the most contagious and deadly pandemic in the world. The virus promoted severe social, mental, psychological, educational, and economic depression, impacting billions of people around the world because it can be transmitted through respiratory droplets. The research aim is to scrutinise Covid-19 infected people based on their symptoms using different machine learning classifier models, i.e., logistic regression, random forest, k-nearest neighbor, support vector machine (SVM), naïve Bayes, XGBoost classifiers. The KNN classifier models perform well considering all other classification metrics with an accuracy of 97.6%, precision, and recall values of 99.5% and 97.5% respectively. With this research, we can predict whether a person is affected by Covid-19 or not.

CHAPTER 1

Introduction

1.1 Background

The Coronavirus (Covid-19) pandemic is an ongoing widespread contagious ubiquitous health crisis pressed by the novel coronavirus SARS-CoV-2. It was first distinguished in December 2019 in Wuhan, Hubei Province, China, and spread exponentially in the world. Moreover, the Middle East Respiratory Syndrome (MERS) and SARS large scale pandemics were both brought on by coronaviruses in the last two decades. It has been commonly thought that SARS-CoV-2 which was mainly originated or found in bats-could generate future contagious outbreaks [1, 2, 3].

However, the outbreak was begun from the local seafood market in Wuhan, and rapidly spread other locations of China. On 11th January 2020, China reported its first death, and on 20th January 2020, the first cases reported outside of China including the U.S and other Asian countries. On 30 January 2020, The W.H.O. (World Health Organizations) declared a global health emergency of international concern officially, and on 11th March 2020, declared pandemic by the W.H.O [4]. As of 2nd August 2023, over 768.98 million cases and 6.953 million deaths have reported. However, It is worse in Europe region about 275.79 million cases, 205.84 million cases in the Western Pacific including China and Australia, 193.21 million cases in American Continent including the United States, 61.2 million cases in Southeast Asian countries including India, 23.5 million cases in Eastern Mediterranean countries, 9.54 million cases in African continent

according to the W.H.O Covid-19 Dashboard.

1.2 Transmission and Case Fatality rate

The Covid-19 typically transmits through respiratory droplets released when an infected person coughs, sneezes, talks, or breathes heavily. The respiratory droplets usually contain the virus, and when the person contacts other people, the virus is transmitted through their nose, eyes, or mouth, which can lead to infection. Covid-19 transmission can also happen by touching surfaces or objects contaminated by the virus already. As it is a contagious disease, infected people who travel also can transmit the virus from one location to another location. In some situations, the virus can spread through the air. The virus can carry small respiratory particles called aerosols that can persist in the air and be breathed by others especially in closed spaces with limited ventilation. This virus is highly contagious, leading the widespread transmission and outbreaks in communities [3, 4, 5].

According to CDC & NHS, The SARS-Cov-2 has significantly higher potential infective virus that it can not contain across the world, and has resulted in a global pandemic compared to previous genetic coronaviruses such as SARS-CoV and MERS-CoV. The reason behind it is a deadly virus because of human-to-human transmission, asymptomatic and pre-symptomatic transmission, and high reproduction number [6, 7].

1.3 Global Health and Economic impact

The healthcare system around the world has been facing outstanding challenges during the Covid-19 beginning period as well as now including in the U.S and Western countries even the virus has deteriorated Asian and African continents healthcare systems. With the flooding of Covid-19 patients in hospitals, the government had to confirm extra negative rooms in anticipation of the surging virus. They needed to ensure they had additional staff and allocated resources to backup primary supporters to provide enough support to Covid-19 patients. However, these were not enough to tackle the virus because of they needed to pay extra money as well training handle affected person to use covid-19 related testing kits, musk, and addressed the PPE shortages. For severe cases infected people, hospitals needed to invest on buying modern ventilise, and recruiting

more doctors. The African continent was suffered and affected mostly shortages of PPE, Kits, Musk, equipment as well as train people. Therefore, every nation needs to increase budget in health care systems. The severe situation was faced in 2020 in Asian countries including Pakistan, Sri Lanka, Bangladesh, India [9, 10, 11, 12]. However, Europe was significantly affected in the health care system during the outbreak, especially in Italy, Germany, the U.K., and France. The shortages of doctors affected treatment for the people. The escalated budget increase related to Covid-19 and cancellation of outpatient visits lost revenue, efficient processes and essential surgeries hospitals amidst the countries turned to financially devastated [12, 13, 14, 25].

The pandemic damaged 15% of the United States GDP in 2020 which was estimated at almost 3.3 trillion. One study has shown that almost 51% people suffered from the reduction of income from employment or lost employment status. [10, 15]. As the virus originated in China, it had a significant impact on international affairs because China is accountable for almost 13% export market worldwide. Therefore, the countries usually export goods from China, once the Chinese government has imposed lockdown, travel restrictions as well mandatory quarantine, resulting countries did not importing products from China at that period. The government imposed strict travel restrictions from Wuhan to other locations, the people of Wuhan as well as other provinces could not access buses, trains or airline stations. With the limited access, China saw 6.8% GDP shrink during the first quarter of 2020 [10, 16, 17]. The first case of Covid-19 was reported in India on January 30, 2020, and the government imposed a lockdown to most of the affected cities such as Kerala, Delhi , Mumbai. The restrictions on both domestic and international travel, public gatherings, and businesses put the Indian economy in a sluggish situation. These implementations immediately impacted on agriculture, manufacturing, and service sectors [18, 19, 20]. The other big Country Brazil observed a major financial stock fall in Sao Paulo stock market, estimated more than 15% decline, its worsening weekly drop since the last financial crisis, and GDP fell over 11% in the second quarter of 2020 [21].

The economic impact due to Covid-19 in middle -and low-income countries (LMICs) was probably challenging and devastating. The major obstacles were access to PPE, testing kits, clean water, widespread poverty, and access to healthcare as well as implementation of lockdown impacted financially LMICs people without proper support

from their government even though limited lockdowns middle income countries such as Bangladesh, they observed 18% remittances fall, and the Kyrgyz Republic saw more than 35% than the previous year [22, 23].

1.4 Impact on Mental Health

The Covid-19 pandemic has promoted social variation among the world population. It has encouraged panic, fear, anxiety, uncertainty, and widespread hysteria. Anxiety and depression have been rising since the inception of pandemic. A study examined by the CDC observed that 40.9% of associates with minimum one terrible mental illness in the U.S. Between the ages of 18 to 24, were related to substances such as mistreatment, anxiety, depression, trauma, suicidal thinking, and stress-related situations. The health workers including doctors were feeling stigmatized, resulting in individuals being perceived as biased, approached differently as well as they were living separately from their family and society. They were suffering for mentally and emotionally as they mostly witnessed the mortality rate. Most of the time, they needed to wear a face mask. It affected their face and breathing problems on some occasions. Also includes the availability of personal protective equipment (PPE), overextended working hours, impact on decision making and fears of infecting loved ones. The society also fears helping infected people. Therefore, the pandemic destroys the social bonding emotionally [24, 25, 26].

1.5 Symptoms of Covid-19

According to W.H.O, CDC, NHS, the covid-19 symptoms can vary from mild to severe, and sometimes depends on the infected person's age or ethnic. However, It has been recognised some common symptoms include fever or chills, cough, shortness of breath, fatigue, muscle aches, headache, loss of taste or smell, sore throat, congestion, runny nose, nausea, and diarrhea. However, some individuals have asymmetric symptoms which means that there are no visible symptoms, yet they can still unknowingly spread the virus to other people. That kind of symptom was a major threat to control covid-19 early because they inadvertently transmitted the virus [6, 7, 8].

Furthermore, the severity and the combination of symptoms experienced by Covid-19 patients can vary widely. Some individuals may experience mild symptoms resembling a common cold or flu, others may experience severe respiratory distress, which can result in pneumonia or acute respiratory distress syndrome (ARDS). However, old people may endure joint pain, reduced muscle mass, memory lapses, vision changes, hearing loss, fatigue, sleep changes, digestive issues, balance problems, osteoporosis, heart and cardiovascular conditions, declining lung function, and changes in the skin due to Covid-19 [27].

1.6 Early and accurate detection in controlling the spread of the virus

Since the inception of Covid-19 outbreak in December 2019, doctors and researchers have been trying to detect the virus in the human body based on symptoms, epidemiological data and diagnostic tests to recognize infected persons. Recollecting signs such as fever, cough, and sense of smell or taste along with travel history and highly Covid-19 affected areas. These were the most common measurements at that time [28].

However, with the development of accepted diagnostic tests for Covid-19, Medical professionals used these methods to confirm or rule out disease. The reverse transcription-polymerase chain reaction (RT-PCR) test has been extremely crucial in distinguishing Covid-19 by detecting the virus' generic materials respiratory samples. After sample collection, RNA is extracted as well as converted cDNA through reverse transcription. DNA (generic material) is generated by the PCR amplification and fluorescent probes release detectable fluorescence, and software confirms viral presence in real time. Despite its accuracy and discovery of even small RNA amounts, the RT-PCR test requires highly trained personal and modern equipment [29, 30]. While effective but has some limitations including sample quality, timing and false result. Despite some limitations, it roles diagnosing Covdi-19 cases, and aiding health care professionals to give relief [31]. Alongside of this, rapid antigen tests are widely used for identifying the specific viral protein of SARS-CoV-2 virus surface. Serological assessments (antibody test) also used for identifying the virus [32, 33].

1.7 Machine Learning and Potential for COVID-19 Detection

Machine Learning (ML) has shown tremendous promise in the detection of Covid-19, accelerating futuristic techniques to determine, diagnose, and manage the virus. ML algorithms used various disease patterns and approaches from the extensive dataset to boost accuracy of detection systems. In this part we will discuss a few insights regarding uses in detecting Covid-19 patients;

ML algorithms can analyse medical imaging data such as chest X-rays and CT scans to help in identifying Covid-19 related anomalies in the lungs. These algorithms develop abilities from labeled images to determine between healthy and infected individuals pattern, assisting healthcare personnel to make more precise interpretations. However, ML can make decisions based on symptoms of Covid-19 whether a person is affected or needed to test Covid-19 virus adding other factors like demographics and travel history. These forecasts classify testing and prepare for allocation resources particularly during the outbreak. Moreover, ML-powered chatbots can press people to real time prediction based on their symptoms and these decision guide users can concern their health to seek medical advice from the medical professionals along with reducing unnecessary pressure on healthcare systems.

ML models also can propose spread of disease, calculate peak infection rate and guide public health officials based on the real time epidemiological forecasting data. In addition, contact tracing by assessing location data to determine potential exposure worries and implementing necessary action. Furthermore, ML techniques manage to gather data sources from the social media platforms and news, scrutinize public sentiment and detect emerging outbreak. These forecasts help to provide containment approaches.

1.8 Vaccination Efforts and dirtribution

During the early outbreak of Covid-19 pandemic, the developed countries including U.S, China, Germany, U.K, Russia have sought to develop vaccines to contain Covid-19 spread globally even though they funded millions of dollars to the pharmaceutical

companies [34, 39, 36]. Initially, Several companies have responded to progress making vaccines against the virus including Moderna, Pfzer and Biontech, Sinopharm as well as other famous medical researchers from such as Oxford, Cambridge, Harvard , John Hopkins etc, [37, 38]. Moreover, out of the all pharmaceutical companies Moderna (mRNA) vaccine was early pushed to the human body for the first stage trial which was 16th March 2020 accoring to Forbes. However, Pfzer and Biontech co-created mRNA vaccine , received high efficiency on clinical trials, and the U.S government approved emergency authorization across the country to combat the virus. Also, Moderna's mRNA vaccine was also approved at that time and received positive and most efficient results, gaining emergency use approval in the U.S and multiple nations. Later, AstraZeneca and the University of Oxford collaborated on Vaxzevria (AZD1222), Sinopharm and Sinovac, inactivated virus based vaccines approved by the Chinese government to contain the virus in China. Russia created Sputnik V, a viral vaccine, received emergency use in Russia and other collaborating with other countries. However, other companies also developed vaccines to mitigate the spread of virus such as Bharat Biotech's Covaxin, Novavax, Johnson and Johnson's Ad26.COV2.S, a single-dose adenovirus-based vaccine etc,. These governments and the companies really played a pivotal role to contain the virus spread globally and mitigate health risk of Covid-19 infection [37, 38, 39, 40].

After developing vaccines, the advanced nations were focused on distributing vaccines around the globe. It is usually a tenuous task to distribute vaccines in developing or low/mid level countries because of the infrastructure of healthcare system. To curb virus spread and severe illness as well as end the pandemic effectively, countries needed to implement diverse approaches based on factors like vaccine availability because those nations who created vaccines distribute most of the production in their country. However, the common strategies included high risk groups, mass vaccination places, mobile units for remote populations, appointment strategies, educational campaigns, second dose remainders. Moreover, COVAX, the international collaboration of some middle or low income countries, participated in COVAX , to ensure access to an equal number of vaccines based on their population. Furthermore, an effective supply chain of vaccines from other countries low or middle income countries needed to ensure their cold storage system as well as link to their alliance countries to get vaccines and infrastructure. Every nation ensures to give vaccine to their people in local clinics at

least two doses within a time frame, and they run campaigns for public awareness to protect public health. These strategies are revised in response to reduce virus infectivity and development of new strategies and emphasizes staying updated through official health authorities and government sources [41, 42, 43].

CHAPTER 2

Related Works

Significant work has been implemented to predict Covid-19 positive or negative using Machine Learning (ML) algorithms with respect to Covid-19 classifier problem. The ML has the potential to save millions of people, and it has the ability to conclude Covid-19 with confidence and accuracy that can contribute information that can be trained through ML.

Choudary et al. and his team worked with the symptoms based dataset. The authors applied four Machine learning techniques including KNN, XGBoost, Random Forest, and SVM algorithms to reduce transmission of the virus and predict Covid-19 positive or negative. However, the team used SHAP analysis, and the Shapley value can identify the attribute contribution, and determine effective prediction based on each attribute according to the author. In this experiment, SVM outperformed all other algorithms in terms of accuracy (98.38%), and SVM can classify the positive or negative classes but in medical perspective, XGBoost outperforms other algorithms in regards to recall value without compromising accuracy score [44].

During the early outbreak of Covid-19, the most proven PT-CTR had a deficit in developing countries including Brazil. The author Batista and his team, designed a Machine learning model to combat Covid-19 advancement in Brazil. The authors collected data from the Israelita Albert Einstein Medical Center in Sao Paulo to implement the Machine learning model. The studies show that Machine Learning (ML) can be extensively characterised Covid-19 victims, and it was conducted by a task force whose response was to contain the outbreak. The data include clinical information including

hemoglobin, platelets, and red blood cells in both suspected and Covid-19 patients. The Authors applied some ML methods including SVM, Logistic Regression, Random Forest, gradient-boosted trees, and Neural Network. Furthermore, the prediction of Support Vector Machine and Random forest performed effectively and successfully differentiated 85% Covid-19 patients. However, SVM and Random forest had the almost same results (sensitivity of 0.677 and specificity of 0.850) but SVM had a slightly better Brier score of 0.160, and Random forest had 0.161. One group of researchers from Bangladesh also implemented several ML methods with several dividers including Logistic Regression, Multilayer Perceptron (MLP), and XGBoost in the same Brazil hospital database, and they received MLP outperformed other ML algorithms. The accuracy score is 93.13% including Precision, Recall, F_1 score, and AUC value [45].

