Sains Malaysiana 50(8)(2021): 2479-2497
http://doi.org/10.17576/jsm-2021-5008-28

Prediction of COVID-19 Patient using Supervised Machine Learning Algorithm

(Ramalan Pesakit COVID-19 menggunakan Algoritma Pembelajaran Mesin Diselia)

BUVANA, M.* & MUTHUMAYIL, K.

ABSTRACT

One of the most symptomatic diseases is COVID-19. Early and precise physiological measurement-based prediction of breathing will minimize the risk of COVID-19 by a reasonable distance from anyone; wearing a mask, cleanliness, medication, balanced diet, and if not well stay safe at home. To evaluate the collected datasets of COVID-19 prediction, five machine learning classifiers were used: Nave Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbour (KNN), and Decision Tree. COVID-19 datasets from the Repository were combined and re-examined to remove incomplete entries, and a total of 2500 cases were utilized in this study. Features of fever, body pain, runny nose, difficulty in breathing, shore throat, and nasal congestion, are considered to be the most important differences between patients who have COVID-19s and those who do not. We exhibit the prediction functionality of five machine learning classifiers. A publicly available data set was used to train and assess the model. With an overall accuracy of 99.88 percent, the ensemble model is performed commendably. When compared to the existing methods and studies, the proposed model is performed better. As a result, the model presented is trustworthy and can be used to screen COVID-19 patients timely, efficiently.

Keywords: Classifier; COVID-19; machine learning; prediction; supervised learning

ABSTRAK

Salah satu penyakit yang paling simptomatik ialah COVID-19. Ramalan pernafasan berdasarkan pengukuran fisiologi awal dan tepat akan meminimumkan risiko COVID-19 dengan jarak yang munasabah daripada sesiapa sahaja; memakai topeng, kebersihan, ubat-ubatan, diet seimbang dan jika tidak sihat, tinggal di rumah. Untuk menilai kumpulan data ramalan COVID-19 yang dikumpulkan, lima pengkelasan pembelajaran mesin digunakan: Nave Bayes, Mesin Vektor Sokongan (SVM), Regresi Logistik, Jiran K-Terdekat (KNN) dan Pohon Keputusan. Set data COVID-19 daripada Repositori digabungkan dan disemak semula untuk menghapus entri yang tidak lengkap dan sejumlah 2500 kes digunakan dalam kajian ini. Ciri demam, sakit badan, hidung berair, kesukaran bernafas, sakit tekak dan hidung tersumbat, dianggap sebagai perbezaan yang paling penting antara pesakit yang menghidap COVID-19 dan mereka yang tidak. Kami menunjukkan fungsi ramalan lima pengelasan pembelajaran mesin. Satu set data yang tersedia untuk umum digunakan untuk melatih dan menilai model. Dengan ketepatan keseluruhan 99.88 peratus, model ensembel dilakukan dengan terpuji. Jika dibandingkan dengan kaedah dan kajian yang ada, model yang dicadangkan dilakukan dengan lebih baik. Hasilnya, model yang dipersembahkan boleh dipercayai dan dapat digunakan untuk menyaring pesakit COVID-19 tepat pada waktunya.

Kata kunci: COVID-19; pembelajaran mesin; pembelajaran yang diselia; pengelas; ramalan

INTRODUCTION

The coronavirus disease of 2019 (COVID-19) is an infectious disease caused by the virus strain ‘severe acute respiratory syndrome coronavirus 2’ (SARS-CoV-2) (Sharma et al. 2020). The 2019 Coronavirus disease (COVID-19) pandemic that originated in Wuhan, China has had devastating impacts on the global population and has overwhelmed advanced healthcare systems worldwide.

The current rapid and exponential increase in the number of patients required an effective and rapid prediction of an infected patient's potential outcome for adequate care using Machine Learning (ML) techniques.

The model utilizes the training data from symptoms of previous cases of the COVID-19 patient to predict the seriousness of the case and the potential outcome. Generally, the infection is transmitted from one person to another through respiratory droplets formed during coughing and sneezing. In a broad spectrum, the time between exposure and onset of symptoms is between 2 and 14 days, with an average of 5 days. The 2019-2020 coronavirus epidemic was declared a pandemic and a public health emergency of international significance (PHEIC) by the World Health Organization (WHO). Since there is no cure for COVID-19, no drug or vaccine has been developed. Several non-medical organizations have stated that they are working on a vaccine to eradicate the virus. During this pandemic, we lack medical resources, which has resulted in less tests. It is impossible to test all of the people who have symptoms because hundreds to thousands of people are being tested positive every day around the world.

The ML technique begins by gathering information independently, that is, from an assortment of resources (Jackins et al. 2021). From that point onward, the following move is to address the pre-handled information to address information related issues and lessen space size (Furqan et al. 2020) by eliminating invalid document information to choose fascinating information. However, for the system to make decisions on the relevance of the dataset can also be very important, therefore, machine learning algorithms are designed to analyze data and derive useful and new information or unknown patterns or data from past occurrences using other concepts (Jackins et al. 2021). The next step is model performance assessment and, finally, model optimization using fresh datasets and rules to develop the model. In a variety of fields, ML strategies are utilized, including medical services, fabricating, tutoring, development and creation, anticipating, and traffic the board and robotics.

The main objective of this research work was to select from a collection of high-dimensional features which is the most important characteristics and simultaneously increase the efficiency of a classifier with decreased computational time. This paper's primary contribution is outlined as follows. 1. Initially, the disease dataset is taken as an input to the method proposed. Real-time COVID-19 datasets are taken for study, so much information in the COVID-19 patient symptom dataset is linked to patient health care and general information. 2. Data preprocessing is pertained to the input datasets, contributing to a reduction of unwanted data for further study. For efficient data analysis, it helps to divide the training data into 80 percent original data and 20 percent testing data to verify the omitted values and verify the correlation. 3. To estimate the system's output against the input dataset, machine learning algorithms are applied. 4. To evaluate the results produced in objective 1,2,3 and compare these results with existing approaches.

Next section describes the literature survey, subsequent section proposes the materials and techniques used in detail, as well as the description, preprocessing, and classification algorithms used, the following section reflects the results and discussions of the experiments, followed by the suggested metrics with random forest, and last section proposes the conclusion and future work of the current research with references.

LITERATURE REVIEWS

Several researchers have contributed to the fields where the pandemic was predicted. The significant impact that Coronavirus2 caused on China and its further spread in Italy was addressed by Remuzzi and Remuzzi (2020). They developed a predictive model that helps to explain the growth of the patient and this will help medical facilities make choices in turn.

A related study (Roosa et al. 2020) used validated phenomenological models to predict the number of cases registered in the province of Hubei. They used Chinese national data to provide forecasts for days. Their model shows that the transmission is reduced by the containment strategies introduced in China and the pandemic has slowed down periodically. Working in the field, Ayyoubzadeh et al. (2020) carried out a pilot study by using 2019-nCoV occurrence to estimate Trend data on Google in Iran. The number of positive cases from 2019-nCoV was estimated using long-term models. Assessed with the Root Mean Square Error (RMSE) metric and Cross-validation methods with 10 folds were used to create the models. There were 27.187 and 7.562 models for RMSE and Linear Regression, respectively. Furthermore, the analysis predicted the 2019-nCoV outbreak pattern.

These forecasts will aid healthcare managers and planners in allocating effective health-care resources.

Using only clinical variables widely documented, ML models were used to predict the presence of COVID-19. Using these health factors, the author was able to distinguish COVID-19 patients from Influenza and patients who were not infected with COVID-19. These machine learning models were created with the help of publicly available data from published literature. Also, and Co relational research have been carried out on this dataset and calculated. Ud Din Khanday et al. (2020) developed a model based on 212 clinical reports divided into four categories: COVID, ARDS, SARS and both (COVID, ARDS). TF/IDF, bag of words, and other features are extracted. The algorithms used for machine learning are used to classify the clinical records into four distinct groups. After performing the classification, it was logistically shown. The Bayesian classifier offers regression and multinomial Naïve Excellent results by getting 94% accuracy, 96% recall, 95% recall 96.2% f1 score and precision. Various extra machines Random learning algorithms that have shown stronger results have been Forest, improving stochastic gradient, decision trees, and decision trees stimulating. The productivity of models can be increased by increasing the amount of information.

