Early detection of Covid-19 based on preliminary features using Machine learning algorithms

Madhav Sharma[1], Ujjawal Prakash[2], Anshu Kumari[3], Kanika Singla[4*]
Department of Computer Science Engineering
School of Engineering & Technology, Sharda University, Greater Noida, India
[1]2018009090. madhav@ug.sharda.ac.in,
[2] 2018012300.ujjwal@ug.sharda.ac.in,
[3] 2018016097.anshu@ug.sharda.ac.in
[4]kanika.singla@sharda.ac.in*

Abstract. In this paper, early detection of Corona virus has proposed using the some machine learning techniques. Corona virus has been the most exceptionally the infectious and dangerous disease in year 2020. By analyzing the significant clinical symptoms, early detection of this disease in patients can be done. A clinical dataset has been used for classification of the disease using Support Vector Machine, Random Forest algorithms and Neural Network. Neural network has been the most promising algorithm in providing the best performance parameters like F1 score, Recall, precision, confusion matrix etc. Experimental analysis shows that neural network outperforms due to the approximation nature of the activation functions.

Keywords: Covid**,** Covid-19, Symptoms, Supervised learning, Support Vector Machine, Random forest, Artificial Neural Network.

1 Introduction

An infectious disease, COVID-19, has spread across the globe. Back in late 2019, it was first seen in Wuhan, a city in China. WHO named COVID-19 as a "PANDEMIC" on 11 March 2020. Many countries banned the international travel and till now it is banned in many countries .The World trade totally disturbed by this pandemic, and all the major economies went into recession **[1].** Respiration system gets most affected during this infection which can lead to death sometimes, especially in old age group people **[5].**

The symptoms of COVID-19: fever, sore throat, chest pain, dry cough, diarrhea etc. Millions of people have been infected by this virus. The side effects of Covid-19 are comparable to the indications of the common cold or flu. The size of the droplets may be typically between 5-10 microns. Lots of world wide scientific community have been analyzing the effects of covid-19, to forsee the economic situation, proper planning of hospitals and medical facilties , socio-political decion making etc.

As per the World Health Organization (WHO) data, COVID-19 or coronavirus causes respiratory illness and is spread through respiratory droplets and close distance contacts. These infectious droplets may potentially enters into your body. In the present trend the use of ML models has been leveraged to produce better results **[3]** as the COVID-19 has been a irresistible illness through respiratory system.

In this paper, Machine learning models **[2]** was implemented for the early detection of early detection of Covid-19 based on preliminary features.

The paper has been divided into different sections which are as follows:

- Section 1: This section conatins the introduction to the paper, aim, motivation and objectives of the problem statement.
- Section 2: In this section, we discuss the related work or the literature survey of the concepts used during the research.
- Section 3: This section includes methods, tools used to achieve objectives of the research problem. Performance metrics and experimental analysis has been conducted in this section respectively for Support Vector Machine (SVM), Random Forests (RF) & Artificial Neural Networks (ANNs).
- Section 4: Results obtained has been shown in this section.
- Section 5: This section contains the analysis and discussions about the result, contribution of the work to the existing research. Possible future scope has also been discussed in this section.

2 Literature Survey

Machine learning has a great capability of prediction and classification in health care systems **[2][3]** using the forecasting mechanisms by applying an appropriate algorithm prediction of cases, number of deaths can be predicted which helps the health care institution to prepare well for the future and make system more robust from earlier. Another line of work in the health care field using machine learning models **[4]** using SVM , Random forest etc.

Another techniques like Linear Regression , Least Absolute Shrinkage and Selection operator , Support Vector Machine **[5]** have been used in the study for forecasting the threatening factors of COVID-19.

Some state of the art solutions are AI and Big Data to fight against the virus. Generation of big data leads to more accurate condition of the world, by analyzing the big data and incorporating AI techniques such as ML, DL etc. The models can detect the covid positive patients easily **[6].**

Another work in this line of approach compares Polynomial Regression Algorithm with Support Vector Machine Algorithm **[7]** . The former method showed an accuracy of approx. 93% by predicting the surge in cases for the months of July and August.

The intention of Artificial Intelligence (AI) is to facilitate human limits for better results or output. Availability of clinical data AI is getting a standpoint on human administrations **[8].** Need of AI is to fight help the world to fight against the COVID-19 Crisis, also highlighting the application of Big Data. Methods used are like: Neural Networks, SVM, and Edge Significant Learning **[9].**

The most significant research in this area of work has been done by the authors **[17].** In this research, epidemiological, demographic, clinical, laboratory, radio-logical and treatment data from Zhongnan Hospital was analyzed. Radio-logical treatment has been given much leverage.

