Development and Evaluation of Machine Learning Models for Early Detection of Asymptomatic COVID-19 Patients Using Heart Rate and Oxygen Levels

Hafiz Haseeb Tasleem[1], Mueed Ahmed [1], Muhammad Waqar Arshad[1*], and Muhammad Hamza [2]
[1]Faculty of Computing, Riphah International University, Islamabad, Pakistan
[2]Faculty of Computing and IT, University of Sialkot, Sialkot, Pakistan
E-mail: hafizhaseebit@gmail.com, mueedahmed92@gmail.com, waqararshad211@gmail.com, muhammadhamzait9@gmail.com
[*]Corresponding author

Keywords: covid-19, sars-cov-2, asymptomatic patients, artificial intelligence, k-nearest neighbor

Received: November 15, 2024

Since the coronavirus disease 2019 (COVID-19) spread across the world in late December 2019, it has caused significant harm and major challenges in over 190+ countries all over the world. The research that is being done today is getting all the chest X-ray images and lung images. With the help of this, researchers are predicting Covid-19. There is mounting evidence that many COVID-19 patients are asymptomatic or have only minor symptoms but may still spread the virus to others. In the existing system, simple pulse oximeters were used to diagnose infectious diseases early. Oxygen level and heart rate can be used to detect virus-related infections, including asymptomatic patient infections. Screening for asymptomatic infections is difficult, which makes national prevention and control of the outbreak more difficult. In this research, we predict the asymptomatic COVID-19 patients with the help of their oxygen level and heart rate level. To build the machine learning model we use SVM, Naïve Bayes, KNN and logistic regression algorithms on the collected dataset. The model predicts the asymptomatic COVID-19 patients early. The dataset contains 105,609 cases with 16 attributes, including information of patients with COVID-19 RT-PCR test results. There are ten key features to be selected from the given dataset for the experiment. First, we analyze the features of the dataset to find the most important features. Heart rate and SPO2 are the most important features of the dataset for predicting asymptomatic COVID-19 patients. Our machine learning technique uses four ML algorithms. Through feature correlation, we improved accuracy by using ten main features. Following that, we trained and evaluated the data with 80-20% splits. This study compares the results of the model with other studies and finds that our technique achieves the best results from others. The current study's findings show that the model developed with the KNN algorithm is more effective at detecting the likelihood of the infected patients and achieved the highest 98% accuracy, 87% precision, 97% recall, 92% f1 score and making it the best model among those that have been developed with other algorithms such as support vector machine, naïve bayes and logistic regression.

Povzetek: Opisana sta razvoj in vrednotenje modelov strojnega učenja za zgodnje odkrivanje asimptomatskih bolnikov s COVID-19, pri čemer so ključni podatki o srčnem utripu in ravni kisika. Z uporabo algoritmov SVM, Naïve Bayes, KNN in logistične regresije na zbranem naboru podatkov študija ugotavlja, da model KNN dosega najvišjo točnost.

1 Introduction

Infectious pneumonia was reported in the Chinese city of Wuhan in December 2019 [1]. This infection is designated as Corona Virus Disease-19, in short form COVID-19. It has been labelled a pandemic by the World Health Organization (WHO). The disease has covered several countries globally and has contaminated millions of people worldwide, and 4.3 million people died from COVID-19. Unfortunately, COVID-19 has no effective and reliable treatment/cure. Because the nature of COVID-19 is yet unknown, it will take a longer time to produce an effective COVID-19 vaccine. On March 23, 2022, a total of 470 million COVID-19 infected cases were reported, and 6 million deaths all over the world [2]. Fever, cough, and fatigue are all symptoms that are close to the flu and must be recognized for early COVID-19 prediction. Surprisingly, a patient with no symptoms could transmit the COVID-19 virus to other people. COVID-19 affects every country in the world and there is a need to be various advancement, and new technologies are required to handle the different problems caused by the impact of issues in the health system [2]. One of the reasons COVID-19 has spread so quickly is that people infected with the virus may have no symptoms but still be contagious; they don't look or feel sick but can transmit the virus without even realizing it. Asymptomatic transmission refers to the spread of disease without the presence of symptoms [3, 4].

An asymptomatic case is one diagnosed with disease symptoms, whereas an asymptomatic case is one with no apparent symptoms. However, you might have heard a word that confuses you: 'pre-symptomatic.' A pre-symptomatic case seems to be someone who has not had any symptoms yet. It's a little more complex than that since pre-symptomatic can also mean asymptomatic. Physical distancing, social distancing, hand sanitization, use of nose and face masks, and handwashing have all been used to avoid the spread of COVID-19 around the world. However, the poor healthcare systems, financial pressure, overcrowding, community attitudes, poverty, and the COVID-19 preparedness and response plan are the major challenges. The majority of people infected with the COVID-19 virus suffer respiratory symptoms. They begin to feel unwell, fever, cough, sore throat, or sneeze [5]. They may develop gastrointestinal symptoms in some people. Many patients have asymptomatic infections, which means they have no symptoms at all. The role of asymptomatic cases in spreading the pandemic was addressed by Maria Van Kerkhove figure 1 [6]. She made comments about the uncertainty by defining "asymptomatic cases" in a way that excluded people who were genuinely pre-symptomatic. Here's the problem. The people are known as 'asymptomatic' if they test positive for COVID-19 (confirmed case) but show no symptoms. However, if the infected person later shows symptoms, you reclassify their initial infection as 'pre-symptomatic' figure 1 represents the two types of coronavirus cases in the Venn diagram above, while the overlapping part represents various stages of infection. World Health Organization describes asymptomatic patients in a very specific system. Asymptomatic patients are only concerned that they never show symptoms (light blue portion of figure 1). World Health Organization ignores patients who move to symptomatic cases after a pre-symptomatic period in which they are active in asymptomatic transmission of the virus. Asymptomatic cases are those that have no symptoms when they spread Covid-19 (blue plus green over-lapping section).

This [7] study addresses challenges in understanding COVID-19's early spread, emphasizing non-countermeasure factors like population density and mobility. By leveraging machine learning it identifies socio-environmental determinants influencing transmission. The research highlights the difficulty of isolating these factors and the need for holistic strategies to complement traditional interventions, offering valuable insights to enhance public health planning.

The world is suffering from a pandemic disease of virus infection, and no specific or traditional treatment for this infection as of now. Studies have shown that COVID-19 caused an estimated $62 billion loss to China's economy and over $280 billion globally in the first quarter [8]. It has been on people's minds ever since the coronavirus spread. And people did a lot of psychological effects. Due to this people turned to hospitals. That is why it was necessary to develop a system that could predict asymptomatic COVID-19 patients early. Many studies conducted research on chest x-ray images and lung images for detecting or predicting COVID-19 patients. And they only focus on symptomatic patients. We didn’t see many studies on the prediction of asymptomatic COVID-19 patients. Studies conducted on asymptomatic patients consider only the oxygen level of the patient. In past, most studies conducted research on COVID-19 prediction using machine learning algorithms [9-11]. These studies use labelled dataset and only symptomatic patients were considered. There is no study conducted on asymptomatic patients that should include pulse rate and oxygen level for prediction of asymptomatic patients. Asymptomatic patients do not have any common symptoms. They are dangerous as they spread the coronavirus without even knowing that they are infected, and they can also have respiratory issues and severe lung damage. For accurate prediction of an asymptomatic patient’s oxygen level and pulse rate plays an important role. In this study, our aim is to add both data streams pulse rate and oxygen level to improve the reliability of the data, which will help to more precisely early detection of asymptomatic COVID-19.