In Dharshana et al. (2020), the sickness prediction is finished through the usage of EEG signal, seeing that the EEG is economical and environment friendly to analyse. The frequency bands are utilised to pinpoint the hazard of occurrence of the disease. The data is collected and then pre-processed to make it suitable for implementing for the prediction. The pre-processed data will be split based on the frequency bands using Fast Fourier Transform (FFT). Different frequency bands like alpha, beta, theta, delta, and gamma are separated from the preprocessed data. The power spectrum and single-sided amplitude of respective signals are used to segregate the different stages. The optimised features are fed to Convolutional Neural Network (CNN) from which the values are learned by the neural network and it is classified into different stages like Normal, Mild, and Severe (Sarwar et al. 2020). The result of the identification of diabetes showed that the ensemble methods ensured 98.60% accuracy.

While Yan et al. (2020) introduced an LSTM model to predict the country-specific risk of COVID-19, it relies on a country's patterns and weather data to predict the likely spread of COVID-19 in that nation. In Wang et al. (2020), researchers added a mannequin based totally on the Convolutionary Neural Network to become aware of sufferers with COVID-19 with the use of CXR images. They used and educated the mannequin the use of a pre-trained ImageNet Chest X-Ray pictures on an open-source dataset (CXR).

ML techniques and other artificial intelligence methods (Muhammad et al. 2021) have played many significant roles in the prediction, detection, and containment of the COVID-19 pandemic, according to the relevant works examined so far, can help reduce the enormous strain on small healthcare systems. To the best of our knowledge, no work has been done to date; labeled datasets have been reported to establish supervised ML models for COVID-19 infection prediction for positive and negative COVID-19 cases. Therefore, the proposed work aims to investigate these variations.

Materials and Methods

The following packages and libraries are needed for the proposed model: Datetime, Matplotlib Numpy, Scikit Learn, Seaborn, and Pandas. The proposed model was implemented using the Google Colab platform. The strategy on how to improve the supervised Machine learning models for the COVID-19 prediction infection in this work using the COVID-19 dataset (Dataset) has been displayed in Table 1.

In this paper, a proposed method is used in order to classify and predict the COVID-19 and non-COVID-19 patients employing the datasets. The basic structure of the model is shown in Figure 1. In the end, the pre-processing of the data is mandatory. First, the data is normalized by randomly dividing the raw data on maximum value for any raw data. As a result, all data exhibited in a specific range 0 to 1. In the next step, feature selection is carried out. There are some outliers in the dataset. In order to remove the outliers, some features of the dataset are removed. For this reason, ExtraTreeClassifier is used. At this point, the dataset is separated into train dataset and test dataset to maintain a strategic distance from any inclination in training and testing. From the information, 80% of the data was utilized for train the ML model and the excess utilized for testing the performance of the proposed action.

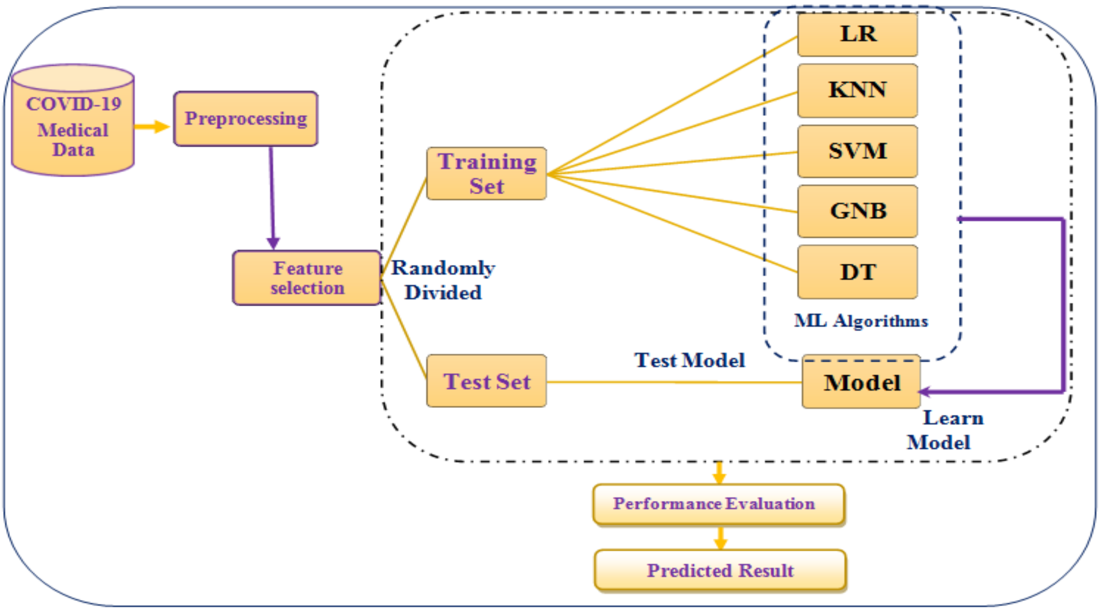

FIGURE 1. Structure of six machine learning model for predicting COVID-19 disease

DATASET

The model utilizes the training data from symptoms of previous cases of corona virus that have occurred in different nations around the world to predict whether an individual is COVID-19 positive or not. The model takes the health specifics of an individual as feedback to predict whether the person is corona-positive or not. The COVID-19 Medical dataset used in the study has been obtained from the GitHub repository provided by (Dataset). The dataset includes research findings in different countries for COVID-19 cases. The dataset includes demographic and clinical data for COVID-19 in patients with a diagnosis of viral respiratory function and contains 2500 cases or records of 12 characteristics. Fever, runny nose, body pain, Runny_nose, Nasal_congestion, sore throat, and severity to breathing were the most well-known manifestations that were seen in patients whose information is accessible in this dataset and are given in Table 1.

TABLE 1. Patients information dataset

Country	Age	Gender	Fever	Bodypain	Runny_ nose	Difficulty_ in_ breathing	Nasal_ congestion	Sore_ throat	Severity	Contact_ with_ covid_ patient	Infected
China	10	Male	102	1	0	0	0	1	Mild	No	0
Italy	20	Male	103	1	1	0	0	0	Moderate	Not known	1

DATA ANALYSIS AND PRE-PROCESSING

The dataset consists of columns of String, Object, and Numeric sort with the results. The categorical factors those are included in the dataset. Since the ML model need all of the information passed as input to be in numerical form, we performed label-encoding of the categorical variables. This assigns a number to each unique categorical value in a table. If the dataset contains many missing values and is passed directly as input, an error occurs. Transformation means shifting the format of data from one type to another, making it more comprehensible by normalising, smoothing, and generalising data aggregation techniques. Country, Gender, Severity, Contact_with_covid_patient are largely categorical features in the dataset. These features are commonly put away as text esteems which address different perceptions. For instance, Gender is depicted as male or Female or transsexual. These sort of feature where the classes are just named with no request for priority or which have some request related with them. Table 2 shows the Categorical-Categorical value of Gender with Contact_with_covidpatient.

TABLE 2. Categorical-Categorical

Gender vs Contact_with_covidpatient	0	1	2
0	502	353	353
1	667	310	280
2	34	0	0

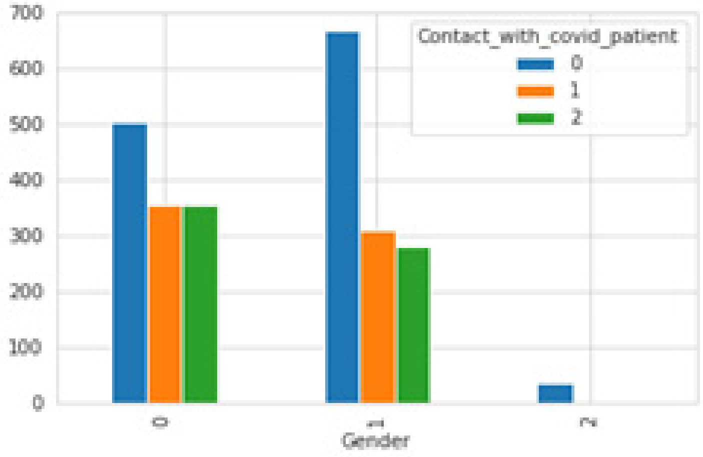

FIGURE 2. Patient's gender frequency

Similarly, we encoded ‘N’:0, ‘Yes’:1, ‘Not known’:2, ‘yes’:1 for for Contact_with_covid_patient for this work across the dataset instances, including the feature Severity was encoded with ‘Mild’:0,’Moderate’:1,’ Severe’:2. The dataset has no missing values. The definition of the dataset is shown in Table 3; the Categorical-Continous for Contact_with_covid_patient and Difficulty_in_breathing is shown in Figures 3 and 4, respectively. The chart presentation of the dataset’s profile information is shown in the chart Figure 5. Figure 6 indicates the patients’ age frequency, Figure 7 indicates the patients’ sex frequency and Figure 6 illustrates the frequency of test results for COVID-19. Table 4 shows the replacement of continuous values Sample of Normalized Covid-19 Dataset.