A new framework has been proposed for detection using the inboard smartphone sensors. Framework will first collect the data from the sensors and predicts the infection of the disease **[18].**

In all the above mentioned work, many prediction based system has been developed using dataset containing images. Not much work has been done on the early detection of preliminary features of this infectious disease. So, this paper is focused on the addressing this limitation. We will be using machine learning techniques to predict COVID-19 with clinical information of patients suffering from COVID-19.

3 Methodology used for early detection of preliminary features

3.1 Techniques

There are number of classification models in machine learning like logistic regression, decision tree, neural network, SVM, Bayes classifier etc . During this work, three supervised classification techniques has been performed and examined:

A. **Support Vector Machines** SVM is one of the most popular Supervised Learning techniques which can be used for both Regression and Classification **[7]**. It has the concept of nonlinear kernels for creating the decision boundary for nonlinear data. Primarily, SVM-C is used for Classification problems. The purpose of the SVM is to establish the boundary of the decision or the best line that can separate n-dimensional space into groups, which provides better intuition to calculate hinge losses between the hyper planes[**11]**. Hyper plane is created with the help of SVM's chosen extreme points/vectors and these extreme cases are known as support a vector which leads to name this algorithm as Support Vector Machine**[10].**

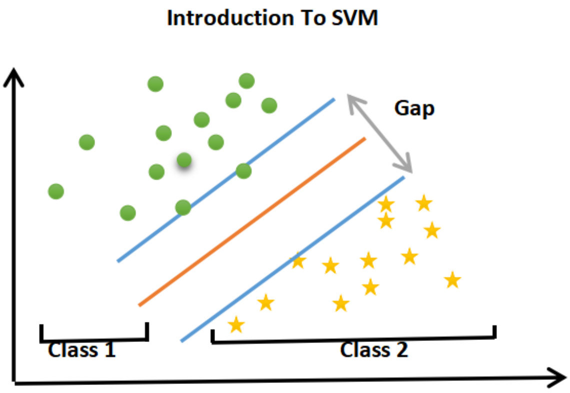


Fig. 1. Support Vector Machine represented in a graphical way.

B. Random Forest. Like SVM, this technique can also be used for both regression &classification **[12].** This technique is based on the basis of Ensemble Learning. Ensemble Learning is combination of multiple classifiers for solving a complex problem and for improving the performance of the model **[7].** The random sampling and ensemble strategies used in RF will help in predicting accurate and better generalization **[16].**

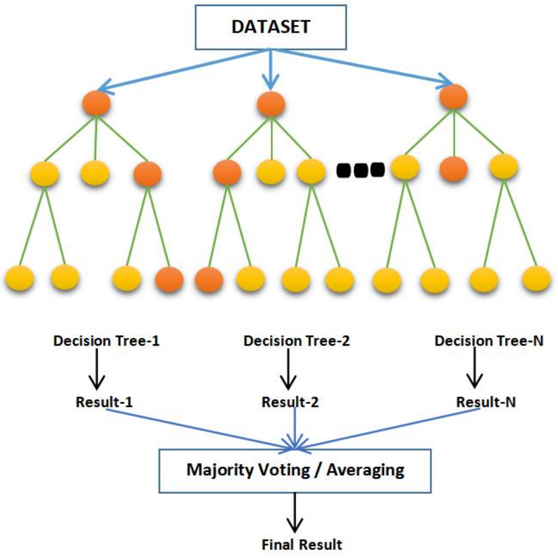


Fig. 2. Visualization of Random Forest techniques.

C. **Artificial Neural Networks.** ANNs are the used to mimic the human brain system. **[7].** It is the kind of framework used to analyze the real world data as the human brain does. It is the foundation of Artificial intelligence(AI) that helps in solving the problems which seems impossible or hard for human brains to solve due to its complexity **[13].**

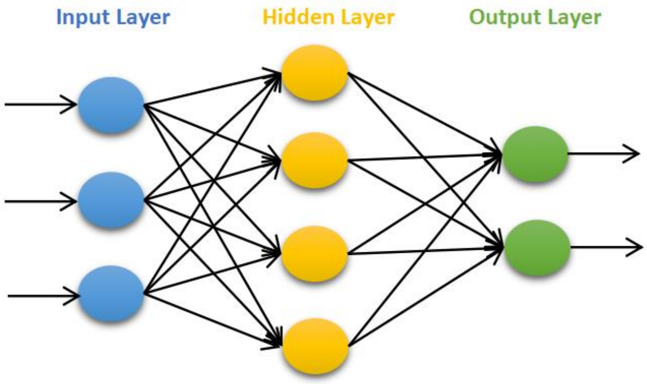


Fig. 3. Basic Artificial Neural Network composition.

