

Article

Not peer-reviewed version

COVID-19 Classification Using Machine Learning

Ayesha Arif and [Manzoor Hussain](#)*

Posted Date: 20 November 2025

doi: 10.20944/preprints202511.1519.v1

Keywords: COVID-19; machine learning; dataset; prediction; classification



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

COVID-19 Classification Using Machine Learning

Ayesha Arif ¹ and Manzoor Hussain ^{2,*}

¹ University of Sialkot, Sialkot, Pakistan

² Indus University, Karachi, Pakistan

* Correspondence: manzoor.hussain@indus.edu.pk

Abstract

COVID-19 is an infectious disorder caused by the SARS-CoV-2 virus, first identified in December 2019. It can cause a range of symptoms such as pneumonia, fatigue, and even death. The study focuses on the COVID-19 dataset, which contains information about patients, symptoms, test results, recovery, and age details. We apply machine learning algorithms, including Random Forest, Decision Trees, KNN, and Naïve Bayes, to predict outcomes of COVID-19, such as test results, recovery time, and hospitalization probability. The dataset undergoes several processing steps, such as handling missing values, normalization, and feature selection. These algorithms are evaluated to determine accuracy, precision, recall, and the confusion matrix to identify the best model.

Keywords: COVID-19; machine learning; dataset; prediction; classification

1. Introduction

COVID-19 is a virus that was first discovered in China in 2019. It quickly spread globally and caused a serious impact on the economy, daily life, and public health. It led to numerous deaths and severe illnesses such as pneumonia, organ failure, and other complications. COVID-19 particularly affects individuals with weak immune systems, especially the elderly.

Scientists developed vaccines, which played an important role in controlling the disease [12]. The dataset provides detailed information about patients, including age, gender, and symptoms like fever, cough, and fatigue, which help in understanding patient conditions [13]. Test results indicate whether a patient was COVID-positive or negative, while hospitalization attributes show whether medical attention was required.

To understand the factors influencing COVID-19, this dataset can be used for predictive analysis. The widespread transmission of COVID-19, caused by the SARS-CoV-2 virus, led to high mortality rates and serious health challenges due to the lack of specific treatments and limited knowledge about the virus's behavior. Scientists investigated how the virus causes infection and interacts with the immune system. They discovered that SARS-CoV-2 uses specific receptors and affects children, adults, and those with organ failure differently. Research also focused on the challenges caused by irregular vaccine distribution worldwide, with regions such as America and Europe being severely affected [14].

The methodology includes steps, algorithms, and techniques for prediction, decision-making, and data analysis. It involves model selection, evaluation, and implementation. Data is preprocessed before evaluation for example, using the "Replace Missing Value" operator in RapidMiner.

RapidMiner is a powerful tool for training and deploying machine learning models. It supports popular algorithms like KNN, Decision Trees, Random Forest, and Naïve Bayes. It can handle missing values, connect to databases, and compute metrics such as accuracy, precision, recall, and F1-score for model evaluation. It also integrates data preprocessing, modeling, evaluation, and deployment in a single platform, with the ability to split data into training and testing sets [15].

2. Literature Review

This section presents a review of the relevant literature. Figure 1 illustrates the overall paper flow.

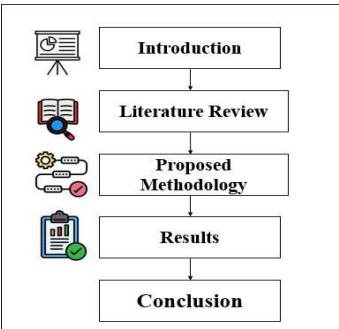


Figure 1. Paper Flow.

Velavan and Meyer (2020) [1] explain how COVID-19 rapidly spread across the world. They discuss its effects on people and how it impacts the human body. They also mention that treatments such as chloroquine, antiretroviral, and antiviral drugs work within the body. Controlling the virus has been particularly difficult in developing countries. The authors emphasize the importance of better treatments and a strong immune system [16].

In 2024, Khan et al. [2] studied how individuals with multiple sclerosis are affected by COVID-19. He explains that the virus particularly affects those with weakened immune systems. Their research highlights the need for better data sharing and the lack of information available for patients with this condition. The study shows that 5.25% of hospitalized individuals provided information about how multiple sclerosis and COVID-19 interact.