One comparative study conducted by Rohini et al. aimed to predict Covid-19 cases to mitigate outbreaks throughout the world, and the authors used other likelihood metrics including geographical location, travel history, health records, etc., to predict the intensity of the case and possible outcome. They found models developed using KNN received the best and most effective result with a prediction 98.34%, Recall of 97% and an F_1 score of 0.97%. The authors used time series data which will assist future outbreaks [46].

In one epidemiology study dealing with COVID-19 cases from Mexico used by author L.J.Muhammad, et. al., and the dataset has attributes such as pneumonia, asthma, hypertension, cardiovascular diseases, and high risk tobacco factors. In regards to accuracy, the model developed with decision trees was the best of all models developed, with 94.99%.In contrast, in terms of sensitivity and specificity SVM and naive Bayes models showed the best model with scores 93.34% and 94.30% respectively [47].

Sharma and his team performed the processed reorganization through support vector machine learning model. A hyperparameter development technique was applied as a reformed cuckoo search method to ameliorate the SVM classifier prediction accuracy. The authors achieve 80.42% using normal SVM, and feature selection SVM executes an accuracy of 85.6%. However, the hyperparameter tuning SVM feature had the best accuracy which was 96.3% [48].

A team from UCLA, and performed ML tasks based on the clinical data, and the dataset included the ethnicity of the U.S. people. They collected data from the UCLA

Health System (Los Angeles, California, USA) to develop the ML model for Covid-19 proxy diagnosis. They considered SARS-CoV-2 all cases which were tested in emergency or inpatient testing background within the UCLA health system between 1 March 2020 and 24 May 2020 including 1455 ancillary laboratory features. They experimented with some ML models and combined these results for the final classification. The method came up with sensitivity and specificity of 0.93 and 0.64 respectively. However, they did not mention which ML method performs best technically [49].

Yinxiaohe Sun and his team - a study was conducted in Singapore at the National Centre for Infectious Diseases (NCID), and they collected electronic medical records, demographic characteristics, exposure risk factors, contact with travelers from China, recent travel history, etc., data and developed four multivariate logistic regression models to predict Covid-19 patients. They found AUC scores 0.91,0.88,0.88,0.65 respectively [50].

Tiwari et al. investigated real-time data that possessed the global record of cases of Covid-19 outbreak. They applied Naïve Bayes, Support Vector Machine (SVM) and Linear Regression, and investigated that among the three methods Naïve Bayes generated potential results to forecast Covid-19 future leads a high accuracy [51].

The rule-based ensemble method was used to determine the covid-19 death rate with multivariate imputation, feature selection, and SMOTE proposes. Rai and his team implemented a voting system to determine the mode, the model XGBoost achieved the highest accuracy and F_1 score which is 86.9% and 71.6% respectively [52].

V.Rani et al. suggested a pre-processd method and implemented a machine learning model that can classify the patients as Covid-19 suspects including individual likelihood. The model performed effectively in the LSTM method with a good accuracy score. By July 14, 2020 there were around 9.36 lakh confirmed cases in India but LSTM model suggested it would be estimated 2.0 Cr cases at the end of August 2020 [53].

However, Some researchers have implemented both Machine Learning and deep learning proposals to predict coronavirus based on image classification.

Radio imagology has been popular with researchers to detect diseases. Tulin and his team suggested that images contain salient information regarding the Covid-19 virus after applying AI techniques. They implemented DarkNet, a neural network framework including 17 convolutional layers, and proposed diverse filtering on each

layer to classify Covid-19 virus using raw chest X-ray images. This approach produced 98.08% and 87.02% accuracy respectively for binary classes and multi-class cases [54].

Hemdan et al. used deep learning models to recognise Covid-19 through X-ray images and recommended a COVIDX-Net model including designing 7 CNN models such as modified Visual Geometry Group Network (VGG19) and the second version of Google MobileNet. Each model can analyse and learn from the X-ray images to classify either positive or negative cases. The Dense Convolutional Network (DenseNet) and VGG19 both model performed well and had similar results in classifying Covid-19 with 0.89 f1-scores and 0.91 normal [55]. Wong and Wang developed a sophisticated Covid-19 identification deep learning model which is COVID-Net. They have investigated 13,975 CXR images across 13,870 patients and achieved 92.4% accuracy in classifying normal, non-COVID pneumonia, and COVID-19 classifications [57]. Ioannis et al. collected of a dataset containing 224 images with confirmed Covid-19 conditions, 714 images including bacteria and viral pneumonia, and 504 images of normal cases. The author and his team applied MobileNet CNN model, and obtained 96.78%, 98.66%, and 96.46% in terms of the best accuracy, sensitivity, and specificity respectively [58]. Sethy and Behera, implemented 12 various CNN architectures to classify Covid-19 virus usage X-ray images with SVM classifiers. According to their research, the ResNet-50 model coupled with SVM classifier demonstrated the most optimal performance [59]. However, several recent research has emphasized detecting COVID-19 virus that implemented numerous deep learning models based on X-ray images [60, 61, 62].

The above literature review concluded that comprehensive studies have been completed in the discipline of COVID-19 classification problems based on symptoms as well as image-based classification using Machine Learning methods and deep learning. However, these potential outcomes are solely based on the dataset including demographic aspects.

CHAPTER 3

Methodology and Materials

This section describes methods, resources, and datasets used in this study to predict the Covid-19 virus.

3.1 Experimental Setup

In this dissertation research, Machine learning classification models are proposed for the COVID-19 virus. Python is the main language to use because it has a wide range of libraries to implement a machine-learning model. The following structure will be used throughout this study: (a) **Software packages:** This experiment used in the Google Colab. It is a web-based interactive computing environment. It is created by Google, and no need to install any packages in the machine separately unless it is not well known. I decided to complete my dissertation using Python programing language because it has extensive machine learning packages including Sci-Kit Learn, Pandas, Numpy, Matplotlib, Seaborn etc.

(b) **Workstation:** A machine with the following configuration was utilised to experiment.

- Operating.System: Windows 11
- Processors: Core i7 CPU @.2.90GHz
- Memory: 16 Giga Bytes of Memory

3.2 Proposed Methodology

The methodology of the study is elaborately organized into different yet interconnected steps, each of which plays an important function in the overall process. The initial process is meticulous dataset collection, acquiring relevant data for subsequent analysis. The second step involves comprehensive preprocessing, data cleaning, normalization, and missing values handling to ensure the integrity of the dataset. The third step focuses focusing exploratory data analysis, and showing insights of the attribute. Finally, we implement the machine learning method, and then evaluate the performance of each model and justify. The depicted below graphical representation can give the basis of the proposed model.

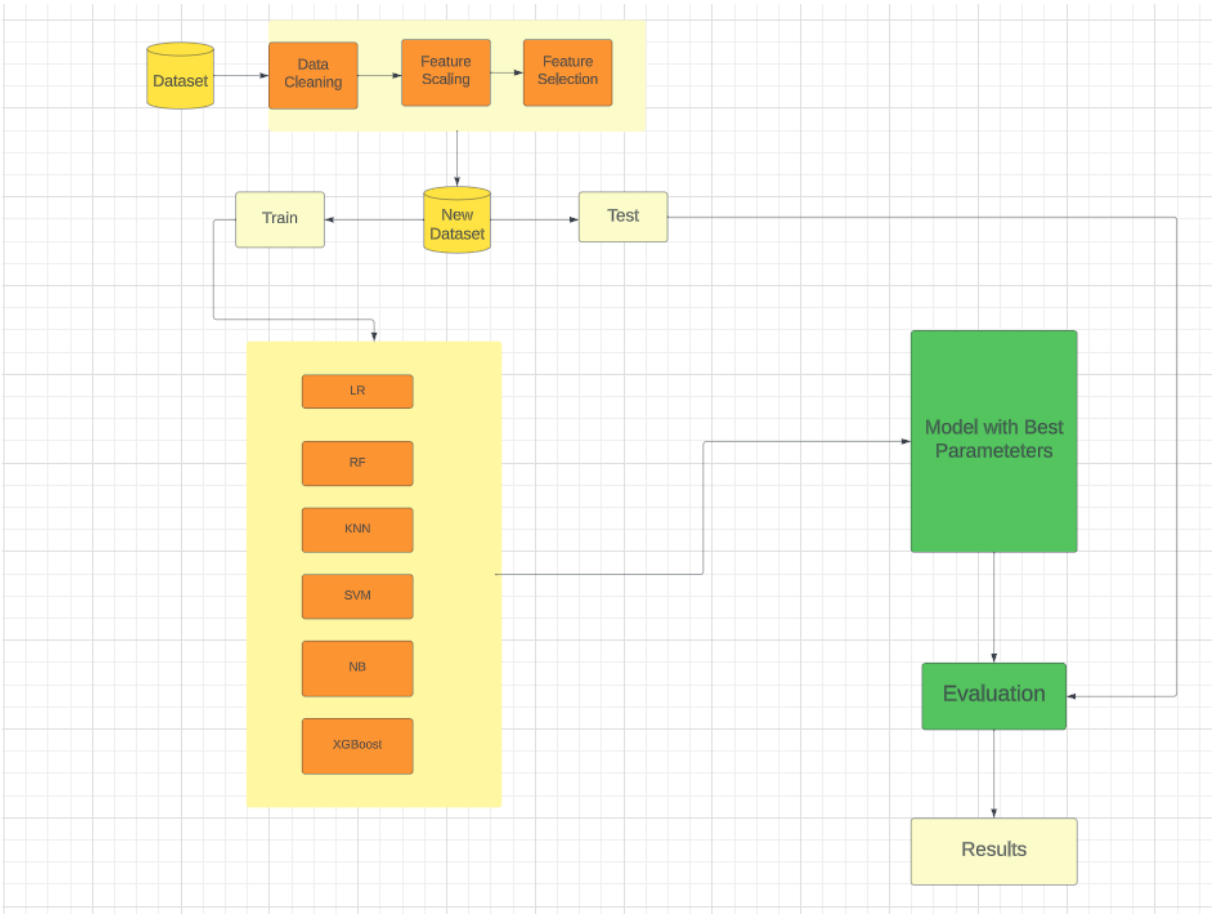

Figure 3.1: Overview of the proposed methodology.

1. **Dataset Description:** As the W.H.O. has declared the coronavirus pandemic a public health crisis, researchers and hospitals have provided open access to data related to the epidemic. The dataset consists of it has 5434 rows and 21

columns. This dataset contains 20 variables that could be determinants in the prediction of COVID-19, as well as one class attribute that defines if COVID-19 is found. The dataset has almost all the attributes available which are the primary symptoms of a COVID-19 patient like we mentioned on the symptoms of Covid-19 sections particularly difficulty breathing, fever, dry cough, sore throat, abroad travel, contact with covid patient, attending large gatherings, and visiting exposed places including other health issues. From the table 3.2, we can clearly see our attribute names and types of data in every attribute. You can find the dataset at: kaggle Symptoms and COVID Presence.

Every attribute in the dataset contains binary data, consisting solely of 'yes' or 'no' values. The dataset comprises 20 independent variables used for predicting whether a person is COVID-19 suspect or not. Additionally, it includes one categorical dependent attribute.

2. **Data Pre-processing:** Data preparation is the process of transforming raw data into a suitable data type. Real-world data can include noise, missing numbers, or be in an unusable format, making it unsuitable with machine learning models. Data preprocessing is an important phase in which we clean the data and make it appropriate for usage in a machine learning model. This improves the model's accuracy and efficiency. Initially, we can observe that we have no missing values in our dataset from the table 3.1.

3. **Removing Features:** From the figure 3.4, we can determine that the attributes "wearing masks" and "sanitization from the market" have only one value, which is 'no', as they do not impact our predictions. We can simply drop those two columns from the dataset even no relation with the correlation matrix.

4. **Features Selection for Implementation:** From the figure 3.5 and 3.6, we can observe that some attribute has no correlation with Covid-19 columns. Therefore, we decided to remove those features, that can overfitting or underfitting our dataset. Now, we have 11 independent variables and one target variable which will be plotted with the independent variables on the x-axis and the target variable on the y-axis.

5. **Encoding Categorical Data:** As we have object values in our all attributes , we

Column	missing_values	percent_missing
Breathing Problem	0	0.0
Fever	0	0.0
Dry Cough	0	0.0
Sore throat	0	0.0
Running Nose	0	0.0
Asthma	0	0.0
Chronic Lung Disease	0	0.0
Headache	0	0.0
Heart Disease	0	0.0
Diabetes	0	0.0
Hyper Tension	0	0.0
Fatigue	0	0.0
Gastrointestinal	0	0.0
Abroad travel	0	0.0
Contact with COVID Patient	0	0.0
Attended Large Gathering	0	0.0
Visited Public Exposed Places	0	0.0
Family working in Public Exposed Places	0	0.0
Wearing Masks	0	0.0
Sanitization from Market	0	0.0
COVID-19	0	0.0

Table 3.1: Missing Values and Percent Missing.

applied label encoding, a widely utilized approach for organizing categorical data in a versatile manner. In this technique, every category is assigned a numerical value based on its alphabetical sequence. As all attributes within our dataset solely comprise either 'yes' or 'no' entries, we have implemented label encoding to translate them into numerical equivalents, specifically 0 and 1. This transformation enhances the dataset's compatibility with machine learning models. Figure 3.4 illustrates the dataset's altered form after undergoing the label encoding process.

Index	Column	Non-Null Count	Dtype
0	Breathing Problem	5434 non-null	object
1	Fever	5434 non-null	object
2	Dry Cough	5434 non-null	object
3	Sore throat	5434 non-null	object
4	Running Nose	5434 non-null	object
5	Asthma	5434 non-null	object
6	Chronic Lung Disease	5434 non-null	object
7	Headache	5434 non-null	object
8	Heart Disease	5434 non-null	object
9	Diabetes	5434 non-null	object
10	Hyper Tension	5434 non-null	object
11	Fatigue	5434 non-null	object
12	Gastrointestinal	5434 non-null	object
13	Abroad travel	5434 non-null	object
14	Contact with COVID Patient	5434 non-null	object
15	Attended Large Gathering	5434 non-null	object
16	Visited Public Exposed Places	5434 non-null	object
17	Family working in Public Exposed Places	5434 non-null	object
18	Wearing Masks	5434 non-null	object
19	Sanitization from Market	5434 non-null	object
20	COVID-19	5434 non-null	object

Table 3.2: Description of Columns.

6. **Splitting the dataset :** The subsequent step in preprocessing data for machine

learning involves dividing the dataset. It's essential to partition a machine learning model's dataset into two distinct segments: one for training and another for testing purposes. In our case, we performed a division of the data using an 80:20 ratio. This signifies that 80% of the data is utilised for training the model, while the remaining 20% is set aside for conducting testing.

7. **Exploratory Data Analysis :**Exploratory data analysis is used to evaluate various datasets in order to simplify them by capturing their primary features. This concise overview can be demonstrated using statistical graphics and various data visualization tools. From the figure 3.2 highlights that the 'COVID-19' column shows 4383 'Yes' values, indicating met conditions related to COVID-19, and 1051 'No' values, representing unmatched conditions. This provides insights into the occurrence frequency. The sample of few attributes table 3.3 describes the scenario of the dataset related to Covid-19 columns.

