# *"SARS-CoV-2 Mortality Risk Prediction Using Machine Learning Technique"*

A PROJECT REPORT

SUBMITTED IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF A DEGREE

OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted By

**Manish Sewariya**

**(2K20/SWE/13)**

Under the supervision of

**Mr. Sanjay Patidar**

**(Assistant Professor)**

**DEPARTMENT OF SOFTWARE ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

**May 2022**

**DEPARTMENT OF SOFTWARE ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

## CANDIDATE'S DECLARATION

I, **Manish Sewariya, 2K20/SWE/13** student of M. Tech (Software Engineering), hereby declare that the project entitled "**SARS-CoV-2 Mortality Risk Prediction Using Machine Learning Technique**" been submitted by me to the Department of Software & Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi. I have done my project in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering and it has not been previously formed the basis for any fulfilment of the requirement in any degree or other similar title or recognition.

This report is an authentic record of my work carried out during my degree under the guidance of **Mr. Sanjay Patidar**.

31/5/22

Place: Delhi

Manish Sewariya

Date: 31st May 2022

(2K20/SWE/13)

**DEPARTMENT OF SOFTWARE ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

## CERTIFICATE

I hereby certify that the project entitled "**SARS-CoV-2 Mortality Risk Prediction Using Machine Learning Technique**" which is submitted by **Manish Sewariya (2K20/SWE/13)** to the Department of Software & Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or diploma to this university or elsewhere.

31/5/2022

Place: Delhi

Date: 31st May 2022

**Mr. Sanjay Patidar**

**(Assistant Professor)**

SUPERVISOR

Department of Software & Engineering

# ACKNOWLEDGEMENT

I am very thankful to **Mr. Sanjay Patidar** (Assistant Professor, Department of Software Engineering) and all the faculty members of the Department of Software Engineering at DTU. They all provided us with immense support and guidance for the project. I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions. I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

Manish
31/5/22

**Manish Sewariya**

**(2K20/SWE/13)**

# ABSTRACT

Coronavirus referred to as COVID-19 has had adverse effects in every possible aspect such as loss of economy, infrastructure, and moreover human life. In the era of growing technology: Artificial intelligence and machine learning can help find a way in reducing mortality, and in the same regard, we have created a prediction model for mortality of in-hospital COVID-19 patients. We used the dataset of 146 countries which consists of laboratory samples of around 2,670,000 confirmed COVID-19 cases. This study presents a Machine Learning model which will assist hospitals and medical facilities in determining who requires immediate attention and who must be given priority for hospitalization when the system is overburdened, or the facility is filled with patients who are not that severe and eliminate any delays in providing needed care to extremely severe patients first. As a result, the overall accuracy of the mortality rate prediction demonstrated is 91.26%. We evaluated different machine learning algorithms namely decision tree (DT), support vector machine (SVM), random forest (RF), logistic regression (LR), and k-nearest neighbor (K-NN) for mortality risk prediction COVID-19 affected patients admitted in hospitals. This proposed research study sheds light upon the identification of most relevant features and concerning symptoms. To perform an indepth examination and assess the results of classifiers, we used different performance measures on the developed model.

# CONTENTS

## List of Tables

## List of Figures

## List of Symbols, Abbreviations

| ABBREVIATIONS | FULL FORM |
|---|---|
| **ML** | Machine Learning |
| **AI** | Artificial Intelligence |
| **ROC CURVE** | Receiver Operating Characteristic Curve |
| **AUC** | Area Under ROC Curve |
| **IAV** | Human adaptable Type A influenza viruses |
| **SARS-COV-2** | Severe Acute Respiratory Syndrome Coronavirus 2 |
| **WHO** | World Health Organization |
| **PHEIC** | Public Health Emergency of International Concern |
| **RF** | Random Forest |
| **SVM** | Support Vector Machine |
| **DT** | Decision Tree |
| **LR** | Logistic Regression |
| **KNN** | K-Nearest Neighbor |
| **CSV** | Comma-separated values file |
| **PCA** | Principal component analysis |
| **TP** | True Positive |
| **FP** | False Positive |
| **FN** | False Negative |
| **TN** | True Negative |
| **TPR** | True Positive Rate |
| **FPR** | False Positive Rate |
| **RBF** | Radial Basis Function |

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

A pandemic refers to an outbreak spreading over international borders, creating social and economic disruptions as well as an increase in worldwide death and morbidity rates. On a worldwide scale, the Spanish flu pandemic, for example, was the leading cause of nearly 100,000 deaths[1]. Seasonal, avian, influenza in animals and pandemic influenza are only a few of the varieties of influenza that have created issues across the world. In the past, there were no reports of the pandemic influenza virus infecting humans. Human adaptable Type A influenza viruses (IAVs) may quickly transmit among individuals, producing transmittable illnesses and posturing a worldwide pandemic hazard. Deep learning and Machine learning technologies have proven to be beneficial in a variety of sectors, including data analysis. In virology, machine learning techniques are a good contender for smoothing the process of identifying viral sequences, particularly IAVs[2]. COVID-19 is related to the ancient influenza viruses in terms of illness presentation. Both of them, in the first instance, produce respiratory disease, which can lead to death. Second, viruses are transmitted by direct touch, droplets, and infected vehicles[3].

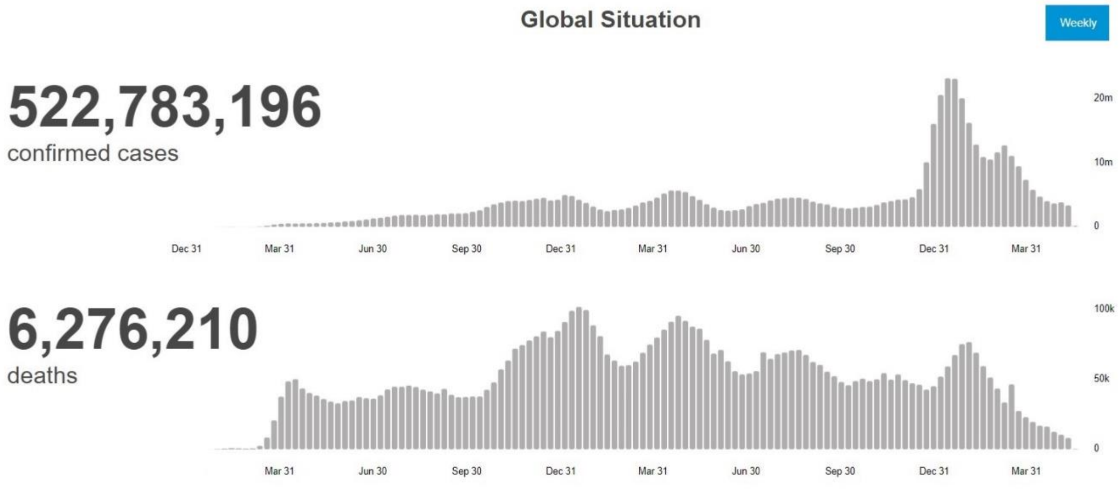


**Figure 1.1.** COVID-19 worldwide confirmed and fatality cases[7].

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), regarded as Coronavirus or COVID-19, began spreading in the Chinese province of Hubei in late 2019 and took many human lives[4]. The WHO classified the emerging coronavirus epidemic as an International Emergency Medical Concern in January 2020.[5]. The communicable disease caused by the above-mentioned SARS-CoV-2 virus was named as COVID-19 (Coronavirus Disease 2019) by WHO in February 2020, and it was declared as a pandemic in March 2020.[6]. As of May 23, 2022, the total number of Coronavirus confirmed cases is stated to be 522,783,196, with a total death toll of 6,276,210 (WHO reports), as shown in figure 1.1[7].

Some coronaviruses may infect animals, and coronaviruses have been seen to move from animal to human populations in rare instances. The new coronavirus may have transferred from an animal species to humans and then started spreading. As a result of the rise in positive cases, when a hospital's capacity is overloaded, the fatality rate rises[8]. To address the existing challenges in effective mortality risk predictions, timely clinical decision-making, and prevention of further fatality rates, machine learning is a tool that could be of great aid.

## 1.2 Aim of the thesis

The intent of the current research work is development of a COVID-19 fatality risk predictor model based on machine learning algorithms using standard clinical data of patients. We want answers to the below-mentioned questions:

- What are the most important determinants of admitted COVID-19 patients' mortality at high risk in hospitals?
- What is the most effective machine learning algorithm for creating a mortality prediction model?

## 1.3 Motivation

COVID-19's rapid growth has caused a significant scarcity of medical resources as well as tiredness among frontline healthcare staff. Furthermore, many COVID-19 patients' symptoms worsen quickly following a period of relatively modest symptoms, emphasizing the need for more refined risk classification models. Machine learning prediction models can help recognize individuals at an increased danger of fatality and give treatment to help them live longer. As a result, in order to relieve the strain on the existing health care system/setting and offer top-notch treatment possible to patients, it is vital to anticipate illness and properly treat censoriously sick patients.

From the start, we were skewed towards working on developing a model for COVID-19 mortality rate prediction to address the existing challenges. This study was exciting because it will help future researchers forecast the fatality rate of coronavirus patients and the treatment decisions that must be made. The key contributions of this study, will include:

- Provide insight into Spatio-temporal COVID mortality dynamics.
- Focus on COVID-19 mortality rate as a crucial variable for better decision-making.
- Add knowledge to the current literature and create a baseline for future research work.

## 1.4 Machine Learning

As a sub-classification under the domain of Artificial Intelligence (AI), Machine Learning (ML) allows for the extraction of high-quality prediction models from massive raw information[9]. Relation between AI, ML and DL can be seen in figure 1.2. It's a useful technique that's increasingly being used in medical research to enhance predictive modeling and uncover novel elements that influence a certain goal outcome[9, 10,11]. By providing evidence-based medication for risk evaluation, screening, prediction, and treatment planning, Machine learning algorithms can minimize uncertainty and ambiguity. They promote trustworthy medical management/decision-making and expect to enhance patients' quality of care and results [12, 13]**.**

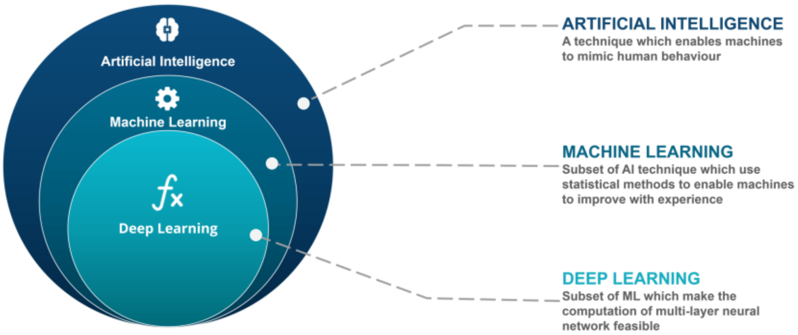

Artificial Intelligence
Machine Learning
$f_x$
Deep Learning
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

**Figure 1.2**. AI and Machine Learning [10].

## 1.4.1 Type of Machine Learning techniques

Generally, ML Techniques are diversified in three different types mentioned below also is represented in figure 1.3[10].

### 1.4.1.1. Supervised Learning

The design of algorithms that exhibit the capability of yielding broad patterns and imaginative concepts using external contexts for estimation of the future of future conditions is known as machine-readable learning. The goal of monitored machine classification algorithms is to classify data based on prior knowledge. In data science challenges, segregation has been widely used. Other supervised learning samples include random forest, KNN, decision tree, log retrospective, and more.

### 1.4.1.2. Unsupervised Learning

The capacity to read & edit information without delivering an error signal to test the proposed response is called unregulated reading. In unsupervised learning, the absence of a learning algorithm guidance can sometimes be helpful, as it allows the system to research patterns that have not been tested before.

### 1.4.1.3. Reinforcement Learning:

It is a strategy that allows an agent to learn in a collaborative environment by trying and incorrectly using feedback from his or her actions and emotions.