Contributions of this research are as follows:

1. In this research we develop a machine learning technique using two new features (oxygen level and pulse rate) for accurate prediction of asymptomatic COVID-19 patients
2. An early prediction of asymptomatic patients using machine learning techniques reduces the massive load on the healthcare sector all over the world.
3. The performance of the machine learning model is analyzed by comprehensive experiments with a baseline and two main proposed features.

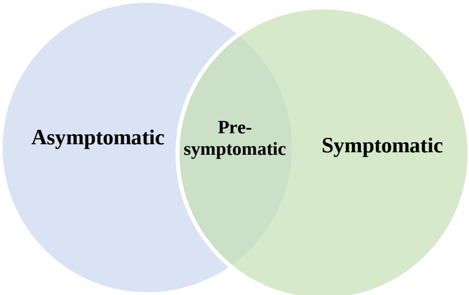


Figure 1: Differentiation of asymptomatic, pre-symptomatic and symptomatic

COVID-19 has become most hot topic of current research. We find a lot of research about COVID-19 to find the solution for early detection of viral disease. A brief overview of related studies about COVID-19 is presented in section 2, section 3 presents the methodology of asymptomatic COVID-19 prediction, section 4 presents the results and validation, section 5 presents the comparative study with the existing study and section 6 presents the conclusion and future work.

2 Related work

For predicting and analyzing the COVID-19 patients [9] study used different machine learning regression techniques. The study used a dataset that is available in

Kaggle for predicting and analyzing the patients. The author used polynomial and linear regression models for the prediction of future confirmed cases. The aim of this study [10] was prediction of infected patients of COVID-19 using ML techniques. From March 17 to March 30, 2020, data was collected through 235 patients at the Hospital of Brazil, of which 102 had 43% positive COVID-19 diagnosis from RT-PCR testing. 70% of the patient's dataset was randomly selected and used to predict diseases with different machine-learning techniques. Their performance was tested on 30% of the dataset. The support vector machines method produced the best predicting result AUC 85%, Sensitivity 68%, Specificity 85%. COVID-19 is an infectious disease that infects billions of people over the world. In this paper [12] authors proposed a forecasting model using LSTM, Bi-LSTM and SVR. The Bi-LSTM provides the best and better results for the prediction of COVID-19 and also better management and planning. COVID-19 affects people in the age categories of 20-30, 30- 40, and 40-50, according to the experiments [13]. The correlation matrices are used to figure out how the attributes of the datasets are related. The importance of the attributes is calculated for the machine learning classifiers that are developed. In which Random Forest Classifier gives higher performance than other machine learning models.

Utilizing the epidemiology labelled dataset [14] of COVID-19 patients, for detection of COVID-19 different ML techniques were used. The training dataset was used to train the models for 80% of the data, and the test dataset was used for the remaining 20%. Performance evaluation of the model, the machine learning decision tree algorithm achieved the best accuracy of 94% according to the results. This study used ML model for predicting the COVID-19 patient's invasive mechanical ventilation within 24 hours of their initial stage for evaluation of the performance of the model [15]. Five US hospitalized infected COVID-19 patients' dataset were used in this research. A total of 197 infected patients' data were used for the forecasting using machine learning techniques. The method had a much higher sensitivity 90% than the previous algorithm which had a sensitivity of 78%, while also having a significantly higher specificity ($p<0.05$). Machine learning algorithms are used in this study [16], and two different solutions are discussed: one is for calculating the likelihood of becoming COVID-19 infected patient, and the other one is forecasting the COVID-19 positive cases. Different algorithms were tested, and the one that produced the most accurate results was chosen.

In this study [17], a prediction approach depends on patient characteristics tracked while quarantine duration at home was used for the prediction and outcome of COVID-19 patients. 287 COVID-19 samples from patients at Saudi Arabia's King Fahad University Hospital were used in the study. Three classification techniques were used to examine the data. For data partitioning, 10-k cross-validation was used, and SMOTE was used to correct the data imbalance. Experiments were carried out with 20 clinical variables that were found to be relevant in predicting survival versus death in COVID-19 patients. Having an accuracy of 95%, the conclusion of the study showed that random forest performs better than other classifiers. An online questionnaire was created as a data collection for this investigation [18]. This information was fed into a different machine learning prediction. Based on their indications and symptoms, these models were used to predict prospective COVID-19 patients. In comparison to the other models, the MLP has achieved an accuracy of 91% among the top accuracy percentage. The AdaBoost method is used to enhance a fine-tuned Random Forest model in this work [19]. In this study author used geographic, demographic, health and travel data for infected patients of COVID-19. For predicting the COVID-19 patients using a dataset from South Korea, data mining models were designed [20]. This study helps to predict the likelihood of recovery from disease or death and severe cases of illness. The dataset contains males and females both genders and the majority have ages between 20 to 70 years. By applying data mining techniques, the model achieved 94% accuracy and 86% F1-socre. Using the programming language, different machine learning classification algorithms were applied directly to the dataset to generate the models. The results of the current study show that a model built using the decision tree machine learning technique is more effective at predicting the COVID-19 with a 99% accuracy.

Because predicting cardiac disease is a difficult task, it is necessary to automate the process to avoid the risks connected with it and to inform the patient well in advance. This paper [21] makes use of the UCI machine learning repository's heart disease dataset. In this study author classifies the risk level of the patients and predicts heart disease likelihood. The results of the model show the random forest achieved the highest accuracy of 90% than other ML algorithms. In the healthcare industry, data mining techniques are typical for predicting and analyzing a large amount of dataset. The model is based on supervised ML methods and present numerous attributes associated with heart disease. The author used UCI based dataset for the prediction of infectious disease from the database of Cleveland of heart disease patients [22]. There are 303 instances and 76 attributes in the collection. Only 14 of the 76 attributes are tested, despite their importance in proving the efficacy of different algorithms. The basic reason for this study is to predict the likelihood of patients having cardiac disease. The results show that K-nearest neighbor achieves the highest accuracy score.