Index	Breathing Problem	Fever	Dry Cough	Sore throat	Abroad travel	COVID-19
40	Yes	Yes	Yes	Yes	No	Yes
41	Yes	Yes	Yes	No	Yes	Yes
42	Yes	Yes	Yes	Yes	No	Yes
43	Yes	Yes	Yes	Yes	Yes	Yes
44	Yes	Yes	Yes	Yes	Yes	Yes
45	Yes	Yes	Yes	No	No	Yes
46	Yes	Yes	Yes	No	No	Yes
47	Yes	Yes	Yes	No	Yes	Yes
48	Yes	Yes	Yes	Yes	No	Yes
49	Yes	Yes	Yes	Yes	No	Yes
50	Yes	Yes	Yes	Yes	No	Yes

Table 3.3: Sample Data Table

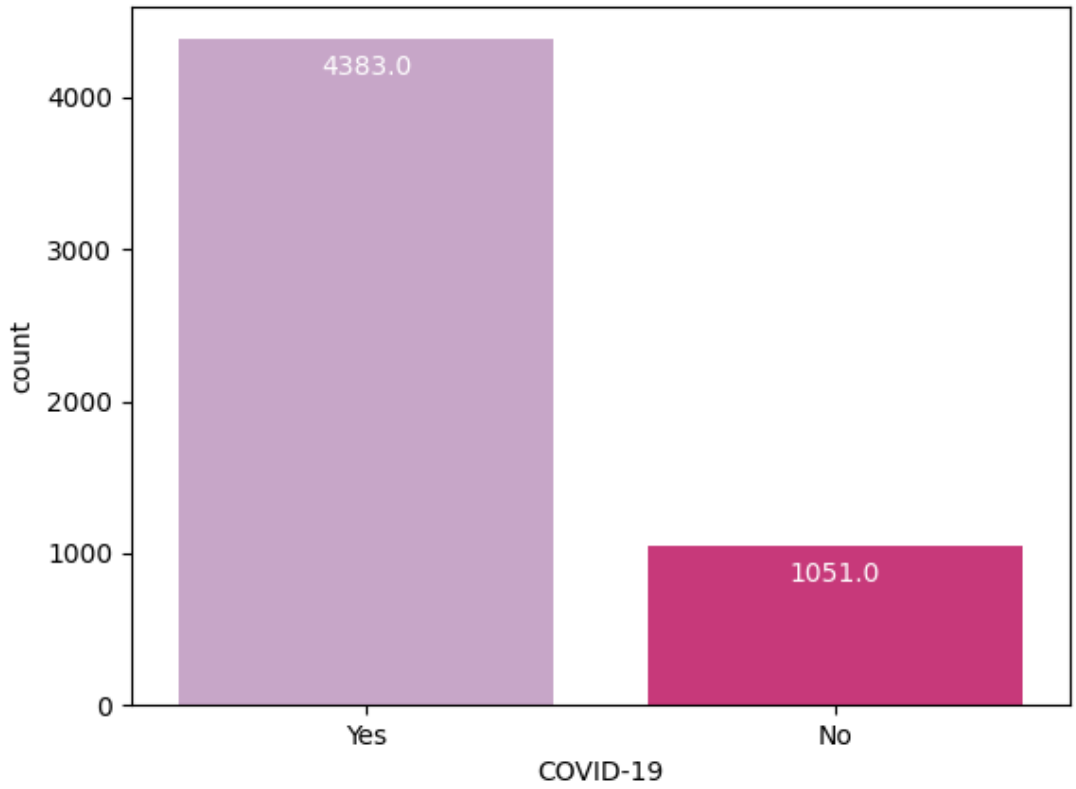

Figure 3.2: COVID-19 Columns Bar Chart.

However, The pie chart shows the column has 80.7% 'Yes' values, and 19.3% 'No' values.

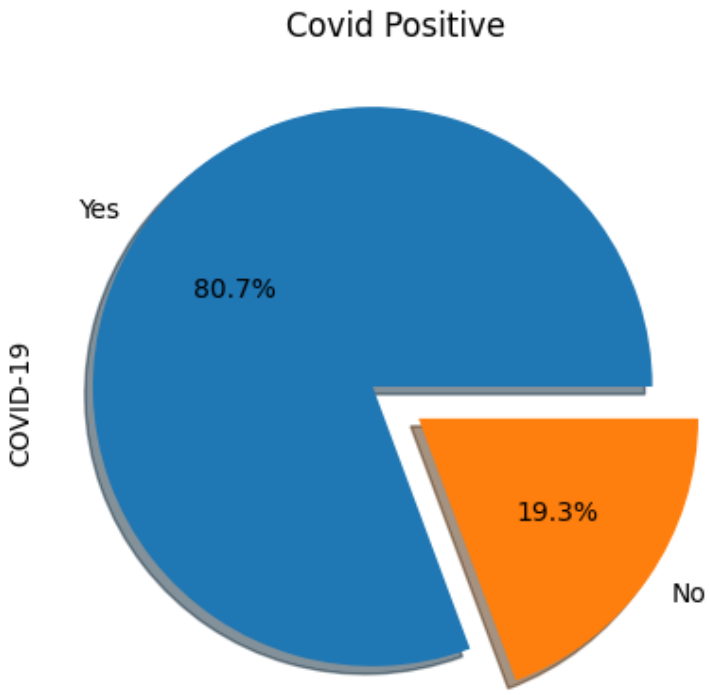

Figure 3.3: COVID-19 Columns pie Chart.

The figure 3.4 illustrates encoded histogram visual representations. However, the 3.4 represents the actual values of all attributes. The value 1 and 0 represented 'Yes', and 'No' respectively after encoding the categorical values to numerical.

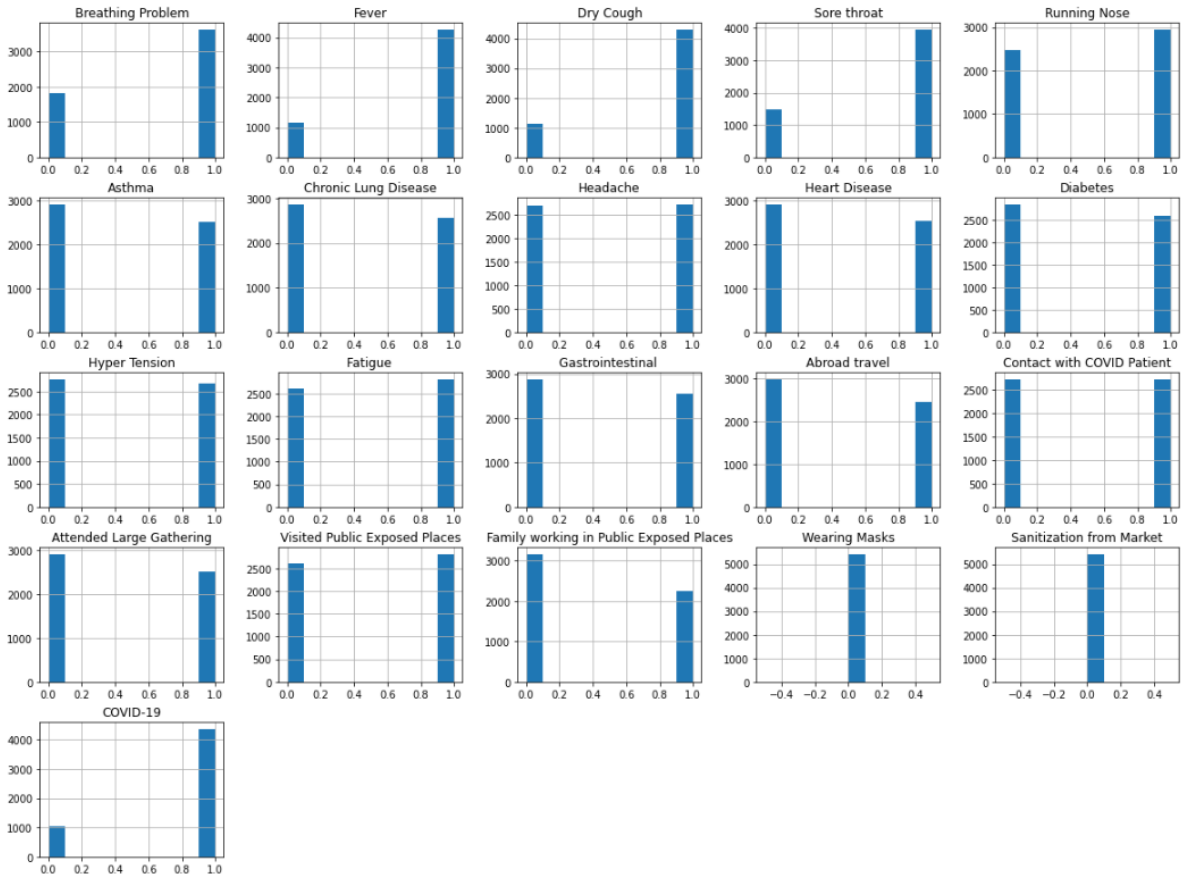

Figure 3.4: Histogram of all the attributes.

The correlation analysis provides intriguing observations about the connections between different attributes and the presence of COVID-19. From the figure 3.6 illustrates that breathing problems, fever, dry cough, and sore throat all exhibit significant positive correlations, meaning that people who have these symptoms are more likely to have COVID-19. Furthermore, Hypertension, foreign travel, and contact with COVID patients all demonstrate moderate positive correlations, indicating a significantly greater chance of COVID-19 in these situations. However, running nose, chronic lung disease, headache, and fatigue exhibit weak negative correlations, suggesting a slight tendency for these attributes to be less associated with COVID-19 from the figure 3.5. The correlation values are generally consistent with expectations, presenting important insights about potential relationships between symptoms and COVID-19 presence.

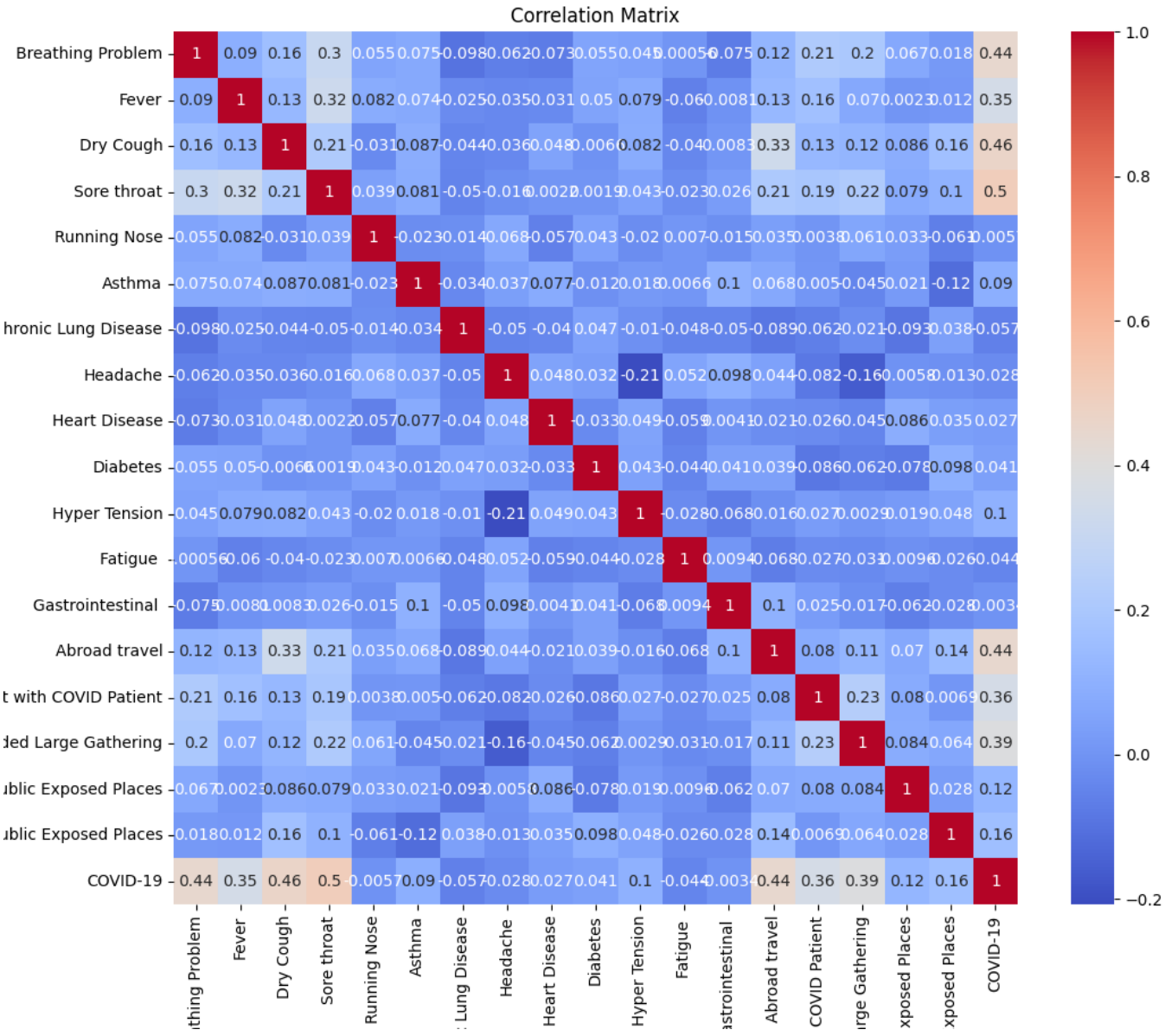

Figure 3.5: Correlation matrix heatmap with COVID-19.

