# Classification of COVID-19 Disease: A Machine Learning Perspective

**Kinza Sardar**

Department of Software Engineering, The University of Lahore, Lahore, Pakistan

**Abstract**
Nowadays the deadly virus famous as COVID-19 spread all over the world starts from the Wuhan China in 2019. This disease COVID-19 Virus effect millions of people in very short time. There are so many symptoms of COVID19 perhaps the Identification of a person infected with COVID-19 virus is really a difficult task. Moreover it's a challenging task to identify whether a person or individual have covid test positive or negative. We are developing a framework in which we used machine learning techniques..The proposed method uses DecisionTree, KNearestNeighbors, GaussianNB, LogisticRegression, BernoulliNB , RandomForest , Machine Learning methods as the classifier for diagnosis of covid ,however, 5-fold and 10-fold cross-validations were applied through the classification process. The experimental results showed that the best accuracy obtained from Decision Tree classifiers. The data preprocessing techniques have been applied for improving the classification performance. Recall, accuracy, precision, and F-score metrics were used to evaluate the classification performance. In future we will improve model accuracy more than we achieved now that is 93 percent by applying different techniques

***Keywords:***
*Covid-19, Machine Learning Techniques, Classification, Dimensionality Reduction*

## 1. Introduction

The most famous epidemic of COVID-19 disease, 2019.That is the reason can of severe respiratory disease in individuals that is the prospective threat to individual's health throughout worldwide, after the severe acute respiratory syndrome (SARS) pandemic 2003[1]. Machine learning (ML) exists as a subfield of artificial intelligence (AI).it is also known as predictive analysis. Hence machine learning approaches are supervised learning and unsupervised learning. Main focus of machine learning is on predictions that is based on training data in which properties have been learned from the given training data. Basically machine learning involves creating a model then the model is trained with some training data and then used for predictions with some additional data. So there are number of models that can be used for this purpose i.e. neural networks, Decision trees, Support vector machines, Regression analysis and Bayesian networks. These machine learning models required large amount of data for performing well. Different programing languages supports machine learning framework and libraries. As well as for programming machine learning preferred language that is used is python. The novel coronavirus named COVID-19 pandemic seemed in 2019 in Wuhan china on December 31, 2019 and it's effected the whole world [2]. Moreover [3] human deaths and disease are caused with respiratory disease (SARS-CoV) Severe acute respiratory syndrome Coronavirus and (MERS-CoV) Middle East respiratory syndrome Coronavirus. [4]Hence COVID-19 clinical features are temperature, nuisance, cough and breathe shortness. For the deduction of COVID-19 CT method is used as a screening tool[5].COVID-19 Radiologic images provides meaningful information for diagnostics[6].In the field of medical, machine learning methods are used for automatically diagnosis of the diseases and it's become popular day by day[7].Machine learning provides a way for analysis data. Now a days there are number of algorithms that are available for extracting hidden information on various biomedical datasets such as Neural networks, Decision Trees, Fuzzy Logic Systems, Naive Bayes, SVM, logistic regression[8]. At a large scale many problems have been using classification models for evaluation and then predict accuracy[9].

## 2. Literature Review

[10]Presented X-Ray COVID-19 screening, classic approach based on HoG and feature selection and performing well on deep learning methods. [11]Proposed Size Aware Random Forest method (iSARF) ,by using random forests classifications on x-rays images of COVID-19 patients and obtaining accuracy of 0.879.[12] Proposed a model on COVID-19 patients by using their abdominal Computed Tomography (CT) images and performed Support Vector Machines (SVM) classification and obtained accuracy of 99.68 percent with 10 cross fold validation and GLSZM. [13]Proposed a deep learning framework that diagnose COVID-19 in x ray images and provided a study of various deep learning architectures comprising of VGG19, DenseNet201, ResNetV2, InceptionV3, InceptionResNetV2, Xception and MobileNetV2 [14]Proposed Semi-supervised Open set Domain Adversarial network (SODA) to detect COVID-19