Table 1: Critical evaluation table

Ref	**Dataset (Size & Source)**	**ML Technique**	**Best Metric**	**Key Strength**	**Key Weakness**
[9]	Kaggle (COVID-19)	Polynomial, Linear Reg.	AUC: 85%	Simple models, public data	Moderate sensitivity
[10]	Brazil hospital (235 patients)	SVM	AUC: 85%	Real-world data	Small dataset
[12]	Not specified	LSTM, Bi-LSTM	Best: Bi-LSTM	Strong temporal prediction	Lacks dataset & metrics details
[13]	COVID-19, demogra	Random Forest	High performance	Attribute importanc	Metrics vague

	phic focus		(unspecified)	e analyzed	
[14]	Epidemiology dataset	Decision Tree	Accuracy: 94%	High accuracy	Feature engineering not detailed
[15]	US hospitals (197 patients)	Various ML algorithms	Sensitivity: 90%	Improved sensitivity	Small dataset
[16]	Epidemiological dataset	ML algorithms	Not detailed	Dual-purpose prediction	Ambiguous reporting
[17]	Saudi Arabia (287 patients)	Random Forest, SMOTE	Accuracy: 95%	Balanced data, robust validation	Small dataset
[18]	Online questionnaire	MLP	Accuracy: 91%	Innovative data collection	Limited demographic clarity
[19]	South Korea demographic dataset	Random Forest, AdaBoost	Accuracy: 94%, F1: 86%	Effective AdaBoost enhancement	Limited generalizability
[20]	South Korea (20–70 years, gender mix)	Decision Tree	Accuracy: 99%	Exceptional accuracy	Overfitting risk
[21]	UCI Heart Disease (303 patients)	Random Forest	Accuracy: 90%	Effective cardiac risk prediction	No imbalance handling
[22]	Cleveland Heart Dataset (303 patients)	KNN	Best for dataset	Simplicity	No full metrics reported

Table 1 shows the studies using machine learning for COVID-19 and heart disease prediction, leveraging datasets like Kaggle, UCI, and hospital datasets. ML techniques such as Random Forest, SVM, and LSTM achieved accuracies up to 99%, but limitations included small datasets and unclear metric details in some cases of the literature.

According to the research that has been done so far, diagnosis and prediction of the COVID-19 pandemic with ML techniques have played an important role, which can be used help to ease the burden on healthcare systems. There is no work has been published in Pakistan that uses labelled datasets of COVID-19 asymptomatic cases to create supervised ML techniques for the prediction of COVID-19 patients. Also, there is no work has been published on asymptomatic COVID-19 patients using two main data streams heart rate and oxygen level of the patients. As a result, the purpose of this study is to look into these gaps.

3 Method

In this study, the process for developing a ML method for the prediction of infected COVID-19 patients using a labelled dataset is illustrated in Figure 2.

3.1. Dataset

A dataset of about 0.1 million COVID-19 suspected cases from the Hospital located in Pakistan, is used for this study. The dataset includes various types of details like age, gender, blood pressure, cholesterol, pulse rate, and oxygen level for each patient's situation. Our work focuses on the patient's heart rate and oxygen level. The results of COVID-19 cases in Pakistan are included in the dataset. The dataset contains 105,609 cases with 16 attributes, including information of patients with COVID-19 RT-PCR test results.

3.2. Data processing and analysis

This is the very first step in the diagnostic procedure. It consists of three steps: handling null values, data redundancy removal, and selection of important features. After testing the whole patient's age group, cholesterol, and blood pressure, the missing value of a particular attribute is substituted. If the majority of a patient's attribute values match, the value is substituted in the same place. By removing redundant (irrelevant) attributes, redundancy removal reduces the size of the data. By mapping the heat map and checking the association between the features, the significant features for analyzing the suspected asymptomatic COVID-19 patients are chosen. After the original dataset is prepared, only important features are extracted, and null values are removed.

There are 0.1 million instances in the extracted dataset with ten attributes including city, age, sex, heart rate and spo2 (oxygen level), bp_systolic_mmHg, bp_diasystolic_mmHg, cholesterol, glucose and smoke. We consider only two cities of the dataset Islamabad and Rawalpindi in Pakistan. The City and Sex attributes have object datatype and Age, Heart Rate and SpO2 have integer datatype. Table 2 shows some sample instances of the extracted dataset. The city instance has 2 cities located in Pakistan Islamabad and Rawalpindi. The dataset is taken from these two cities to use in this study. Age instances have different numbers of ages. Heat Rate shows whether a patient has a high heart rate and SpO2 shows the patient's oxygen level.

There are no missing values in the extracted dataset. Figure 3 shows the gender frequency in the dataset. In this study males and females both included in experiments. There are 68654 male patients and 36954 female patients in the dataset. Figure 4 shows the original dataset before normalize and after normalize the dataset.

Table 2: Sample of dataset

Sr. No	Attribute	Datatype
1	City	Object
2	Age	Int
3	Sex	Object
4	Heartrate	Int
5	SpO2	Int
6	bp_systolic_mmHg	Int
7	bp_diastolic_mmHg	Int
8	Cholesterol	Int
9	Glucose	Int
10	Smoke	Int

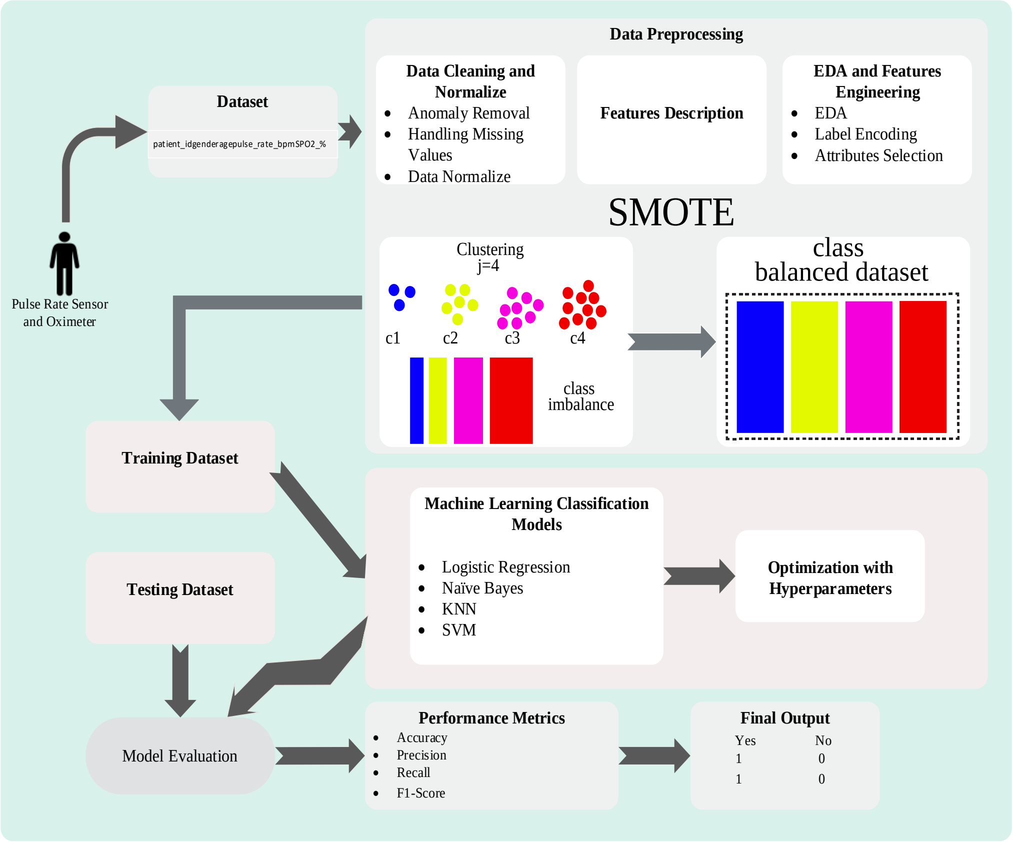


Figure 2: Proposed framework for prediction of asymptomatic COVID-19 Patients

We have used minmax scaling to normalize the dataset. In minmax scaling every value in the column with a new value using the following formula.

$$m = \frac{X - Xmin}{Xmax - Xmin}$$

To normalize the data, we use the normalize function. It takes an array as input and converts the values to a range between 0 and 1. The output array has the same dimensions as the input array. Normalized data helps to convert the numeric column values in the dataset to a standard scale while keeping variations in ranges of values in the dataset. We need to normalize the data because dataset have different ranges of features.