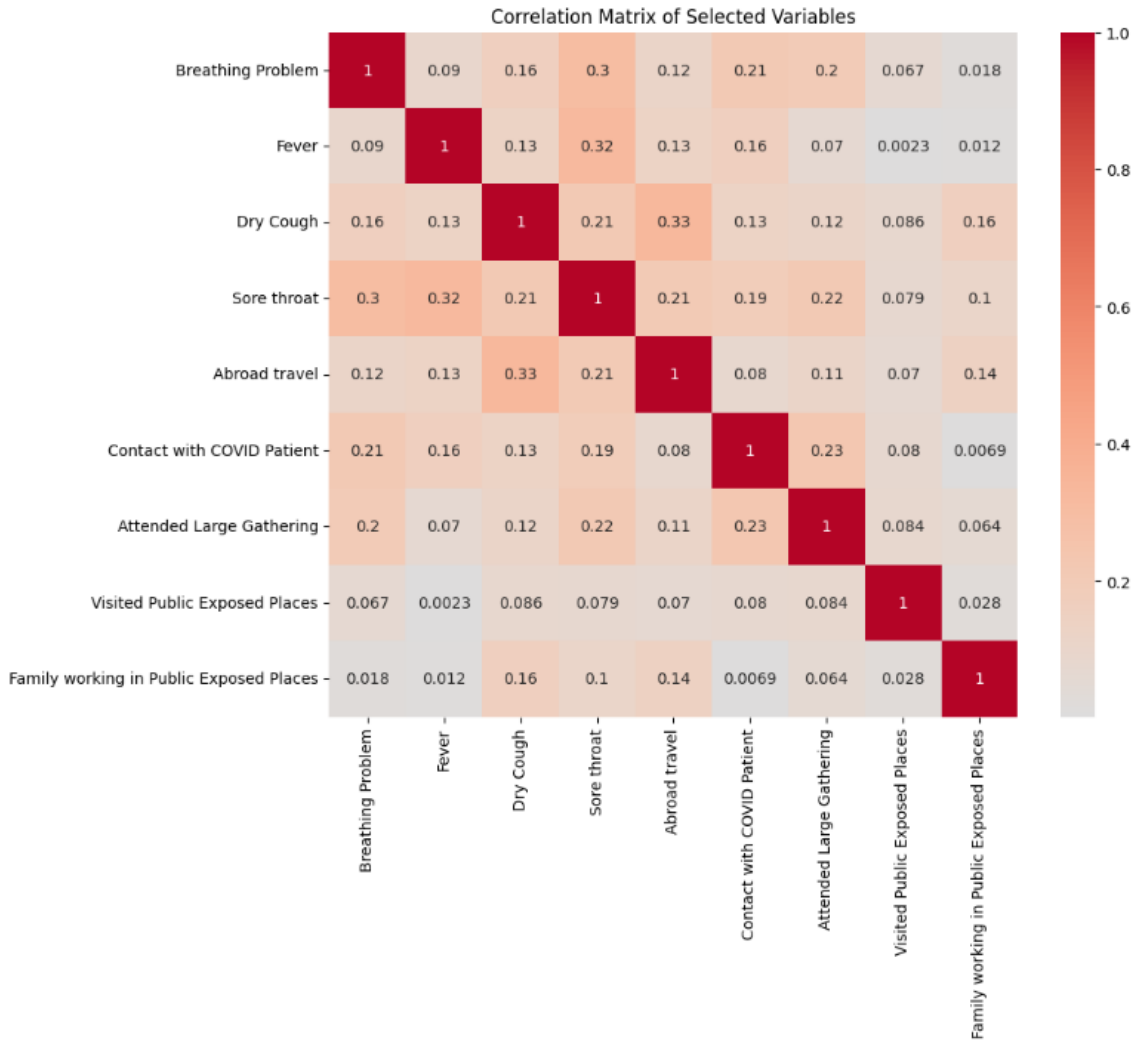

Figure 3.6: Selected variables correlation matrix.

3.3 Classification Metrics

The proposed Machine Learning approach will be evaluated using multiple performance metrics. Evaluating a ML system using numerous metrics can provide an extensive overview of its capabilities, this includes F_1 Score, accuracy, precision, recall, and Area Under Curve (AUC). This section has processed into details of these metrics.

3.3.1 Accuracy

The accuracy metric is the proportion of correct prediction (both true positives and true negatives) implemented by the model among the number of cases examined. The accuracy metrics are a very useful Machine Learning evaluation method and a well-

known binary classification test that correctly identifies or excludes a condition. The accuracy metric is crucial when classes are well-balanced in the dataset which means that each instance has a similar number of classes [63]. It is calculated using the below formula;

$$\text{Accuracy } = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.1}$$

where, TP (True positive) : Total Covid-19 patients precisely classified as Covid-19 patients; TN (True Negative): Total non-Covid-19 patients precisely classified as non-Covid-19; FP (False Positive): Total Covid-19 patients classified as non-Covid-19 patients; FN (False Negative): Total non-Covid-19 patients classified as Covid-19.

3.3.2 Sensitivity or Recall

Sensitivity, also known as recall or true positive outcome rate, is an essential evaluation metric significantly used in classification tasks. It determines mainly the proportion of exact positive instances (or true positives) that the model correctly computes [64].

The formula of sensitivity or recall is given by;

$$\text{Sensitivity or Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

3.3.3 Precisions

Precision is another evaluation matric computes the ratio of True positive predictions to the total positive predictions generated by the model; only positive instances. It is mainly the proportion of True positive outcomes and the total number of True and False positive outcomes. A low false positive rate means the model is making fewer mistakes in classifying instances that are negative as positive, leading to a higher number of true positive predictions [65]. The formula of Precision is given by;

$$\text{Precision } = \frac{TP}{TP + FP} \tag{3.3}$$

3.3.4 Specificity

Specificity, also known as True negative rate outcome, it computes the ratio of the proportion of True negative prediction (exact negative instances prediction) and among

all actual negative instances [64]. The formula of Specificity is given by;

$$\text{Precision } = \frac{TN}{TN + FP} \tag{3.4}$$

3.3.5 Area Under the ROC Curve (AUC-ROC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used evaluation metric in machine learning for classification problems. It has the ability to differentiate between positive and negative instances across various thresholds. An absolute model has an AUC-ROC of 1 which means that the model completely distinguishes between positive and negative outcomes. While 0.5 is the random guessing AUC-ROC score. The higher the rate chance to model fit and strongly discriminate outcome. The AUC-ROC curve is particularly useful in classification [66, 67].

3.3.6 ROC Curve

The receiver operating characteristics (ROC) graph is a strategy visualizing, organizing and choosing classifiers based on the model performance. ROC graphs have been used to depict True and False rates of classification tasks as well as have long been used to analyze behaviors of diagnostic systems. This has also been used extensively in decision making for diagnostic testing. In graphically, a perfect classifiers model would provide straight up along with Y-axis (True positive rate (TPR) =1), and progress horizon along with X-axis (False positive rate). However, a random classifier would provide an equal line (TPR=FPR) [68, 69].

3.3.7 Precision-Recall curve

The Precision-Recall (PR) curve assesses a model performance in text classification by calculating precision against recall at various probability thresholds. It is understanding the trading between measuring positive outcomes correctly and mitigating false positives. In the curve, precision values plotted on the Y-axis while recall values are plotted on the X-axis. The precision will change according to recall values, and the model performance can display over different thresholds [70]. A good PR curve shows high precision and high recall at various thresholds , and indicates the model can identify

the positive instances while minimising false instances [71].

3.3.8 Confusion Matrix

The Confusion Matrix is a table able to determine the effectiveness of a classification model in machine learning. It illustrates the model predictions compared to the actual class labels for a classification task. The matrix presents details concerning how accurate the model is and the types of errors it generates. However, if we are dealing with classification problem then the matrix is a 2×2 [72].

The matrix that provides the outcomes of true positive (TP), true negatives (TN) , false positive (FP) and false negatives (FN) as follows;

TP	FN
FP	TN

3.3.9 F_1 Score

The F_1 score is a common method to evaluate a ML classification model performance. It is a harmonic mean of precision and recall. Although it is not as simple to understand as accuracy, the F_1 Score is frequently more helpful than accuracy, especially if any dataset contains inconsistent class labels [69]. The formula of F_1 score as follows;

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3.5}$$

CHAPTER 4

Machine Learning Algorithms for Classification

Machine Learning algorithms for classification are used to classify the data points based on their attributes into predefined classes or labels. According to researchers, there are several classification ML algorithms available, we will discuss their theoretical aspects in this section;

4.1 Logistic Regression

Logistic Regression is a statistical and machine learning algorithm that is extensively used in classification problems where the predefined class or instance has only two possible outcomes. In logistic regression, if the response or dependent variables y_i can take one of two classes. The predictor variables x_i represent the independent variables. To construct the mathematical equation, we use a mathematical function called the logistic function or sigmoid function.The logistic function will take any real-world value and will transform it into 0 or 1. The transform value illustrates the particular category. The logistic regression model presumes a linear relationship between the independent variable and the log-odd response variable. The log-odds (also known as the logit) is the logarithm of the odds ratio [73]. The logistic regression model can be written as;

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \tag{4.1}$$

where:

$\text{logit}(p)$ is the log-odds of the response variable,

p is the probability of the event occurring,

β_0 is the intercept term,

$\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients associated with the predictor variables $X_1, X_2, \ldots, X_k$.

4.2 K Nearest Neighbors

K Nearest Neighbors (KNN) is a popular and simple machine learning algorithm used for classification and regression tasks. It is a non-parametric and instance-based learning method, meaning it makes predictions based on data points which is similar to each other. KNN is widely used in spam filtering, medical diagnosis, sentiment analysis, and forecasting stock market due to its simplicity and efficiency [74].

4.3 Random Forest Classifier

The Random forest is a powerful ensemble machine-learning algorithm used for classification and regression problems. The random forest classifier is executed by the synthesis of tree classifiers, each of which is constructed applying a random vector sampled independently from the input vector, and the final prediction is obtained through most votes, reducing overfitting and improving robustness [75].

The random forest classifier is employed to construct a mixture of features randomly selected at each node to expand a decision tree. Bagging is used to lead training datasets by randomly selecting instances from the original training set and replacing them with new examples. The forest's tree predictors classify instances by acquiring the majority vote from all trees. The Gini Index is used to assess the attributes impurity relating to classes by using selection metrics. As random forest trees are grown to maximum depth without pruning, a key advantage of using Random forest classifiers over decision tree classifiers. Researchers recommend that the preference of the pruning method, not the selection method, can change the tree-based classifier performance as the number of trees increases without pruning, the generalization error consistently approaches convergence because of the Strong Law of Large Numbers. All features are used at each

node and the number of trees in the forest are determined by the users. Classification is distributing each case to all trees, and forest selecting the class with the most votes out of all [76]. The figure 4.2 is taken from the Covid-19 prediction journals [82].

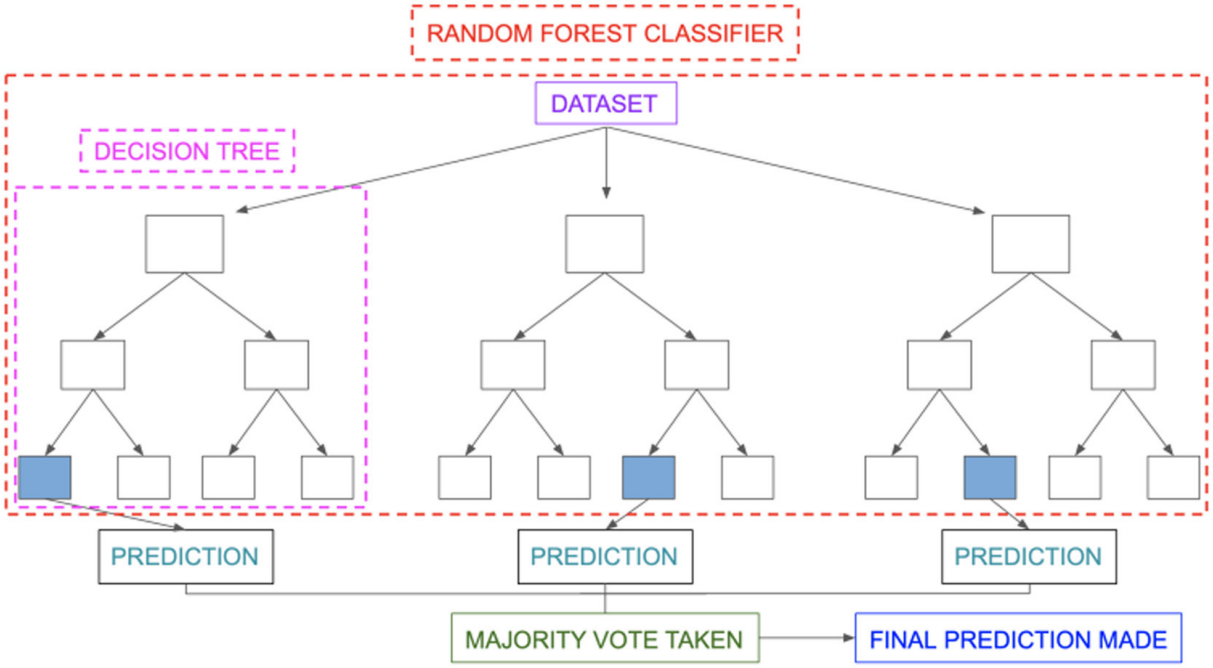

Figure 4.1: Random forest Classification structure.

4.4 Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm based on Bayes's theorem. In NB, Bayes's theorem provides the update of the probability based on the input features, and it assumes that all the features are conditionally independent given the instances or class labels. That means the presence or absence of one attribute does not affect the other feature's presence or absence. These assumptions elaborate the calculation and allow the algorithms to work well in high dimensional datasets where other algorithms may struggle [77].

The advantage of NB is that it provides faster training and prediction times. This approach takes only a modest quantity of training data to predict the classification parameters, and the time complexity NB is linear with respect to the number of instances and training. Moreover, due to its simplicity, NB tends to have a lower risk of overfitting because it can mitigate or avoid capturing the complexity of other features interaction

which may mitigate the chance of fitting noise in the data. However, because of its simplicity and can avoid complexity it is widely used in spam filtering, sentiment analysis, text classification, medical diagnosis, and market segmentation [78].

4.5 Support Vector Machine (SVM) classifiers

SVM is based on statistical learning theory and powerful supervised ML algorithms used for both classification and regression. The aim of the SVM is to separate classes determine the decision boundaries. In a two-class classification where instances are linearly separable, SVMs select the on classes leaving the greatest margin between two classes. SVMs create a hyperplane between two classes, there is a gap between the decision boundary (line) and hyperplane, and it is called margin [79].

The hyperplane is choosen to maximize the gap between the hyperplane and the nearest data points, also known as Support vectors. The SVM only depends on these support vectors (data points) but not other observations. The idea behind decision boundaries with large margins can tend to have lower generalization error whereas small margins can lead to overfitting. SVM is strongly effective in high dimensional space, making it suitable for complex datasets. The flexibility of the kernel trick allows SVM to handle non-linear datasets [79, 80]. The figure 4.2 is taken from the Covid-19 prediction journals [82].

4.6 XGBoost classifiers

XGBoost is an ensemble learning and optimized implementation of gradient descent features. The approach of the algorithm numerous trees grow and subsequently each tree aims to reduce the errors of the previous trees. Each tree learns from its predecessors, and effort to mitigate prediction errors. The decision tree anticipated the features and thresholds, particularly the best branch, and constructed the split method. The final outcome becomes consistent and reliable in regard to diagnosis [89, 90].