TABLE 3. Definition of COVID-19 dataset

#	Feature	Non-Null Count	Dtyp
0	Country	2499 Non-Null	object
1	Age	2499 Non-Null	int64
2	Gender	2499 Non-Null	object
3	Fever	2499 Non-Null	int64
4	Bodypain	2499 Non-Null	int64
5	Runny_nose	2499 Non-Null	int64
6	Difficulty_in_breathing	2499 Non-Null	int64
7	Nasal_congestion	2499 Non-Null	int64
8	Sore_throat	2499 Non-Null	int64
9	Severity	2499 Non-Null	object
10	Contact_with_covid_patient	2499 Non-Null	object
11	Infected	2499 Non-Null	int64

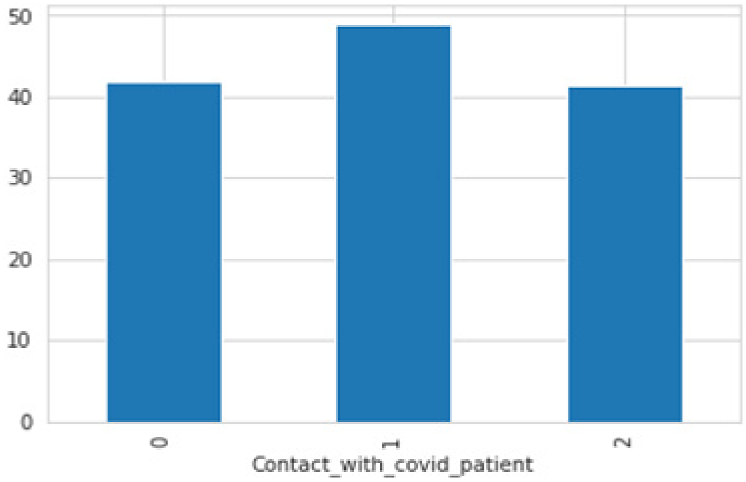

FIGURE 3. Categorical-Continous for Contact_with_covid_patient

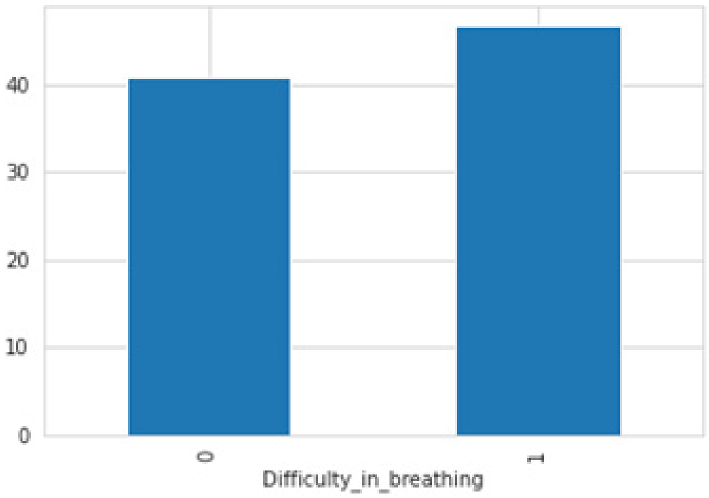

FIGURE 4. Categorical-Continous for Difficult_in_breathing

TABLE 4. Sample of normalized COVID-19 dataset

	Country	Age	Gender	Fever	Bodypain	Runny_ nose	Difficulty_ in_ breathing	Nasal_ congestion	Sore_ throat	Severity	Contact_ with_ covid_ patient	Infected
0	16	10	1	102	1	0	0	0	1	0	0	0
1	41	20	1	103	1	1	0	0	0	1	2	1
2	37	55	2	99	0	0	0	1	1	2	0	0
3	69	37	0	100	0	1	1	0	0	0	1	1
4	27	45	1	101	1	1	1	1	0	1	1	1