Manuscript received March 5, 2024
Manuscript revised March 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.3.13

disease by using chest x-ray images. [15]Adapted transfer learning respectively for extracting important biomarkers that are related to COVID-19 disease by using x ray images in which confirmed Covid-19 images are 224, images that are confirmed bacterial and pneumonia and images with normal conditions are 714 and 504, hence the proposed model achieved 96.78 percent accuracy, 98.66 percent sensitivity as well as obtained 96.46 percent specificity. [16] Mobile Net used for the classification task with a dataset of 3905 X-ray images based on 6 diseases that is used for training MobileNet v2 hence achieved 99.18% accuracy, 97.36% Sensitivity, and 99.42% Specificity for the finding of COVID-19 disease. For detection of COVID-19 disease, DarkNet model (YOLO) real time object detection system is used for classification with chest x ray images hence this model achieved binary classes accuracy of 98.08 percent and multi-class cases accuracy 87.02% [17]. [18]proposed a system based on deep learning that quantified and segmented the infected regions of chest CT images.

## 3. Dataset and Features

Dataset covers the clinical characteristics of those patient taken a COVID-19 test with HIPAA Privacy Rules De Identification standard. All batches have separate CSV file in Google Sheets tab with 1611 rows and 45 columns.

**Preprocessing:**

**Handling Null Values:**

There is always possibilities that datasets contains null (represented in python NaN) values.so either it's a classification or regression problem independent of this we have to handle such type of data. There are numerous ways to handle it. Drop those rows that contain null values for avoiding error. Initially our datasets contains 1611 rows and 45 columns. After removing missing values from our dataset now our dataset contains 1145 rows and 17 columns.

**Handle Imbalanced Data:**

After removing null values from dataset now dataset have 1145 rows and 17 columns. In 1145 rows only 77 rows belongs to positive class so data is not balance. To balance the dataset append positive rows due to imbalanced datasets. Our interested class is positive class. Now we are checking the records of that class 77 rows and 17 columns. Then append the positives and dataset becomes 1761 rows × 17 columns.

**Changing Data Types** :

After that we changed data types of few attribute values like asthma, diabetes, chd, htn, copd, autoimmune_dis and ctab . because we cannot process string data effectively and floating point coputation is very expensive resource wise.

**Feature Scaling: Standardization** is necessary at this step because each feature follow a different unit system. There are two ways of feature scaling 1) standardization and 2) normalization. Standardization technique is very useful to the data transformation. Standardize Numeric Attributes with different scales cannot participate in the same way for analysis since there are some chance it might create bais.

**Dimension Reduction:** Hence after analysis of dataset three algorithms of dimensionality reduction can be used 1) PCA generating new axis to get fewer dimensions also it identifies the most important features 2) LDA used for supervised leaning in which finding the correlation between feature and target class and 3) SVD for the very large amount of data that is generating in a very high speed and for the different types of data that is collected. For such type of data if preprocessing cannot be done then the model will be complex. If dimensions are in large number then automatically complexity will increases, also it take a lot of computation time and meanwhile it can also see that the chances of model over fitting will be higher.so that's why it would be a good option to reduce model complexity by applying dimensionality reduction techniques and then train a model and save computation time.so it can be observed when we need to process such kind of data that have properties of big data that data being generating in a great speed in great amount and in a high speed so to manage datasets dimensionality reduction techniques can be performed.it can also observed that if input good data to machine learning algorithm shown in fig below good output can be seen and vice versa. Applied PCA to datasets becomes 1761 rows x 3 columns (PCA_Component_1 PCA_Component_2 labels).

Moreover, performing PCA analysis, where PCA dimensional reduction technique is used to reduce the dimension of the datasets. And then train models and check the accuracy while during analysis it has been observed that before applied PCA dimensional reduction technique and after applied PCA, results of accuracy remains the same but processing time it takes will reduce.