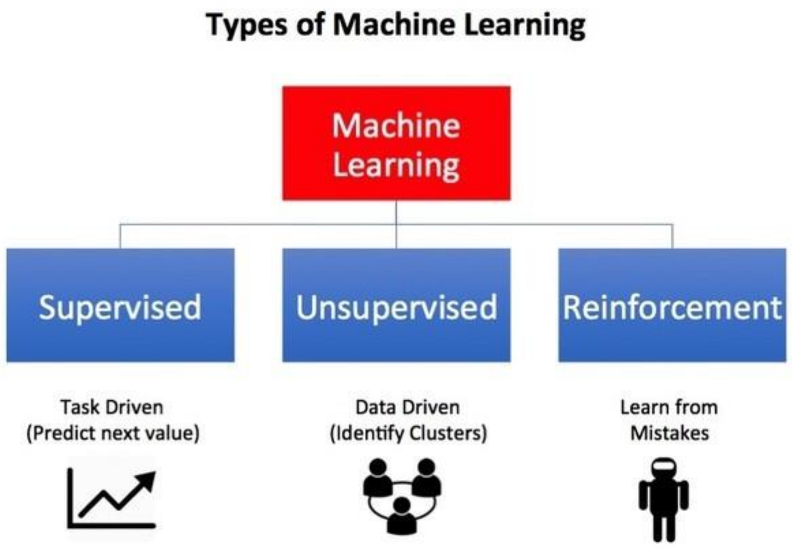


**Figure 1.3.** Type of Machine learning techniques[10].

### 1.4.2 Machine learning techniques used

We evaluated a few classified machine learning algorithms, and the respective accuracies were found on the dataset namely Decision Tree (DT) (Accuracy - 89.10%), Support Vector Machine (SVM) (Accuracy - 89.87%), random forest (RF) (Accuracy - 91.26% %), Logistic Regression (LR) (Accuracy - 89.24%), and K-Nearest Neighbor (K-NN) (Accuracy - 89.98% %) for forecasting the hazard of death of patients with coronavirus. The above-mentioned machine learning techniques are explained in detail in chapter 3.

## 1.5 Thesis outline

The dissertation aims to evaluate/estimate the accuracy and find the best algorithm from several machine learning algorithms in predicting COVID-19 patient fatality rates. The results obtained from our studies will help to improve the prediction in future. In order to do this, we evaluated a few classified machine learning algorithms, namely decision tree (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K-nearest neighbor (K-NN) for predicting the mortality fatality risk in patients infected with coronavirus. These techniques are applied on the dataset of 146 countries which consists of laboratory samples of around 2,670,000 confirmed covid-19 cases, where 307,382 labelled samples encompassing both female and male patients. These techniques are the modern trends in the field of prediction models. To evaluate the results performance metrics which included accuracy, f1-score, precision, AUC, recall, ROC, and Confusion Matrix have been considered.

The thesis is well organized and divided into 5 chapters. The synopsis of the thesis is mentioned below precisely.

**Chapter 1** encloses the origin of COVID-19, aims of the current study, motivation of the present work, and briefly discusses/mentions the various machine learning techniques utilized in the current dissertation work.

**Chapter 2** presents the recent developments related to the work done by various eminent scientists around the globe. It focuses on the importance of machine learning being a vital tool in the prediction of COVID-19 mortality rates.

In **chapter 3**, all the terminologies used are described that includes datasets, and machine learning models that have been used as the base learners in our prediction model and validated using performance metrics. It discusses the research work and used methodology describing the flow of our proposed study.

In **chapter 4**, the results after the utility of different ML algorithms have been briefly explained. The results showed that how we were able to achieve an exceptional 91.26 % accuracy using the Random Forest algorithm.

**Chapter 5** envelopes the current research study's conclusion and future work. Finally, we've listed all of the sources we utilized in our thesis.

# CHAPTER 2

# LITERATURE REVIEW

Scientists have been actively involved in publishing useful reports by utilizing ML algorithms concerning COVID-19. Some of the useful studies are summarized below.

In 2021, Zhendong Xiao published that the exploratory research findings indicate that, due to COVID-19, the senior age category (>80) had the maximum risk of death, whereas the category from 70 to 79 had the second-highest risk. This was discovered after testing a variety of ML techniques, including logistic regression, neural networks , light GBM, and decision trees. The data for this study came from the Kaggle website, and it included all confirmed cases of COVID-19 in the USA from 31 December 2019 to December 1, 2020. As a conclusion, it indicates that older people are more likely to die[14]. In 2021, Sowmya Sundari L K and colleagues stated that under the effect of the worldwide pandemic epidemic, cardiovascular disease is a serious issue in the medical community. 17.9 million fatalities have been documented, and this number has risen due to co-morbidities in the global coronavirus outbreak. The study is to provide a strategy for dealing with a predicted cognitive approach to identifying and evaluating the risks of heart disorders. For decision assistance via risk factor validation, the approach is supplemented by an SVM-based classifier. For this study, the American Heart Association provided a dataset of 500 individuals. As a result, the proposed study states that predictions made are able to construct a more accurate clinical decision diagnostic assistance system which avoids costly medical tests and checks, the patient saves money and time, and the patient may plan for appropriate treatment at the earliest stages of the condition as a preventative step [15]. In 2021, Quazi Adibur Rahman Adib and colleagues published that the impact of the virus on pregnant women and newborns has become a major concern among citizens and public health officials around the world. This work generated a model that predicts the mortality of COVID-19 infected mothers based on reported symptoms such as rhinorrhea, dyspnea, arthralgia, cough, and pneumonia. Gradient boosting, support v vector machine, random forest, artificial neural network, and decision tree are the machine learning models employed in this research. As a result, the best models are gradient boosting and artificial neural networks (ANN), and based on forecasts, health care providers can take better measures and provide intensive care to pregnant women [16]. In 2021, Safynaz Abdel-Fattah Sayed and colleagues proposed that to help healthcare professionals anticipate the seriousness and fatality risk of a patient, efficient and accurate Intelligent Systems tools are required. By offering early pharmaceutical interventions, earlier detection of patient severity might assist save hospital resources and reducing the number of patients who die. X-ray pictures are being utilized to detect COVID-19 patients as early signs. Cohen JP[17] generated a publicly available dataset that was used in this study. COVID-19 and most cases of pneumonia are included in the database. It includes patient information, such as his/her age, type of disease, X-ray image, state of recovery, and whether the patient was admitted to the ICU. As a result, a prediction model has been constructed for the COVID-19 treatment based on X-ray scans utilizing machine learning techniques such as Random Forest (RF), KNN, XG boosting, Bagging, Extra Tree, and SVM[18]. In 2021, Nikhil and colleagues stated that the outbreak of COVID-19 continues to rise with increasing deaths globally. In addition to the amazing technological breakthroughs in our daily lives, especially in machine learning and in-depth learning, AI has also helped people in this difficult COVID-19 controversy. This study demonstrates the effect of a polynomial-based linear regression model that predicts future conditions based on current conditions using data from the past few months. Data used in this study were obtained from the European Center for Disease Prevention and Control. As a result of this study, the use of Machine learning and artificial intelligence in the Coronavirus outbreak projecting infection rates, diagnosing using photos, and the significance of machine learning in vaccine development[19].

I would like to put on record that we have published a review article entitled “ parameter based literature survey of COVID-19 mortality dynamics using machine learning techniques” in the 6th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) in 2021[20]. The paper discusses current improvements in COVID-19 mortality rate forecasts and aids in getting an understanding of Spatio-temporal mortality dynamics, which will be useful in the implementation of future control strategies. COVID-19 mortality rate projections based on several factors were explored in this paper. Xu and colleagues published that Predictable mortality tests were performed using the blood sample of 485 infected people in Wuhan, China. Scientists used the XG boost classifier to predict the death of each patient 10 days before using three biomarkers: lymphocyte, lactic dehydrogenase, and high-sensitivity C-reaction protein. They found that having high levels of LDH was responsible for a large portion of people who needed immediate medical attention. This might be related to tissue disintegration caused by excessive LDH levels, which can occur in a variety of conditions including pulmonary disorders (pneumonia). For the research investigation, they created an XG Boost machine-learning model. From 10 January to 18 February 2020, medical data was collected for the model's development. Overall, they proposed an implementable decision rule for forecasting the most at-risk individuals, resulting in a lower fatality rate. The projections were 90 percent accurate[21]. Zheyong, Bijie, and colleagues analyzed those 248 cases of relapse of COVID-19 patients in China were studied in laboratory, clinical and radiological aspects. To investigate the risk of hospital deaths, the authors used a reversal and inconsistent regression. The biomarker D-dimer was discovered to be the primary cause of mortality in the research. Finally, the study discovered that D-dimer is only high in COVID-19 patients and is linked to in-hospital mortality[22]. Wu and team reported that various variables influence COVID-19 morbidity rates in different nationalities. Machine learning algorithms, specifically linear regressions [23], was employed on a cross-sectional dataset of 169 nations to predict death rates. The information for this project came from the website "Worldometer: Coronavirus." Case numbers, test numbers, and death numbers were all included in the data collected from over 200 nations. The findings revealed that the mortality dynamics of COVID-19 was significantly worse, with the COVID-19 test rate per 100 people, the number of beds in the medical clinic, the government's efficacy score, and the population aged 65 and up scoring positively on transport infrastructure quality. According to the study, low-income countries, low effective government scores, young people, and fewer hospital beds had higher mortality rates and screening rates. This can be enhanced in order to lower the COVID-19 fatality rate in the long run. Atlam and colleagues reported that Climate data (seasonal patterns) have been identified as a determinant in the transmission, incidence, and type of human coronavirus illness [24]. Coronavirus is expanding at an alarming rate over the world, posing a serious hazard. As COVID-19 is not yet fully known or tested, climate factors should be considered and researched with the aim of better understanding the virus' status and the real influence of climate limitations on viral spread in individuals. Several regressor ML models, such as linear models (Linear Regression, Ridge Regression[25], Automatic Relevance Determination, Bayesian Ridge, Huber Regressor) and learning-based models (Random Forest, Gradient Boosting Regressor, etc.) are used to extract the relationship between the different climatic conditions, namely humidity and temperature, census, and health centre resources, in the transmission of COVID-19[26] , XG boost, Light GBM[27], support vector machine (SVM)[28], k-nearest neighbors regressor (KNN), and Decision Tree[29] are some of the other algorithms investigated. The data for this investigation came from Kaggle and the Johns Hopkins Center, while the meteorological data came from a historical climate database[30], [31]. Climate characteristics are substantially more potent in predicting death rates than numerous statistical elements such as population, age, and urbanisation, according to the study. We may deduce from this study that humidity and temperature are the most important factors in COVID-19 mortality rate forecasting, with higher temperatures indicating fewer illness cases[24]. On 10th, July 2020, According to World Health Organization data, COVID-19 confirmed cases totalled over 12 million, including 549,247 mortality cases (WHO). For the current COVID-19 pandemic, many variables that may increase mortality risk in people are being assessed. Comorbidities such as diabetes, hypertension, and coronary heart disease are the key concerns for greater mortality predictions, according to recent data analysis[32], [33]. Licia Iacoviello *et al.* stated that retrospective investigations were conducted on 3894 patients. From 19 February to 23 May 2020, individuals with coronavirus infection were admitted and disseminated around Italy in 30 clinical facilities. The authors employed the Cox survival analysis, which is a machine learning approach based on a random forest algorithm[34]. Medical Research data came from a web-based website with 61.7 percent male members (average age 67) and a follow-up period of 13 days. Based on the findings, aging, high C-protein, and kidney function appear to be the most common causes of high mortality. During the studies, the inter-variable and mortality relationships were homogeneous. Zhang and team showed that COVID-19 prognosis and severity have been associated to thrombo-inflammatory biomarkers. The FAD-85 score was created via the support of univariate analysis and multivariable logistic regression, as well as a nomogram based on three factors: ferritin, D-dimer, & age and was detected by patients admitted to hospital with a 28-day death predictor. This research assists in determining the severity of COVID-19 disease and in making appropriate treatment decisions for COVID-19 patients with coagulation abnormalities and inflammatory diseases[35]. Gerard Torrats-Espinosa evaluated that the trend in COVID-19 fatality throughout the United States is attributed to racial residential segregation. The author(s) have made use of double-lasso regression to identify the utmost relevant parameters on large databases, including population, political, social capital, air pollution, ideology, overcrowding, critical industrial activities, and the opportunity to engage with society. The Harvard data verse contains the datasets utilised in this study. The author found that blacks (8%) and more segregated counties have higher death rates as a result of this study[36]. In 2021, Gulcin *et al.* reported that using a machine learning approach called logistic regression, researchers looked at the relationship between coronavirus fatality and cancer or non-cancer patients. They concluded that cancer patients had a greater death rate than non-cancer patients. Electronic clinical records were used to collect data for the study. The researchers also utilised logistic regression to discover other risk variables linked to cancer patient death rates[37].