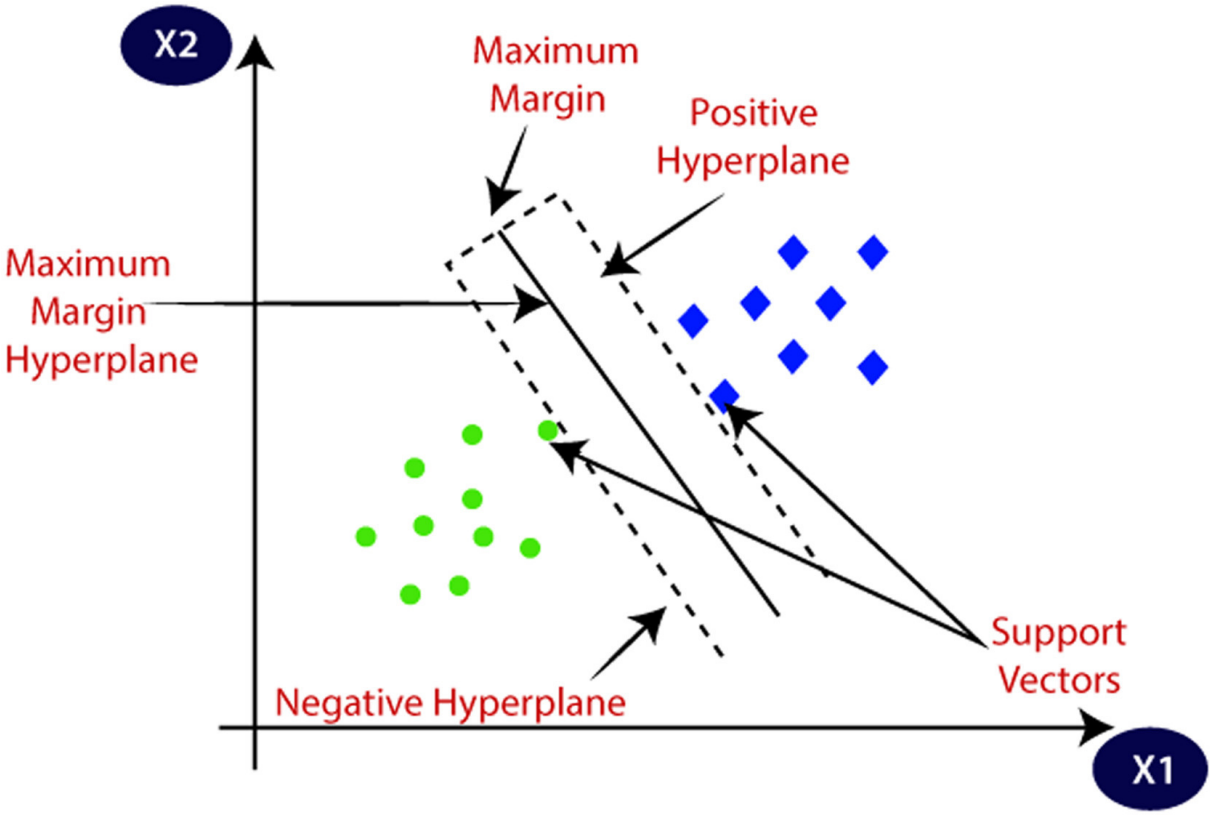

Figure 4.2: Visual Representation of SVM architecture.

CHAPTER 5

Result and Discussion

To assess the performance of the machine learning algorithms employed in this study, we chose to utilise accuracy, precision, recall, specificity, and F_1 measure. These metrics are commonly employed in the evaluation of machine learning classification tasks.

Table 5.1: Model Performance Metrics.

Model	Recall	Precision	Accuracy	F1-Score	Specificity
Logistic Regression	0.990950	0.975501	0.972401	0.983165	0.891626
Random Forest	0.989819	0.983146	0.977921	0.986471	0.926108
SVM	0.986425	0.983089	0.975161	0.984754	0.926108
KNN	0.975113	0.995381	0.976081	0.985143	0.980296
Naïve Bayes	0.725113	1.000000	0.776449	0.840656	1.000000
XGBoost	0.989819	0.983146	0.977921	0.986471	0.926108

Table 5.1 illustrates that among the implemented models, all of them display impressive results across different measures. The logistic regression model shows excellent recall, precision, and accuracy, along with a well-balanced F_1 score of roughly 0.983. However, its specificity score of approximately 0.891 suggests limitations in its ability to correctly identify negative instances. The Random forest and XGBoost models consistently yield similar recall, precision, accuracy, F_1 score, and specificity values which are 0.989, 0.983, 0.977, 0.986, 0.986 respectively, indicating their ability to detect both positive and negative cases. The SVM classifiers perform similarly with regard

to random forest and XGBoost classifiers. On the other hand, the KNN model recognizes itself by achieving perfect precision 0.995, showcasing its exceptional ability in detecting positive instances. In addition, it achieves better specificity which is 0.980 than previously mentioned classifiers. However, it has 0.976 accuracy score and 0.975 recall score respectively. The Naïve Bayes model struggles with recall, resulting in a suboptimal balance between precision and recall despite achieving perfect precision and specificity.

5.1 Execution Time Comparison

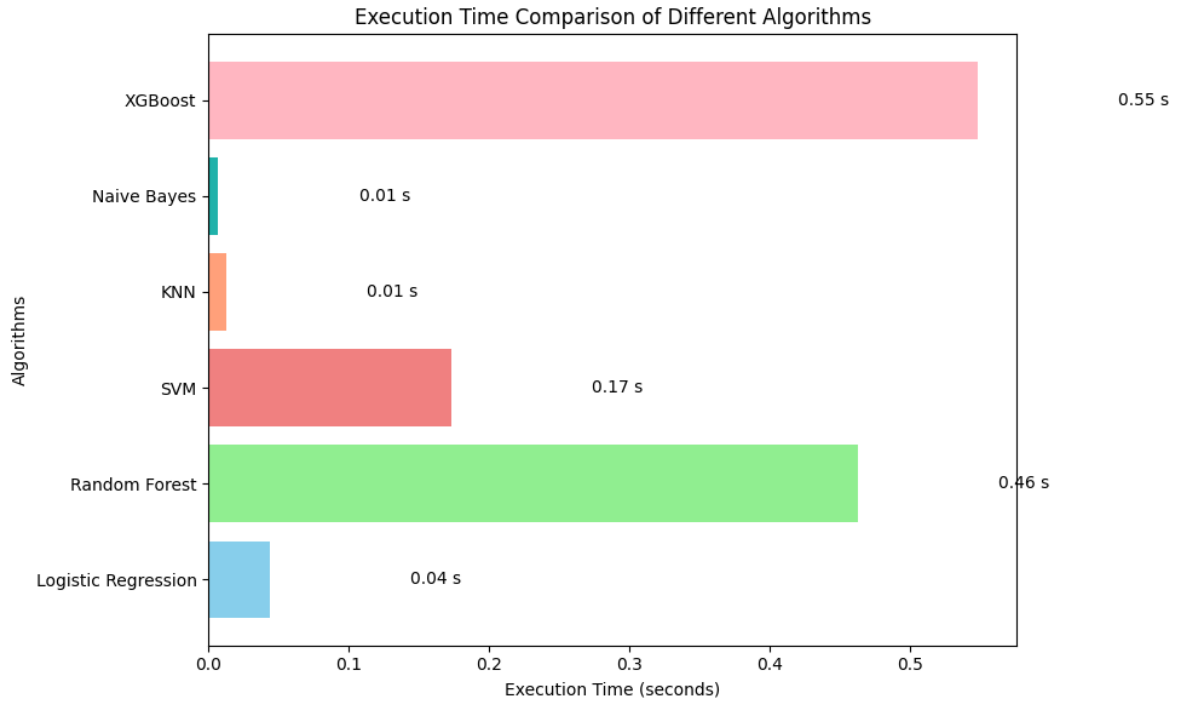

Figure 5.1: Execution time comparison of different algorithms.

The figure 5.1 highlights each algorithm's execution time. Logistic Regression execution time 0.04s, makes it a favourable choice for classifier problems involving small datasets but has problems maintaining accuracy. Furthermore, Random Forest and SVM classifiers perform very well but take more execution times 0.46s and 0.17s respectively. Although both algorithms can maintain good performance. Moreover, KNN and Naive Bayes algorithms are taking only 0.01s completion time despite Naive Bayes maintaining poor performance. However, the completion times of KNN could diminish with larger

datasets. Finally, XGBoost classifiers perform well in regards to performance despite its taking longer completion times. The choice of algorithms of XGBoost classifiers is still effective, probably large datasets but has to consider resources.

5.2 ROC AUC score comparison

From the table 5.2, the ROC AUC scores for the different models indicate their strong ability to classify. The logistic regression model identifies classes with the accuracy of 0.995, distinguishing between the positive and negative classes. The random forest and XGBoost both have an exceptionally high ROC AUC score of around 0.998. This suggests that the model is extremely accurate in classifying while SVM proves its strength in classifying instances and dividing classes with a score of 0.992. At 0.994, Naïve Bayes performs well, demonstrating the usefulness of simpler probabilistic models. Meanwhile, KNN also performs well in ROC AUC scores, stating that it can also be well suited for classification tasks with a score of 0.996. The below ROC curve figure 5.2 shows that all the models perform significantly well in this dataset.

Table 5.2: ROC AUC Scores for different models.

Model	ROC AUC Score
Logistic Regression	0.9950
Random Forest	0.9982
SVM	0.9920
KNN	0.9967
Naive Bayes	0.9942
XGBoost	0.9980

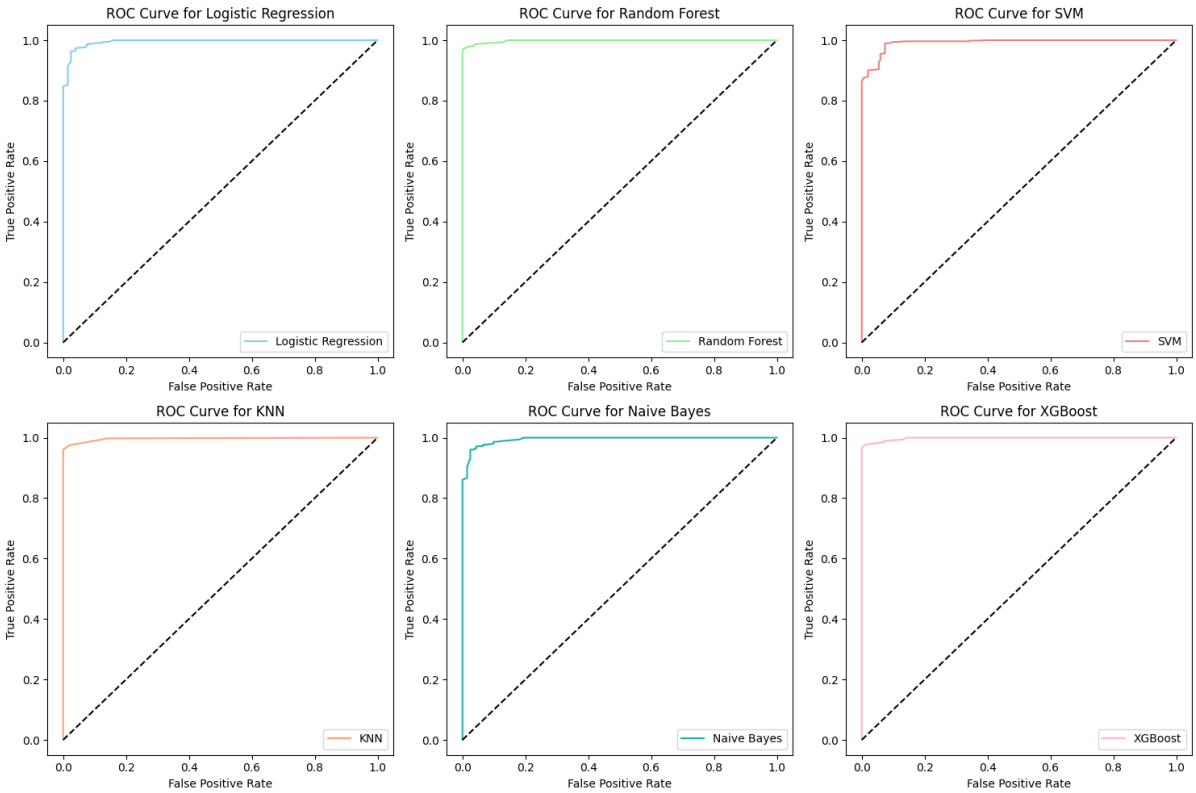

Figure 5.2: ROC Curve for all models.

5.3 Precision-Recall Curve

From the figure 5.3, we can demonstrate that logistic regression low precision point is somewhere between 0.98 to 0.96 while the recall value is approximately 0.8. It is suggested that logistic regression can identify the Covid-19 individuals. In the random forest, XGBoost models slightly fall from 0.994 precision to 0.98 close recall value, meaning that both models can identify the symptom-based Covid-19 dataset and forecast effectively. The SVM models precision falls close to 0.82 recall value. On the other hand, Naïve Bayes precision value falls close to 0.83 recall threshold. However, KNN models outperform all other ML methods but the precision value falls to 0.975 recall value. The model is extremely effective in this dataset based on the graph.

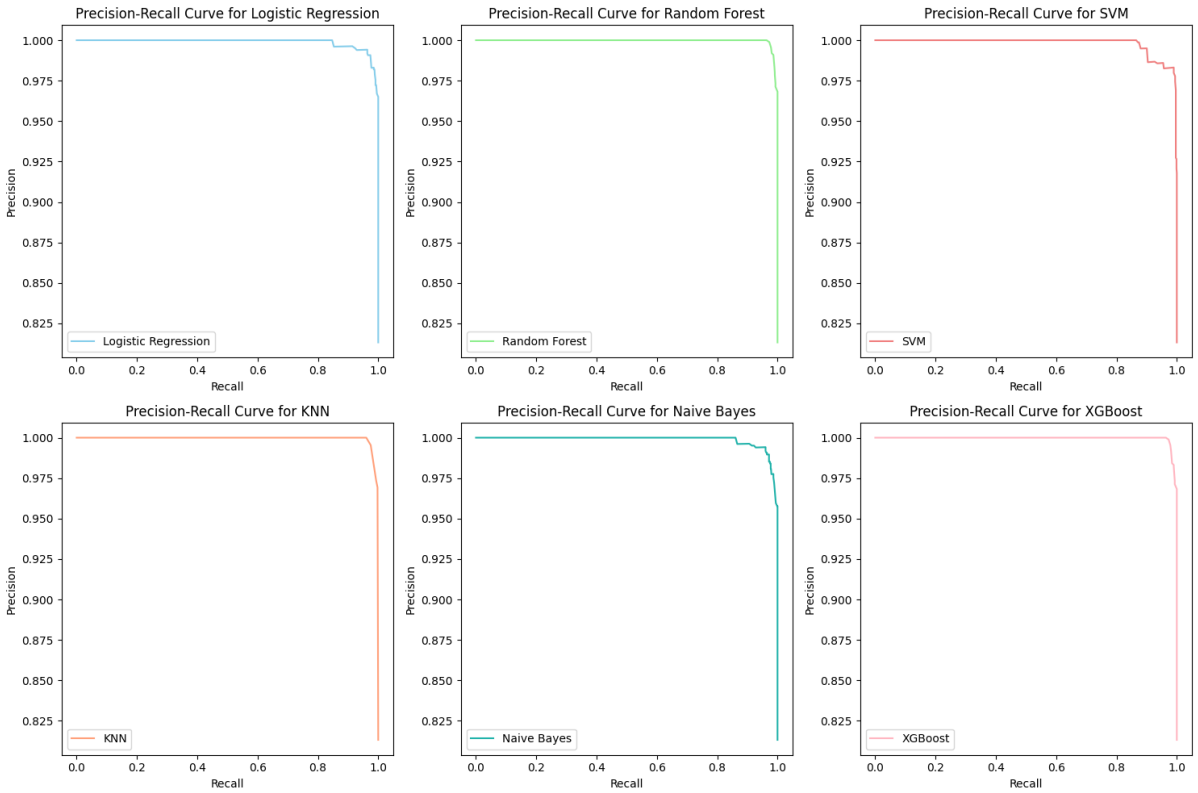

Figure 5.3: Precision Recall Curve for all models.

5.4 Analysis our model using SHAPLEY value and Permutation features

SHAP (SHapley additive exPlanations) values are used in ML for measuring attribute contribution during final prediction. SHAP values are extremely helpful in understanding feature selection and valuable for model debugging, and transparency [44]. The permutation feature is a technique that can correct model selection because it can provide column contribution based on the prediction. It can detect attribute bias, and help to make decision-accurate ML models based on the dataset [87, 88].