## 4. Methods and Techniques

So now for this work, used different classifier for classification task for evaluating our dataset and then find out the accuracy of the model.

Machine learning techniques:

**Decision Tree:**

Decision Tree algorithm belongs to the parts of supervised learning algorithms. In contrast to other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The aim of using a "Decision Tree" is to create a training

model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In "Decision Trees", for predicting a class label for a record we start from the base of the tree. Welook at the estimations of the root attribute with the record's attribute. On the basis of comparison,we follow the branch corresponding to that value and jump to the next node.

**KNN:**
The k nearest neighbors (KNN) technique is a very straightforward, managed supervised machine learning calculation that can be utilized to take care of both arrangement and relapse issues. anything but difficult to execute and see, yet has a significant disadvantage of turning out to be fundamentally eases back as the size of that data being used develops. KNN works by finding the distances between a query and all the models in the datasets, selecting the predetermined number of models (K) nearest to the question, at that point votes in favor of the most incessant class (in the case of classification) or standard the classes (in the case of regression). The training duration of K nearest neighbor classification is a lot faster as compare to other classification algorithms. There is no more need to train a specific model for generalization, so that is why KNN is also called as the simple and instance-based learning approach . KNN can be useful in case of nonlinear data.

**Bernolli NB:**
This is a probabilistic classification technique. This classifier has been executed best result when applied to huge datasets. NBclassifier processes posterior probability by using the formula
posterior probability …….(1)
Equivalently,
P(Class i |z)= …….(2)
It utilizes Naive Bayes (NB) for a multivariate Bernoulli distribution of data. Consequently, every class needs samples, which must be characterized into binary values. Additionally, BernoulliNB can change over contributions of some other sort of information into binary form. The BernoulliNB standard is clarified as:
P(xi/y) = P(xi/y) xi+(1-p(i/y)) (1-xi) …… (3)
The distinction between Bernoulli Naive Bayes and Multinomial NB standard in that it unequivocally Punishes the non-occurrence of a feature that is a identify for class y, where the Multinomial variation would essentially overlook a non-occurrence of the features.

**Logistic Regression:**
This is Regression Model that is used for the classification purpose. LR is commonly used to relate only a categorical dependent variable to at least one autonomous variables. LR endeavors to discover a hyper-plane which expands the division hole between the classes. Logistic Regression uses a very complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function.
Sigmoid function S(z)= ……(4)
Random Forest:

RF is a combination method. Each and every classifier use in the "Random Forest" that is call"Decision Tree" classifier. RF classifier assembles a lot of decision trees from the training Dataset [Coronavirus Disease 2019 (COVID-19) Clinical Data Repository]. After fetching the votes from the non-identical
decision trees, it chooses the final class or label of the test entity. The estimated parameters values ofthe RF classifiers are tuned as: n estimators = 150, max depth = 07.

**TP**: TRUE POSITIVE, **FP**: FALSE POSITIVE, **TN:** TRUE NEGATVE and **FP:** FALSE NEGATIVE.
Precision =TP/(TP+FP)
Recall =TP/(TP+FN)
F1 score=2. (Precision.Recall)/( Precision+Recall)

**Research Methodology**

Layered architecture representation shown below is followed during this study in which we are working on multi layered architecture e.g. 1) Data source layer, 2) Data Staging: in this step data is filtered and preprocessing of datasets is performed, 3) Reconciled layer 4) loading: Machine learning classifier implemented at this stage and 5) Analyzing the result