The literature review shows the “well-documented” evidence of the evolution and applications of machine learning in the regime of COVID-19 pandemic. The mentioned literature reports refer to the various factors spotted by different research groups responsible for COVID-19 infection and the commonality among them. We observed that scientists have picked several factors and applied machine learning algorithms to the respective dataset(s) to predict the COVID-19 detection risk. We could also see that there are limited reports of the utility of ML in COVID-19 mortality dynamics compared to that of COVID-19 risk prediction models. Thus, this was an additional reason biasing us to create a model for mortality prediction working on the dataset of combined patient information across all regions/international boundaries. Also, scientists have restricted themselves to specific demographic regions while performing studies which further motivated us to expand the dataset on a global scale. Inspired from the recent developments in this field, we have developed a mortality risk prediction model on COVID-19 patient working on the in-hospital COVID-19 patients’ data across the globe.

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Proposed Architecture

Herein, the proposed planning of the model is explained i.e., how the model will flow and follow all the steps using the dataset and machine learning techniques used to make the predictions shown in figure 3.1. depicts the linkages and information flow between the tasks we employed to obtain the prediction outcomes. Data Pre-processing is done, then features are extracted, and various techniques are applied.

In the Proposed Model the implementation has been carried out in following different steps:

1) In the first step the dataset collection is being performed.
2) Next on the collected data we would be performing data preprocessing such as dealing with data cleaning, noisy data, data transformation and data reduction.
3) To convert the text into matrix we would perform One hot encoding.
4) For better results we applied feature engineering using feature extraction using PCA.
5) Then we would train the data to the machine learning model. Here we have used five ML classification techniques.
6) Each of the ML model is trained multiple times on feature list [2,5, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 50, 1000, 1050, 1100 and 1150] no. of features. In this way cross validation of no. of features is done to find best performing model.
7) At the end we have to check how each model is performing so in order to assess a model we would calculate various evaluation measures to know the performance of each model and do comparison for to find the best one out.

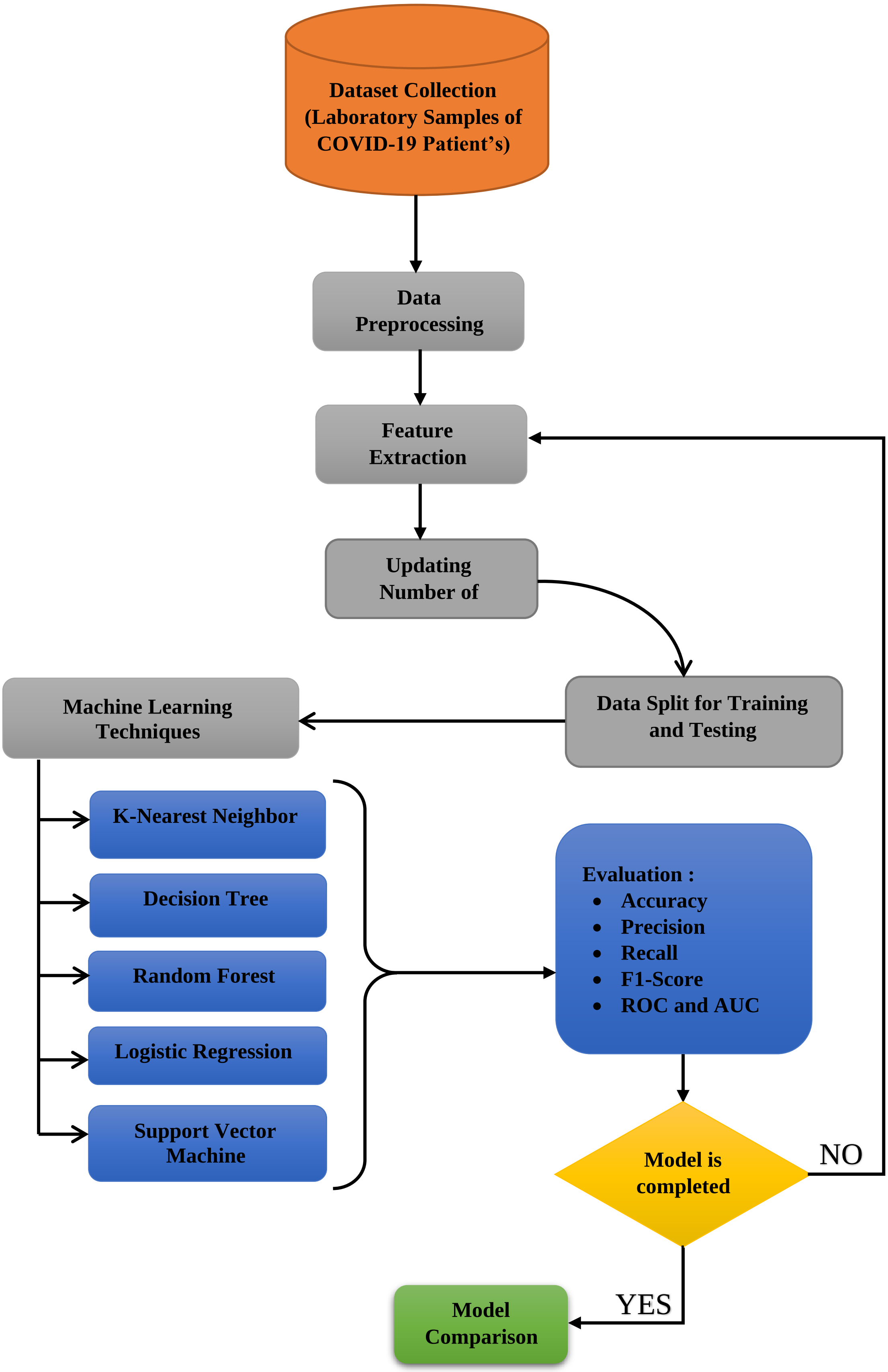


**Figure 3.1.** System Architecture

## 3.2 Dataset used

In this research, we used the dataset of 146 countries which consists of laboratory samples of around 2,670,000 confirmed covid-19 cases, where 307,382 labeled samples encompassing both female and male patients with 44.75 years of average age[38]. The presence of viral nucleic acid confirms the illness. Each patient had 32 data items in the original dataset, comprising demographic and physiological information which is represented in table 3.1. We deleted irrelevant and redundant data items during the data cleaning step. The unlabeled data samples have also been deleted. In order to make model more efficient the data has been normalized. We also balance our dataset in order to obtain an accurate and impartial model.

To train and evaluate our model, we constructed a balanced dataset with an adequate ratio of recovered and dead patients. The training dataset's data samples (patients) were chosen at random and are fully independent of the testing data. table 3.1 represents the initial features in the dataset

**Table 3.1.** The characteristics employed in the machine learning techniques.

| **Symptoms:** | | **Pre-Existing Conditions**: | **Demographics**: |
|---|---|---|---|
| • Anorexia<br>• Fever<br>• Shortness of Breath<br>• Chest Pain<br>• Gasp<br>• Somnolence<br>• Chills<br>• Headache<br>• Sore Throat<br>• Conjunctivitis<br>• Kidney Failure<br>• Sputum<br>• Dizziness<br>• Myalgia<br>• Cold<br>• Expectoration<br>• Myelofibrosis | • Fatigue<br>• Cough<br>• Lesions on Chest Radiographs<br>• Septic Shock<br>• Dyspnea<br>• Obnubilation<br>• Cardiac Disease<br>• Eye Irritation<br>• Respiratory Distress<br>• Rhinorrhea<br>• Diarrhea<br>• Hypertension<br>• Emesis<br>• Pneumonia<br>• Hypoxia<br>• Heart Attack | • Diabetes<br>• COPD<br>• Hypertension<br>• Parkinson's Disease<br>• Chronic Kidney Disease<br>• Asthma<br>• Cerebral Infarction<br>• HIV Positive<br>• Cardiac Disease<br>• Dyslipidemia<br>• Hypothyroidism<br>• Cancer<br>• Chronic Bronchitis<br>• Any Chronic Disease<br>• Prostate Hypertrophy<br>• Coronary Heart Disease<br>• Travel History<br>• Hepatitis B | • Age<br>• Country<br>• Province<br>• Gender<br>• Tuberculosis<br>• City |

## 3.3 Data Preprocessing

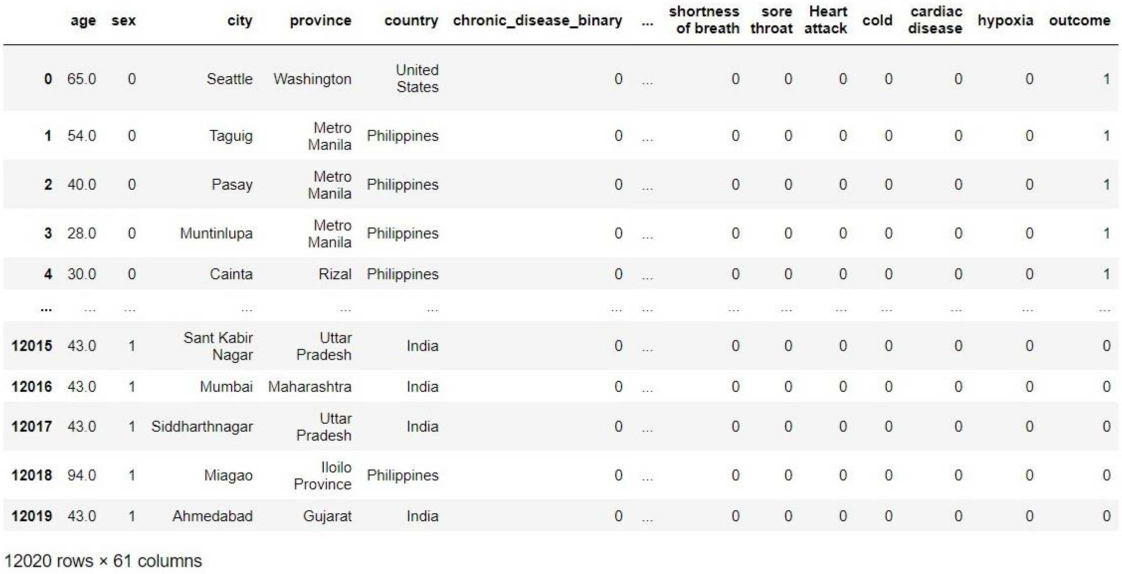
Data preprocessing is the way of preparing raw information to be used in a learning system. It's the most important phase in building a prediction model based on machine learning. If we are working on some kind of machine learning project, we do not really have access to clean and refined data. Also, before engaging in any data action, the data must be cleaned and formatted. As a result, we use a data pre-processing function to do this.

### 3.3.1 Import libraries:

Pandas and Numpy are two important data-loading libraries. The Pandas programming language was utilised to perform database operations and data processing. For computing, Numpy is utilized. For visualizing, Matplotlib and Seaborn were utilised. The libraries for using machine learning algorithms are imported from sklearn.