3.2 Dataset

The experiments were conducted on the clinical dataset. The dataset has been a public dataset [**18].** The data set contains information about hospitalized patients with COVID-19.

Table 1: Composition of Dataset

Number of feature vectors	***21***
Numberof observations/examples	**5434**
No of Missing cells	**0**
Missing cells (%)	**0.0%**
Memory size	**891.6 KB**
Average record size in memory	**168.0 B**

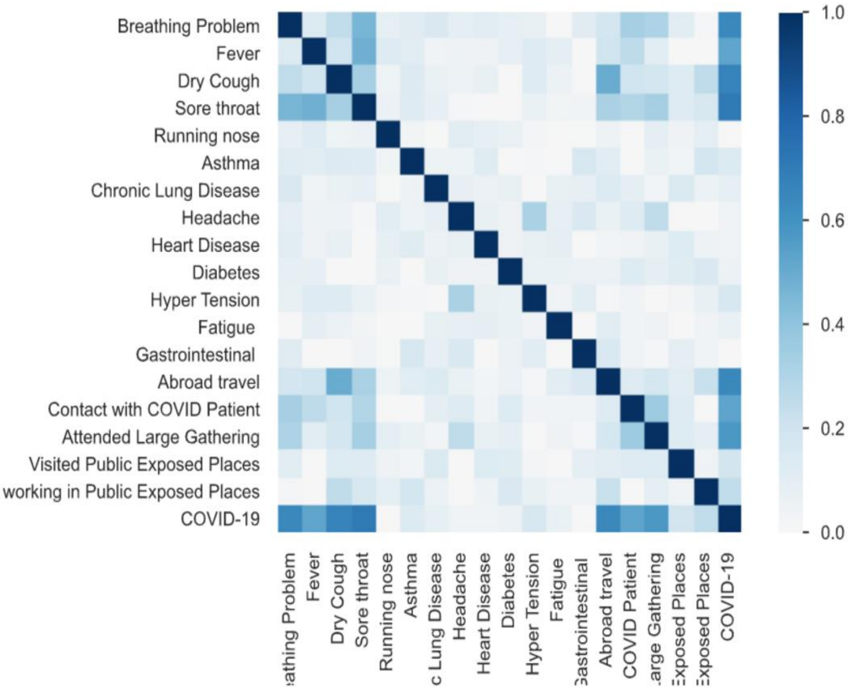


Fig 4: Correlation between the feature vectors

3.3 Flow of the work

Basic machine learning algorithms contains few basic steps as pipeline. In medical and healthcare application, data processing is very crucial to get the accurate results, as wrong predictions may lead to loss of human life. Extracting the essential features from the datset before training the training of the model will give better results.

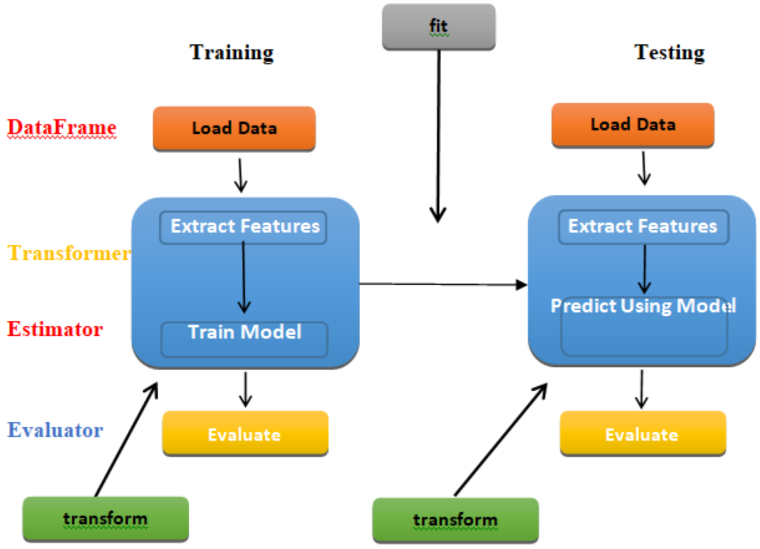


Fig 5: Process of Solving the Problem

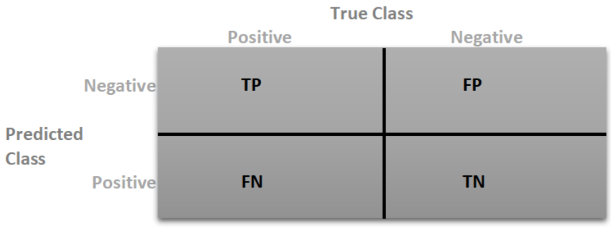
3.4 Performance Metrics

a. **Accuracy** . Accuracy is used for evaluation of classification models. For binary classification accuracy can also be calculated in terms of positives and negatives:

Accuracy= (TP+TN)/ (TP+TN+FP+FN) .eq(1)

Where TP = True Positives, TN = True Negatives, FP = False Positives, & FN = False Negatives.

b. **Confusion Matrix**. The Confusion matrix is a table frequently used to describe classification model's results.



		True Class	
		Positive	Negative
Predicted Class	Negative	TP	FP
	Positive	FN	TN

Fig 6: Confusion Matrix

i. **True Positive (TP)**

If the value which has been estimated, is same as the real value. The actual value was favorable; a positive value was also expected by the model.

ii. True Negative (TN)

When predicted value comes to be same as the actual value. Actual value was negative; model also predicted a negative value

iii. False Positive (FP)

When predicted value comes to be falsely predicted. The model predicted a positive value but actual value was negative. It is also known as "**Type 1 error**"

iv False Negative

It is also known as Type-2 error. It is the error when model estimated false, and the actual result is also.

c. Classification Report

i. **Precision:** A classifier's ability not to mark a positive example that is actually negative.

Precision = TP/(TP + FP) .eq(2)

ii. **Recall**: The capacity of a classifier to classify all positive examples.

Recall = TP/(TP + FN) .eq(3)

iii. **F1 score**: Weighted harmonic mean of precision and recall. 1.0is the best score and 0.0 is the worst.

F1 score = 2*(Recall * Precision) / (Recall + Precision) .eq(4)

iv. **Support**: The number of actual class occurrences in a dataset that is listed. In the training results, the imbalanced support could indicate systemic flaws in the classifier's recorded scores and may indicate the need for re-balancing.

4 Experiment Results and discussions

This section presents the results that are obtained from the experiment. All the performance metrics that has been explained in the above section has been utilized to evaluate the performance. All the methods mentoned in Section 3.1 has been exploited over this dataset.

4.1 Results based on SVM and Random Forest

The result section has been divided in two parts. This result section will give the result for SVM and Random forest.

The next result section will be based on neural network. Below is the table for comparison

i. Accuracy for SVM. and Random-forest

Table 2: Accuracy of SVM and Random-Forest

SVM.	*Random Forest*
97.79	**98.16**

ii. Confusion Matrix of SVM

Table 3: Confusion Matrix of SVM

		Positive	*Negative*	
Actual Values	**Positive**	**TP=179**	**FN=13**	**Sensitivity =0.93**
	Negative	**FP=11**	**TN=884**	**Specificity =0.98**
		Precision= 94.2	**Negative prediction value=0.98**	**Accuracy= 0.977**

Confusion Matrix of Random-Forest

Table 4: Confusion Matrix of Random-Forest

		Positive	*Negative*	
Actual Values	**Positive**	**TP=184**	**FN=8**	**Sensitivity =0.95**
	Negative	**FP=12**	**TN=883**	**Specificity =0.95**
		Precision= 93.8	**Negative prediction value=0.99**	**Accuracy= 0.981**

iii. Classification report of SVM

Table 5: Classification Report of SVM

	Precision	*Recall*	*F1-score*	*Support*
0	**0.94**	**0.93**	**0.94**	**192**
1	**0.99**	**0.99**	**0.99**	**895**

Accuracy			0.98	1087
Macroavg	0.96	0.96	0.96	1087
Weighted-avg	0.98	0.98	0.98	1087

Classification report of Random Forest

Table 6: Classification Report of Random Forest

	Precision	*Recall*	*F1-score*	*Support*
0	0.94	0.96	0.95	192
1	0.99	0.99	0.99	895
Accuracy			0.96	1087
Macroavg	0.96	0.97	0.97	1087
Weighted-avg	0.98	0.98	0.98	1087

4.2 Results based on ANN Model

Due to the presence of ReLu activation function which actually helps in breaking the linearity of the data. **The accuracy is 98.62.**

i. Confusion Matrix

Table 7: Confusion Matrix of ANN

		Positive	*Negative*	
Actual Values	**Positive**	**TP=190**	**FN=2**	**Sensitivity =0.98**
	Negative	**FP=13**	**TN=882**	**Specificity =0.98**
		Precision= 93.5	**Negative prediction value=0.99**	**Accuracy= 0.987**

ii. Classification Report

Table 8: Classification Report of ANN

	Precision	*Recall*	*F1-score*	*Support*
0	0.94	0.99	0.96	192

1	**1**	**0.99**	**0.99**	**895**
Accuracy			**0.96**	**1087**
Macroavg	**0.96**	**0.97**	**0.98**	**1087**
Weighted-avg	**0.98**	**0.98**	**0.99**	**1087**