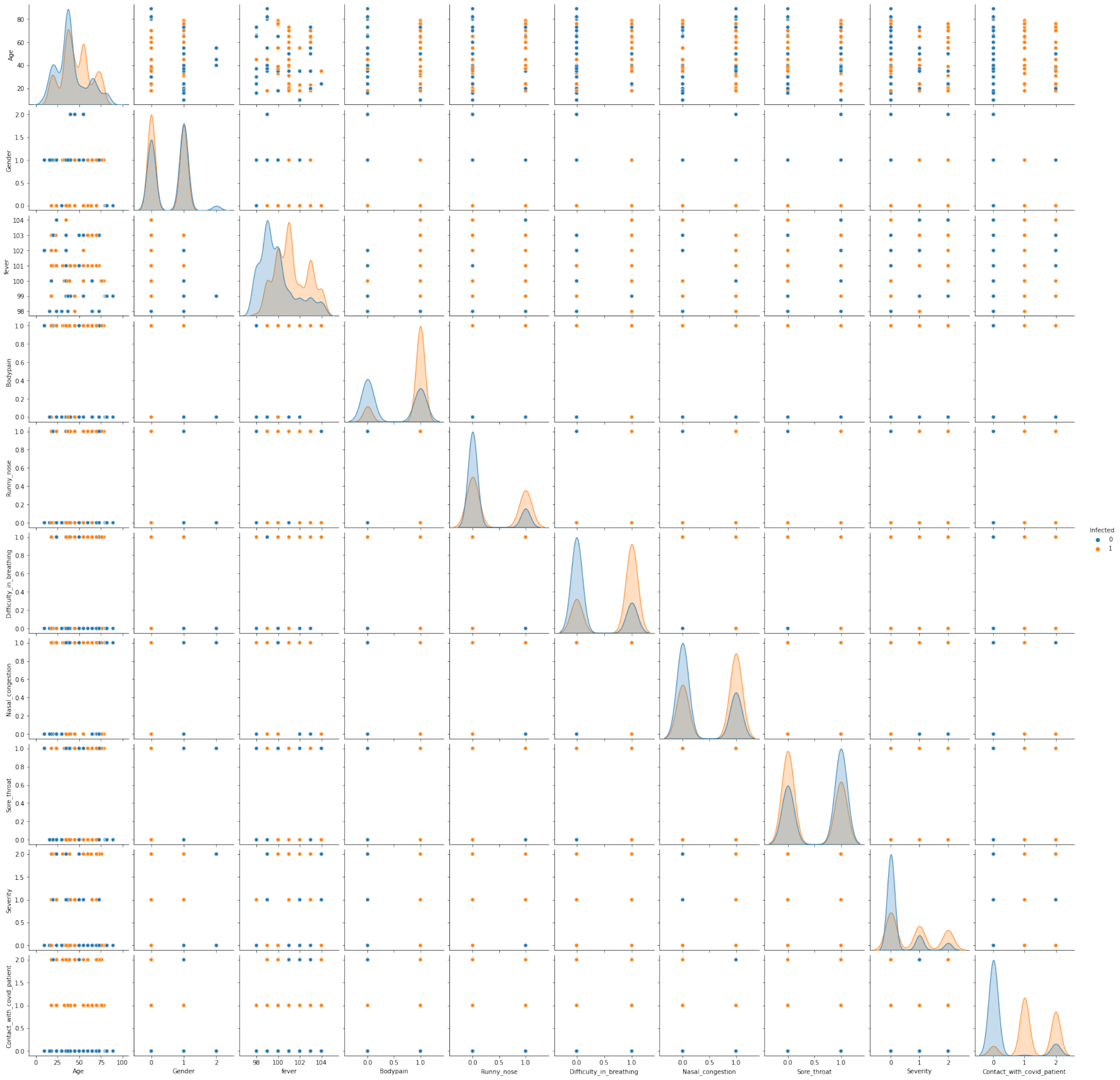

FIGURE 5. Pairplot presentation of the profile information of the dataset

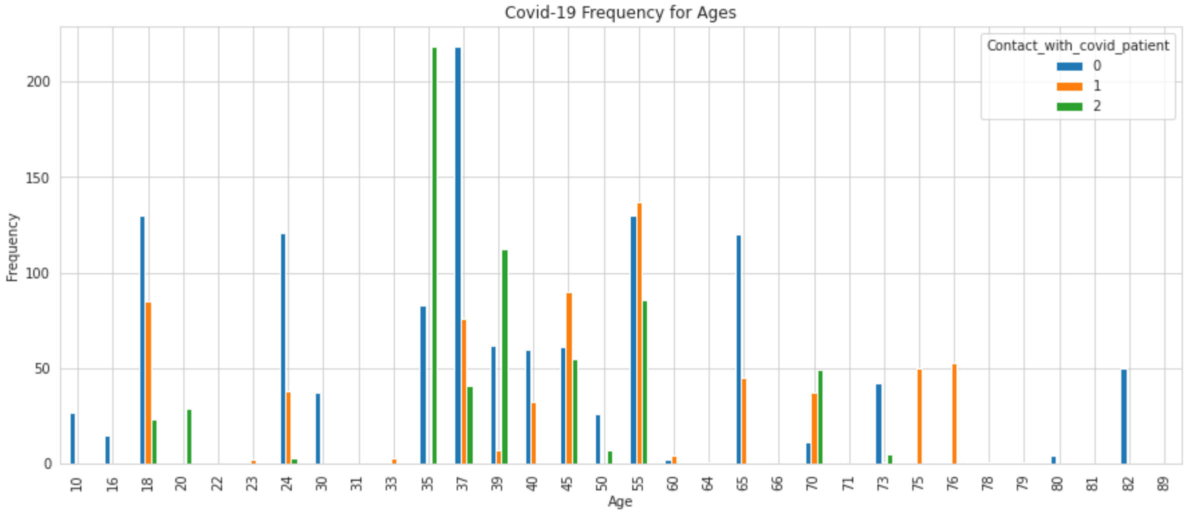

FIGURE 6. Age frequency of the COVID-19 patients

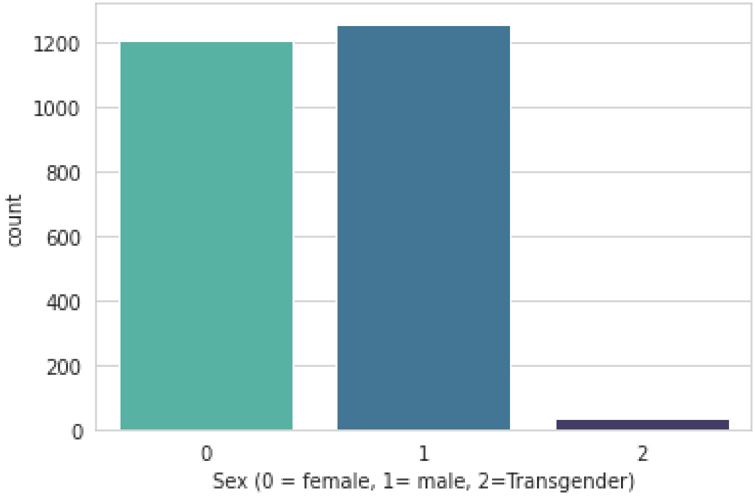

FIGURE 7. Gender wise count

FEATURE SELECTION

This step involves extracting and selecting features to get the best results from our model. Getting nice and best features helps us to demonstrate the information's underlying structure. ExtraTreesClassifier (ExtraTreesClassifier) starts with all features contained in the dataset. Then, runs a model and calculates the importance level associated with each feature. By utilizing that significance of highlights may have various qualities in view of the arbitrary idea of highlights. Each component is requested in sliding request as per the Importance of each element and afterward we choose the top highlights like Contact_with_covid_patient, Difficulty_in_breathing, Bodypain, and fever. The most important and best feature is to determine the output label according to the ExtraTreeClassifier is the 'Contact_with_covid_patient'. The best function to get more precise prediction was established in order to draw better conclusions from this dataset (Table 5).

TABLE 5. Feature importance

9	Contact_with_covid_patient	2210.09
0	Age	570.8576
4	Runny_nose	371.1826
8	Severity	347.1238
5	Difficulty_in_breathing	272.0271
6	Nasal_congestion	120.0951
7	Sore_throat	105.311
3	Bodypain	95.85706
1	Gender	28.9237

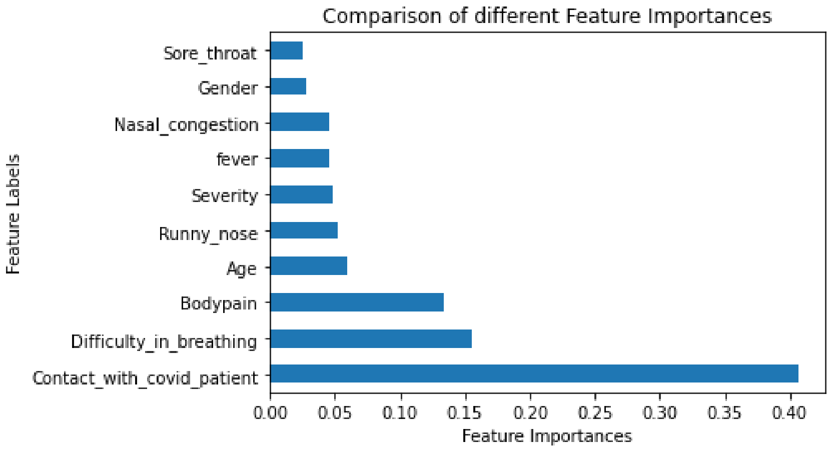

FIGURE 8. COVID-19 dataset feature selection

The data is being part into 80:20 proportions where 80% data is being utilized for training or preparing the model and 20% is utilized for testing the model. The classification was accomplished utilizing Machine learning algorithms. We break our underlying dataset into various training and test subsets to investigate the speculation of our model from preparing information to concealed information and lessen the danger of over fitting. At that point; the standardized information is arranged by a proposed method. A 10-cross validation technique is utilized for the training set to assess the best classifier and to help the precision of the model, and the test set information isn't utilized during the training process. The training various models of ML classification utilized in this exploration consists of: Logistic Regression, SVM classifier, DT Classifier, GNB Classifier, and KNN Classifier. Since the dataset we utilized could be an imbalanced dataset, the essential measurement for examination is F1 Score, Accuracy, Precision and Recall.

CORRELATION COEFFICIENT ANALYSIS

Using the Pearson heat map the correlation between the data set features is shown in Figure 8, which clearly indicates a comparatively stronger positive correlation between the fever, Body Pain, Sore Throat when the symptoms were first felt and visited the hospital, and infected. The correlogram depicts the entire set of variables and correlations. Positive and negative statistics, as well as correlational statistics, are depicted in a variety of colors. The correlation and color intensity are proportional. Finally, an indirect relationship between two variables implies that the variables are changing in opposite directions, i.e. if one increases, the other decreases, and *vice versa.*

The correlation coefficient can be anywhere between -1 and 1. If the fad has not collapsed between the suggest value then there is be some error within system. A correlation value of -1 denotes a basic indirect correlation. The correlation value 1 means that there is a clear direct correlation. Age, Gender, fever, Bodypain, Runny_nose, Difficulty_in_breathing, Nasal_congestion, Sore_throat, Severity, Contact_with_covid_patients, and Infected are representing the different parameters. A good positive correlation is also found between Fever and Body Pain and Runny Nose and Difficult to Breath. Table 6 and Figure 9 show the relationship value of each dependent feature versus the independent feature.