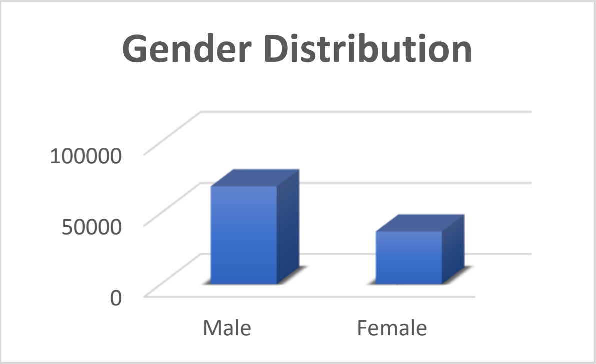


Figure 3: Gender distribution in the dataset

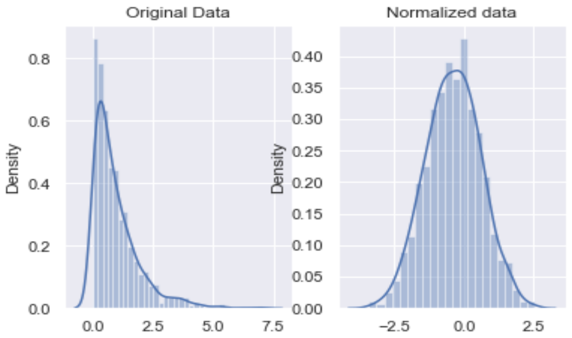


Figure 4: Before normalize and after normalize the dataset

3.3. Handling imbalanced data

Data imbalance is one of the most difficult problems to solve in data analysis, and it frequently leads to overfitting the model. The dataset to be used in this thesis is also unbalanced, as shown in Figure 5. The total number of entries in the suspected category is 13864, whereas the total number of records in the good health category is 91744.

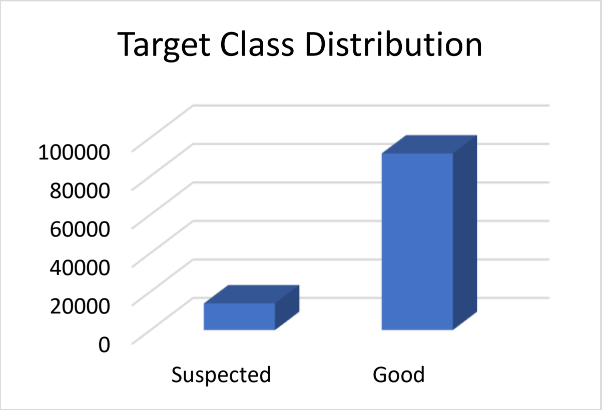


Figure 5: Target class distribution of the dataset

To address the data imbalance, a synthetic minority oversampling technique (SMOTE) based on K-nearest neighbor (KNN) was applied. SMOTE is a machine learning technique created by [23] to overcome the problem of imbalanced datasets. First, we divided the COVID-19 data into two parts: 80% training data and 20% test data. The training dataset was then oversampled using SMOTE, and the oversampled data was utilized to train machine learning models. We calculated the cross-validation score and the test score on the test dataset. The k-nearest neighbor (KNN) technique is utilized in the SMOTE algorithm to determine the Euclidean distance between minority class instances in order to create new class labels in the neighborhood. Using SMOTE the minority class set B, the k-nearest neighbors of x are calculated by computing the Euclidean distance between x and each other sample in set B for each x € B. The unbalanced proportion sets the sampling rate S. S samples (s1, s2,...sn) are randomly selected from their k-nearest neighbors for each x € B, and they form the set B1. The following formula is used to generate a new example for each xk € B1 (k=1, 2, 3...n): x' = x + rand(0, 1) * |x - xk|, where rand(0, 1) denotes a random number among 0 and 1 and xk denotes Euclidean distance with neighbors. Jupyter Notebook and the sklearn library were used to implement the models in Python. The data was partitioned using a 10-fold cross-validation technique.

3.4. Machine learning classifiers

This research used the preprocessed data to build a predictive model for our prediction framework. The aim of this model is to predict how likely a patient is to become infected with COVID-19. We trained the model on dataset which we collected form the hospital. Classification algorithms are used to determine if the patient has a high risk of COVID-19 infection. The prediction can be done using different classification algorithms. KNN, Naïve Bayes, SVM, and Logistic Regression are selected for COVID-19 prediction due to their complementary strengths. KNN classifies patients by similarity, Naïve Bayes handles probabilistic reasoning, SVM excels in high-dimensional spaces and Logistic Regression offers interpretable binary classification. Together, they ensure accurate and efficient identification of high-risk patients.

3.4.1. Naïve Bayes (NB)

In the Naive Bayes algorithm, the predictor values on a particular class are independent of other values of predictors. For the classification of learning tasks in which features of the COVID-19 dataset are differentiated based on a specified instance, the machine learning NB algorithm is used [24]. Class conditional independence is the term for this assumption. For classification assignment, NB is a Bayes theorem-based probabilistic classifier [25]. The Bayes theorem is as follows:

$$P\ (A|B) = \frac{P\ (B|A)\ P\ (A)}{P(B)}$$

Input:

Dataset of Training T,

T= (t1, t2, t3, .. , tn) : in the testing dataset, the value of the predictor variable.

Output:

Evaluate with test data.

Step:

1. Look at T, the training dataset.
2. Determine of calculate the mean and standard deviation of the predictor variables for each class.
3. Repeat again and again
   Calculate the likelihood of zi in each class.
   Until all predictor variables' probabilities (t1, t2, t3,..., tn) have been calculated.
4. For each class, calculate the prediction with the help of bayes theorem.

Get the most accurate prediction

3.4.2. Support vector machine (SVM)

A supervised learning approach [26] that is based on ML algorithm that can be helpful in handling regression and classification problems. In order to do a classification task in SVM, it is necessary to test and train data that contains some features of the dataset. Because each feature in the dataset includes more than one target value, the basic objective of SVM is to develop a model that can predict the target values in the dataset.

Following is the procedure for applying SVM

Input Data

d* variables and a binary result are included in this data set..