Ndwandwe and Wiysonge (2021) [3] discuss the urgent need for vaccine development to reduce the death rate from COVID-19. They analyze the benefits and limitations of various vaccines, such as viral vector and nucleic acid vaccines. In low-income developing countries, only 1% of people had access to vaccines, although over 3 billion doses were distributed worldwide [17–19]. The use of new technologies like mRNA vaccines helped combat the virus, although unequal distribution remained a major challenge [4].

Shie et al. (2020) examine how COVID-19, caused by a highly infectious virus, became a global threat with many symptoms. They point out that early in the pandemic, we lacked knowledge about how the virus spread, how it affected the body, and how to control it. This paper reviews how China attempted to manage the outbreak, as well as the development of vaccines and the immune response. The study of SARS helps explain why the virus spread so rapidly, even though its death rate was lower than some earlier outbreaks.

Suryasa et al. (2021) [5] studied the global impact of the COVID-19 pandemic. They focused on the effects of irregular vaccine distribution worldwide, noting that America and Europe were especially affected by the virus.

Daniel (2020) [6] discusses how COVID-19 transformed global education, leading to a shift toward online learning. Many schools were unprepared and lacked the necessary resources to meet students’ needs. Studies showed that educational programs, online tests, and short video lessons helped students learn effectively. The pandemic accelerated the development of digital education and demonstrated how quickly the education system can adapt.

The widespread transmission of COVID-19, caused by the SARS-CoV-2 virus, resulted in a high death rate and posed serious health risks due to the lack of specific treatments and limited understanding of how the virus functions. Scientists examined how the virus infects the body and interacts with the immune system [7]. They found that SARS-CoV-2 uses specific receptors, which may explain why the virus affects children, adults, and individuals with organ failure differently.

Singh and Singh (2020) [8] discuss how lockdowns and social distancing impacted mental health and the economy. They emphasized the lack of awareness around loneliness and anxiety. Using data from their report, they explored how the pandemic led to increased mental health issues and economic decline. The impact of the lockdowns was significant, highlighting the need for comprehensive solutions.

Sepandi et al. (2020) [9] highlight risk factors such as age, gender, and pre-existing health conditions—that influence COVID-19 mortality. They argue that these factors need to be studied together to guide effective treatment strategies. Their analysis found that heart and respiratory diseases, older age, and being male increased the risk of death. The study emphasizes the importance of targeted treatment for high-risk groups.

Nalbandian et al. (2021) [10] investigated the long-term effects of COVID-19, including fatigue, breathing difficulties, and mental illness. They also highlighted the lack of comprehensive information on how COVID-19 affects different organs. Their study found that symptoms vary depending on age and pre-existing health conditions.

3. Proposed Methodology

The proposed methodology provides a framework designed to solve a specific problem using machine learning techniques[20–23]. It outlines the steps, algorithms, and processes used for prediction, decision-making, and data analysis. This methodology includes:

- i. Model selection
- ii. Evaluation
- iii. Implementation

Before evaluation, the dataset is carefully prepared. We use the “Replace Missing Value” operator in RapidMiner, a powerful platform for training and deploying machine learning models. RapidMiner supports widely used algorithms. Table 1 shows the featured attributes of the dataset. Figure 2 shows the methodology.

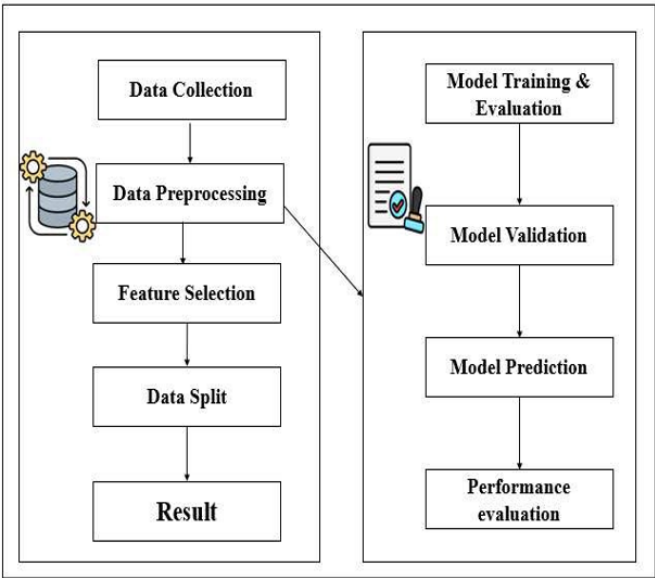


Figure 2. Framework of our Proposed Work.

Table 1. Featured Attributes.