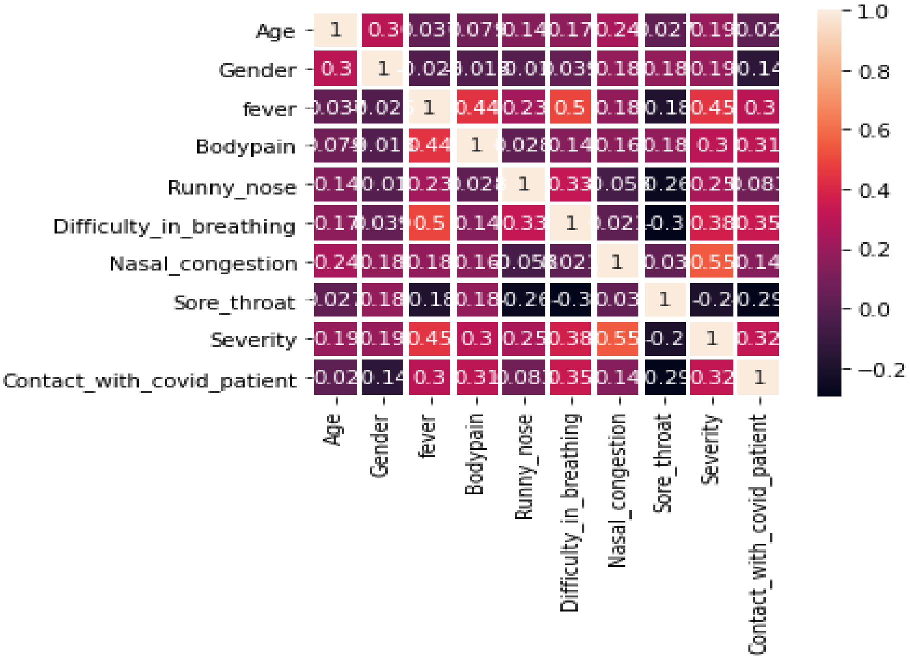

FIGURE 9. Correlation matrix of the dataset features

TABLE 6. Values of correlation coefficient

	Age	Gender	fever	Bodypain	Runny_ nose	Difficulty_ in_breathing	Nasal_ congestion	Sore_ throat	Severity	Contact_ with_ covid_ patient	Infected
Age	1.0	0.3	0.0	0.1	0.1	0.2	0.2	0.0	0.2	0.0	0.2
Gender	0.3	1.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	-0.1	-0.1
Fever	0.0	0.0	1.0	0.4	0.2	0.5	0.2	-0.2	0.5	0.3	0.4
Bodypain	0.1	0.0	0.4	1.0	0.0	0.1	0.2	0.2	0.3	0.3	0.4
Runny_ nose	0.1	0.0	0.2	0.0	1.0	0.3	-0.1	-0.3	0.2	0.1	0.3
Difficulty_ in_ breathing	0.2	0.0	0.5	0.1	0.3	1.0	0.0	-0.3	0.4	0.3	0.5
Nasal_ congestion	0.2	0.2	0.2	0.2	-0.1	0.0	1.0	0.0	0.6	0.1	0.3
Sore_ throat	0.0	0.2	-0.2	0.2	-0.3	-0.3	0.0	1.0	-0.2	-0.3	-0.2
Severity	0.2	0.2	0.5	0.3	0.2	0.4	0.6	-0.2	1.0	0.3	0.4
Contact_ with_ covid_ patient	0.0	-0.1	0.3	0.3	0.1	0.3	0.1	-0.3	0.3	1.0	0.7
Infected	0.2	-0.1	0.4	0.4	0.3	0.5	0.3	-0.2	0.4	0.7	1.0

METHODS

The research is about novel corona virus which also known as COVID-19 predictions. A current possible danger the humanity has been proven by COVID-19. It kills millions of people every year, and the death rate is growing all over the world. In order to contribute to this pandemic situation, this study aims to predict COVID-19 patients for speedy recovery and the identified patients who undergo for medication.

SUPERVISED MACHINE LEARNING MODELS

When an unknown input example is supplied, a supervised learning model is constructed to make a prediction. The learning algorithm therefore takes a dataset with input instances along with their corresponding classification model in this learning technique. For predictive models' development, this learning method can use classification algorithms.

In this study of COVID-19 prediction, five classification models were used: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor, Gausian NB, and Decision Tree.

LOGISTIC REGRESSION

Logistic regression (Li et al. 2020) is a classification algorithm used to assign observations based on one or more predictor variables to a different group (or category) of individuals (x). A binary outcome is used to model a variable that can have only two possible values: 0 or 1, yes or no, diseased or non-diseased, true or false, negative or positive. Unlike linear regression, which produces continuous number values, logistic regression transforms the output using the logistic sigmoid function to return a probability value that can then be mapped to two or more discrete groups.

To convert expected values to probabilities, we use the sigmoid function. The function converts a real value between 0 and 1 to a number between 0 and 1. We use sigmoid to map predictions to probabilities in machine learning. The mathematical representation of the logistic regression algorithm is shown herewith. Equation (1) is used to measure the relationship between the dataset's dependent and independent attributes or features (Naw Safrin et al. 2019).

$$s(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Here, s(z) denotes an output between 0 and 1 (probability estimate); z denotes the input to thefunction, such as mx + b; and e denotes the base of natural log.

K-NEAREST NEIGHBOR

The supervised learning technique's K-Nearest Neighbor algorithm assumes correlations between new case data with the existing data and categorizes it into the most similar available group. The K-NN algorithm stores existing data and categorizes new data based on similarities. K-NN is used in both classification and regression problems but mostly with classification problem. It is a non-parametric algorithm and does not make any assumptions with underlying data, it's also known as the lazy learner algorithm because it doesn't learn from the training set, instead storing and acting on the dataset during classification. KNN algorithm uses similarity of the feature for predicting the values of new data points and assign them with values on analysing how closely the data points match the training set points. The following steps are taken into account in KNN algorithm: Train data and test data is loaded into code; K value nearest to the data point is chosen; Euclidean, Manhattan, and Hamming distance methods are used to measure the distance between test data and each row of training data; The measured distances are sorted in descending order, and the algorithm picks the top K rows; and On analysing the frequent classes in the row, a test data point is assigned with class.

SUPPORT VECTOR MACHINE

Classification and regression problems are solved using the Support Vector Machine, one of the Supervised Learning machine learning methods. SVM create the decision boundary which will separate n-dimensional space into classes so as to place the new upcoming data into a place or point in its related category in future. The decision boundary known as a hyperplane. A hyperplane can be created by finding out the correct extreme vector points. The nearest data points or vectors to the hyperplane that influence the hyperplane's location is referred to as Support Vectors. There can be many decision boundaries for classes' segregation in n-dimensional space. An efficient data boundary should be found so as to classify the data points. Dataset features determine the hyperplane dimension. They have the presence of maximum margins in hyperplane is used for the indication of data points with maximum distance.

DECISION TREE

The decision tree is one amongst the prognostic modeling approaches used in machine learning. The target variable in each tree model will have a different set of values. The branches denote choice combinations, while the leaves represent group names. The measure used on decision tree are entropy and information gain. The values of (C), D, and E are formed by the knowledge set attribute partition of the numerical data type (B) z, where z is the value of the B domain for the entire categorical attribute of the data kind partition C, and E is a restricted set of attributes (B).The pruning technique method for the last word tree construction once adult is utilized to eliminate noise from the dataset.

GNB

For classification, GaussianNB implements the Gaussian Naive Bayes algorithm. It involves calculating the classes in the dataset prior and posterior probability and the test data given to a class, respectively.

$$Prior\ Probality(c) = \frac{\text{No. of Class c instances}}{\text{Total No. of Dataset Instances}} \quad (2)$$

Prior probabilities of all classes are determined with the aid of (2).

EVALUATION STUDY

Using decision trees, logistic regression, KNN, Gaussian naive Bayes, and SVM machine learning algorithms, and a dataset of positive and negative COVID-19 cases from several countries, supervised machine learning models for COVID-19 infection were created. Before the model's case, a correlation analysis of the various dependent and independent options was performed to ensure a clear association between each dependent feature and the dataset's independent feature. The relationship value of each dependent characteristic is shown by Table 4 against the independent characteristic of the dataset. All the dependent characteristics have a positive relationship coefficient of correlation with the dataset's independent function. It is, however, a weak positive relationship coefficient of correlation that all dependent characteristics have an independent function. There are a lot of parameters that can be evaluated and compared after a number of different classification models have been developed. It uses the following parameters to compare the classifiers.