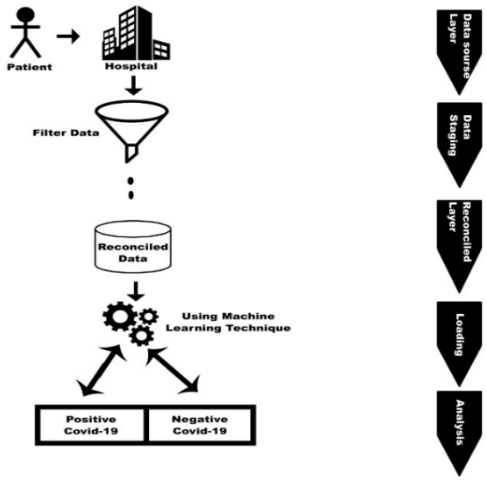


**Figure 1:** Multilayered Architecture Diagram

**Algorithm 1:** Prediction of COVID-19

**Begin**
Fetch from Excel sheets data is retrieved
For Data , Do:
Pre-processing procedures:
Eliminate Null values
Append positive rows : Over Sampling
Change datatype
Feature Scaling Standardize data
Perform dimension reduction
End Procedure
Classify using machine learning Techniques
End Until
**End**

The Algorithm followed for this study as shown below: as in step 1 preprocessing of data have be done through multiple techniques and then applied different classifier and performed classification and then evaluated results for the prediction of COVID-19.

## 5. Experiments/Results/Discussion

These steps have been taken for the machine learning.
Step I Data Preprocessing
Step II Test Train Split Prepare for Machine learning
Step III Train Model
Step IV Evaluate Results

The results of the machine learning model that trained using different classifiers models are presented in this section. Preprocessing is done on the labeled dataset. DecisionTree, KNeighbors, GaussianNB, LogisticRegression , BernoulliNB , RandomForest machine learning classifiers used for this experiment.

**Table 1:** 5-Fold Cross-Validations Scores

| Algorithm | Score1 | Score2 | Score3 | Score4 | Score5 |
|---|---|---|---|---|---|
| DecisionTree | 0. | 0. | 0. | 0. | 0. |
| KNearestNeighbors | 0. | 0. | 0. | 0. | 0. |
| GaussianNB | 0. | 0. | 0. | 0. | 0. |
| LogisticRegression | 0. | 0. | 0. | 0. | 0. |
| BernoulliNB | 0. | 0. | 0. | 0. | 0. |
| RandomForest | 0. | 0. | 0. | 0. | 0. |

**Table 2:** 10-FoldCross-Validations Scores

| Algorithm | Score | Score | Score | Score | Score | Score | Score | Score | Score | Score |
|---|---|---|---|---|---|---|---|---|---|---|
| DecisionTree | 0.96 | 0.94 | 0.97 | 0.95 | 0.95 | 0.954 | 0.90 | 0.97 | 0.97 | 0.97 |
| KNearestNeighbors | 0.90 | 0.85 | 0.89 | 0.88 | 0.88 | 0.818 | 0.90 | 0.90 | 0.92 | 0.86 |
| GaussianNB | 0.41 | 0.44 | 0.43 | 0.43 | 0.42 | 0.443 | 0.41 | 0.46 | 0.42 | 0.43 |
| LogisticRegression | 0.66 | 0.69 | 0.67 | 0.70 | 0.65 | 0.664 | 0.64 | 0.66 | 0.62 | 0.66 |
| BernoulliNB | 0.61 | 0.60 | 0.59 | 0.65 | 0.64 | 0.590 | 0.63 | 0.62 | 0.6 | 0.64 |
| RandomForest | 0.96 | 0.93 | 0.97 | 0.95 | 0.92 | 0.965 | 0.94 | 0.90 | 0.92 | 0.86 |

DecisionTree,KNN,GaussianNB,LogisticRegression ,

BernoulliNB , RandomForest machine learning classifiers predicted accuracy given in the table.Decision Tree classifier obtained better accuracy when compared with other classifiers results.

**Table 3:** Accuracy of Classifiers

| Algorithm | Accuracy |
|---|---|
| DecisionTree | 0.93 |
| KNearestNeighbors | 0.84 |
| GaussianNB | 0.63 |
| LogisticRegression | 0.65 |
| BernoulliNB | 0.65 |
| RandomForest | 0.88 |

**Table 4:** Evaluating Results

| Classifier | Precision | Recall | F-Measure |
|---|---|---|---|
| DecisionTree | 0.93 | 0.95 | 0.9 |
| KNeighbors | 0.85 | 0.87 | 0.8 |
| GaussianNB | 0.61 | 0.61 | 0.6 |
| LogisticRegression | 0.62 | 0.60 | 0.6 |
| BernoulliNB | 0.66 | 0.66 | 0.6 |
| RandomForest | 0.87 | 0.88 | 0.88 |

Graphical representation of different machine learning algorithms predicted accuracy in the given table. All classifier result comparison showed that decision tree classifier gives better accuracy as compared to other.

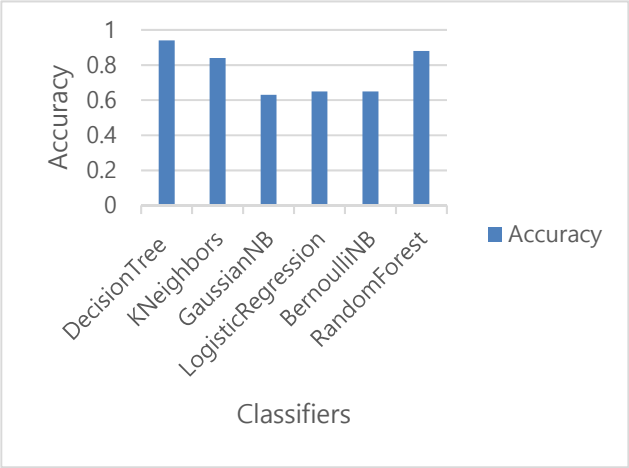


**Figure 2:** Proposed Classification Model

Figure graphically represents the percentage of the precision, recall, and F-measure for each machine learning classifier.

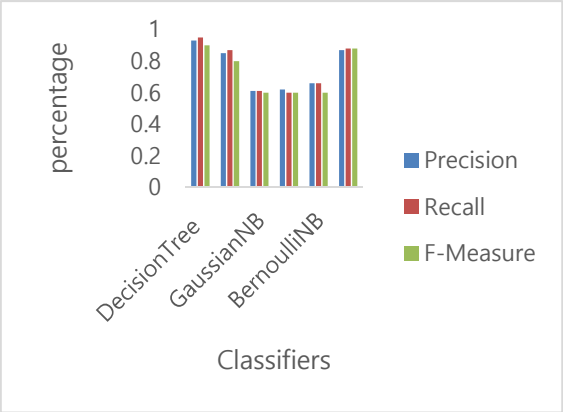


**Figure 3: Evaluating Results**

## Before Balancing: BB After Balancing: AB

In the above table Before Balancing result indicates with" BB" and After Balancing result indicated with "AB". As shown in the table it can be observed in the case of k nearest neighbor. The accuracy is above 90 percent. The negative class precision is 0.92 and the positive class precision is 0.00. The recall of the negative class is 1.00 that is perfect. But the recall of the positive class is 0.00 it can be observed that it would be the worst case. Therefore, we can determine that our model is not performing well **for positive class**. The F1 score is 0.92 for the negative class. Although the model accuracy is good but when we observed each class result than it represented that our model is not working correctly. So that we can say that it's an inadequate model.so after analyzing such type of results we again train our model with balance datasets and achieved good results. All the results shown in table.

**Table:** Accuracy Comparison of Imbalanced dataset with balanced dataset

| Algorithm | Balanced dataset Accuracy | Imbalanced dataset Accuracy |
|---|---|---|
| DecisionTree | 0.93 | 0.86 |
| KNearestNeighbors | 0.84 | 0.92 |
| LogisticRegression | 0.65 | 0.91 |
| BernoulliNB | 0.65 | 0.91 |
| RandomForest | 0.88 | 0.92 |
| SVM | 0.75 | 0.92 |

As seen in the table when imbalance datasets used for model training purpose, good accuracy results achieved. But it can also observed that whether we are achieving good accuracy by using such type of imbalanced data, we achieved good accuracy but our model is baised.it can also see in table results evaluation that F1 score value is very small nearest to zero for almost all classifiers. For instance decision tree classifier show 0.06 F1 score in the case of minority class.as expected it is zero. It means our model is not working properly in the given situation where dataset is not balance, vice versa it can be observed in other case when our dataset is balanced than results indicates that our model is working properly.