We analysis Logistic regression, SVM, Random forest, and XGBoost using SHAPLEY values. And we applied permutation features to analyse KNN and Naïve Bayes that how well our attribute worked during final prediction.

5.4.1 Logistic Regression

We used SHAP values to interpret the behaviour of the logistic regression. From the figure 5.4, we can observe that public gatherings and overseas traveling accounted more than any attribute. However, symptoms like sore throat, breathing issues, cough, fever, and contact with COVID patients also contributed final outcome of logistic regression models.

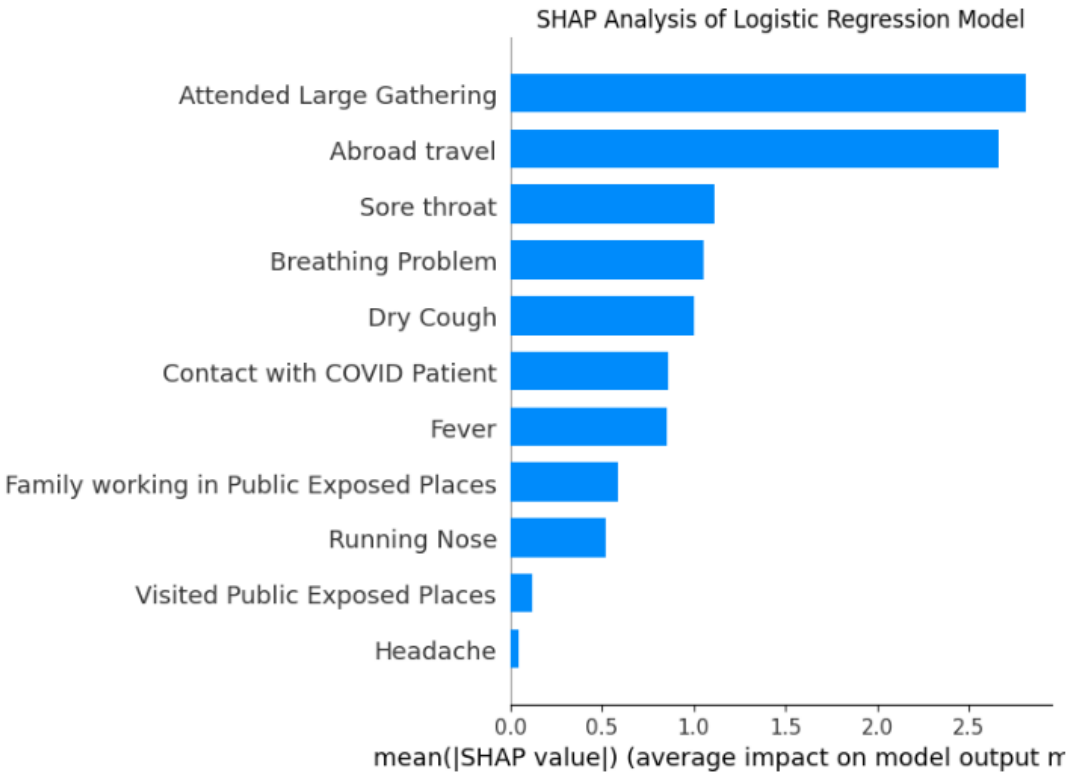

Figure 5.4: Attributes analysis of logistic regression performance.

5.4.2 Random Forest and Support Vector Classifiers

From figure 5.5 and 5.6, during the final prediction of random forest and SMV, dry cough, public gathering, sore throat, and contact with infected people are considered mostly. Family working and fever significantly participate in the final model forecast. Individuals who visited exposed places and had headaches contributed less than 0.2. Running nose and family working in infected areas also contributed 0.5.

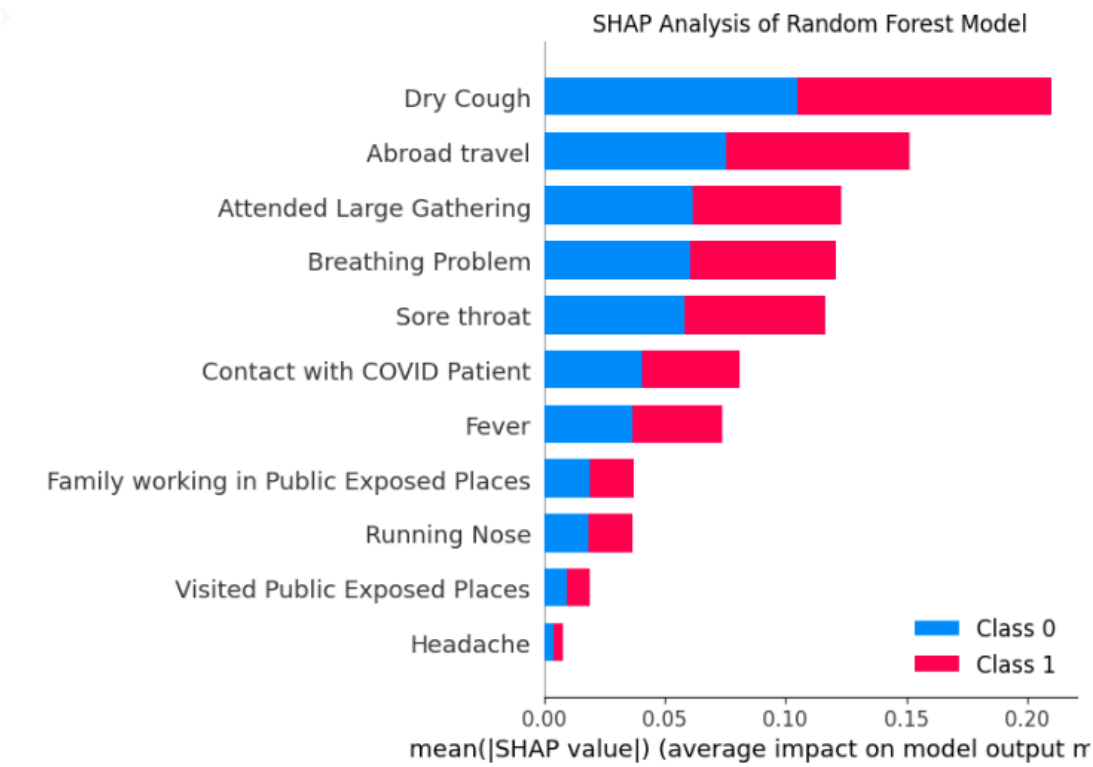

Figure 5.5: Attributes analysis of random forest performance.

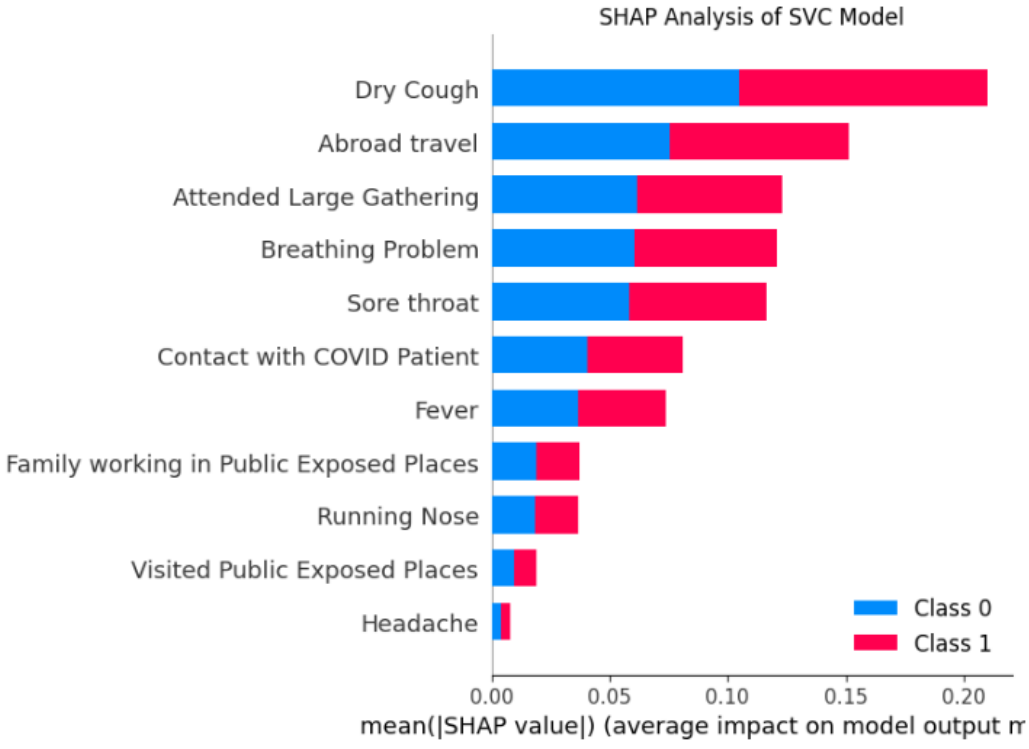

Figure 5.6: Attributes analysis of SVM performance.

5.4.3 XGBoost Classifiers

Traveling around the globe and gatherings are significantly promoted to predict in this model based on the figure 5.7. As mentioned previously three models dry cough, sore throat, breathing, and contact with virus virus-affected person have been pressed during prediction.

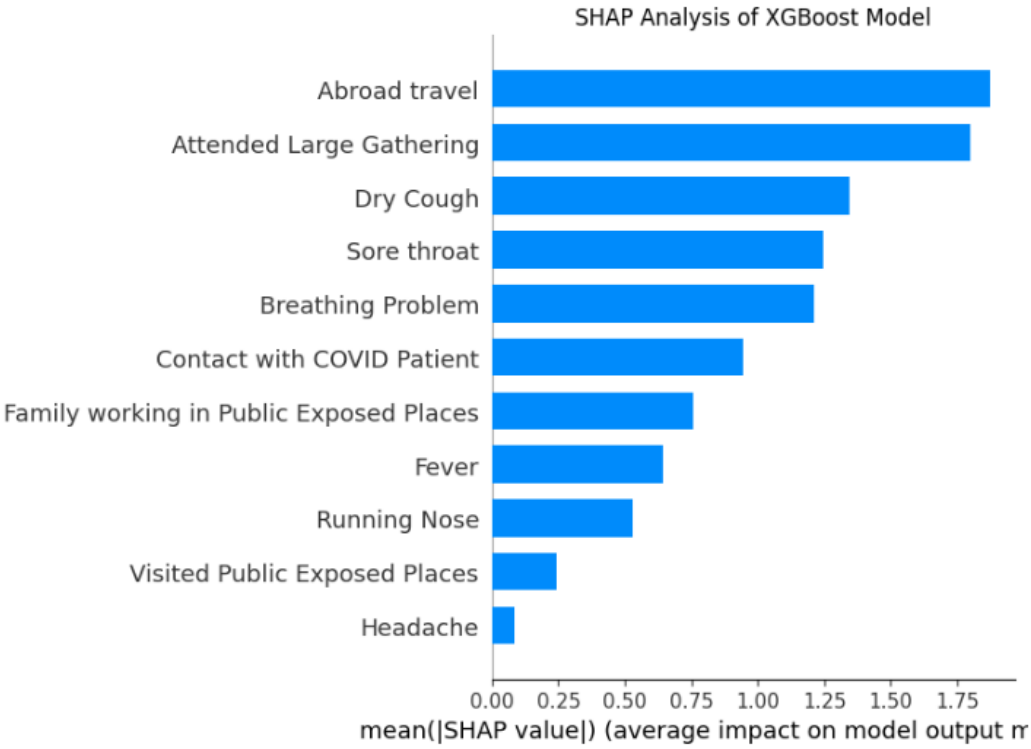

Figure 5.7: Attributes analysis of XGBoost performance.

5.4.4 KNN Classifiers and Naïve Bayes

In both models, we applied permutation performance to analysis attribute contribution. From the figure 5.8, we can state that common symptoms such as sore throat, dry cough, breathing, travel, and gathering boosted the model outcome. However, with regard to the figure it is clearly observed that each attribute contributed significantly. For this reason, we achieved remarkably good results in each evaluation metric. However, Naïve Bayes ML methods did not consider other attributes except abroad travel and public gatherings. For the instances, Naïve Bayes did not predict recall value correctly, and underperformed compared to other models we can see in the figure 5.9.

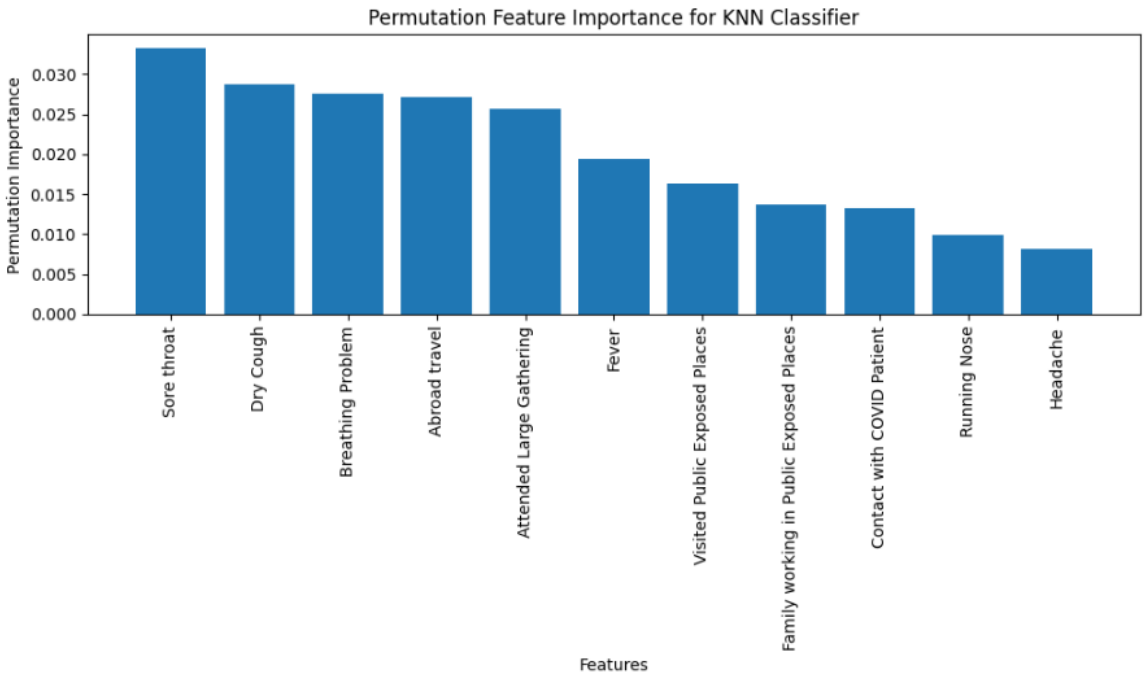

Figure 5.8: Attributes analysis of KNN performance.

