# COVID-19 Prediction using Machine Learning Algorithms

**Saily Suresh Patil**

104, Leelawati Residency, Behind Chruch, Jalgaon, Maharashtra, India
7304702728
Email: *saily.s.patil[at]gmail.com*

**Abstract:** *Population aging and the increase of chronic conditions incidence and prevalence produce a higher risk of hospitalization or death due to COVID-19. This is particularly high for patients with multimorbidity leading to great resource consumption. Most important challenge is to identify possible high-risk patients in order to improve health care service provision and also to reduce costs. Nowadays, population health management, based on intelligent models, can be used to assess the risk and identify these "complex" patients infected with COVID-19. Main focus of project is to implement machine learning algorithms SVM, Knn, MLP & Decision tree to predict the risk of hospitalization or death starting from administrative and socio-economic dataset. Thus the training would be provided on training dataset which would help in predicting results for testing phase dataset provided.*

**Keywords:** data mining, machine learning

## 1. Introduction

Coronavirus Disease (COVID-19) is an infectious disease caused by a novel coronavirus that originated in Wuhan China last December (2019). This disease will affect the respiratory system of a person, and some people will eventually get better without having special treatment, especially for those who have a strong immune system [1]. For others, though, it may be different—old persons are more vulnerable, including those with existing comorbidities such as cardiovascular disease, diabetes, respiratory disease and cancer. COVID-19 is not just a respiratory disease, it is multisystemic. Recent studies determined that this virus affects almost all the organs of the body, which is stimulated by its widespread inflammatory response [2]. Moreover, about 10–15% of COVID-19 patients develop severe symptoms; these individuals may experience long COVID-19, which may cause complications to the heart, lungs, and nervous system [3]. COVID-19 can spread because this virus is transmissible by droplets into the air from the infected person through speaking, coughing, and sneezing, or even touching some contaminated objects or areas.

The World Health Organization (WHO) stated that frequent handwashing, disinfecting, social distancing, wearing masks and not touching your face can protect one from being infected. The WHO listed several symptoms and emphasized that fever, a dry cough, and tiredness were the most common, while less common symptoms were headaches, sore throat, diarrhea, conjunctivitis, loss of smell, and rashes, and serious symptoms were breathing problems, chest pain and loss of speech and movement. As of 29 June 2021, there were 182,333,954 COVID-19 cases and 3,948,238 deaths worldwide [4], and this disease had mutated into several variants documented in countries such as the United Kingdom, South Africa, the United States, India and Brazil, which brings increased severity to the disease, as well as quicker transmission, a higher death rate and reduced effectivity of vaccines [5].

As the virus keeps on spreading despite the efforts of the community to contain the virus, an outbreak can lead to increased demands in hospital resources, and shortages of medical equipment, healthcare staff and of course COVID-19 testing kits [6]. Limited access to COVID-19 testing kits can hinder the early diagnosis of the disease, and giving the best possible care for the suspected COVID-19 patients can be burdensome. Consequently, an automatic prediction system that aims to determine the presence of COVID-19 in a person is essential. Machine learning classification algorithms, datasets and machine learning software are the necessary tools for designing a COVID-19 prediction model.

## 2. Related Work

Machine learning can be categorized as supervised, unsupervised, and reinforcement learning. Supervised machine learning is an approach that trains the machine using labeled datasets, wherein the examples are correctly labeled according to the class to which they belong [7]. The machine will analyze the given data and will eventually predict new instances based on information learned from the past data. Unlike supervised machine learning, the unsupervised machine learning learns by itself without the presence of the correctly labeled data. In unsupervised machine learning, the machine will be fed by the training samples, and it is the job of the machine to determine the hidden patterns from the dataset. For the reinforcement learning, the machine acts as an agent that aims to discover the most appropriate actions through a trial-and-error approach and observation in the environment [8]. Every time the machine successfully performs a task, it will be rewarded by increasing its state; otherwise, it will be punished by decreasing its state, and this approach will be repeated several times until the machine learns how to perform a specific task correctly. Reinforcement learning is used in training robots in how to perform human-like tasks and personal assistance.