Output

Variables are ranked in order of their relevance. Find the best values for the SVM model's tuning parameters; then the SVM model will work properly

```
    d ← d*
while d ≥ 2 do
    SM M_d ← SVM using the d variables optimal tuning parameters and in the data observations
    w_p ← calculate the vector's weight of the SM M,, (w,,1, ... , w,,,, )
    criteria for ranking ← ( $w_{d1}^2$ , ... , $w_{dd}^2$)
criteria for minimum rank ← In the rank criterion, the variable with the lowest value vector
```

```
        Data that have minimum rank requirements should be removed.
        $Rank_p$ ← criteria for minimum rank
        D ← d - 1
end
        $Rank_1$ ← a data variable ∉ ($Rank_2$, ... , $Rank_{d*}$ )
return ($Rank_1$ , ... , $Rank_{d*}$)
```

3.4.3. K nearest neighbor (KNN)

KNN is a simple ML algorithm, which is very helpful to apply on both regression and classification tasks [27]. The KNN algorithm uses a similarity measure to store and classify new data. A distance function in KNN indicates how new data in the dataset is grouped into related classes by a majority vote of its neighbors. The distance between training and testing data is calculated using Euclidean distance.

```
Input
        A n * n distance matrix D(1 ….. n,1….n) and an index s of the starting point.
Output
        calculate the distance and predict the results.
        Set the value of K
(Estimating the nearest neighbor)
        for n = 1 to d
          calculating the distance from x to $j_i$
          if <= k
            then add $j_i$ to E
          else if $j_i$ is closer to x than any preceding nearest neighbor
              after that, delete the outermost neighbor
              and contain $j_i$ in the set E
```

3.4.4. Logistic regression (LR)

LR is a supervised ML algorithm. It performs a regression function. Its main applications include forecasting and finding relationships between variables. Logistic Regression is a machine learning algorithm for predicting the outcome of classification learning tasks. This algorithm is utilized to define the relationship between categorical dependent features and independent features [28, 29]. When dependent features of the datasets have binary values, this algorithm can be used [30]. The dependent variable (y) value is estimated using logistic regression based on the independent variable (x) value. As a result of this technique, a relationship between x (input) and y (output) is detected (output).

$$P = \frac{e^{a+bx}}{1+e^{a+bx}}$$

```
Input
Training data
1.      For i ← 1 to k
2.      For each training data instance $d_i$
    3.      Set the target value for the regression to
                $z_i \leftarrow \frac{y_i - P(1 \mid d_j)}{[P(1 \mid d_j) * (1 - P(1 \mid d_j))]}$
    4.      Initialize the weight of instance $d_j$ to P (1 | $d_j$) * (1 - P (1 | $d_j$)
    5.      finalize a ($f_j$) to the data with class value ($z_j$) & weights ($w_j$)
Classification label decision
        Assign (label class:1) if P (1 | $d_j$) > 0.5, otherwise (label class: 2).
```

3.5. Training and testing

We performed our experiment using an 80% training dataset and a 20% testing dataset for each of the four machine-learning algorithms. The accuracy of different machine learning algorithms in predicting COVID-19 asymptomatic patients is shown in Figure 6. Furthermore, as shown in figure 6, the KNN provides the highest level of accuracy, i.e. 98% accuracy. Logistic Regression performs the lowest in terms of accuracy, with a score of 92%.

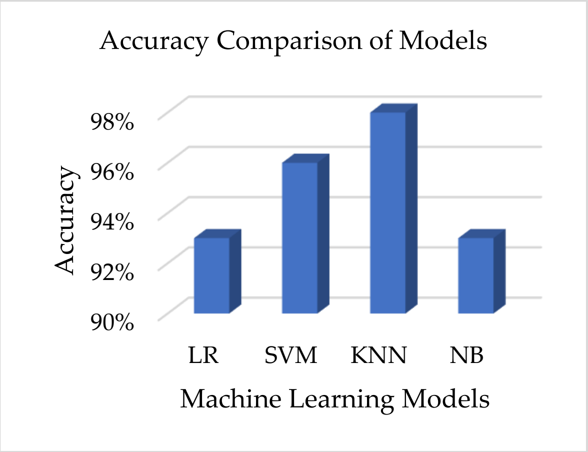


Figure 6: Accuracy comparison of four machine learning algorithms

Figure 7 shows the confusion matrix of four machine learning algorithms using a training dataset of 80% and a testing dataset of 20%.

True positive (TP)

The number of positive cases that have been labelled and are still positive.

False positive (FP)

The number of instances that were labelled as positive but were actually negative using the predictive model.

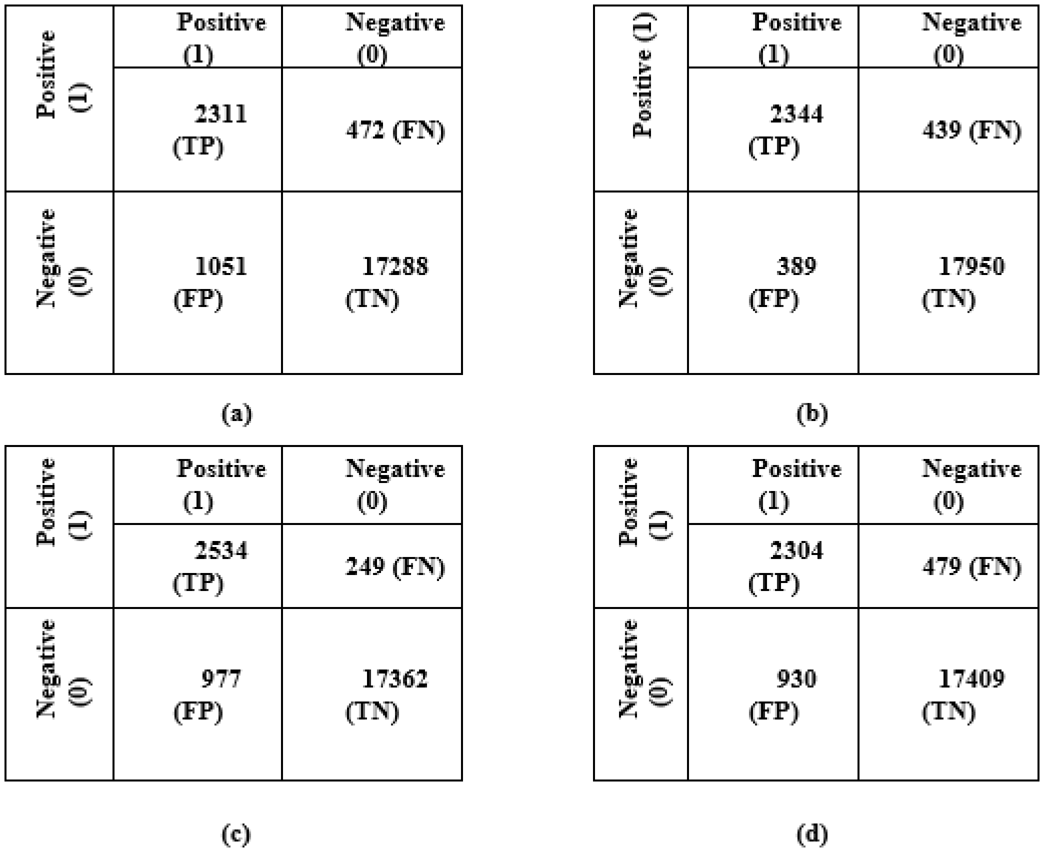
False negative (FN)

This is the number of instances that were labelled as negative by the predictive model but are actually positive.