For both positive and negative COVID-19 scenarios, the dataset was divided into training and test sets. The model was then trained with 80% of the training data before being double-checked with the remaining 20% of the dataset. Five different models were developed to predict whether a patient is infected with COVID-19 or not.

CONFUSION MATRIX

Evaluating the quality of every prediction system depends on the number of instances that are correctly and wrongly predicted. The parameters required were shown in the form of a matrix known as the confusion Matrix. It defines the efficiency of the classification algorithm using the parameters shown in Table 3. Based on some basic parameters, it gives the specifics and quality of the classification algorithm, such as true positive rate, true negative rate, false positive rate and false negative rate. The confusion matrix aids in the prediction of classification issues. The total number of exact predictions for a class is entered into the estimated row for that class value in the projected class. Similarly, the cumulative number of incorrect class predictions is entered into the estimated row for that class value and the predicted column's class value. As shown in Figure 10, the confusion matrix containing details or data about real and expected classifications completed by a classification procedure. The output is assessed based on the data in the matrix (Table 6). The table shows the confusion matrix for a two-class classifier.

As shown in Figure 10 in Confusion Matrix, first matrix showed the Logistic Regression, 839 COVID-19 Negative patients of a total of 1000 COVID-19 Negative patients were classified in this category and the rest were not included in this category. Accordingly, 83.9% of them were classified in the class, while 16.1% did not fall into this category. On the other hand, only 778 COVID-19 Positive patients of the 990 patients in the COVID-19 were classified in this category and 212 patients were not included in this category. In other words, 78.5% of COVID patients were in this class, while 21.5% were not categorized. Accordingly, 1607 COVID-19 Positive and COVID-19 Negative patients were classified in the category (83.5%) and 393 were classified wrongly in this category, which is 24.45%. Similarly, all other models' matrices are shown in Figure 10 and Table 7.

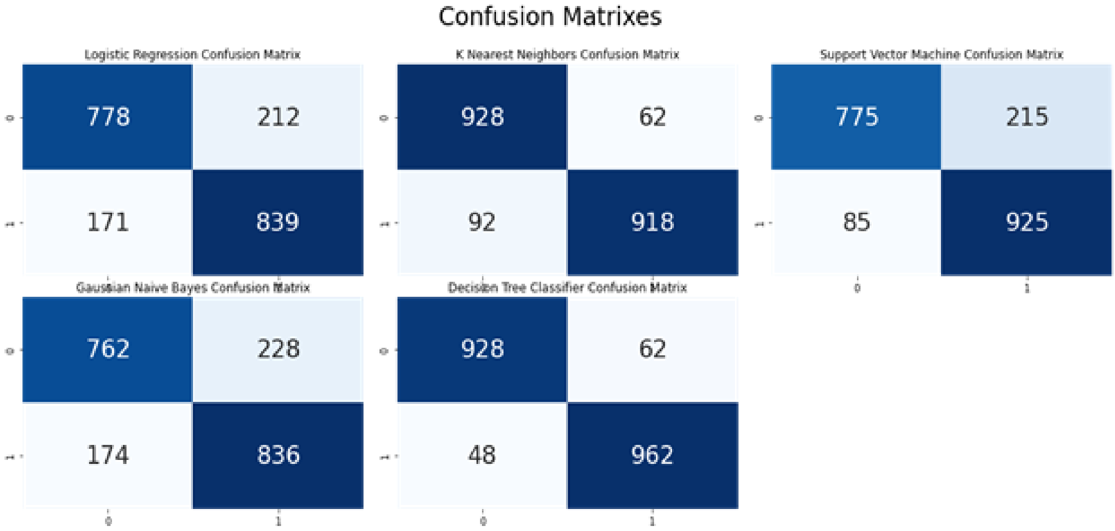

FIGURE 10. Confusion matrix for COVID-19 of the dataset features

TABLE 7. Five models True negative(tn), False Negative(fn), True Positive(tp), False positive(fp) values

	tn	fn	tp	fp
GNB	762	174	836	228
LG	778	171	839	212
SVM	831	120	890	159
KNN	928	92	918	62
DT	928	48	962	62

The results show that the classifier has very good classification accuracy, with an overall score of 94.60%. It also has a high precision (97.52%) for the positive class, as well as a better recall for the positive class (83.10%). When it comes to negative classes, the classifier has a high precision (93.67%) as well as a high recall (99.10%).

The following parameters used to evaluate the proposed model.

$$\text{Accuracy} = \frac{(\text{TruePositive+TrueNegative})}{(\text{TruePositive+TrueNegative+FalsePositive+FalseNegative})} \quad (3)$$

$$Recall = \frac{(\text{TruePositive})}{(\text{TruePositive+FalseNegative})} \quad (4)$$

$$Recall = \frac{(\text{TruePositive})}{(\text{TruePositive+FalseNegative})} \quad (5)$$

$$Percision = \frac{\text{TruePositive}}{\text{TruePositive+}FalsePositive} \quad (6)$$

The Accuracy, recall, precision, F1 score are the primary metric for determining the number of correctly identified COVID-19 patients.