This study is mainly focused on predicting the presence of COVID-19 in a person; thus, a supervised machine learning model had to be developed. Several machine learning methods were utilized in building disease prediction models (e.g., coronary artery disease, respiratory disease, breast cancer, diabetes, dementia, and fatty liver disease [9–14]). The researchers devised a list of published disease prediction studies that utilized supervised machine learning algorithms, such as J48 Decision Tree (J48 DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (k-NN), and Naïve Bayes (NB). More algorithms were utilized, such as Multi-layer Perceptron (MLP), Logistic Regression (LR) and Artificial Neural Network (ANN). displays a list of disease prediction studies, machine learning algorithms utilized, the best algorithms according to the experiments performed, and the accuracy obtained.

## Need of System

The need of this project is to describe a framework for an iterative approach to event discovery high risk patient's that is based upon a clinical data repository and natural language processing techniques. We summarize our definitions and assumptions and the literature describing feature extraction, feature selection, classification and we discuss performance evaluation related to the approaches.

## Application of System

1) Proposed system can be implemented for detecting high risk patients in pandemic like Covid-19
2) It can also be implemented for detection of life threatening diseases with high severity

## 3. System Design

### Introduction to proposed system:

For this paper, we will adopt the following error related definitions. A medical error is defined as the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim [1]. An adverse event is defined as an injury caused by medical management rather than by the underlying disease or condition of the patient [1]. An adverse outcome is defined as an undesirable and unintended outcome of care such as prolonged hospitalization, disability, or death at the time of discharge [2]. A near miss (as defined by the Agency for Healthcare Research and Quality's Center for Quality Improvement and Patient Safety) is an event in which the unwanted consequences were prevented because there was a recovery by identification and correction of the failure. The recovery might be planned or unplanned.

### Objectives

- Medical error accompanied by an adverse outcome (e.g., a drug rash due to prescribing a medication to which the patient is known to be allergic);
- Near miss, which is a medical error from which there has been a recovery (e.g., if the pharmacist catches a prescription to a medication to which the patient is known to be allergic);
- Medical error without a recovery but without an adverse outcome because of luck or the robustness of human

physiology (e.g., a medication to which the patient is known to be allergic is prescribed, dispensed, and taken, but the patient has no reaction);
- Adverse outcome without an error (e.g., an allergic reaction to a medication for which there was no known allergy).

## 4. Algorithms

### k-nearest neighbor algorithm

The k-nearest neighbor algorithm is a powerful non-parametric technique [13] for density estimation [20] and classification [11]. It is a instance-based learning or lazy learning algorithm, where the approximation happens only locally and all the computation is done after classification. [14]

### Advantages:

1) Ease of implementation and debugging.
2) Some Noise reduction techniques can improve the accuracy of the algorithm. [15]
3) Case-retrieval Nets can improve the run-time performance for large data sets. [16]

### Support vector machines

Support vector machine is a discriminative classification algorithm defined by separating hyperplane. It is a supervised learning algorithm in which the training data set is labeled which outputs optimal hyperplane categorizing new datasets.

### Advantages:

1) Effective in large scale regression problems. [17]
2) Effective in high dimensional space. [18]
3) Effective In cases number of samples is smaller than number of dimensions.
4) Memory efficient due to use of support vectors with slight modifications. [19]

## 5. Conclusion

Thus the implementation of machine learning algorithms SVM, Knn, MLP & Decision tree to predict the risk of hospitalization or death starting from administrative and socio-economic dataset has been implemented and tested with standard dataset. Thus the training was provided on training dataset which has helped in predicting results for testing phase dataset provided.

Our future work includes further extending the algorithms, so that it can handle big database with large number of values.

One more future study can be implemented for testing performance of system by using combination of multiple algorithms and comparing their results.

# References

[1] World Health Organization (WHO). Coronavirus 2021. Available online: https://www.who.int/health-topics/coronavirus (accessed on 23 May 2021).