## 7. Conclusion/Future Work

COVID-19 was initially come across at Wuhan, China and badly effecting individual's health, business as well as world economy. The symptoms of this virus shows are similar like viral pneumonia. So that, the ratio of this virus spreading is very high and almost now it's uncontrollable. In this paper, for the machine learning classification purpose COVID-19 disease datasets is used. All The dataset has been preprocessed by using preprocessing techniques. Machine learning algorithm DecisionTree, K Nearest Neighbors, GaussianNB, LogisticRegression, BernoulliNB, RandomForest used for this experiment. For this study performance parameters like accuracy, precision, recall and F-Measure have been calculated. 5-fold cross validation, 10-fold cross validation have been calculated for estimating the generalization accuracy of the model. Hence the results suggested that the best accuracy obtained from the classifier decision tree.while GaussianNB lowest accuracy Moreover PCA dimension reduction technique analysis have been analyzed that showed that accuracy remains the same but processing time it takes will reduce. In future, we will definitely increase accuracy of our model by applying different techniques.

## 8. References


[1] Who, "Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003." WHO Geneva, Switzerland, 2003.

[2] K. Roosa *et al.*, "Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020," *Infect. Dis. Model.*, vol. 5, pp. 256–263, 2020.

[3] W. Kong and P. P. Agarwal, "Chest imaging appearance of COVID-19 infection," *Radiol. Cardiothorac. Imaging*, vol. 2, no. 1, p. e200028, 2020.

[4] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *Indian J. Pediatr.*, pp. 1–6, 2020.

[5] E. Y. P. Lee, M.-Y. Ng, and P.-L. Khong, "COVID-19 pneumonia: what has CT taught us?," *Lancet Infect. Dis.*, vol. 20, no. 4, pp. 384–385, 2020.

[6] J. F.-W. Chan *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster," *Lancet*, vol. 395, no. 10223, pp. 514–523, 2020.

[7] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Comput. Methods Programs Biomed.*, vol. 161, pp. 1–13, 2018.

[8] P. Herron, "Machine learning for medical decision support: evaluating diagnostic performance of machine learning classification algorithms," *INLS 110, Data Min.*, pp. 1–16, 2004.

[9] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

[10] D. Gil, K. Díaz-Chito, C. Sánchez, and A. Hernández-Sabaté, "Early Screening of SARS-CoV-2 by Intelligent Analysis of X-Ray Images," *arXiv Prepr. arXiv2005.13928*, 2020.

[11] F. Shi *et al.*, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification," *arXiv Prepr. arXiv2003.09860*, 2020.

[12] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," no. 5, pp. 1–10, 2020, [Online]. Available: http://arxiv.org/abs/2003.09424.

[13] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv Prepr. arXiv2003.11055*, 2020.

[14] J. Zhou, B. Jing, and Z. Wang, "SODA: Detecting Covid-19 in Chest X-rays with Semi-supervised Open Set Domain Adaptation," 2020, [Online]. Available: http://arxiv.org/abs/2005.11003.

[15] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, 2020, doi: 10.1007/s13246-020-00865-4.

[16] I. D. Apostolopoulos, S. I. Aznaouridis, and M. A. Tzani, "Extracting possibly representative COVID-19 Biomarkers from X-Ray images with Deep Learning approach and image data related to Pulmonary Diseases," *J. Med. Biol. Eng.*, p. 1, 2020.

[17] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, p. 103792, 2020.

[18] F. Shan *et al.*, "Lung infection quantification of covid-19 in ct images with deep learning," *arXiv Prepr. arXiv2003.04655*, 2020.