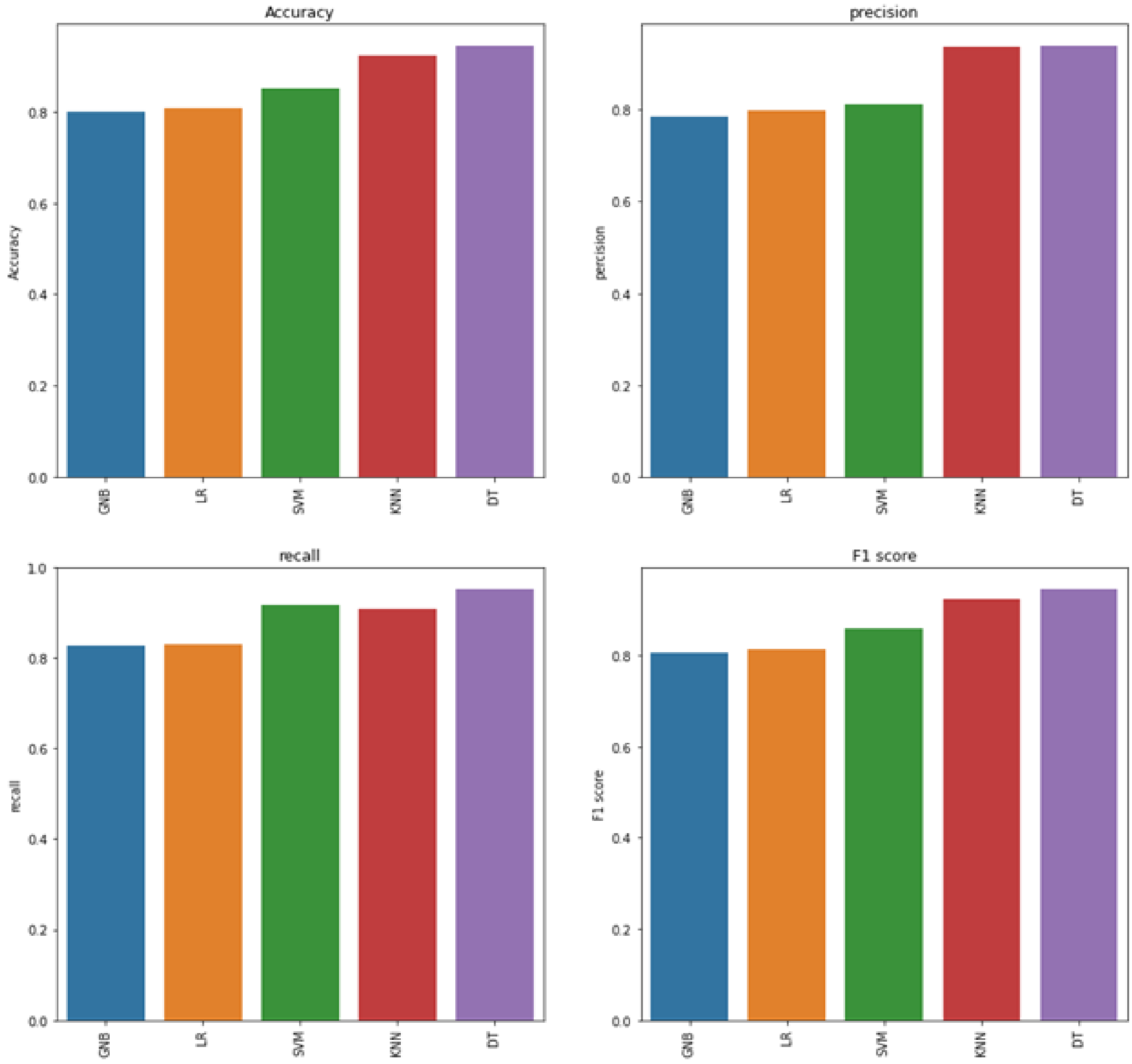

FIGURE 11. Performance evaluation of Accuracy, recall, Precision, F1 score

Results and Discussion

Early COVID-19 prediction will help reduce the huge burden on healthcare facilities by assisting in the diagnosis of COVID-19 patients. This research used a dataset of positive and negative COVID-19 cases from various countries to develop supervised learning classification models for COVID-19 infection prediction using decision trees, logistic regression, KNN, GNB, and SVM. From the analysis we made using accuracy, F1 score, Precision and recall for the algorithmic techniques of Logistic Regression, K-Nearest Neighbor, Decision tree, Gaussian NB and support vector machine we obtained the following values with the Decision Tree algorithm having high accuracy, precision, recall and F1 score. The performance of various models was evaluated and shown in Table 8 and Figure 11.

TABLE 8. Result of algorithm performance over the dataset

ML Algorithms	Accuracy	precision	Recall	F1 score
GNB	0.799	0.786	0.828	0.806
LR	0.809	0.798	0.831	0.814
SVM	0.850	0.811	0.916	0.860
KNN	0.923	0.937	0.909	0.923
DT	0.945	0.939	0.952	0.946

Thus, we evolved from our predictive analysis emphasizing that Decision tree gives high prediction of acquiring COVID-19 patients based on different parameters with the Accuracy of DT, which is 94.5%, is the highest among them, followed by KNN, SVM, LR, and GNB. The DT precision, accompanied by KNN, SVM, LR and GNB, is 93.9%, which is the maximum. The recall of DT is 95% which is high followed by KNN, SVM, GNB, and LR. Also, we can get F1 score for DT is high, which is 94.5% along with KNN, SVM, LR, and GNB as interpreted from Figure 12 and Table 8. Also, draw from Figure 12 that the GNB, KNN, and DT corresponding AUCs are 0.4950, 0.4536, and 0.6463, respectively.

We have also fitted our data with ML classification models and we constructed a computational model that successfully identified COVID-19 patients with high recall (sensitivity) and precision (specificity). Using GNB, LR, SVM, KNN and DT, we obtained an AUC of 89%, 92%, 90%, 97%, and 98%, respectively, and then attempted to distinguish between patients COVID-19 Negative and COVID-19 positive patients. The recall was around 82.8% with GNB, and the precision was about 78.6%. The recall was 83.1% with LR, and the precision was about 79.8%. The GNB method does not have great strength, however, based on the AUC curve. With SVM, the recall was about 91.6%, while the precision was around 81.1%.

The decision tree model showed that the most significant feature of all the dependent features of the dataset, including the clinical features, is in Contact_with_Covidpatient feature. The model shows that, relative to people of lower ages and not in touch with COVID-19 patients, most people over the age of 52 are likely to be infected with COVID. Similarly, individuals with signs of difficulty-in-breathing, runny-nose, fever are more likely to be COVID-19 compromised. As far as gender is concerned, males are more vulnerable than females to COVID-19 infection, and those who smoke tobacco are more likely to be infected than smokers of non-tobacco. We accept that our model exhibited the feasibility of using machine learning algorithm to advise diagnostic decisions for COVID-19. The model will assist health professionals with the identification of suspected patients with COVID-19, and this will assist those taking the chest X-Ray examination to reduce the immense strain on healthcare systems (Figure 13).