True negative (TN)

This is the number of instances that have been labelled as negative yet are actually negative.

True Positive indicates that the percentage of patients known in the database as COVID-19 affects patients and is the best predicted of all the dividers. However, a False Negative class was created for COVID-19



	Positive (1)	Negative (0)
Positive (1)	2311 (TP)	472 (FN)
Negative (0)	1051 (FP)	17288 (TN)

(a)

	Positive (1)	Negative (0)
Positive (1)	2344 (TP)	439 (FN)
Negative (0)	389 (FP)	17950 (TN)

(b)

	Positive (1)	Negative (0)
Positive (1)	2534 (TP)	249 (FN)
Negative (0)	977 (FP)	17362 (TN)

(c)

	Positive (1)	Negative (0)
Positive (1)	2304 (TP)	479 (FN)
Negative (0)	930 (FP)	17409 (TN)

(d)

Figure 7: Confusion matrices. (a) Naïve Bayes. (b) SVM. (c) KNN. (d) Logistic regression

with affected patients, but because it was predicted wrongly, they were unable to be attacked. In the same way, the True Negative class category correctly predicted all of the patients who were not attacked. Also, the False Positive category is incorrectly reported.

4 Results and validation

While prediction of asymptomatic infected patients of COVID-19, early prediction of infected patients helps to minimize the huge load on the health-care system.

Table 2: Summary of results

Models	Accu	Precision	Recall	F1-Score
NB	93%	%71	%82	76%
SVM	96%	86%	84%	85%
KNN	98%	%87	97%	92%
LR	93%	%69	%83	75%

Using a dataset that is collected from a hospital in Pakistan, logistic regression, SVM, Naïve Bayes and KNN model for the prediction of asymptomatic COVID-19 patients were used. Accuracy, precision, recall, and F1-score parameters were used for the assessment of the efficiency of all machine learning models. Table 2 shows the machine learning model performance results.

We validate the results of our training model by testing unseen dataset. The training-set accuracy score of using KNN is 0.9606 while the test-set accuracy to be 0.9420. These values are quite comparable. So, there is no question of overfitting the model. Figure 8 shows the train and test scores.

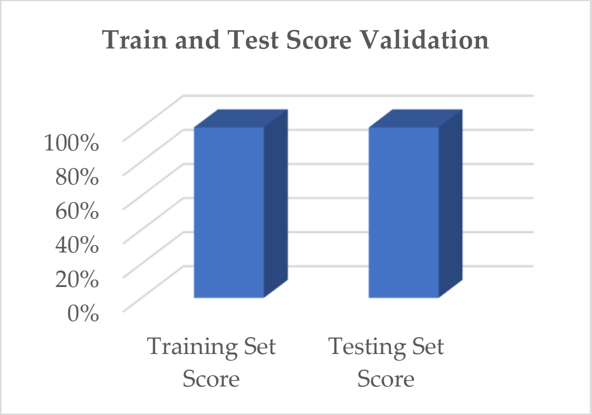


Figure 8: Train and test score validation

4.1. K-Fold Cross-validation

When using cross-validation, the dataset is randomly divided into 'k' groups. One group serves as the test set, while the others serve as the training set. On the training set, the model is trained, and on the test set, it is scored. The method is then repeated until each distinct group has been employed as a test set. In this paper, we used 10-fold cross-validation, which divides the dataset into 10 groups and trains and tests the model 10 times, giving each group a chance to be the test set.

The 'holdout' method is the same as the train-test-split method we used earlier. Because the holdout method's score is dependent on how the data is split into train and test sets, cross-validation is preferable to it. Cross-validation allows us to test the model on multiple splits in order to obtain a better understanding of how it will perform on unknown data. The results of cross-validation and test scores are shown in figure 9.

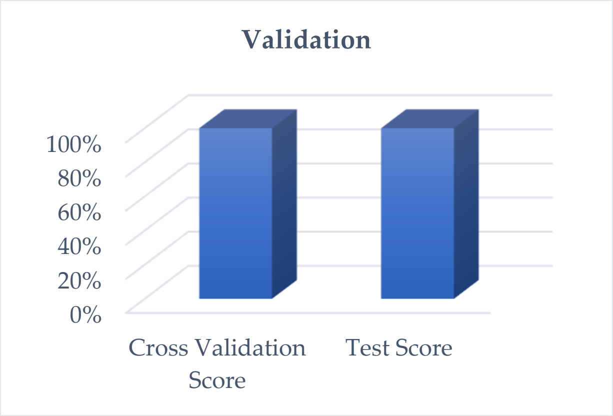


Figure 9: Comparison of 10-fold cross-validation and test score

We'll use the 'cross-val score' function with a cross-validation value of 10 to train and test our model using cross-validation. Our k-NN model and data are passed as arguments to the 'cross-val score.' Then it divides our data into ten groups and fits and scores it ten times, each time saving the accuracy score in an array. The accuracy scores will be saved in the 'cross_val_score' variable.

Our mean cross-validation score is around 89.6%. This is a more accurate portrayal of how our model will perform on unseen data than the holdout method we used previously.

5 Comparative study of our with the existing methods

The performance of machine learning classifiers is evaluated using three based papers as well as our proposed technique using new features and with the results being critically evaluated. The accuracy of the proposed technique experiment was used to evaluate their performance. The goal of this experiment is to look at how proposed new features interact with based-paper techniques for early prediction of COVID-19 patients. These are the paper [13, 14, 18, 31] considered for comparison with our machine learning technique. The comparison of machine learning technique results is shown in table 3. With the comparison of three base papers, our proposed research achieved the best results with the use of KNN algorithm on the accuracy performance measure