### 3.3.2 Read data:

The raw data is given in a CSV file that may be read using the pandas_pd_read.csv() function. figure 3.2 illustrates the Data view.



| | age | sex | city | province | country | chronic_disease_binary | ... | shortness of breath | sore throat | Heart attack | cold | cardiac disease | hypoxia | outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 65.0 | 0 | Seattle | Washington | United States | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **1** | 54.0 | 0 | Taguig | Metro Manila | Philippines | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **2** | 40.0 | 0 | Pasay | Metro Manila | Philippines | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **3** | 28.0 | 0 | Muntinlupa | Metro Manila | Philippines | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **4** | 30.0 | 0 | Cainta | Rizal | Philippines | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **12015** | 43.0 | 1 | Sant Kabir Nagar | Uttar Pradesh | India | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12016** | 43.0 | 1 | Mumbai | Maharashtra | India | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12017** | 43.0 | 1 | Siddharthnagar | Uttar Pradesh | India | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12018** | 94.0 | 1 | Miagao | Iloilo Province | Philippines | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12019** | 43.0 | 1 | Ahmedabad | Gujarat | India | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

12020 rows × 61 columns

**Figure 3.2.** COVID-19 Patient's Dataset view

### 3.3.3 Checking for missing and values:

To locate the missing values, use a data set labeled '0' and '1'.

### 3.3.4 Checking for categorical and variable data:

In the classification, the target variable was identified here. Techniques for feature dispersal have been used in the Seaborn display().

### 3.3.5 Splitting dataset into training and test set:

We separated the data in two sets: a testing set and a training set, in order to prepare it for machine learning. This is a vital activity in data preparation since it improves our machine learning model's capabilities. figure 3.3 shows a split visual of a dataset in any prediction model.

Dataset
Training Set
Test Set

**Figure 3.3.** Dataset Splitting

#### 3.3.5.1 Training Set:

A subset of the dataset used to train the prediction model has already revealed the outcome.

#### 3.3.5.2 Test set:

The predictive model is tested on a subset of the dataset, and the test set is utilised to forecast the outcome.

We ensure that our dataset was balanced in order to achieve an accurate and impartial model. To train and evaluate our model, we built a balanced dataset with an equal sample sizes for both revived and died patients. we split the data into 70% training and 30% testing.

### 3.3.6 Normalization

There are a number techniques to create normal machine readings, but the most frequent is to re-measure the data such that the values fall within a certain range, generally between 0 and 1. This is commonly accomplished by multiplying each value by a very big integer in the database. The standard deviation and setting (re-measuring data such that the average is 0 and the standard deviation is 1) are two more approaches of normalizing (re-measuring data so that the minimum value is 0 and the maximum value is 1). Here we have performed normalization of data by min-max method.

$$z = \frac{(x - \min(x))}{\max(x) - \min(x)} \qquad (3.1)$$

**Where :**

z = standard score of sample x.

x = sample.

min(x) = min value in sample.

max(x) = max value in sample.

### 3.3.7 Feature Extraction

When the original raw data collection is compressed to a small set of features that nevertheless retain crucial information from the real data, feature extraction is a technique of decreasing size. This may be accomplished by a number of methods, including key independent component analysis, component analysis, and non-negative matrix factorization. In this case, we utilised PCA. The patient's health state was explained by numerous values on the outcome label. Patients who were done with treatment from the hospital or who were in a good state with no further symptoms were classified in the recovered category. At the moment of admittance to the hospital, healthcare personnel documented the symptoms. For better results in the prediction, we applied One-Hot Encoding to the dataset and after that, we applied the feature extraction using PCA. We retrieved features that include symptoms, physicians' medical reports, demographics, and physiological parameters from the original dataset after extracting. To ensure that all-important characteristics were retrieved, we spoke with a medical team.

### 3.3.8 One-Hot Encoding

It is a data pre-processing and modification method that helps our Models comprehend the input better. It has its own set of pros and cons. keep in mind the sort of data it will be processing while concentrating on the conventional model's end conclusions and output. These newly discovered representation techniques have been critical in better characterising data and improving the accuracy and comprehension of the models being produced[39]. This is the case with algorithms like Deep Learning and Machine Learning. since machines can only read numbers and cannot comprehend speech in the first place. One hot code coding is the process of converting classification data into machine and in-depth learning methods, that improves model guessing and segment accuracy. A common strategy for prepping section features before applying them to machine learning models are one hot encoding. This encoding assigns a value of 1 to every sample that corresponds to the original category and forms a new binary feature for every feasible class. One of the most important aspects of the feature engineering process in learning methods training is hot encoding. If our variables were colours and the labels were "red," "green," and "blue," each of these tags might be encoded as a three-element binary vector like this: [1, 0, 0], [0, 1, 0], and [0, 0, 1] are the colour codes for red, green, and blue, respectively. Categorical data must be converted to numerical form as part of the processing procedure. One-hot encoding is frequently used to encode data in its integer form. The integer encoded variable is erased, and every unique integer value is shown by a new binary variable. Following a column with category data that was labelled encoded throughout the method, it separates the next column into numerous columns. The numbers are altered at random with 1s and 0s according to whatever column has the greatest value.

After applying a one-hot encoding on our dataset, a total of 1261 features were created and, on those features, we applied the feature extraction using PCA.

### 3.3.9 PCA

PCA (Principal component analysis), is the dimensionality-reduction method that decreases the dimensionality of huge datasets by reducing a large collection of elements into a smaller group that retains the bulk of the content in the larger set.

PCA is an uncontrolled learning approach for reducing size in machine learning. It is a mathematical procedure that converts the visibility of a connected element into a set of irrelevant feature lines via orthogonal conversion. Newly changed characteristics are key components. It's one of the most extensively used tools for testing data analysis and predictive modelling. It is a method of decreasing variability in order to derive solid patterns from a database. PCA seeks a low-dimensional environment in which to handle high-resolution data. Because the top characteristic reveals the appropriate differences across classes, PCA minimises the size by assessing the variability of each feature. Real-world PCA applications include image processing, movie recommendation systems, and improved power-sharing across various communication channels. It keeps key variables while rejecting less relevant ones since it is a way of extracting a feature. The PCA method is based on the mathematical ideas listed below.

- Covariance and Diversity
- Eigenvectors and Eigenvalues

The key components used for PCA in our model are listed in 25 different values as: [2,5, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1050, 1100, 1150]. Every algorithm is being cross validated on these set of features.

## 3.4 Predictive Analytics Algorithm

We built a prediction model using multiple machine learning methods after the Feature extraction process. Support vector machine, random forest, decision tree, logistic regression, and K-nearest neighbor were among the techniques utilized in this study.

### 3.4.1 Support Vector Machine

SVM (support vector machine) is a machine learning approach that is typically used for regression and classification, although it is also used for pattern recognition and data analysis[40]. The SVM approach is used to determine the border with the maximum width. The model generated here is a classification model that is shown as a boundary where information of diverse attributes is assigned in one area. The input data must be seen as a high-dimensional feature space in nonlinear classification, which may be done efficiently using kernel approaches[41]. In this study kernel used is "rbf" and the training and testing accuracy was found to be 90.49% and 89.87% respectively at 350 features. The accuracy vs number of feature graph for SVM is represented in figure 3.4. This figure displays cross validation of SVM model on different number of features. This graph has a relation between train test accuracies and number of features taken, which shows that how SVM algorithm performed on different number of features. Initially we observed that model underfits, probably due to low amount of features that have low variance, but that is solved as we increase the number of features. Also, the confusion matrix of optimal SVM which is obtained can be seen in figure 3.5.

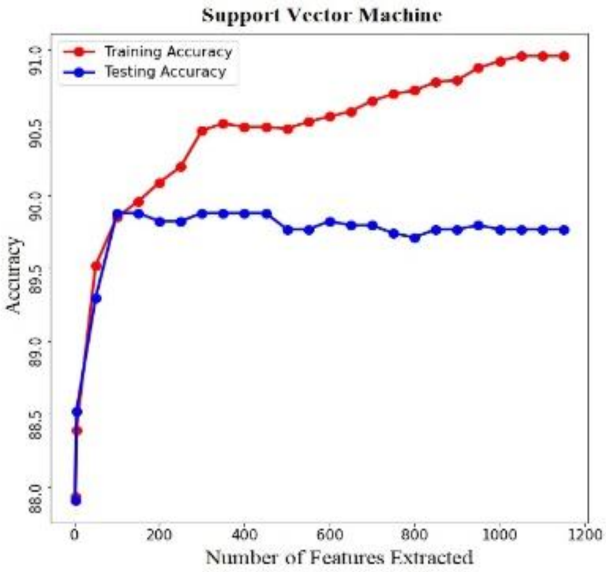


**Figure 3.4.** Accuracy vs Number of feature graph of Support Vector Machine.

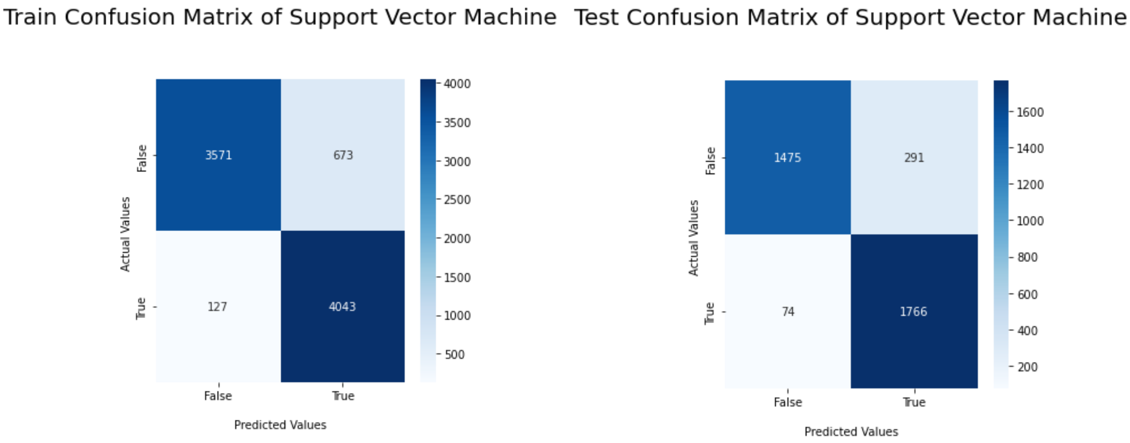


**Figure 3.5.** Training and Testing Confusion matrix of Support Vector Machine.

### 3.4.2 Decision Tree

Decision tree learning method of machine monitoring classification and retrieval problems. Decision tree imposes that the data set must be labelled. The classification is done using a set of conditions using the method of the decision tree. In the decision tree, the node represents the element, the branch characterizes the process, and the leaf node represents the conclusion. It is a tree-like structure, that provides more stability and precision. In the first stage, a tree will be built with input features as its nodes. It will then select a feature for the next stage from the input features.

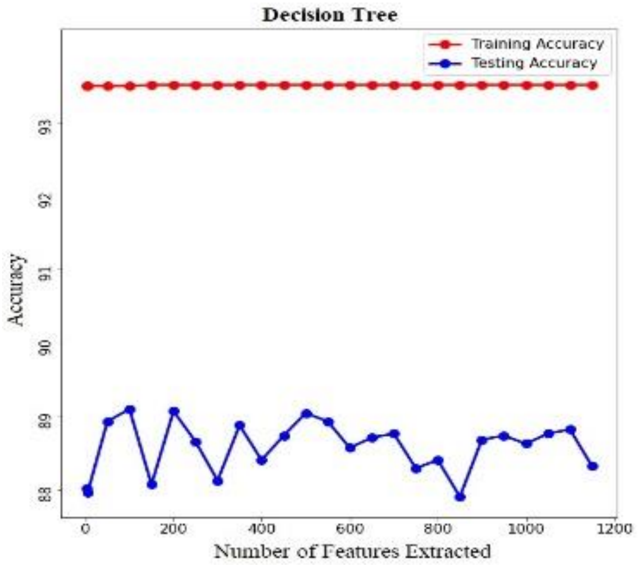


**Figure 3.6.** Accuracy vs Number of feature graph of Decision Tree.

It will then select a feature from the input features to predict the output, resulting in the greatest growth in knowledge Apply the procedures outlined above to the Subtrees are created by utilizing characteristics that are available and were never used before[42].

In this study, the training and testing accuracy was found to be 93.49% and 89.10% respectively at 100 features. The accuracy vs number of feature graph for decision tree is indicated in **Figure 3.6**. This figure displays cross validation of DT model on different number of features. This graph has a relation between train test accuracies and number of features taken, which shows that how DT algorithm performed on different number of features Also, the confusion matrix of obtained for decision tree is viewed in **Figure 3.7**.

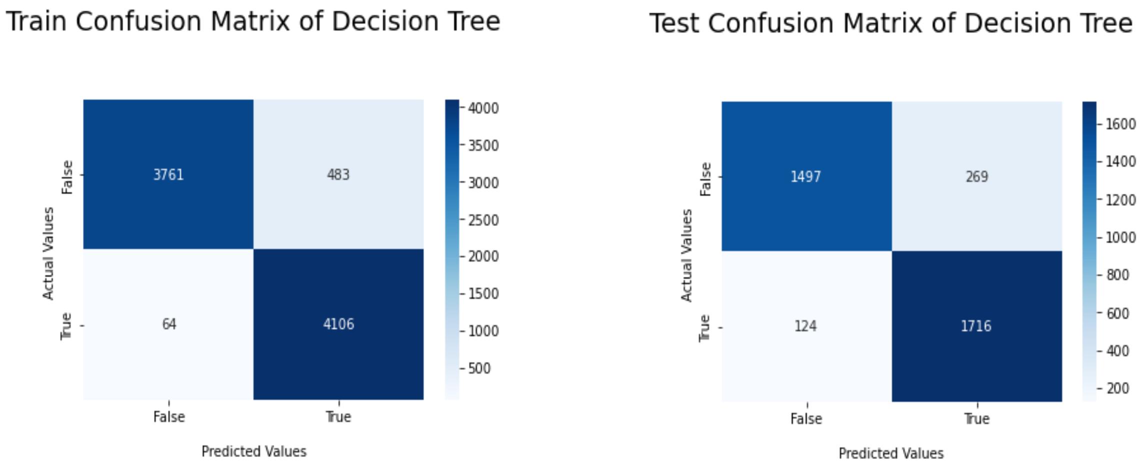


**Figure 3.7.** Training and Testing Confusion matrix of Decision Tree.

### 3.4.3 Logistic Regression

In statistics, the generalized linear regression model which is logistic regression used for the prediction with the target class having 2-level events likely alive or dead, true or false, lose or win[43].

To put it another way, chances are the proportion of the likelihood of an event occurring divided by the likelihood of it not occurring. If p stands for probability, then the equation $\frac{p}{1-p}$ stands for odds. The natural logarithm of the chances is used as a predictive factor in the logistic regression model, and the equation expressed mathematically is[14]:

$$L = \log_b \frac{p}{1-p} \qquad (3.2)$$

In this study, the training and testing accuracy was found to be 89.70% and 89.24% respectively at 450 features. The accuracy vs number of feature graph for logistic regression is depicted in figure 3.8. This figure displays cross validation of RF model on different number of features. This graph has a relation between train test accuracies and number of features taken, which shows that how RF algorithm performed on different number of features. Initially we observed that model underfits, probably due to low amount of features that have low variance, but that is solved as we increase the number of features. Also, the confusion matrix of obtained for logistic regression is exhibited in figure 3.9.

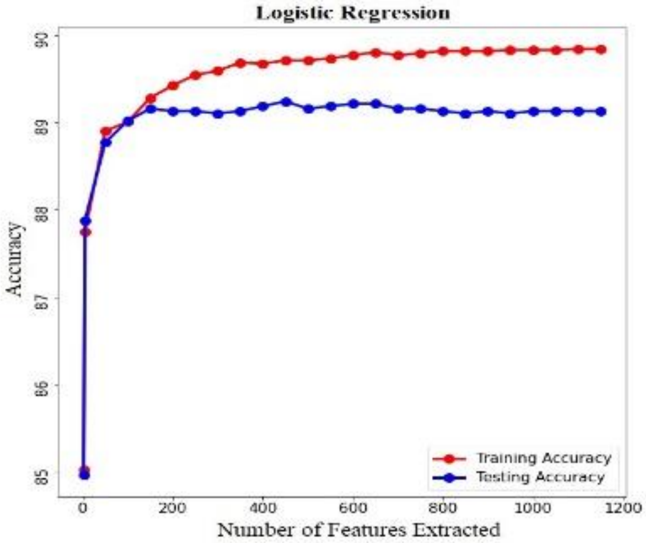


**Figure 3.8.** Accuracy vs Number of feature graph of Logistic Regression.

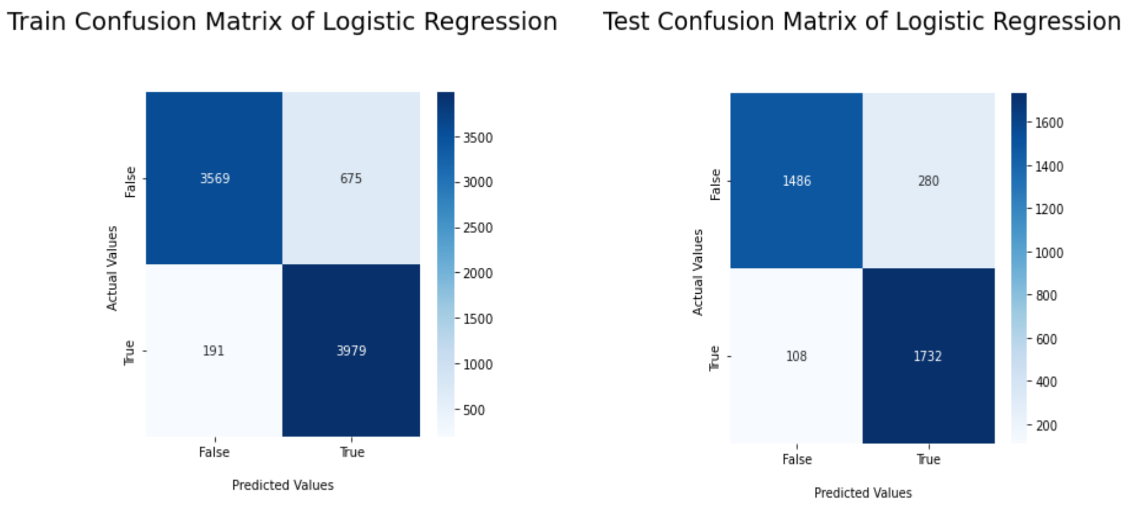


**Figure 3.9.** Training and Testing Confusion matrix of Logistic Regression.

### 3.4.4 Random Forest

A machine learning ensemble model called a random forest that helps in solving both regression and classification problems. It is an ensemble model, i.e., it combines numerous ML techniques to outperform others. By randomly picking a subset of the training data set, random forest builds different decision trees. Using decision trees, it will predict the class of test class objects[44].

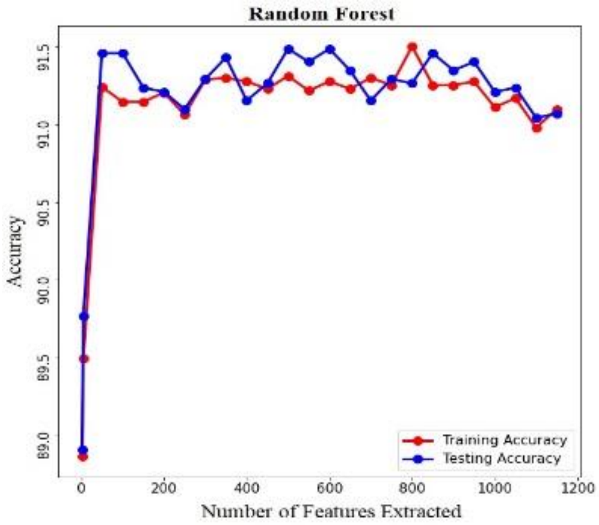
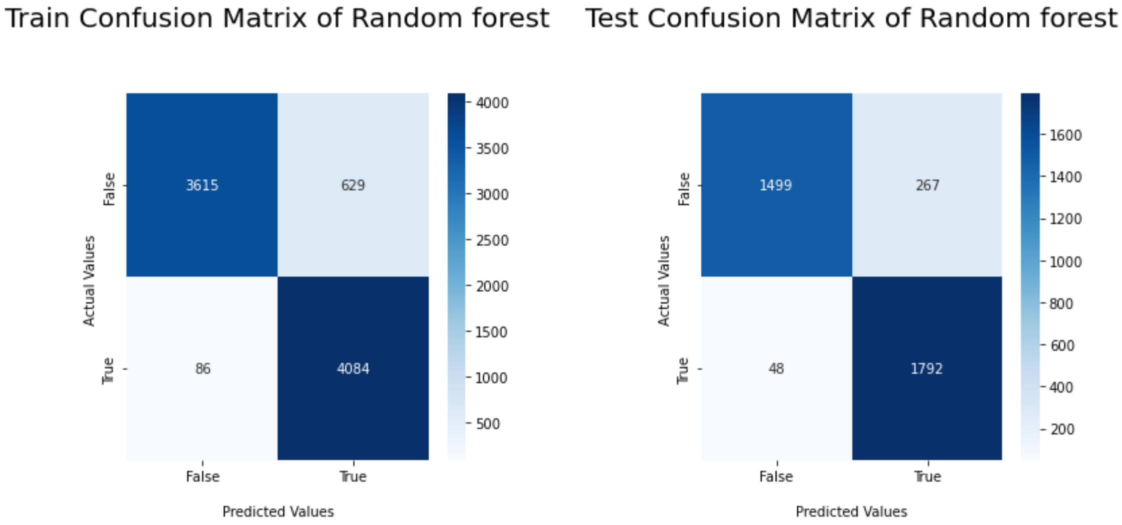


**Figure 3.10.** Accuracy vs Number of feature graph of Random Forest.

In this study, the training and testing accuracy was found to 91.50% and 91.26% respectively at 800 features. The accuracy vs number of feature graph for random forest is displayed in figure 3.10. This figure displays cross validation of RF model on different number of features. This graph has a relation between train test accuracies and number of features taken, which shows that how RF algorithm performed on different number of features. Initially we observed that model underfits, probably due to low amount of features that have low variance, but that is solved as we increase the number of features.  Also, the confusion matrix of obtained for random forest can be viewed in figure 3.11.



Train Confusion Matrix of Random forest

Test Confusion Matrix of Random forest

**Figure 3.11.** Training and Testing Confusion matrix of Random Forest.

### 3.4.5 K-Nearest Neighbor

For both classification and regression tasks, the simplest supervised learning approach, K-nearest neighbor, is used. As it is a controlled ML approach, the data must have both input and output parameters in order for the algorithm to be trained. For a given value of k, the K-NN method will locate the k closest data points. The data point's class will then be decided using the class of the largest set of data items with almost the same class. It employs and calculates the K nearest neighbours using either the similarity metric or Euclidean distance. The Euclidean distance formula is:

$$d(x,y) = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2} \qquad (3.3)$$

**Where :**

x, y are two point in Euclidean k-space.

$x_j$, $y_j$ Euclidean vectors.

After that, the most likely class is allocated to the data point. The probability may be denoted as follows:

$$P(Y = j|X = x) = \frac{1}{k}\sum\nolimits_{y\in A} I(y^i = j) \qquad (3.4)$$

The methodology is the same for regression problems, except instead of neighbor classes, target values are used.

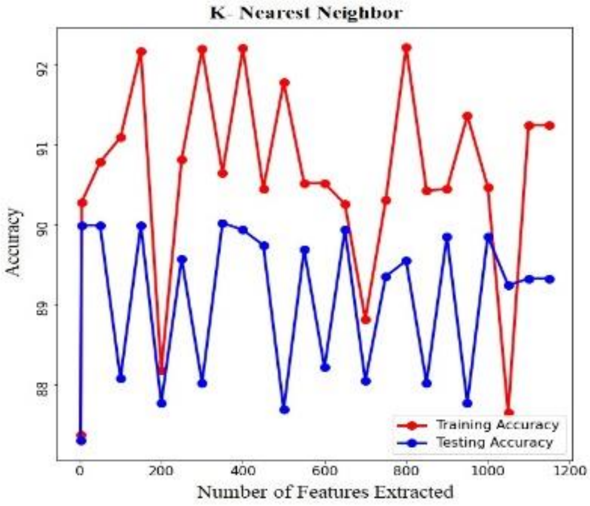


**Figure 3.12.** Accuracy vs Number of feature graph of K-Nearest Neighbor.

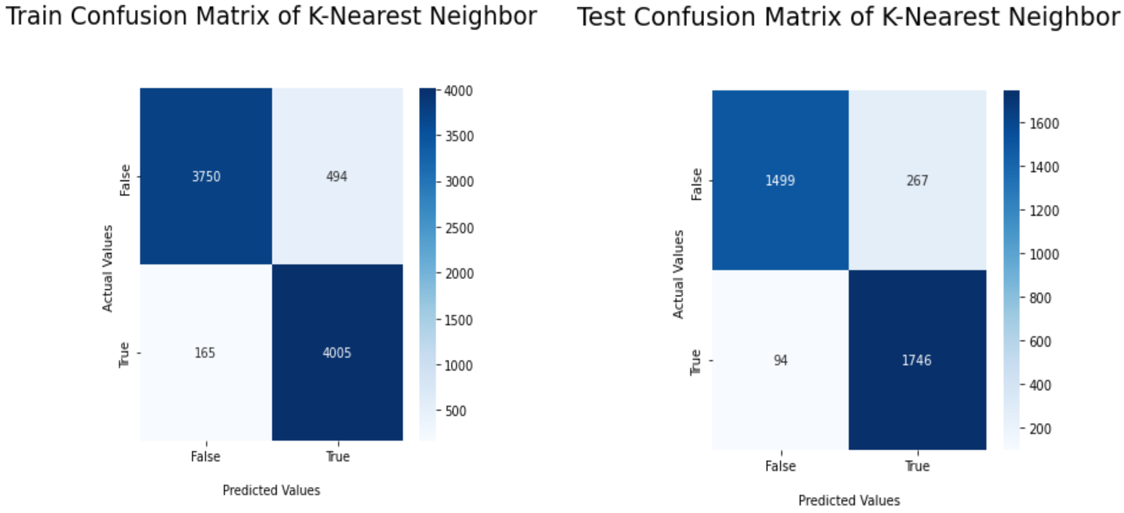


**Figure 3.13.** Training and Testing Confusion matrix of K-Nearest Neighbor.

Choosing an appropriate k is one of the most difficult tasks in KNN. The choice border will be more irregular if k is less, while a greater value of k will result in a smoother decision boundary[45].

In this study value of k=5 and the training and testing accuracy was found to be 92.16% and 89.98% respectively at 150 features.

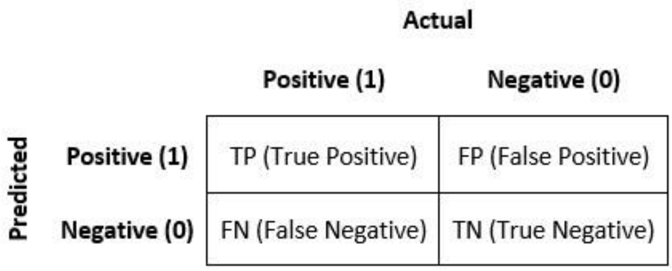
The accuracy vs number of feature graph for K-nearest neighbor is depicted in figure 3.12. This figure displays cross validation of KNN model on different number of features. This graph has a relation between train test accuracies and number of features taken, which shows that how KNN algorithm performed on different number of features. We observed that model underfits few times, probably due to low amount of features that have low variance, but that is solved as we increase the number of features. Also, the confusion matrix of obtained for K-nearest neighbor can be viewed in figure 3.13.

## 3.5 Performance Metrics

Following the normal Feature Engineering, feature extraction, and, of course, building a model and receiving some output in the form of a probability or a class, the very next step is to determine the model's effectiveness using test datasets based on some performance measure.

### 3.5.1 Confusion Matrix

The Confusion Matrix is a rational and straightforward (unless you aren't confused) indicators of the model's accuracy and correctness. It's used to solve classification issues whenever the outcome can be classified into two or more classes.



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

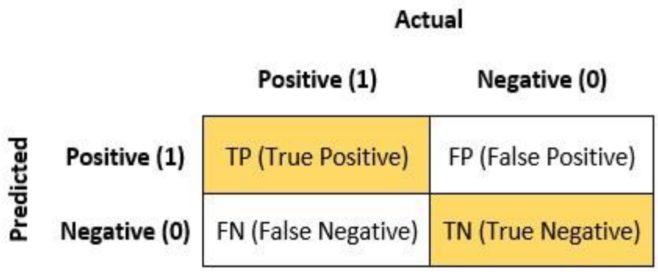
**Figure 3.14.** Confusion Matrix

- **True Positive** (TP) = The no. of instances in which the method has correctly identified them as positive.
- **False Positive** (FP) = The no. of cases in which the method incorrectly classifies a case as positive.

- **False Negative** (FN) = The no. of cases that the method incorrectly classifies as negative.
- **True Negative** (TN) The no. of occurrences in which the method correctly classifies them as negative.

### 3.5.2 Accuracy

In classification tasks, accuracy refers to the number of correct predictions by the model across all sorts of predictions. Our accurate forecasts (True positives and True Negatives) (highlighted in yellow in the picture) are in the numerator, while all other predictions generated by the algorithm are in the denominator (There are both correct and incorrect answers).



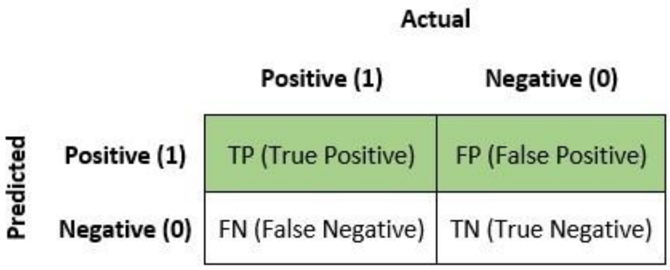
| | | Actual | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted | Positive (1) | TP (True Positive) | FP (False Positive) |
| | Negative (0) | FN (False Negative) | TN (True Negative) |

**Figure 3.15.** Accuracy

$$\textbf{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + FN + TN}} \tag{3.5}$$

### 3.5.3 Precision

Precision is a metric that informs us what percentage of cancer patients we diagnosed truly had cancer. People who are expected to be malignancy (TP and FP) and cancer patients are both TP.



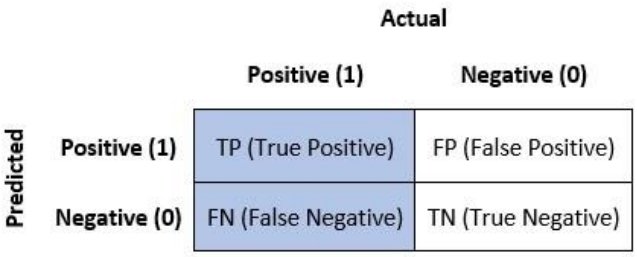
| | | Actual | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted | Positive (1) | TP (True Positive) | FP (False Positive) |
| | Negative (0) | FN (False Negative) | TN (True Negative) |

**Figure 3.16.** Precision

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{3.6}$$

### 3.5.4 Recall

Recall is a degree that informs us about the correctly diagnosed proportion of cancer patients investigated by the algorithm. Actual positives (cancer patients are TP and FN) and model-diagnosed cancer patients are both TP. (Note: FN is included since the Person was diagnosed with cancer against the model's forecast.)



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

**Figure 3.17.** Recall

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3.7}$$

### 3.5.5 F1-score

Tackling a classification challenge when building a model with carrying both Precision and Recall is cumbersome. So, it is feasible to combine Precision(P) and Recall(R) into a single score (R). Computing their arithmetic mean is one method for achieving this (P + R) / 2, with precision denoted by P and recall denoted by R. However, this might be troublesome in particular situations.

$$\textbf{F1} - \textbf{Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.8}$$

### 3.5.6 ROC

The true positive rate (TPR) is often represented on the Y axis, while the false positive rate (FPR) is typically plotted on the X axis. The "ideal" spot, with a false positive rate of zero and a genuine positive rate of one, appears to be in the top left corner of the picture. Although this isn't exactly correct, it does indicate that a larger area under the curve (AUC) is usually preferred. The "steepness" of ROC curves is particularly important since the goal is to maximise the genuine positive rate while minimising the false positive rate.

In binary classification, ROC curves are widely used to analyse a classifier's output. To broaden the ROC curve and ROC area for multi-label classification, the output must be binarized. For each label, a single ROC curve may be created. Each member of the label indicator matrix, on the other hand, may be seen as a binary prediction (micro-averaging).

$$\textbf{TPR (True Positive Rate)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (3.9)$$

$$\textbf{FPR (False Positive Rate)} = \frac{\boldsymbol{FP}}{\boldsymbol{TN+FP}} \qquad (3.10)$$

### 3.5.7 AUC

The area under the ROC Curve is known as the AUC. We essentially want to maximise this area so that we can have the highest TPR and lowest FPR for some threshold. This area is always represented as a value between 0 and 1 (just as both TPR and FPR can range from 0 to 1), and we essentially want to maximise this area so that we can have the

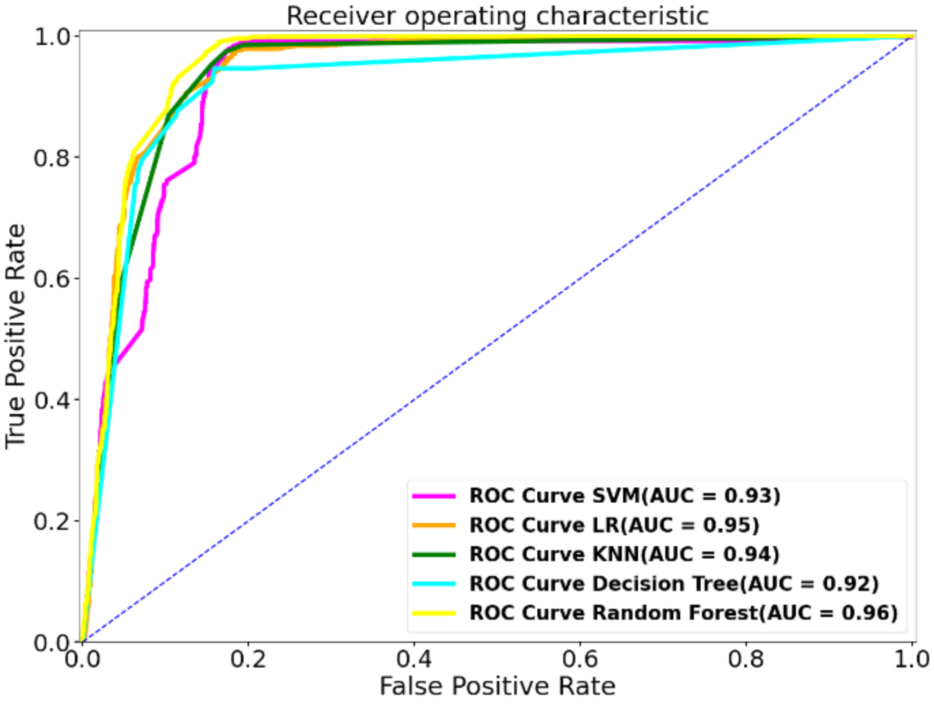


**Figure 3.18.** ROC Curve for all the algorithms.

highest TPR and lowest FPR for some threshold AUC may alternatively be calculated as the likelihood that a classifier would score a randomly selected positive instance higher than a randomly selected negative instance.

As a result, an AUC of 0.5 indicates that the chance of a positive case ranking higher than a negative instance is 0.5, and so random. With an AUC of 1, a perfect classifier always would score a positive occurrence higher than a negative one. AUC, ROC curve for all the machine learning techniques used in our model in shown in figure 3.18. Where we can clearly see that random forest obtained the highest AUC score of 0.96. and outperforms all the all the algorithms.

# CHAPTER 4

# RESULTS AND DISCUSSION

## 4.1 Model Evaluation

After performing normalization on all models, they were trained and tested by cross validating the number of features in a range.

**Table 4.1.** Shows the accuracies computed by models on each iteration.

| | SVM | | LR | | KNN | | DT | | RF | |
|---|---|---|---|---|---|---|---|---|---|---|
| **No. of Features** | **Train Acc.** | **Test Acc.** | **Train Acc.** | **Test Acc.** | **Train Acc.** | **Test Acc.** | **Train Acc.** | **Test Acc.** | **Train Acc.** | **Test Acc.** |
| **2** | 87.94 | 87.91 | 85.02 | 84.97 | 87.37 | 87.30 | 93.50 | 88.02 | 88.86 | 88.91 |
| **5** | 88.39 | 88.52 | 87.75 | 87.88 | 90.28 | 89.99 | 93.50 | 87.96 | 89.49 | 89.77 |
| **50** | 89.52 | 89.30 | 88.90 | 88.77 | 90.78 | 89.99 | 93.50 | 88.94 | 91.24 | 91.46 |
| **100** | 89.85 | 89.88 | 89.01 | 89.02 | 91.09 | 88.08 | **93.50** | **89.10** | 91.15 | 91.46 |
| **150** | 89.96 | 89.88 | 89.28 | 89.16 | **92.17** | **89.99** | 93.51 | 88.08 | 91.15 | 91.24 |
| **200** | 90.09 | 89.82 | 89.42 | 89.13 | 88.17 | 87.77 | 93.51 | 89.07 | 91.21 | 91.21 |
| **250** | 90.19 | 89.82 | 89.54 | 89.13 | 90.81 | 89.57 | 93.51 | 88.66 | 91.06 | 91.10 |
| **300** | 90.44 | 89.88 | 89.59 | 89.10 | 92.19 | 88.02 | 93.51 | 88.13 | 91.29 | 91.29 |
| **350** | **90.49** | **89.88** | 89.68 | 89.13 | 90.65 | 90.02 | 93.51 | 88.88 | 91.30 | 91.43 |
| **400** | 90.47 | 89.88 | 89.67 | 89.18 | 92.20 | 89.93 | 93.51 | 88.41 | 91.28 | 91.15 |
| **450** | 90.47 | 89.88 | **89.71** | **89.24** | 90.44 | 89.74 | 93.51 | 88.74 | 91.23 | 91.26 |
| **500** | 90.46 | 89.77 | 89.71 | 89.16 | 91.78 | 87.69 | 93.51 | 89.05 | 91.31 | 91.49 |
| **550** | 90.50 | 89.77 | 89.73 | 89.18 | 90.52 | 89.68 | 93.51 | 88.94 | 91.22 | 91.40 |
| **600** | 90.54 | 89.82 | 89.77 | 89.21 | 90.52 | 88.21 | 93.51 | 88.57 | 91.28 | 91.49 |
| **650** | 90.58 | 89.79 | 89.80 | 89.21 | 90.25 | 89.93 | 93.51 | 88.71 | 91.23 | 91.35 |
| **700** | 90.65 | 89.79 | 89.77 | 89.16 | 88.82 | 88.05 | 93.51 | 88.77 | 91.30 | 91.15 |
| **750** | 90.69 | 89.74 | 89.79 | 89.16 | 90.30 | 89.35 | 93.51 | 88.30 | 91.25 | 91.29 |
| **800** | 90.72 | 89.71 | 89.81 | 89.13 | 92.22 | 89.55 | 93.51 | 88.41 | **91.50** | **91.26** |
| **850** | 90.78 | 89.77 | 89.81 | 89.10 | 90.42 | 88.02 | 93.51 | 87.91 | 91.25 | 91.46 |
| **900** | 90.79 | 89.77 | 89.81 | 89.13 | 90.44 | 89.85 | 93.51 | 88.69 | 91.25 | 91.35 |
| **950** | 90.87 | 89.79 | 89.83 | 89.10 | 91.36 | 87.77 | 93.51 | 88.74 | 91.28 | 91.40 |
| **1000** | 90.92 | 89.77 | 89.83 | 89.13 | 90.47 | 89.85 | 93.51 | 88.63 | 91.11 | 91.21 |
| **1050** | 90.96 | 89.77 | 89.83 | 89.13 | 87.65 | 89.24 | 93.51 | 88.77 | 91.17 | 91.24 |
| **1100** | 90.96 | 89.77 | 89.84 | 89.13 | 91.24 | 89.32 | 93.51 | 88.82 | 90.98 | 91.04 |
| **1150** | 90.96 | 89.77 | 89.84 | 89.13 | 91.24 | 89.32 | 93.51 | 88.33 | 91.10 | 91.07 |

The models were cross validated on 2,5, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1050, 1100 and 1150 features which can be seen in table 4.1. The documentation of all the training and testing accuracies obtained by all models has been done on above mentioned points. The highest obtained accuracy has been highlighted in the table 4.1.

After this, the optimal value of a number of features was found. Also, the best training and testing accuracy comparison graph for all the machine learning algorithms is represented in figure 4.1.

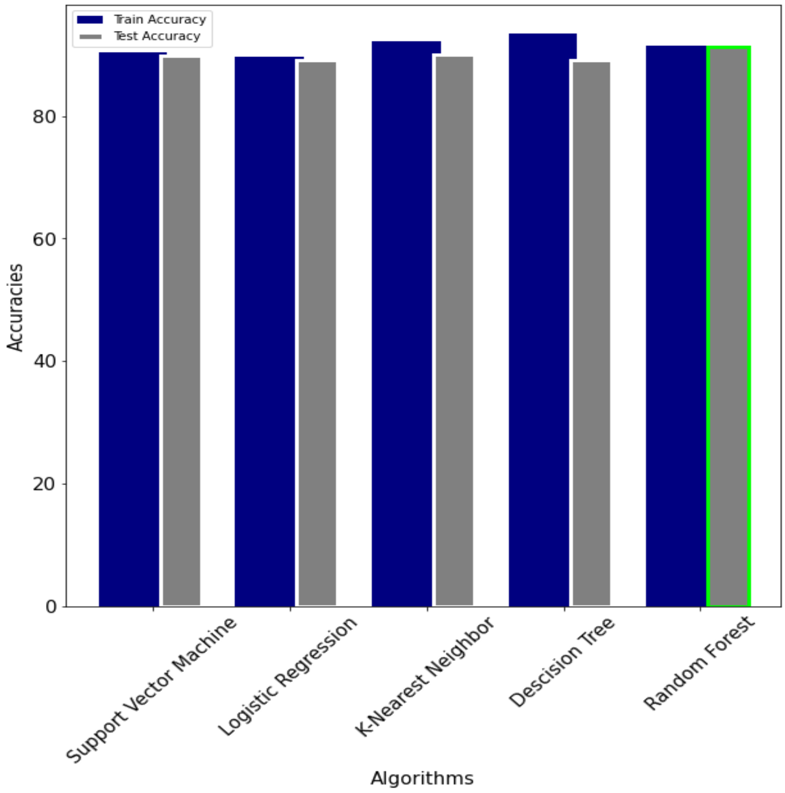


**Figure 4.1.** Test Vs Train Accuracy for All Machine Learning techniques

## 4.2 Results

In present work, the coronavirus infection rate was compared to the five models' predictions. Decision tree (DT), support vector machine (SVM), random forest (RF), logistic regression (LR), and K-nearest neighbor algorithms were implemented to assess the created model, which included recall, accuracy, f1-score, precision, AUC, ROC, and Confusion Matrix (KNN). table 4.2 summarizes all of the findings, demonstrating the accuracy of several machine learning algorithms in predicting death of coronavirus infected patients. Random forest was inferred as the most accurate and precise algorithm/technique displaying a rating of 91.26 %. Each machine learning method's ROC curves and AUC are displayed and compared in this study in figure 3.17. The Random Forest method classifier's performance is described and shown using a confusion matrix (figure. 3.11), which also provides insight into what the model misclassifies. And the final comparison of all the algorithms is shown in figure 4.2. Where it shows the random forest outperforms all the algorithms by giving the accuracy of 91.26% which is highest amongst other techniques.

**Table 4.2.** Results of all the techniques based on different performance measures.

| **S.no.** | **Algorithm/ Measures** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC** |
|---|---|---|---|---|---|---|
| **1.** | **Support Vector Machine (SVM)** | 89.87% | 90.44% | 89.87% | 89.82% | 92% |
| **2.** | **Logistic Regression (LR)** | 89.24% | 89.24% | 89.58% | 89.20% | 95% |
| **3.** | **K- Nearest Neighbor (KNN)** | 89.98% | 90.34% | 89.98% | 89.95% | 94% |
| **4.** | **Decision Tree (DT)** | 89.10% | 89.33% | 89.10% | 89.07% | 93% |
| **5.** | **Random Forest (RF)** | 91.26% | 91.86% | 91.26% | 91.22% | 96% |

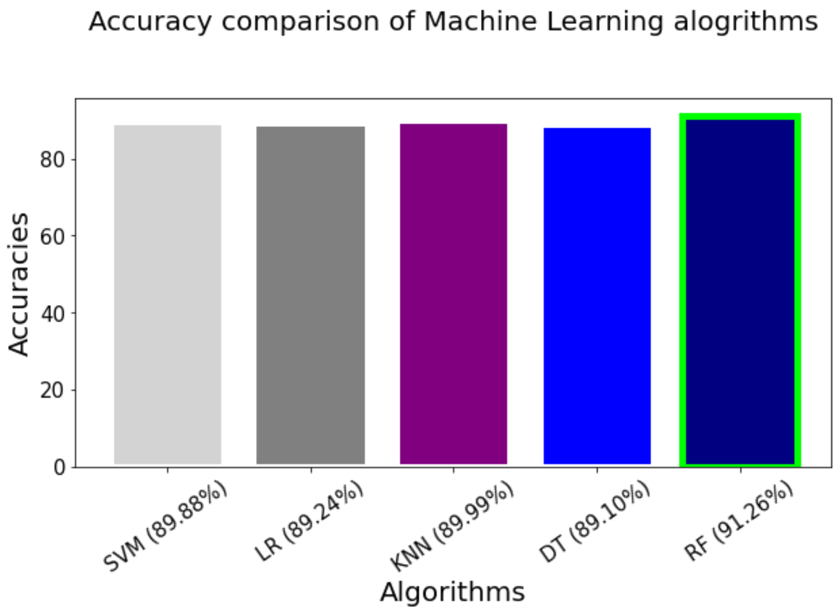


**Figure 4.2.** Algorithm Accuracy Comparison

## 4.3 Discussion

For estimation of COVID-19 patients' death rate around the globe, we enveloped cutting-edge machine learning methods using a huge dataset of positive cases gathered from all across the globe. Several performance indicators were used to assess the created algorithms. The assessment findings show that the generated models are very accurate and effective. Other research has demonstrated that from clinical data and blood test results, it is possible to estimate the death rate in COVID-19 patients[46]. However, we concentrated on demographic information, physiological information, patient symptoms, and pre-existing disorders. Using the Random Forest model, we were able to achieve an exceptional accuracy of 91.26 %. Furthermore, past research has mostly focused on data gathered in China[46], [47]**.** Rather than being trained on data from a particular region, we aggregated hospital data from across the world to create a more comprehensive model that is applicable to the entire world's population.

# CHAPTER 5

# CONCLUSION AND FUTURE SCOPE

## 5.1 Conclusion

- In the present work carried out, we utilized a prediction algorithm that would help the healthcare system/setting in increasing the survival percentage by offering an accurate and precise tool for the improvement of decision-making in clinics and better prioritize COVID-19 infected patients during the pandemic.
- Our approach can accurately screen or evaluate the death risk of COVID-19 patients on the basis of their physiological states, symptoms, pre-existing scenarios, and demographic data.
- This method will be of great aid to hospitals, and medical institutions in determining the patients who needs to be treated first before others, prioritize the critical patients when the system is overcrowded, and shorten wait times for needed care.
- The work might be expanded to cover additional diseases, empowering the healthcare setting to react faster in the case of an epidemic or pandemic.

## 5.2 Future Work

- Further, we can add more features in this dataset or club a new dataset which is much richer than current dataset for more better results.
- Various ensemble ML algorithms/techniques may be used to achieve the maximum efficacy of the results.
- Also, hyperparameter tuning can be done in the current model for better accuracy results.

# REFERENCE

[1] A. Erkoreka, “The Spanish influenza pandemic in occidental Europe (1918-1920) and victim age,” *Influenza and Other Respiratory Viruses*, vol. 4, no. 2, pp. 81–89, Mar. 2010, doi: 10.1111/j.1750-2659.2009.00125.x.

[2] J. Li *et al.*, “Machine Learning Methods for Predicting Human-Adaptive Influenza A Viruses Based on Viral Nucleotide Compositions,” *Molecular Biology and Evolution*, vol. 37, no. 4, pp. 1224–1236, Apr. 2020, doi: 10.1093/molbev/msz276.

[3] "World Health Organization. Coronavirus disease (COVID-19): Similarities and differences with influenza". Accessed on: 17 March 2020.[Online].Available:https://www.who.int/emergencies/diseases/novelcoronavirus-2019/question-and-answers-hub/q-a-detail/coronavirusdisease-covid-19-similarities-and-differences-with-influenza.

[4] Q. Li *et al.*, “Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia,” *N Engl J Med*, vol. 382, no. 13, pp. 1199–1207, Mar. 2020, doi: 10.1056/NEJMoa2001316.

[5] I. I. Bogoch, A. Watts, A. Thomas-Bachli, C. Huber, M. U. G. Kraemer, and K. Khan, “Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel,” *Journal of Travel Medicine*, vol. 27, no. 2, p. taaa008, Mar. 2020, doi: 10.1093/jtm/taaa008.

[6] World Health Organization. (Jan 2020). Statement on the second meeting of the International Health Regulations Emergency committee regarding the outbreak of novel coronavirus (2019-nCoV), world health organization (WHO). Archived from the original on 31 january 2020. WHO.

[7] https://covid19.who.int/.

[8] S.-M. Hyun, T.-H. Hwang, and K. Lee, “The Prediction Model for Classification of COVID-19 Infected Patients Using Vital Sign,” in *2021 International Conference on Information and Communication Technology Convergence*

*(ICTC)*, Jeju Island, Korea, Republic of, Oct. 2021, pp. 678–681. doi: 10.1109/ICTC52510.2021.9620902.

[9] M. Shanbehzadeh, Dept. of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran., R. Nopour, Dept.of Health Information Technology,School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran., H. Kazemi-Arpanahi, and Dept. of Health Information Technology, Abadan Faculty of Medical Sciences, Abadan, Iran., "Comparison of Four Data Mining Algorithms for Predicting Colorectal Cancer Risk," *J Adv Med Biomed Res*, vol. 29, no. 133, pp. 100–108, Feb. 2021, doi: 10.30699/jambs.29.133.100.

[10] https://ospreydata.com/wp-content/uploads/AI-vs-ML-vs-Deep-Learning.png.

[11] D. F. Hernandez-Suarez *et al.*, "Machine-Learning-Based In-Hospital Mortality Prediction for Transcatheter Mitral Valve Repair in the United States," *Cardiovascular Revascularization Medicine*, vol. 22, pp. 22–28, Jan. 2021, doi: 10.1016/j.carrev.2020.06.017.

[12] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," *J. Thorac. Dis*, vol. 11, no. S4, pp. S574–S584, Mar. 2019, doi: 10.21037/jtd.2019.01.25.

[13] Y. Gao *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nat Commun*, vol. 11, no. 1, p. 5033, Dec. 2020, doi: 10.1038/s41467-020-18684-2.

[14] Z. Xiao, "COVID 19 Mortality Rate Prediction based on Machine Learning Methods," in *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, Fuzhou, China, Sep. 2021, pp. 169–177. doi: 10.1109/CEI52496.2021.9574541.

[15] S. S. L K, S. T. Ahmed, K. Anitha, and M. K. Pushpa, "COVID-19 Outbreak Based Coronary Heart Diseases (CHD) Prediction Using SVM and Risk Factor Validation," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Kuala Lumpur, Malaysia, Nov. 2021, pp. 1–5. doi: 10.1109/i-PACT52855.2021.9696656.

[16] Q. A. R. Adib, S. T. Tasmi, S. I. Bhuiyan, M. S. Raihan, and A. B. Shams, “Prediction Model for Mortality Analysis of Pregnant Women Affected With COVID-19,” in *2021 24th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, Dec. 2021, pp. 1–6. doi: 10.1109/ICCIT54785.2021.9689824.

[17] R. Mostafiz, M. S. Uddin, N.-A.- Alam, Md. Mahfuz Reza, and M. M. Rahman, “Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features,” *Journal of King Saud University - Computer and Information Sciences*, p. S1319157820306182, Dec. 2020, doi: 10.1016/j.jksuci.2020.12.010.

[18] S. A.-F. Sayed, A. M. Elkorany, and S. Sayed Mohammad, “Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity,” *IEEE Access*, vol. 9, pp. 135697–135707, 2021, doi: 10.1109/ACCESS.2021.3116067.

[19] Nikhil, A. Saini, S. Panday, and N. Gupta, “Polynomial Based Linear Regression Model to Predict COVID-19 Cases,” in *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, Bangalore, India, Aug. 2021, pp. 66–69. doi: 10.1109/RTEICT52294.2021.9574032.

[20] M. Sewariya and R. Katarya, “Parameter Based Literature Survey of COVID-19 Mortality Dynamics Using Machine Learning Techniques,” in *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, Bangalore, India, Aug. 2021, pp. 141–145. doi: 10.1109/RTEICT52294.2021.9573980.

[21] L. Yan *et al.*, “An interpretable mortality prediction model for COVID-19 patients,” *Nat Mach Intell*, vol. 2, no. 5, pp. 283–288, May 2020, doi: 10.1038/s42256-020-0180-7.

[22] Y. Yao *et al.*, “D-dimer as a biomarker for disease severity and mortality in COVID-19 patients: a case control study,” *j intensive care*, vol. 8, no. 1, p. 49, Dec. 2020, doi: 10.1186/s40560-020-00466-z.

[23] L.-L. Liang, C.-H. Tseng, H. J. Ho, and C.-Y. Wu, “Covid-19 mortality is negatively associated with test number and government effectiveness,” *Sci Rep*, vol. 10, no. 1, p. 12567, Dec. 2020, doi: 10.1038/s41598-020-68862-x.

[24] Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, “Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches,” *Chaos, Solitons & Fractals*, vol. 138, p. 110137, Sep. 2020, doi: 10.1016/j.chaos.2020.110137.

[25] Md Ehsanes Saleh AK, Mohammad Arashi BGK. Introduction to ridge regression. In:Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.; 2019.

[26] O. Steinki and Z. Mohammad, “Introduction to Ensemble Learning,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 2634092, Aug. 2015. doi: 10.2139/ssrn.2634092.

[27] W. Liang, S. Luo, G. Zhao, and H. Wu, “Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms,” *Mathematics*, vol. 8, no. 5, p. 765, May 2020, doi: 10.3390/math8050765.

[28] I. Gad and D. Hosahalli, “A comparative study of prediction and classification models on NCDC weather data,” *International Journal of Computers and Applications*, vol. 44, no. 5, pp. 414–425, May 2022, doi: 10.1080/1206212X.2020.1766769.

[29] D. Kumar, “Decision tree classifier: a detailed survey,” p. 24.

[30] F. Zhou *et al.*, “Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study,” *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020, doi: 10.1016/S0140-6736(20)30566-3.

[31] Z. Zheng *et al.*, “Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis,” *Journal of Infection*, vol. 81, no. 2, pp. e16–e25, Aug. 2020, doi: 10.1016/j.jinf.2020.04.021.

[32] W. Tian *et al.*, “Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis,” *J Med Virol*, vol. 92, no. 10, pp. 1875–1883, Oct. 2020, doi: 10.1002/jmv.26050.

[33] A. Mantovani, C. D. Byrne, M.-H. Zheng, and G. Targher, “Diabetes as a risk factor for greater COVID-19 severity and in-hospital death: A meta-analysis of observational studies,” *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, no. 8, pp. 1236–1248, Jul. 2020, doi: 10.1016/j.numecd.2020.05.014.

[34] A. Di Castelnuovo *et al.*, “Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study,” *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, no. 11, pp. 1899–1913, Oct. 2020, doi: 10.1016/j.numecd.2020.07.031.

[35] J. Wang *et al.*, “Thrombo-inflammatory features predicting mortality in patients with COVID-19: The FAD-85 score,” *J Int Med Res*, vol. 48, no. 9, p. 030006052095503, Sep. 2020, doi: 10.1177/0300060520955037.

[36] G. Torrats-Espinosa, “Using machine learning to estimate the effect of racial segregation on COVID-19 mortality in the United States,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 7, p. e2015577118, Feb. 2021, doi: 10.1073/pnas.2015577118.

[37] G. S. Erdal *et al.*, “The mortality rate of COVID-19 was high in cancer patients: a retrospective single-center study,” *Int J Clin Oncol*, vol. 26, no. 5, pp. 826–834, May 2021, doi: 10.1007/s10147-021-01863-6.

[38] B. Xu *et al.*, “Epidemiological data from the COVID-19 outbreak, real-time case information,” *Sci Data*, vol. 7, no. 1, p. 106, Dec. 2020, doi: 10.1038/s41597-020-0448-0.

[39] https://analyticsindiamag.com/when-to-use-one-hot-encoding-in-deep-earning/.

[40] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[41] X. Yan and N. A. Chowdhury, “A comparison between SVM and LSSVM in mid-term electricity market clearing price forecasting,” in *2013 26th IEEE*

*Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, May 2013, pp. 1–4. doi: 10.1109/CCECE.2013.6567685.

[42] A. Z. Woldaregay *et al.*, “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes,” *Artificial Intelligence in Medicine*, vol. 98, pp. 109–134, Jul. 2019, doi: 10.1016/j.artmed.2019.07.007.

[43] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, Oct. 2002, doi: 10.1016/S1532-0464(03)00034-0.

[44] S. Liu, H. Li, Y. Zhang, B. Zou, and J. Zhao, “Random forest-based track initiation method,” *The Journal of Engineering*, vol. 2019, no. 19, pp. 6175–6179, 2019, doi: 10.1049/joe.2019.0180.

[45] J. Huang, Y. Wei, J. Yi, and M. Liu, “An Improved kNN Based on Class Contribution and Feature Weighting,” in *2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Changsha, Feb. 2018, pp. 313–316. doi: 10.1109/ICMTMA.2018.00083.

[46] L. Yan *et al.*, “A machine learning-based model for survival prediction in patients with severe COVID-19 infection,” Epidemiology, preprint, Mar. 2020. doi: 10.1101/2020.02.27.20028027.

[47] R.-H. Du *et al.*, “Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study,” *Eur Respir J*, vol. 55, no. 5, p. 2000524, May 2020, doi: 10.1183/13993003.00524-2020.

# PUBLICATION

1. **Paper 1**

“**Parameter Based Literature Survey of COVID-19 Mortality Dynamics Using Machine Learning Techniques**”

**Presented and published** in:

*6th INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ELECTRONICS, INFORMATION & COMMUNICATION TECHNOLOGY RTEICT2021*, August 2021.

**Date of Conference**: 27-28 Aug. 2021

**ISBN Information**:

**Electronic ISBN**:978-1-6654-3559-8

**Print on Demand (PoD) ISBN**:978-1-6654-0254-5

**INSPEC Accession Number**: 21412057

**DOI**: 10.1109/RTEICT52294.2021.9573980

**Publisher**: IEEE

**Conference Location**: Bangalore, India

2. **Paper 2**

**“SARS-CoV-2 mortality risk prediction using machine learning algorithms to aid medical decision making”**

**Accepted** in *4th International Conference on Inventive Computation and Information Technologies ICICIT 2022.*