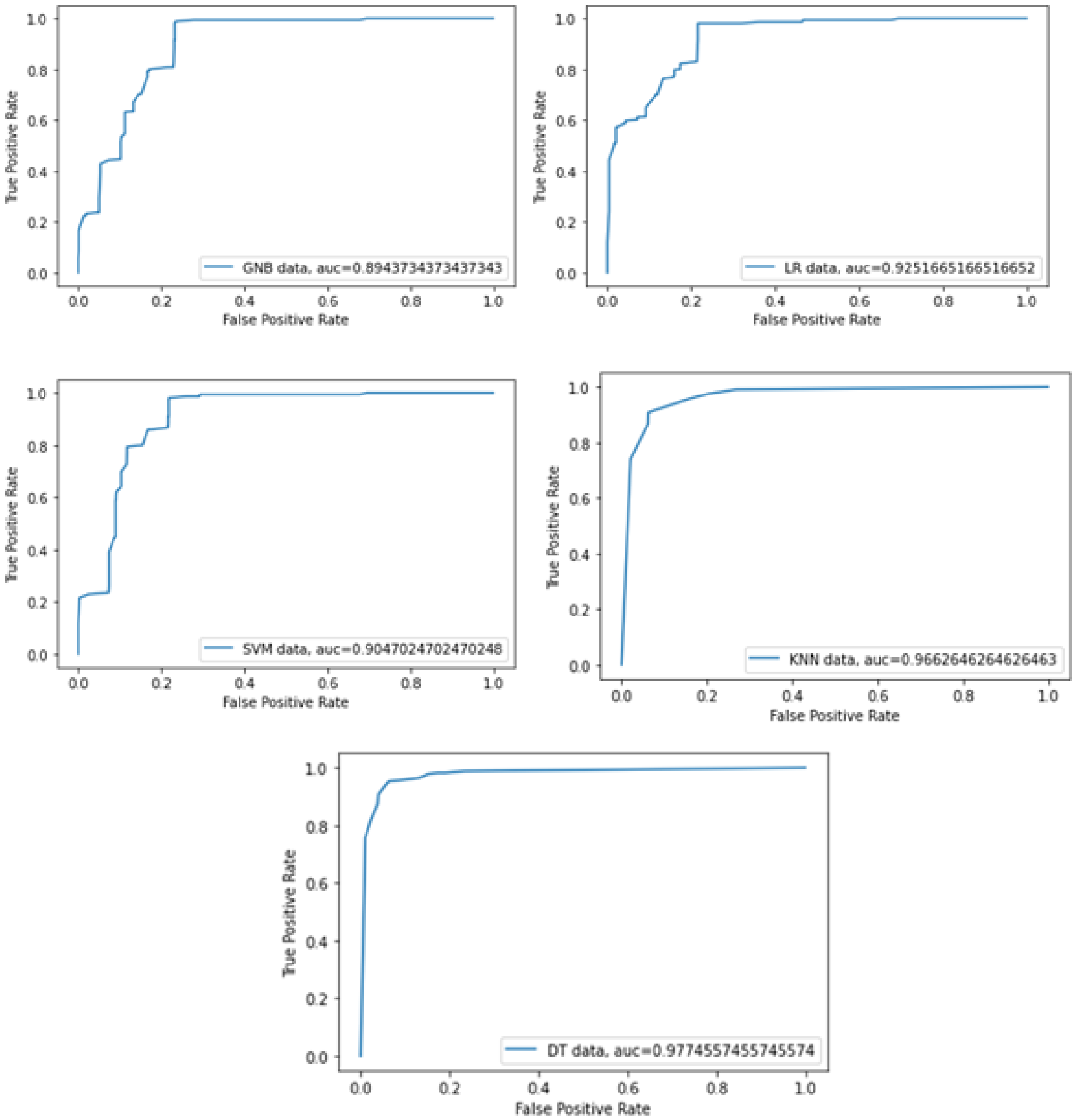

FIGURE 12. AUC curve of five models

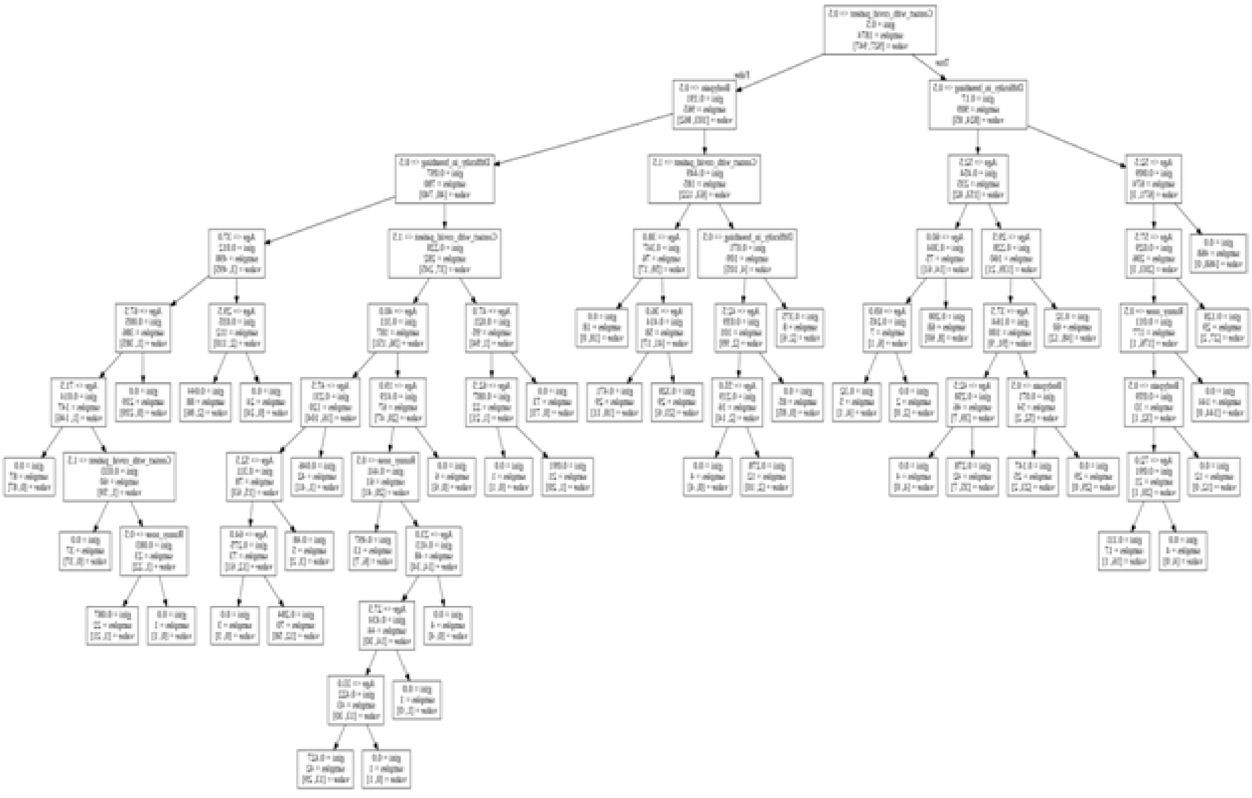

FIGURE 13. Decision tree of the proposed model

Conclusion

Corona virus has stunned the world because of its non-accessibility of antibody or medication. Different scientists are working for overcoming this dangerous virus. We used 2500 clinical reports which are labeled. Various features are being extracted and validated from these reports. In conclusion, we exhibited the utilization of ML models utilizing just broadly recorded clinical factors to anticipate COVID-19 presence. In this paper, comparative analyses of different classifiers were done for the prediction of the COVID-19 dataset for positive and negative diagnosed people. The algorithms were used K- Nearest Neighbor (K-NN), Naive Bayes, SVM, and Decision Table (DT) classifiers. Subsequent to performing classification, it was uncovered that decision tree classifier gives astounding outcomes by having 94% precision, 96% recall, 95% f1 score and accuracy 96.2% which beat other used references utilized in this paper. Also, the result shows AUC value of 98%. Increasing the amount of data in a model will increase its performance. Also, the disease can be categorized based on gender, so we can learn whether males or females are more affected, or whether transgender people are more affected. It can also be classified based on age. For better results, further feature engineering is needed, and a deep learning approach may be used in the future.

References

Ayyoubzadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M. & R Niakan Kalhori, S. 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health and Surveillance* 6(2): e18828. doi.org/10.2196/18828.

COVID-19. *Dataset*. https://github.com/Simranpandey16/Covid-19-prediction.

COVID-19 Public Health Emergency of International Concern (PHEIC). Global research and innovation forum. https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum.

Dharshana Deepthi, L., Shanthi, D. & Buvana, M. 2020. An intelligent Alzheimer's Disease prediction using convolutional neural network (CNN). *International Journal of Advanced Research in Engineering and Technology (IJARET)* 11(4): 12-22.

ExtraTreeClassifier. https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/.

Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam & Gyu Sang Choi. 2020. COVID-19 future forecasting using supervised machine learning models. *IEEE Access* 8: 101489-101499.

Jackins, V., Vimal, S., Kaliappan, M. & Lee, M.Y. 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing* 77: 5198-5219. https://doi.org/10.1007/s11227-020-03481.

Li, W.T., Ma, J-Y., Neil, S., Grant, C., Jaideep, C., Tsai, J., Apostol, L., Honda, C., Xu, J-Y., Wong, L., Zhang, T-Y., Lee, A., Gnanasekar, A., Honda, T., Kuo, S., Yu, M., Chang, E., Rajasekaran, M.R. & Ongkeko, W. 2020. Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. *BMC Medical Informatics and Decision Making* 20: 247. DOI. 10.1186/s12911-020-01266-z.

Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C. & Mohammed, I.A. 2021. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput. Sci.* 2(1): 11. https://doi.org/10.1007/s42979-020-00394-7.

Naw Safrin Sattar, Shaikh Arifuzzaman, Minhaz F. Zibran & Md Mohiuddin Sakib. 2019. Detecting web spam in webgraphs with predictive model analysis. *2019 IEEE International Conference on Big Data (Big Data)*. pp. 4299-4308. doi: 10.1109/BigData47090.2019.9006282.

Remuzzi, A. & Remuzzi, G. 2020. COVID-19, and Italy: What next? *The Lancet* 395(10231): 1225-1228.

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P. & Chowell, G. 2020. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th. *Infectious Disease Modelling* 5: 256-263.

Sarwar, A., Ali, M., Manhas, J. & Sharma, V. 2020. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *Int. J. Inf. Tecnol.* 12: 419-428.

Sharma, A., Tiwari, S., Deb, M.K. & Marty, J.L. 2020. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): A global pandemic and treatment strategies. *International Journal of Antimicrobial Agents* 56(2): 106054. https://doi.org/10.1016/j.ijantimicag.2020.106054.

Ud Din Khanday, A.M., Rabani, S.T., Khan, Q.R., Rouf, N. & Ud Din, M.M. 2020. Machine learning based approaches for detecting COVID-19 using clinical text data. *Int. J. Inf. Tecnol.* 12: 731-739 https://doi.org/10.1007/s41870-020-00495-9.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M.J., Yang, J.Y., Li, Y.D., Meng, X.F. & Bo, Xu. 2020. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *Eur. Radiol.* 31(8): 6096-6104.

Yan, L., Zhang, H-T., Goncalves, J., Xiao, Y., Wang, M-L., Guo, Y-Q., Sun, C., Tang, X-C., Jin, L., Zhang, M-Y., Huang, X., Xiao, Y., Cao, H., Chen, Y-Y., Ren, T-X., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N-N., Jiao, B., Zhang, Y., Luo, A-L., Mombaerts, L., Jin, J-Y., Cao, Z-G., Li, S.S., Xu, H. & Yuan, Y. 2020. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *medRxiv* https:// doi.org/10.1101/2020.02.27.20028 027.

Buvana, M.*
Department of Computer Science and Engineering
PSNA College of Engineering and Technology
Tamil Nadu
India

Muthumayil, K.
Department of Information Technology
PSNA College of Engineering and Technology
Tamil Nadu
India

*Corresponding author; email: buvana@psnacet.edu.in

Received: 7 April 2021
Accepted: 27 June 2021