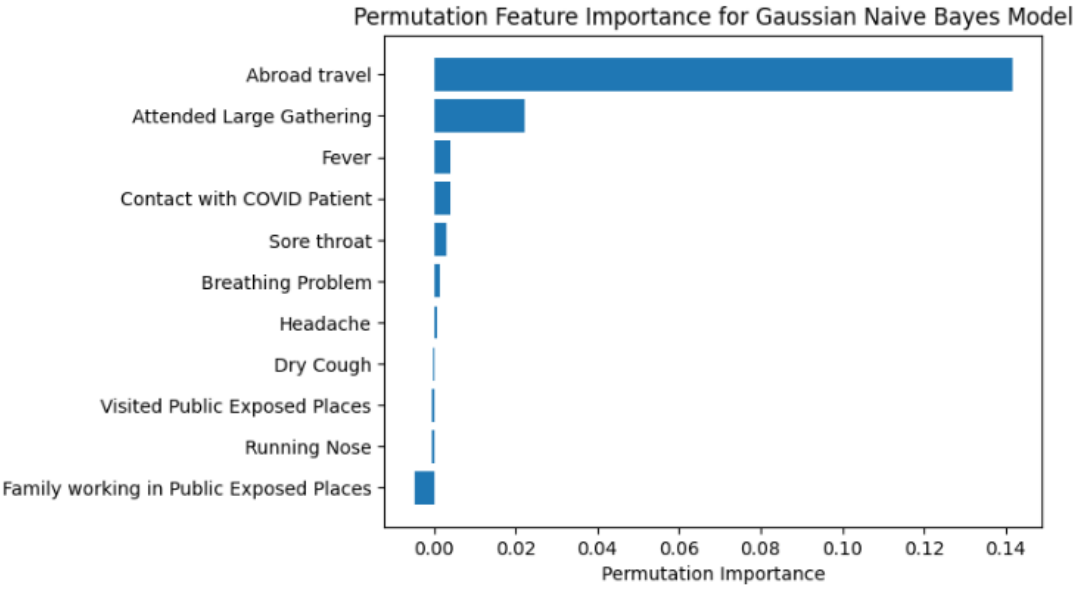

Figure 5.9: Attributes analysis of NB performance.

5.4.5 Confusion Matrix

To conclude the figure 5.10, we can observe that the logistic regression model correctly predicted 876 cases as Covid-19 positive, and 181 correctly predicted as true negative

Covid-19. It also predicted false positives 15 cases and false negatives 8 cases. The random forest and Xgboost classifiers predicted the same true positive and negative cases respectively 875 and 188 as well as predicted false positive and negative respectively 15 and 9 Covid-19 cases. On the other hand, Naïve Bayes predicted 0 false positive Covid-19 cases even though predicted true negative than any models but predicted mostly false negative Covid-19 cases. Meanwhile, SVM classifier predicted 872 true positive cases and 188 true negative cases. In addition, KNN predicted 862 positive Covid-19 cases and 199 true negative cases as well as predicted false positive and negative respectively 4 and 22 cases.

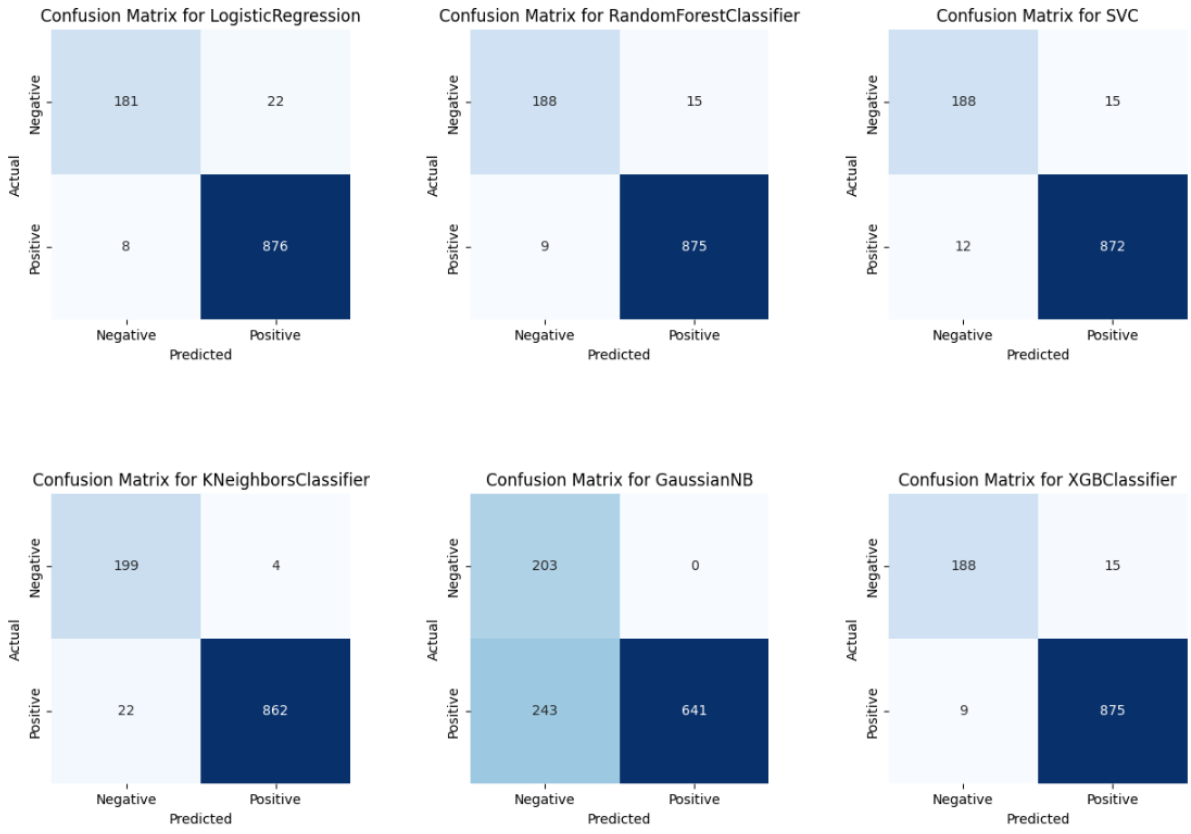

Figure 5.10: Confusion matrix graph for all models.

5.5 Strengths and Limitations

The Covid-19 prediction is comprehensive and well-balanced. The dataset has no missing values, and as every column is categorical we had to convert it into numerical format to apply exploratory data analysis as well as ML models. To get accurate predictions, we applied machine learning model and obtained the possible best result through all

the attributes, and we also applied the SHAPLEY value and permutation feature to understand the attribute contribution during the final prediction. The comprehensive assessment which we analysed using different metrics including confusion matrices, classification reports, ROC curves, and precision-recall curves, shows a clear picture of how well each model is effective.

The dataset is imbalanced, it can lead to overfitting or underfitting the model performance. Some models perform extremely well based on symptoms-based medical diagnosis but models such as XGBoost or deep learning are less interpretable as well as need more resources and time.

5.6 Future Work

In pursuit of improving imbalance data regarding COVID-19 prediction, we can consider resampling methods to handle imbalance datasets like cost-sensitive learning, Smote oversampling, and removing attributes not related to the target attribute. Furthermore, exploring sophisticated preprocessing techniques like feature scaling, and dimensionality reduction can improve the imbalance dataset. Using the hyperparameter tuning gridSearchCV and validation test can help to make more appropriate decisions based on the models [83, 84]. In addition, exploring ensemble techniques like bagging, boosting, and stacking could improve accuracy and generalization [85, 86]. However, developing a web application for users can input symptoms, and can access real-time health risk predictions, supported by continuous monitoring to keep models updated with changing data trends.

CHAPTER 6

Conclusions

The main motive of this work was to predict Covid-19 presence using machine learning classifier methods based on symptoms including likelihood metrics like geographical location, travel history, health records etc. Based on the symptoms provided by the W.H.O, CDC, NHS, and our dataset attribute relates to symptoms we concluded that our all machine learning classifiers perform very well, and provide insightful information regarding the COVID-19 virus. The model performance was assessed in a comparative analysis. With due respect to the other machine learning method, the KNN machine learning classifiers perform well considering the fact of precision, recall, specificity value, and the score respectively 0.995, 0.975, 0.980, having accuracy score 0.976 even considering runtime for training for this dataset as well as permutation analysis have shown that primary symptoms columns contributed mostly in the final prediction.

This research can be determined as an effective tool for medical professionals. With the implemented models, individuals can take advantage of assisting their health support from the doctors or hospitals. Also, they can decide whether they can transmit viruses to society. Initially, medical practitioners can assess the test as the primary tool to detect Covid-19. Businesses can mitigate interaction physically with clients who are at risk of COVID-19 infection.

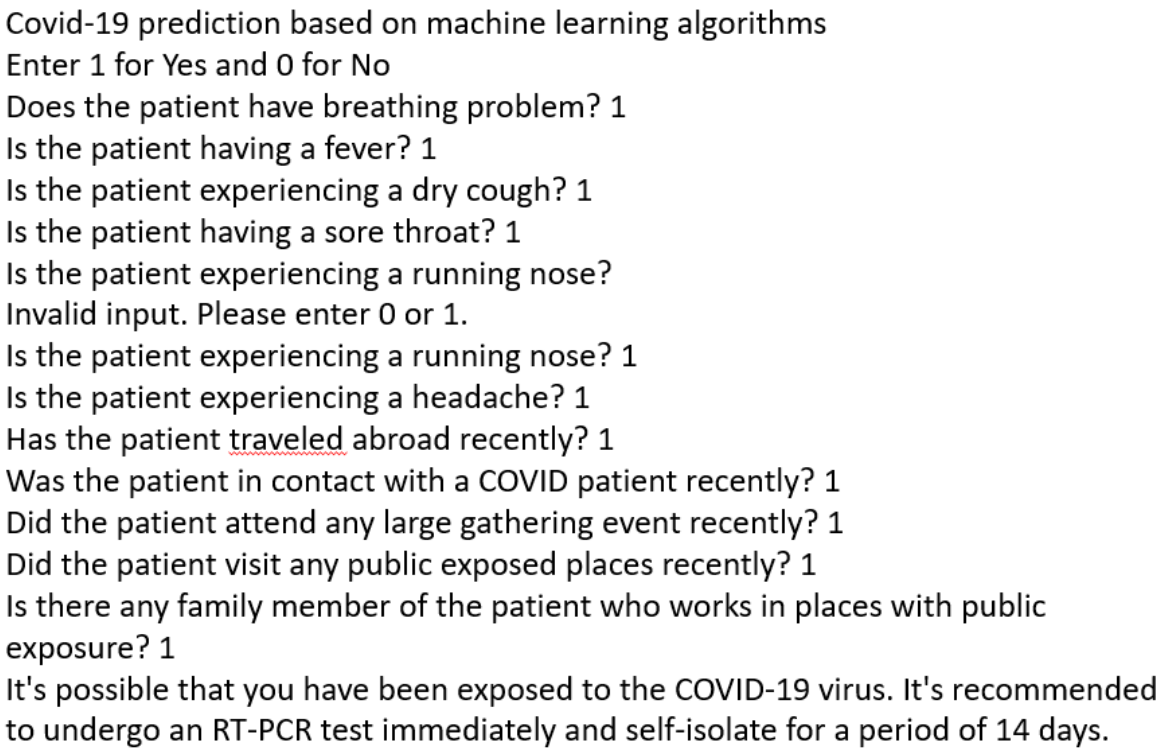

Figure 6.1: Prediction model takes input and gives the result of COVID Positive.

The figure 6.1, shows that based on the symptoms provided by the W.H.O, CDC, NHS, the machine learning model can predict Covid-19 positive. It also can provide awareness to the people regarding Covid19 virus as well as figure 6.2 shows a prediction of COVID-19 negative.

Covid-19 prediction based on machine learning algorithms
Enter 1 for Yes and 0 for No
Does the patient have breathing problem? 0
Is the patient having a fever? 0
Is the patient experiencing a dry cough? 0
Is the patient having a sore throat? 0
Is the patient experiencing a running nose? 0
Is the patient experiencing a headache? 1
Has the patient traveled abroad recently? 0
Was the patient in contact with a COVID patient recently? 1
Did the patient attend any large gathering event recently? 0
Did the patient visit any public exposed places recently? 1
Is there any family member of the patient who works in places with public exposure? 0
You are not showing any signs of COVID-19. It's advisable to remain at home and prioritize your safety.

Figure 6.2: Prediction model takes input and gives the result of COVID Negative.

Bibliography

[1] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang,Ben Hu, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*,2020,1-5.

[2] Heng Li, Shang-Ming Liu, Xiao-Hua Yu, Shi-Lin Tang, Chao-Ke Tang , Coronavirus disease 2019 (COVID-19): current status and future perspectives. *Elsevier*,2020,1-3.

[3] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang , A new coronavirus associated with human respiratory disease in China. *Nature*, 2020, 1-3.

[4] Derrick Bryson Taylor , A Timeline of the Coronavirus Pandemic. *The New York Times*, 2021.

[5] Mahesh Jayaweera, Hasini Perera, Buddhika Gunawardana, Jagath Manatunge, Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environmental research*, 2020, vol 188.

[6] National Health Service, https://www.nhs.uk/conditions/covid-19

[7] Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/index.html

[8] The World Health Organization, https://www.who.int/health-topics/coronavirus.

[9] Shajeea Arshad Ali, Mariam Baloch, Naseem Ahmed, Asadullah Arshad Ali, Ayman Iqbal. The outbreak of Coronavirus Disease 2019 (COVID-19) An emerging global health threat. *ScienceDirect; Journal of Infection and Public Health*, 2020, Volume 13, Issue 4, p 644-646.

[10] Alan D. Kaye, Chikezie N. Okeagu, Alex D. Pham, Rayce A. Silva, Joshua J. Hurley, Brett L. Arron, Noeen Sarfraz, Hong N. Lee, G.E. Ghali, Jack W. Gamble, Henry Liu, Richard D. Urman, Elyse M. Cornett, Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives *ScienceDirect*, 2021, Volume 35, Issue 3, p 293-306.

[11] Dominic A. Fitzgerald, Gary W.K. Wong, COVID-19: A tale of two pandemics across the Asia Pacific region, Paediatric Respiratory Reviews, *ScienceDirect*, 2020, Volume 35, Pages 75-80.

[12] Shabir Ahmad Lone, Aijaz Ahmad. COVID-19 pandemic-an African perspective, *PMID*, 2020, doi: 10.1080/22221751.2020.1775132. PMID: 32458760; PMCID: PMC7473237.

[13] Johannes Korth, Benjamin Wilde, Sebastian Dolff, Olympia E. Anastasiou, Adalbert Krawczyk, Michael Jahn, Sebastian Cordes, Birgit Ross, Stefan Esser, Monika Lindemann, Andreas Kribben, Ulf Dittmer, Oliver Witzke, Anke Herrmann, SARS-CoV-2-specific antibody detection in healthcare workers in Germany with direct contact to COVID-19 patients, *Journal of Clinical Virology*, 2020, Volume 128, https://www.sciencedirect.com/science/article/pii/S1386653220301797.

[14] Aristodemou, K., Buchhass, L. and Claringbould, D. The COVID-19 crisis in the EU: the resilience of healthcare systems, government responses and their socio-economic effects, *Eurasian Economic Review*, 2021, pp.251-281.

[15] CBO's current projections of output, employment, and interest rates and a preliminary look at federal deficits for 2020 and 2021 | Congressional budget office. 2020. https://www.cbo.gov/publication/56335.