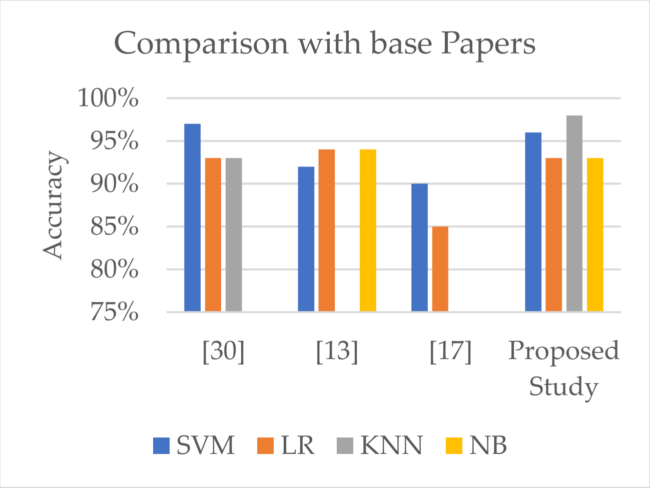


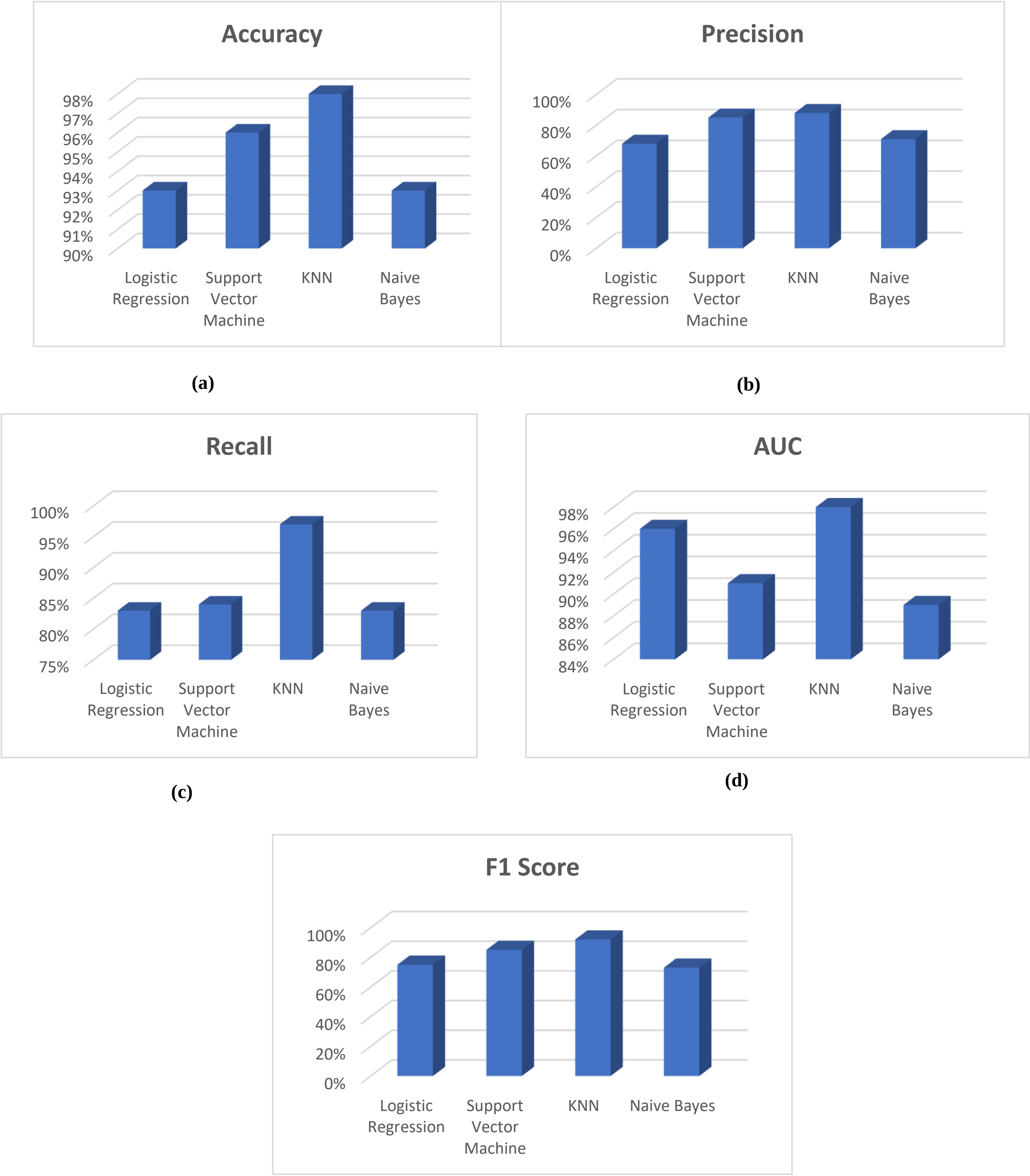
Figure 10: Accuracy comparison of proposed research

In figure 10 we have compared accuracy with based papers and achieved the highest accuracy 98% of the model.

Table 3: Comparison of proposed research with based papers

Based and Proposed Methods	SVM	LR	KNN	NB
A. F. de Moraes Batista [31]	97%	93%	93%	
L. Muhammad, E. A. Algehyne [14]	92%	94%		94%
E. Fayyoumi, S. Idwan [18]	90%	85%		
Proposed Study	96%	93%	98%	93%

The work of this research presents the early prediction of asymptomatic COVID-19 patients. There are four ML algorithms including support vector machine, logistic regression, KNN and naïve bayes used in this research. With the comparative analysis our proposed technique for the prediction of asymptomatic COVID-19 patients using main features of the dataset like heart rate and spo2 and achieved the best accuracy of 98% (Table 3). The protocol of COVID-19 asymptomatic patients such as heart rate and spo2 are the most important features for the prediction of patients. The best performance based on accuracy KNN perform well and achieves best accuracy. Figure 11 shows the results of our model in graphical representation, in graph KNN performs better than other machine learning algorithms.



(a)

(b)

(c)

(d)

(e)

Figure 11: Performance measures of four machine learning algorithms (a) Accuracy (b) Precision (c) Recall (d) AUC (e) F1 Score

6 Conclusion and future work

The COVID-19 pandemic, like other infectious diseases such as tuberculosis, HIV/AIDS, hepatitis, and measles now appears to be endemic. It has infected over 223 countries and territories throughout the world, as well as this disease infects people all around the world. This disease spreads through sneezing and coughing after close contact with an infected patient. No medically approved vaccination or therapy is available yet for COVID-19. A symptomatic case is one diagnosed with disease symptoms, whereas an asymptomatic case is one with no apparent symptoms. Developing the machine learning system in the medical healthcare system provides effective emergency services to patients. E-health application is also used in the different medical fields, i.e. early detection and monitoring of medical issues and emergency notification.

As an alternative approach, COVID-19 pandemic cases are being diagnosed using supervised ML algorithms. The use of ML algorithms for COVID-19 prediction in patients reduces the load of work on sufficient healthcare sources across the world. With the help of machine learning models, we are able to early predict asymptomatic COVID-19 patients. The models of machine learning were trained using 80% of the dataset and then tested dataset with 20%. In terms of accuracy, the KNN model was the best of all the models developed, with a score of 98%. In terms of accuracy, precision, and recall the KNN model shows top results among all models, with 98%, 88% and 97%, respectively.

We have a small amount of data, but with the help of a significant amount of data, the same mechanism can be used with more deep-learning models for better results. In the proposed research, there are two main variables used for the asymptomatic COVID-19 patients but as the data is increasing due to new variants of the virus, in the future more variables like lungs and heart damage can be added for better predictions.

7 Declarations

Ethics approval and consent to participate

The dataset used in this study was provided by a government hospital (PIMS) to our university under a formal Memorandum of Understanding. The dataset was anonymized and all patient-identifying information was removed prior to access by the research team.