Attribute	Description
Patient ID	Unique identifier for each patient
Gender	Male or Female

Age	Patient’s age
Symptoms	Fever, cough, fatigue
Hospitalized	Whether the patient was hospitalized
Test Results	COVID-19 test result (Positive/Negative)
Recovery Time	Time taken to recover (in days or weeks)

4. Results

This section presents the results of the proposed methodology, including evaluation using a confusion matrix and comparison of algorithm accuracy. The data was collected from Kaggle for COVID-19 prediction and processed using RapidMiner. Figure 3 shows the final confusion matrix.

Table 1. Confusion Matrix.

	True Female	True Male	Class Precision
Pred. Male	831	783	51.49%
Pred. Female	693	693	50.00%
Class recall	54.53%	46.95%	

The confusion matrix helps us evaluate the performance of classification models based on predicted and actual gender classes. Precision and recall are calculated per class. From Table 3, it can be observed that Naïve Bayes achieved the highest accuracy among the applied classifiers. Figure 3 shows the overview of the rapid miner. Figure 4 shows the data split flow, illustrating the separation of training and testing datasets. Table 4 shows comparison with benchmark.

Table 3. Accuracy Comparison of Algorithms.

Algorithm	Accuracy (%)
KNN	50.57%
Decision Trees	50.00%
Random Forest	49.00%
Naïve Bayes	50.80%

Table 4. Comparison with Other Studies.

Author	Year	Dataset	Classifier	Accuracy (%)
N. David Yanez	2020	COVID-19	RM	87.4%
Mojtaba Sepandi	2020	COVID-19	REF	66%
Hamza Khan et al.	2020–2024	COVID-19	VSM	76%
Y. Shi et al.	2020	COVID-19	REF	56%
J. Singh	2020	COVID-19	VSM	66%

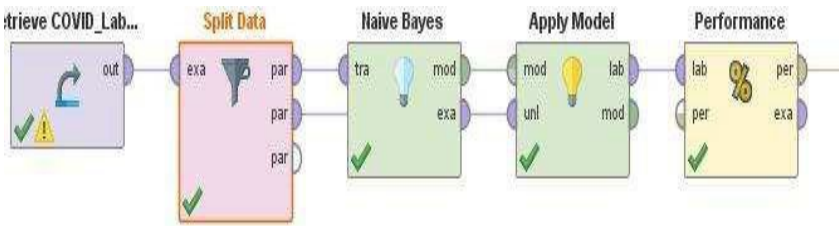


Figure 3. Overview of the Model in Rapid Miner.

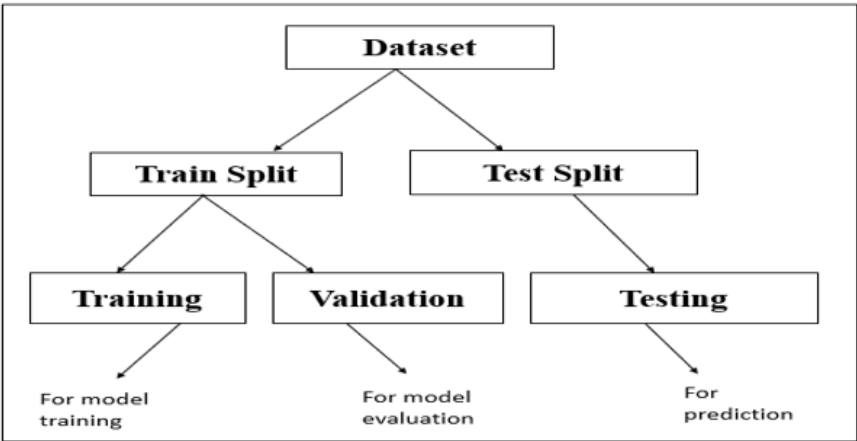


Figure 4. Data Split Flow.

5. Conclusions

This paper discusses the significance of COVID-19, which was a major global issue that quickly spread across the world. It caused severe damage, including deaths, economic losses, and a negative impact on education. We applied different algorithms such as KNN, Decision Trees, Random Forest, and Naïve Bayes to achieve better accuracy. This study helps in understanding the factors that affect COVID-19, and by identifying these factors, we can work to reduce the death rate and economic loss.

References

1. T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical Medicine & International Health*, vol. 25, no. 3, pp. 278–280, 2020. doi: 10.1111/tmi.13383
2. D. Ndwandwe and C. S. Wiysonge, "COVID-19 vaccines," *Current Opinion in Immunology*, vol. 71, pp. 111–116, 2021. doi: 10.1016/j.coi.2021.07.003
3. