Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age.

The best way to prevent and slow down transmission is to be well informed about the disease and how the virus spreads. Protect yourself and others from infection by staying at least 1 metre apart from others, wearing a properly fitted mask, and washing your hands or using an alcohol-based rub frequently

The virus has hit 213 countries as of the third week of June 2022, infecting about 545,788,600 people (with active cases exceeding 17,758,752) and leading to a death toll of over 6,343,588 persons. Besides this, thanks to the limited resources available and the increased burden on the healthcare staff, patients infected with other diseases experienced delayed treatments. All these factors once again made us question even the nations with the best healthcare services around the world.

We have performed a comparative analysis on the six models trained and chosen the best one giving the results with high accuracy. Thus, selected model can be implemented at a good scale for effective screening and prioritization of testing for the virus in the general population.

Related Work

The human body is guarded by the immune system, but sometimes this system alone is not capable of preventing our body from diseases. Environmental conditions and living habits of individuals are the cause of many diseases

that are the main reasons for a huge number of deaths in the world, and diagnosing these diseases sometimes becomes challenging. We need accurate, feasible, reliable, and robust systems to diagnose diseases in time so that these can be properly treated. With the expansion of medical data, many researchers are using these medical data and a few machine learning algorithms to help the healthcare communities in the diagnosis of many diseases. Numerous studies have been done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. Continuous growth in medical data gave us how to extract the required information to predict the disease. Health data collected from patients can be used to predict various diseases with the help of modern techniques of data science and big data. These disease prediction models are vital to knowing the presence of disease. Various machine learning techniques like supervised, semi-supervised, unsupervised learning, etc. and raw medical data are required to detect different diseases. This data could easily be obtained from famous government hospitals. Machine learning techniques can use the data for the learning process and based on that learning they can predict the disease later. There are many literature reviews available in Disease Prediction. A study from Infectious Diseases of Poverty shows that machine learning techniques are often used to predict the severity of COVID-19, thereby enabling providers to optimize care. [1] AI and machine learning are often used to examine a person for COVID-19 as an alternative to traditional time- consuming and expensive methods. Although there are several studies on COVID-19, this study concentrated on the use of machine learning in forecasting COVID-19 cases and diagnosing patients for COVID-19 infection through their symptoms. Various machine learning models and apps have been made in the field of covid-19 prediction.

Mikko Vihtakari (May 2020)[6] made an app licensed by MIT License that illustrates how COVID-19 infection could develop in your country and why the drastic measures to fight the outbreak are justified. The model uses simple exponential math, and median estimates and ignores an entire lot of important parameters, like reporting error, development of immunity, population density, demography, variation, and uncertainty. Consequently, the model isn't accurate but gives an idea of how the outbreak could develop during the uncontrolled start phase most European countries have been going through in March 2020. The model parameters are adjusted for the situation in Norway 2020-03-17.

In [7], the authors developed App-based COVID-19 syndromic surveillance and prediction of hospital admissins: The COVID Symptom Study Sweden. The app-based COVID Symptom Study was launched in Sweden in April 2020 to contribute to real-time COVID-19 surveillance. Data from 19,161 self-reported PCR tests were wont to create a symptom-based model to estimate the individual probability of symptomatic COVID-19, with an AUC of 0.78 (95% CI 0.740.83) in an external dataset. The individual probabilities were wont to estimate daily

regional COVID-19 prevalence, which was successively used together with current hospital data to predict next week's COVID-19 hospital admissions. It had been found that this hospital prediction model demonstrated a lower median absolute percentage error (MdAPE: 25.9%) across the five most populated regions in Sweden during the primary pandemic wave than a model based on case notifications (MdAPE: 30.3%). Identical error rates were found during the second wave.

Ramesh Kumar Mojjada, Arvind Yadav, A.V. Prabhu and Yuvaraj Natarajanc (2020)[8] made a study showing the power to predict the number of individuals who are affected by COVID-19 as a potential threat to human beings by ML modelling. During this analysis, the danger factors of COVID-19 were exponential smoothing (ES). The Lower Absolute Reductor and Selection Operator, (LASSo), Vector Assistance (SVM), and four normal potential forecasts, like Linear Regression (LR)). Each of those machine-learning models has three distinct kinds of predictions: the number of newly infected COVID 19 people, mortality rates and therefore the recovered COVID- 19 estimates in the next 10 days.

Wie Kiang H. (2020)[9] in his article focussed on how machine learning is often used to study the spread of covid-

19. The dataset was retrieved from the official repository of Johns Hopkins University. This data consists of daily case reports and daily statistic summary tables. Within the study, they need selected time-series summary tables in CSV format having three tables for confirmed, death, and recovered cases of COVID-19 with six properties. In concurrent, state-of-the-art mathematical models were chosen to support machine learning for a computational process to predict the spread of the virus, for instance: Support Vector Regression (SVR), Polynomial Regression (PR) and Deep Learning regression models. It also involved Artificial Neural Network (ANN) and Recurrent Neural Networks (RNN) using Long STM (LSTM) cells.

Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R. Sujatha, JyotirMoy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai and Ohyun Jo (July 2020) [10] worked on COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. They used boosted random forest for prediction. Boosted Random Forest is an algorithm, which consists of two parts; the boosting algorithm: AdaBoost and therefore the Random Forest classifier algorithm, which successively consists of multiple decision trees. The model uses the COVID-19 patient's geographical, travel, health, and demographic data to predict the severity of the case and therefore the possible outcome, recovery, or death. The model gave results with an accuracy of 94% and an F1 score of 0.86 on the dataset used. It was observed that the patients gender and deaths are positively correlated and that the bulk of patients is aged between 20 and 70 years.

M. Shobana, S. Vaishnavi, C. Gokul Prasad, P. Poonkodi,

R. Sabitha, and S. Karthik (2022)[12] published a piece of

writing in which they worked on Relating Design Thinking Framework in Predicting the Spread of COVID in Tamilnadu Using ARIMA. ARIMA (Auto Regressive Integrated Moving Average) models are often used in forecasting the spread of COVID with the previous datasets extracted from Kaggle. They have considered the info from March 2020 to June 2021 and predicted the COVID cases for the next one month July 2021. In specific, it's concentrated in one particular state Tamilnadu from INDIA.

Ismail Kirbas, A. SÃ¶zen, Azim Dou Tuncer, F. . Kazancolu (June 2020)[13] performed a Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approach. during this study, confirmed COVID-19 cases in Denmark, Belgium, Germany, France, UK, Finland, Switzerland and Turkey were modelled with Auto- Regressive Integrated Moving Average (ARIMA), Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM) approach. Six model performance metrics were wont to select the most accurate model (MSE, PSNR, RMSE, NRMSE, MAPE and

SMAPE). consistent with the results of the first step of the study, LSTM was found to be the foremost accurate model. Within the second stage of the study, the LSTM model was provided to form predictions in a 14-day perspective that is yet to be known. Results of the second step of the study show that the entire cumulative case increase rate is expected to decrease slightly in many countries.

Yasminah Alali, Fouzi Harrou & Ying Sun (February 2022)[14] develop an assumption-free data-driven model to accurately forecast the COVID-19 spread. They started with Bayesian optimization to tune the Gaussian process regression (GPR) hyperparameters to develop a GPR-based model to forecast the recovered and confirmed COVID-19 cases in two highly impacted countries, India and Brazil. However, machine learning models don't consider the time dependency in the COVID-19 data series. Here, dynamic information has been taken under consideration to alleviate this limitation by introducing lagged measurements in constructing the investigated machine learning models. They also assessed the contribution of the incorporated features to the COVID-19 prediction using the Random Forest algorithm. Their results highlighted the superior performance of the dynamic GPR compared to the opposite models (i.e., Support vector regression, Boosted trees, Bagged trees, Decision tree, Random Forest, and XGBoost) and procured an averaged mean absolute percentage error of around

0.1%. They provided the arrogance level of the predicted results based on the dynamic GPR model and showed that the predictions are within the 95% confidence interval.

Methology

**1. Data Preprocessing**

This consists of two steps, i.e., Data Collection and Data Pre-processing. Data can be referred to as the raw material. Therefore, the first step in the development of COVID-19 applications is data collection. Multiple datasets are put online in regards to COVID-19. Most if not all of those datasets are open source meaning that they are free for anyone to download and use. The dataset that we are using is formed using 8 attributes that were noted in 278848 Israeli patients and was gathered by the Israeli Ministry of Health. The dataset contains initial records, on a day to day basis, of all the residents who were tested for COVID-19 nationwide. The dataset's attributes include cough, fever, pharyngitis, shortness of breath, headache, corona result, age 60 and above, gender, and test indication. The corona result tells whether or not people may have the coronavirus in their bodies. The dataset's majority of variables are in binary format. If a feature variable's value is "1," it signifies that a specific symptom is present; if it is "0," there's no symptom. The symptoms taken are supported by guidelines given by the World Health Organization (WHO) and the Ministry of Health and Family Welfare, India.

**The following attributes describe each of the datasets features used by the model:**

1.Sex- This feature is assessed into three categories- male, fmale and none. After data pre- processing, the specific data has been converted to numerical data. All the male categories are replaced with 0.0 and feminine ones with 1.0. The third category, i.e., none has been replaced with the amount 3.0.

2.Cough\_symptoms-This feature is assessed into two categories, presence and absence.0 indicates the absence of cough and 1 indicates the presence of cough. After data pre-processing 0 is replaced with 0.0 and 1 is replaced with 1.0.

3.Fever–This feature is assessed into two categories, presence and absence of fever .0 indicates the absence of fever and 1 indicates the presence of

4.fever. After data pre-processing, 0 is replaced with 0.0 and 1 is replaced with 1.0.

5.Sore\_throat-This feature is assessed into two categories, presence and absence of sore\_throat .0 indicate the absence of sore\_throat and 1 indicates the presence of sore\_throat. After data pre-processing, 0 is replaced with 0.0 and 1 is replaced with 1.0.

6.Shortness\_of\_breath -This feature is assessed into two categories, presence and absence of shortness\_of\_breath.0 indicates the absence of it and 1 indicates presence. After data preprocessing, 0 is replaced with 0.0 and 1 is replaced with 1.0.

7.headache -This feature is assessed into two categories, presence and absence of headache.0 indicates the absence of head\_ache and 1 indicates the presence of

headache. After data preprocessing, 0 is replaced with 0.0 and 1 is replaced with 1.0.

8.Age\_60\_and\_above -This feature is assessed into three categories -None No, Yes. No indicates that the age is below 60 and Yes indicates that the age is above 60.

9.known\_contact-This feature is assessed into three categories-Other, Abroad, Contact with Confirmed. This indicates whether an individual has come into contact with covid positive, he/she has come from abroad or there's any other reason for existing symptoms

10.Corona-This feature is assessed into three categories -Positive, Negative, and Other. Positive indicates that the individual is covid positive, negative indicates that the individual isn't covid positive, other indicates that there's no surety about the result it can be some other allergy also.

Data preprocessing can be defined as a process of preparing the raw data and making it suitable for a machine learning model. it's the first and crucial step while creating a machine learning model.

Steps that were followed during data pre-processing are:

**Getting the Dataset**

To create a machine learning model, we require a dataset as a machine learning model completely works on data. The data that was used to train and test the machine learning models was retrieved from the website of the Israeli Ministry of Health. Furthermore, the machine learning models were tested on another set of data collected through a survey done on the Indian population.

**Importing the libraries**

Various predefined python libraries were used for data pre- processing. Some of the libraries used are Numpy, Pandas, Seaborn, Sklearn, pickle etc.

**Importing the Datasets**

For performing on datasets collected for machine learning models, the present directory was set to the working directory. Then the datasets were imported. To import the dataset, the read\_csv() function of the pandas library was used, which may read a CSV file and perform various operations on it. With this function, CSV files are often read both locally as well as through URL

**Handling Missing data**

If the dataset contains some missing data, then it's going to create a huge problem for our machine learning model. Therefore, it's required to handle missing values present in the dataset. The next step that was followed in data pre- processing was handling the missing data. The process that

was followed was deleting the rows or columns having null values. If columns have quite half of the rows as null, then the entire column can be dropped. The rows which are having one or more column values as null also can be dropped.

ENCODING CATEGORICAL DATA

If there are categorical variables, it can cause trouble in building the model because the machine learning model completely works on mathematics and numbers. Therefore, the specific variables were encoded into numbers using replace function. The categorical variables were converted to the following numerical values

"No" to value=0.0 "Yes" to value=1.0 "0"to value=0.0 "1", value=1.0 "Male" to value=0.0

"Female" to value=1.0

"Other" to value=1.0 "Abroad", value=2.0

"Contact with confirmed" to value=3.0 other to value 2.0

**Splitting the Dataset into the Training set and Test set**

Splitting the dataset into a training and test set is important because if we train the models on a certain dataset and test the models on a completely different dataset then it will be difficult for our model to understand the correlations between the models. If we train our model alright and its training accuracy is also very high, but we offer a new dataset to it, then it'll decrease the performance. So we always attempt to make a machine learning model which performs well with the training set and also with the test dataset.

The Israeli dataset[3] was divided into 75% of the dataset as the training set and 25% as the test set. Training Set can be described as a subset of dataset to coach the machine learning model, and we already know the output. Test set can be defined as a subset of the dataset to check the machine learning model, and by using the test set, the model predicts the output.

Development of the machine learning models

Random Forest Classifier : It's a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and supports the majority votes of predictions, it predicts the ultimate output.

K Nearest Neighbor : It's one of the simplest Machine Learning algorithms based on the Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that's most similar to the

available categories. The K-NN algorithm is usually used for Classification problems. K-NN may be a non- parametric algorithm, which suggests it does not make any assumption on underlying data. It's also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

Gradient Boosting: This algorithm is one of the most powerful algorithms in the field of machine learning. The errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize the bias error of the model. t is often used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it's used as a classifier then the cost function is Log loss.

Logistic Regression : It's one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. it's used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression can work on categorical variables. The results are often either Yes or No, 0 or 1, True or False, etc. but rather than giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. The values from 0.5 to 1 are

often considered as 1 and 0 below 0.5.

Naive Bayes : It's one of the fast and easy ML algorithms to predict a class of datasets. Naive Bayes may be a generative model. It's a probabilistic classifier, which suggests it predicts based on the probability of an object. We've used Gaussian Naive Bayes in our work. (Gaussian) Naive Bayes assumes that every class follows a Gaussian distribution.

Support Vector Machine : It's a very popular Supervised Learning algorithm; it is employed for Classification as well as Regression problems. The goal of the SVM algorithm is to make the best line or decision boundary that can divide n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is named a hyperplane. SVM chooses the acute points/vectors that help in creating the hyperplane. These extreme cases are called support vectors.

Dataset was splitted in 75% and 25% for training and testing purposes respectively. After training the models they were also tested on the surveyed data that is on Indian population. Predictions were made using the model giving highest accuracy on Indian data.

Evaluation of the machine learning models

The models are evaluated using accuracy score, confusion matrix and classification report back to assess the reliability of the proposed machine learning models. The results of these measures are compared and the model with the best results in all the aspects has been chosen for the integration purpose in the android application. These metrics are calculated on the idea of the following:

True Positives(TP)- This is often the portion of the dataset in which the patients who were covid positive were correctly identified by the model.

True Negatives(TN)- this is often the portion of the dataset in which the patients who were covid negative were correctly identified as negative by the model.

False Positives(FP)- This is often the portion of the dataset in which the patients who were covid negative were incorrectly identified as positive by the model.

False Negatives(FN)- This is often the portion of the dataset in which the patients who were covid positive were incorrectly identified as negative by the model.

Confusion matrix

A confusion matrix may be a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. From our confusion matrix, we can calculate five different metrics measuring the validity of our model.

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN

Precision (true positives / predicted positives) = TP / TP

+ FP

Sensitivity/ Recall (true positives / all actual positives) = TP / TP + FN

Specificity (true negatives / all actual negatives) =TN / TN + FP

A classification report is also a performance evaluation metric in machine learning. It's used to show the precision, recall, F1 Score, and support of the trained classification model.

Accuracy score: This measures the share of correctly identified cases relative to the entire dataset. The ML algorithm performs better if the accuracy is higher.

Precision: This metric measures the exactness, which may be computed as the ratio of true positives to the sum of true and false positives.

Recall: This metric may be a measure of completeness, which may be computed as the ratio of true positives to the sum of true positives and false negatives.

F1 Score: It is often described as the weighted harmonic

mean of precision and recall. The closer the worth of the F1 score is to 1.0, the higher the expected performance of the model is.

Support: It is often described as the weighted harmonic mean of precision and recall. The closer the worth of the F1 score is to 1.0, the higher the expected performance of the model.

Results