In Pakistan, there are currently no specific regulations governing the use of human data in research. However, to ensure strict compliance with ethical standards and to safeguard patient privacy, the hospital shares such data exclusively with universities under formal agreements. These agreements ensure that the data is used responsibly and ethically. Universities, in turn, provide access to students under strict supervision and after ensuring full observance of professional conduct, thereby minimizing the risk of misuse. Given that the dataset was anonymized and shared under these regulated conditions, specific ethical approval from an Institutional Review Board was not required. The study adhered to the principles of the Declaration of Helsinki, institutional guidelines and the provisions of the data-sharing agreement between the hospital and the university.

Availability of data and material

The datasets generated and analyzed during the current study are not publicly available but available in the manuscript due to privacy restrictions but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Hafiz Haseeb Tasleem played a key role in outlining the research methodology and identifying the research problem through an extensive review of the literature. Mueed Ahmed significantly contributed by designing and conducting the experiments and analyzing the results. Muhammad Waqar Arshad provided detailed explanations and insights into the experiments conducted by Mueed Ahmed. Hamza Ansari constructed the abstract and conclusions, synthesizing the research findings effectively. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to express their gratitude to Professor Dr Muhmmad Bilal Basheer Iqra University Islamabad for their support and guidance during this research.

References

[1] W. H. Organization, "Novel Coronavirus (2019-nCoV): situation report, 11," 2020.

[2] WHO, "https://covid19.who.int/." vol. Accessed on 7th Nov 2020.

[3] www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/quarantine.html. Accessed July 04, 2020.

[4] M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish, "Industry 4.0 technologies and their applications in fighting COVID-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 14, no. 4, pp. 419- 422, 2020. https://doi.org/10.1016/j.dsx.2020.04.032

[5] WHO, "https://www.emro.who.int/health-topics/corona-virus/questions-and-answers.html," 2021.

[6] R. Li *et al.*, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science,* vol. 368, no. 6490, pp. 489-493, 2020. https://doi.org/10.1126/science.abb3221

[7] X. Chen, M. Ma, and A. Liu, "Dynamic power management and adaptive packet size selection for IoT in e-Healthcare," *Computers & Electrical Engineering,* vol. 65, pp. 357-375, 2018. https://doi.org/10.1016/j.compeleceng.2017.06.010

[8] C. Sitaula, A. Basnet, A. Mainali, and T. Shahi, "Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets," *Computational Intelligence and Neuroscience,* vol. 2021, 2021. https://doi.org/10.1155/2021/2158184

[9] M. S. Rahman, N. C. Peeri, N. Shrestha, R. Zaki, U. Haque, and S. H. Ab Hamid, "Defending against the Novel Coronavirus (COVID-19) outbreak: How can the Internet of Things (IoT) help to save the world?," *Health policy and technology,* 2020. https://doi.org/10.1016/j.hlpt.2020.04.005

[10] A. Kumari and A. K. Mehta, "Effective prediction of COVID-19 using supervised machine learning with Ensemble Modeling," in *Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences*, 2022, pp. 537-547: Springer. https://doi.org/10.1007/978-981-16-5747-4_45

[11] L. Bai *et al.*, "Chinese experts' consensus on the Internet of Things-aided diagnosis and treatment of coronavirus disease 2019 (COVID-19)," *Clinical eHealth,* vol. 3, pp. 7-15,2020.https://doi.org/10.1016/j.ceh.2020.03.001

[12] R. S. Abirami and G. S. Kumar, "Comparative Study Based on Analysis of Coronavirus Disease (COVID-19) Detection and Prediction Using Machine Learning Models," *SN Computer Science,* vol. 3, no. 1, pp. 1-8, 2022.https://doi.org/10.1007/s42979-021-00965-2

[13] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. Pawan, "Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms," *International Journal,* vol. 8, no. 5, pp. 2199-2204, 2020. https://doi.org/10.30534/ijeter/2020/117852020

[14] L. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset," *SN computer science,* vol. 2, no. 1, pp. 1-13, 2021. https://doi.org/10.1007/s42979-020-00394-7

[15] H. Burdick *et al.*, "Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial," *Computers In Biology And Medicine,* vol. 124, pp.1- 6,2020. https://doi.org/10.1016/j.compbiomed.2020.103949

[16] D. Painuli, D. Mishra, S. Bhardwaj, and M. Aggarwal, "Forecast and prediction of COVID-19 using machine learning," in *Data Science for COVID-19*: Elsevier, 2021, pp. 381-397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

[17] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients," *Scientific Programming,* vol. 2021, pp. 1-10, 2021. https://doi.org/10.1155/2021/5587188

[18] E. Fayyoumi, S. Idwan, and H. AboShindi, "Machine learning and statistical modelling for prediction of novel COVID-19 patients case study: Jordan," *Machine Learning,* vol. 11, no. 5,pp. 3- 11,2020. https://dx.doi.org/10.14569/IJACSA.2020.0110518

[19] C. Iwendi *et al.*, "COVID-19 patient health prediction using boosted random forest algorithm," *Frontiers In Public Health,* vol. 8, pp.1- 9,2020. https://doi.org/10.3389/fpubh.2020.00357

[20] L. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery," *SN Computer Science,* vol. 1, no.4,pp. 1- 7, 2020. https://doi.org/10.1007/s42979-020-00216-w

[21] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Heart disease prediction using machine learning," *International Journal of Research and Technology,* vol. 9, no. 04, pp. 659 6- 2, 2020. http://dx.doi.org/10.17577/IJERTV9IS040614

[22] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science,* vol. 1, no. 6, pp. 1- 6, 2020. https://doi.org/10.1007/s42979-020-00365-y

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002. https://doi.org/10.1613/jair.953

[24] A. P. Genoud, Y. Gao, G. M. Williams, and B. P. Thomas, "A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals," *Ecological Informatics,* vol. 58, pp. 1-12, 2020. https://doi.org/10.1016/j.ecoinf.2020.101090

[25] G. R., "Naive Bayes classifier, towards data science. 2018," *https ://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c,* no. Accessed 25 Apr 2020., 2020.

[26] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," *Machine Learning,* vol. 62, no. 03, pp. 233-244,

2020. https://doi.org/10.1109/ICSET53708.2021.9612526

[27] O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," *https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761*, vol. Data & Machine Learning, 2018.

[28] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study," *JMIR public health and surveillance*, vol. 6, no. 2, p. e18828, 2020. https://doi.org/10.2196/18828

[29] F. Ishaq, L. Muhammad, B. Yahaya, and Y. Atomsa, "Data mining driven models for diagnosis of diabetes mellitus: a survey," *Indian J Sci Technol*, vol. 11, p. 42, 2018. https://dx.doi.org/10.17485/ijst/2018/v11i42/132665

[30] E. Ong, M. U. Wong, A. Huffman, and Y. He, "COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning," *Frontiers in immunology*, vol. 11, p. 1581, 2020. https://doi.org/10.3389/fimmu.2020.01581

[31] A. F. de Moraes Batista, J. L. Miraglia, T. H. R. Donato, and A. D. P. Chiavegatto Filho, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," *medRxiv*, 2020. https://doi.org/10.1101/2020.04.04.20052092