[16] Gan Y, Ma J, Wu J. Immediate and delayed psychological effects of province-wide lockdown and personal quarantine during the COVID-19 outbreak in China, *Psychol Med*, 2020, https://doi.org/10.1017/S0033291720003116.

[17] Yuan Z, Xiao Y, Dai Z. Modelling the effects of wuhan's lockdown during covid-19, China. *Bull World Health Organ*, 2020, https://doi.org/10.2471/BLT.20.254045.

[18] Varghese GM, John R. COVID-19 in India: Moving from containment to mitigation. *Indian J Med Res*, 2020.

[19] Kaur S, Sonali S. India fights COVID-19. *Psychol trauma theory*, 2020, doi:10.1037/tra0000615.

[20] Rakshit B, Basishtha D. Can India stay immune enough to combat COVID-19 pandemic? An economic query. *Public Aff*, 2020. https://doi.org/10.1002/pa.2157.

[21] Melo, Cristiane & Souza-Silva, Guilherme & Melo, Alanne Rayssa & Freitas, Antonio. COVID-19 pandemic outbreak: the Brazilian reality from the first case to the collapse of health services.. *Pub med*, 2020. doi: 10.1590/0001-3765202020200709.

[22] IMF blog. COVID-19: without help, low-income developing countries risk a lost decade e IMF blog. 2020.https://www.imf.org/en/Blogs/Articles/2020/08/27/blog-covid-19-without-help-low-income-developing-countries-risk-a-lost-decade.

[23] Bong CL, Brasher C, Chikumba E, McDougall R, Mellin-Olsen J, Enright A. The COVID-19 Pandemic: Effects on Low- and Middle-Income Countries, Anesth Analg. *Pub Med*, 2020. doi: 10.1213/ANE.0000000000004846.

[24] Pearman A, Hughes ML, Smith EL, Neupert SD. Mental health challenges of United States healthcare professionals during COVID-19. *Frontiers in Psychology*, 2020,11:2065. doi: 10.3389/fpsyg.2020.02065. PMID: 32903586; PMCID: PMC7438566.

[25] Lupu, Dan,Tiganasu, Ramona. COVID-19 and the efficiency of health systems in Europe. *Health Economics Review*, 2022, 1-5.

[26] Ermal Bojdani, Aishwarya Rajagopalan, Anderson Chen, Priya Gearin, William Olcott, Vikram Shankar, Alesia Cloutier, Haley Solomon, Nida Z. Naqvi, Nicolas Batty, Fe Erlita D. Festin, Dil Tahera, Grace Chang, Lynn E. DeLisi, COVID-19 pandemic: impact on psychiatric care in the United States. *Psychiatry research*, 2020, 1-3.

[27] Yuki, K., Fujiogi, M. and Koutsogiannaki, S. COVID-19 pathophysiology: A review. *Clinical immunology*, 2020, 1-3.

[28] Quer, G., Radin, J.M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E., Kheterpal, V., Topol, E.J. and Steinhubl, S.R. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature*, 202, p.73-77.

[29] Emery SL, Erdman DD, Bowen MD, Newton BR, Winchell JM, Meyer RF, Tong S, Cook BT, Holloway BP, McCaustland KA, Rota PA, Bankamp B, Lowe LE, Ksiazek TG, Bellini WJ, Anderson LJ., Real-time reverse transcription-polymerase chain reaction assay for SARS-associated coronavirus. *Emerg Infect Dis: Pub Med*, 2004, doi: 10.3201/eid1002.030759.

[30] Stang, A., Robers, J., Schonert, B., JÃ¶ckel, K.H., Spelsberg, A., Keil, U. and Cullen, P. The performance of the SARS-CoV-2 RT-PCR test as a tool for detecting SARS-CoV-2 infection in the population. *Journal of Infection*, 2021,83(2), pp.237-279.

[31] Afzal, A. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of advanced research*, 2020,26, pp.149-159.

[32] Mak, G.C., Cheng, P.K., Lau, S.S., Wong, K.K., Lau, C.S., Lam, E.T., Chan, R.C. and Tsang, D.N., Evaluation of rapid antigen test for detection of SARS-CoV-2 virus. *Journal of Clinical Virology*, 2020, 129, p.104500.

[33] Li, Z., Yi, Y., Luo, X., Xiong, N., Liu, Y., Li, S., Sun, R., Wang, Y., Hu, B., Chen, W. and Zhang, Y., Evaluation of rapid antigen test for detection of SARS-CoV-2 virus. *Journal of medical virology*, 2020, 92(9), pp.1518-1524.

[34] Frank, R.G., Dach, L. and Lurie, N., It was the government that produced COVID-19 vaccine success. *Health Affairs Forefront*, 2021.

[35] Le, T.T., Andreadakis, Z., Kumar, A., RomÃ¡n, R.G., Tollefsen, S., Saville, M. and Mayhew, S., The COVID-19 vaccine development landscape. *Nat Rev Drug Discov*, 2020, 19(5), pp.305-306.

[36] Wouters, O.J., Shadlen, K.C., Salcher-Konrad, M., Pollard, A.J., Larson, H.J., Teerawattananon, Y. and Jit, M., Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment. *The Lancet*, 2021, 397(10278), pp.1023-1034.

[37] Meo, S.A., Bukhari, I.A., Akram, J., Meo, A.S. and Klonoff, D.C., COVID-19 vaccines: comparison of biological, pharmacological characteristics and adverse effects of Pfizer/BioNTech and Moderna Vaccines. *Eur Rev Med Pharmacol Sci*, 2021, pp.1663-1669.

[38] Self, Wesley H and Tenforde, Mark W and Rhoads, Jillian P and Gaglani, Manjusha and Ginde, Adit A and Douin, David J and Olson, Samantha M and Talbot, H Keipp and Casey, Jonathan D and Mohr, Nicholas M and others, Comparative effectiveness of Moderna, Pfizer-BioNTech, and Janssen (Johnson & Johnson) vaccines in preventing COVID-19 hospitalizations among adults without immunocompromising conditions United States. *Centers for Disease Control and Prevention*, 2021, 70(38), pp.1337.

[39] Le, T.T., Andreadakis, Z., Kumar, A., RomÃ¡n, R.G., Tollefsen, S., Saville, M. and Mayhew, S., The COVID-19 vaccine development landscape. *Nat Rev Drug Discov*, 2020, 19(5), pp.305-306.

[40] Jeyanathan, M., Afkhami, S., Smaill, F., Miller, M.S., Lichty, B.D. and Xing, Z., Immunological considerations for COVID-19 vaccine strategies. *Nature Reviews Immunology*, 2020, 20(10), pp.615-632.

[41] Burgos, R.M., Badowski, M.E., Drwiega, E., Ghassemi, S., Griffith, N., Herald, F., Johnson, M., Smith, R.O. and Michienzi, S.M., The race to a COVID-19 vaccine: Opportunities and challenges in development and distribution. *Drugs in context*, 2021, 10.

[42] Kashte, S., Gulbake, A., El-Amin III, S.F. and Gupta, A., COVID-19 vaccines: rapid development, implications, challenges and future prospects. *Springer*, 2021,34(3), pp.711-733.

[43] Loembé, Marguerite Massinga and Nkengasong, John N., COVID-19 vaccine access in Africa: Global distribution, vaccine platforms, and challenges ahead. *Elsevier*, 2021,1353–1362.

[44] M. N. S. Choudary, V. B. Bommineni, G. Tarun, G. P. Reddy and G. Gopakumar., Predicting Covid-19 Positive Cases and Analysis on the Relevance of Features

using SHAP (SHapley Additive exPlanation). *ieee*, 2021, pp. 1892-1896, doi: 10.1109/ICESC51422.2021.9532829.

[45] André Filipe de Moraes Batista and João Luiz Miraglia and Thiago Henrique Rizzi Donato and Alexandre Dias Porto Chiavegatto Filho., COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *Researchgate*, 2020, doi: 10.1101/2020.04.04.20052092.

[46] M. Rohini, K. R. Naveena, G. Jothipriya, S. Kameshwaran and M. Jagadeeswari, A Comparative Approach To Predict Corona Virus Using Machine Learning. *ieee*, 2021, pp. 331-337, doi: 10.1109/ICAIS50930.2021.9395827.

[47] Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al., Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *Springer*, 2021, https://doi.org/10.1007/s42979-020-00394-7.

[48] Dilip Kumar Sharma, Muthukumar Subramanian, Pacha. Malyadri, Bojja Suryanarayana Reddy, Mukta Sharma, Madiha Tahreem, Classification of COVID-19 by using supervised optimized machine learning technique. *ScienceDirect*, Volume 56, Part 4,2022,Pages 2058-2062,ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.11.388.

[49] Goodman-Meza, David AND Rudas, Akos AND Chiang, Jeffrey N. AND Adamson, Paul C. AND Ebinger, Joseph AND Sun et al., A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PloS one*, 2020, 15. e0239474. 10.1371/journal.pone.0239474.

[50] Sun, Yinxiaohe and Koh, Vanessa and Marimuthu, Kalisvar and Ng, Oon Tek and Young et al., Epidemiological and Clinical Predictors of COVID-19, Clinical Infectious Diseases, Volume 71, Issue 15, 2020, pp.786-792, https://doi.org/10.1093/cid/ciaa322.

[51] Tiwari, D., Bhati, B. S., Al-Turjman, F., & Nagpal, B., Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques.Expert Systems, 39(3), e12714. https://doi.org/10.1111/exsy.12714.

[52] Nishant Rai, Naman Kaushik, Deepika Kumar, Chandan Raj, Ahad Ali., Mortality prediction of COVID-19 patients using soft voting classifier, International Journal of Cognitive Computing in Engineering,Volume 3,2022,Pages 172-179,ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2022.09.001.

[53] V. R. J and A. Jakka., Forecasting COVID-19 cases in India Using Machine Learning Models," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 466-471, doi: 10.1109/ICSTCEE49637.2020.9276852.

[54] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, U. Rajendra Acharya., Automated detection of COVID-19 cases using deep neural networks with X-ray images, Computers in Biology and Medicine, Volume 121, 2020, 103792, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2020.103792.

[55] Ezz El-Din Hemdan and Marwa A. Shouman and Mohamed Esmail Karar., COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images. *arXiv*, 2020, https://doi.org/10.48550/arXiv.2003.11055.

[56] Wang, L., Lin, Z.Q. & Wong, A., A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 10, 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z.

[57] Wang, L., Lin, Z.Q. & Wong, A., A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 10, 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z.

[58] Apostolopoulos, I.D., Mpesiana, T.A., Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 43,(2020). https://doi.org/10.1007/s13246-020-00865-4.

[59] Sethy, P.K.; Behera, S.K., Detection of Coronavirus Disease (COVID-19) Based on Deep Features. Preprints 2020, 2020030300. https://doi.org/10.20944/preprints202003.0300.v1.

[60] Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Wang, R., Zhao, H., Chong, Y. and Shen, J., 2021. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM transactions on computational biology and bioinformatics, 18(6), pp.2775-2780.

[61] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X. and Xu, B., 2021. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). European radiology, 31, pp.6096-6104.

[62] Zheng, C., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W. and Wang, X., 2020. Deep learning-based detection for COVID-19 from chest CT using weak label. MedRxiv, pp.2020-03.

[63] Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006, December. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Berlin, Heidelberg: Springer Berlin Heidelberg.

[64] Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994 Jun 11;308(6943):1552. doi: 10.1136/bmj.308.6943.1552. PMID: 8019315; PMCID: PMC2540489.

[65] Bzdok, D. and Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3(3), pp.223-230.

[66] Brzezinski, D. and Stefanowski, J., 2017. Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. Knowledge and Information Systems, 52, pp.531-562.

[67] Narkhede, S., 2018. Understanding auc-roc curve. Towards Data Science, 26(1), pp.220-227.

[68] Tom Fawcett An introduction to ROC analysis,Pattern Recognition Letters,Volume 27, Issue 8,2006,Pages 861-874, ISSN 0167-8655,https://doi.org/10.1016/j.patrec.2005.10.010.(https://www.sciencedirect.com/science/ar

[69] Chicco, D. and Jurman, G., 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16(1), pp.1-23.

[70] Ozenne, B. and Subtil, F. and Maucort-Boulch, D.,2015 The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases.Journal of Clinical Epidemiology, 68, pp.1545-1569.

[71] Miao, J. and Zhu, W.,2022 Precision-recall curve (PRC) classification trees.Evolutionary Intelligence, 15, pp.855-859.

[72] Caelen, O. A Bayesian interpretation of the confusion matrix,Annals of Mathematics and Artificial Intelligence,pp.429-450,2017,https://doi.org/10.1007/s10472-017-9564-8.

[73] Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression,(Vol.398). John Wiley & Sons.

[74] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," Proceedings of International Conference on Neural Networks (ICNN'96), Washington, DC, USA, 1996, pp. 1480-1483 vol.3, doi: 10.1109/ICNN.1996.549118.

[75] Breiman, L. 1996. Bagging predictors,Machine Learning,26, pp.123-140.

[76] Pal, M., 2005. Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), pp.217-222.

[77] Webb, G.I., Keogh, E. and Miikkulainen, R., 2010. NaÃ¯ve Bayes. Encyclopedia of machine learning, 15(1), pp.713-714.

[78] M. Abd-Elnaby, M. Alfonse, and M. Roushdy., Classification of breast cancer using microarray gene expression data: A survey,Journal of biomedical informatics, vol. 117, p. 103764, 2021.

[79] Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20, pp.273-297.

[80] Keerthi, S.S. and Gilbert, E.G., 2002. Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning, 46, pp.351-360.

[81] Weerts, H.J., Mueller, A.C. and Vanschoren, J., 2020. Importance of tuning hyperparameters of machine learning algorithms. arXiv preprint arXiv:2007.07588.

[82] Sanzida Solayman, Sk. Azmiara Aumi, Chand Sultana Mery, Muktadir Mubassir, Riasat Khan, Automatic COVID-19 prediction using explainable machine learning techniques, International Journal of Cognitive Computing in Engineering, Volume 4, 2023, Pages 36-46, ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2023.01.003.

[83] Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C.O., Chuah, J.H. and Gheni, H.M., 2022. An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods. ieee access, 10, pp.79974-79985.

[84] Alhakeem, Z.M., Jebur, Y.M., Henedy, S.N., Imran, H., Bernardo, L.F. and Hussein, H.M., 2022. Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques. Materials, 15(21), p.7432.

[85] Wen, L. and Hughes, M., 2020. Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. Remote Sensing, 12(10), p.1683.

[86] Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.W., Han, Z. and Pham, B.T., 2020. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. Landslides, 17, pp.641-658.

[87] Ojala, M. and Garriga, G.C., 2010. Permutation tests for studying classifier performance. Journal of machine learning research, 11(6).

[88] Altmann, A.,Toloşi, L., Sander, O. and Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), pp.1340-1347.

[89] Chunjiao Dong, Yixian Qiao, Chunheng Shang, Xiwen Liao, Xiaoning Yuan, Qin Cheng, Yuxuan Li, Jianan Zhang, Yunfeng Wang, Yahong Chen, Qinggang Ge, Yurong Bao, Non-contact screening system based for COVID-19 on XGBoost

and logistic regression, Computers in Biology and Medicine, Volume 141, 2022, 105003, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.105003.

[90] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu and J. Peng, "XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 2018, pp. 251-256, doi: 10.1109/BigComp.2018.00044.