[2] Temgoua, M.N.; Endomba, F.T.; Nkeck, J.R.; Kenfack, G.U.; Tochie, J.N.; Essouma, M. Coronavirus Disease 2019 (COVID-19) as a Multi-Systemic Disease and its Impact in Low- and Middle-Income Countries (LMICs). SN Compr. Clin. Med. **2020**, 2, 1377–1387. [CrossRef]

[3] Ames, H. How Long Does Coronavirus Last in the Body, Air, and in Food? Available online: https://www.medicalnewstoday. com/articles/how-long-does-coronavirus-last (accessed on 11 June 2020).

[4] Worldometer. COVID Live Update, 29 June 2021. Available online: https://www.worldometers.info/coronavirus/ (accessed on 29 June 2021).

[5] Centers for Disease Control and Prevention (CDC). SARS-Cov-2 Variant Classifications and Definitions, 17 May 2021. Available online: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html (accessed on 23 May 2021).

[6] Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Bonten, M.M.J.; Dahly, D.L.; Damen, J.A.; Debray, T.P.A.; et al. Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. BMJ **2020**, 369, m1328. [CrossRef]

[7] Supervised vs. Unsupervised Learning: Key Differences. Available online: https://www.guru99.com/supervised-vsunsupervised-learning.html (accessed on 27 May 2021).

[8] Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement Learning: A Survey. J. Artif. Intell. Res. **1996**, 4, 237–285. [CrossRef]

[9] Abdar, M.; Ksiazek, W.; Acharya, U.R.; Tan, R.-S.; Makarenkov, V.; Plawiak, P. A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput. Methods Programs Biomed. **2019**, 179, 104992. [CrossRef] [PubMed]

[10] Jinny, V.; Priya, R.L. Prediction Model for Respiratory Diseases Using Machine Learning Algorithms. Int. J. Adv. Sci. Technol. **2020**, 29, 10083–10092.

[11] Asri, H.; Mousannif, H.; Al Moatassime, H.; Noel, T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Comput. Sci. **2016**, 83, 1064–1069. [CrossRef]

[12] Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. **2018**, 132, 1578–1585. [CrossRef]

[13] Bansal, D.; Chhikara, R.; Khanna, K.; Goopta, P. Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia Detecting Dementia. Procedia Comput. Sci. **2018**, 132, 1497–1502. [CrossRef]

[14] Rahman, A.S.; Shamrat, F.J.M.; Tasnim, Z.; Roy, J.; Hosain, S.A. A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. Int. J. Sci. Technol. Res. **2019**, 8, 419–422.

[15] Turabieh, H.; Karaa, W.B.A. Predicting the existence of COVID-19 using machine learning based on laboratory findings. In Proceedings of the 2021 International Conference of Women in Data Science at Taif University, Taif, Saudi Arabia, 30–31 March 2021.

[16] Luo, J.; Zhou, L.; Feng, Y.; Bo, L.; Guo, S. The selection of indicators from initial blood routine test results to improve the accuracy of early prediction of COVID-19 severity. PLoS ONE **2021**, 16, e0253329. [CrossRef]

[17] Rangarajan, A.; Krishnaswamy, R.; Krishnan, H. A preliminary analysis of AI based smartphone application for diagnosis of COVID-19 using chest X-ray images. Expert Syst. Appl. **2021**, 183, 1–11. [CrossRef]

[18] Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; et al. An interpretable mortality prediction model for COVID-19 patients. Nat. Mach. Intell. **2020**, 2, 283–288. [CrossRef]

[19] Khalilpourazari, S.; Doulabi, H.H. Robust modelling and prediction of the COVID-19 pandemic in Canada. Int. J. Prod. Res. **2021**, 1–17. [CrossRef]

[20] Majumder, P. Chapter 10-Daily confirmed cases and deaths prediction of novel coronavirus in Asian continent Polynomial Neural Network. In Biomedical Engineering Tools for Management for Patients with COVID-1; Academic Press: Cambridge, MA, USA, 2021; pp. 163–172.