5 Conclusion and Future Scope

In this paper, a well organized literature work has been conducted for the existing algorithms for COVID-19 prediction and classification, but no as such algorithm has been found for the early detection of the preliminary features of the infectious disease. Therefore, few supervised algorithm like Support Vector Machine (SVM), Random Forests (RF) & Artificial Neural Networks (ANNs) were implemented for the clinical dataset. These algorithms were trained on the said dataset. The outcomes shows that Random-Forest accomplished maximum accuracy i.e 98.16 which surpassed SVM's accuracy i: e 97.79. But, implementing ANN model on the same dataset shows that "ANN model surpassed the ML algorithm's accuracy rate by achieving 98.62" which is "even more accurate , which clearly shows model based on ANN is giving accuracy more than ML based models. Moving towards non-linearity i: e from Machine learning algorithms to Deep Learning Algorithms,we can observe more good results.

We hence believe that calibration of ensemble methods and deep learning models and architectures can provide better solutions to the complex datasets which are highly non-linear in nature. The future scope may include finding the more non-linear dataset for COVID-19. Moreover, to identify and help diagnose the disease, various AI based applications can be developed.

References

1. F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.
2. Awatramani, V., & Gupta, D. (2021). Parkinson's Disease Detection Through Visual Deep Learning. In *International Conference on Innovative Computing and Communications* (pp. 963-972). Springer, Singapore.
3. Vashisht, G., Jha, A. K., & Jailia, M. (2021). Predicting Diabetes Using ML Classification Techniques. In *International Conference on Innovative Computing and Communications* (pp. 845-854). Springer, Singapore.

4. Jain, P., Babu, C. A., Mohandoss, S., Anisham, N., Gadade, S., Srinivas, A., & Mohan, R. (2021). A Novel Approach to Classify Cardiac Arrhythmia Using Different Machine Learning Techniques. In *International Conference on Innovative Computing and Communications* (pp. 517-526). Springer, Singapore.
5. E. Gambhir, R. Jain, A. Gupta and U. Tomer, "Regression Analysis of COVID-19 using Machine Learning Algorithms," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 65-71, doi: 10.1109/ICOSEC49089.2020.9215356.
6. M. Jamshidi et al., "Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment," in IEEE Access, vol. 8, pp. 109581-109595, 2020, doi: 10.1109/ACCESS.2020.3001973.
7. S. jie and H. Wankun, "Experimental Results of Maritime Target Detection Based on SVM Classifier," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 2020, pp. 179-182, doi: 10.1109/ICICSP50920.2020.9232038.
8. A. A. Hussain, O. Bouachir, F. Al-Turjman and M. Aloqaily, "AI Techniques for COVID-19," in IEEE Access, vol. 8, pp. 128776-128795, 2020, doi: 10.1109/ACCESS.2020.3007939.
9. E. Casiraghi et al., "Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments," in IEEE Access, vol. 8, pp. 196299-196325, 2020, doi: 10.1109/ACCESS.2020.3034032.
10. R. I. H. Ortiz, J. C. B. Barrera and K. M. B. Barrera, "Analysis model of the most important factors in Covid-19 through data mining, descriptive statistics and random forest," 2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2020, pp. 1-8, doi: 10.1109/ROPEC50909.2020.9258765.
11. E. -S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid and S. E. Hussein, "Novel Feature Selection and Voting Classifier Algorithms for COVID-19 Classification in CT Images," in IEEE Access, vol. 8, pp. 179317-179335, 2020, doi: 10.1109/ACCESS.2020.3028012.
12. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.
13. Bauer, E. and Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1-2), pp.105-139.
14. Liu, W., Zhang, Q., Chen, J., Xiang, R., Song, H., Shu, S., Chen, L., Liang, L., Zhou, J., You, L. and Wu, P., 2020. Detection of Covid-19 in children in early January 2020 in Wuhan, China. New England Journal of Medicine, 382(14), pp.1370-1371.
15. Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA..
16. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., ... & Peng, Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama*, *323*(11), 1061-1069.

17. Maghdid, H. S., Ghafoor, K. Z., Sadiq, A. S., Curran, K., & Rabie, K. (2020). A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: Design study. *arXiv preprint arXiv:2003.07434*.
18. “COVID-19 Symptoms Checker | Kaggle.” https://www.kaggle.com/iamhungundji/covid19-symptoms-checker (accessed Aug. 13, 2020).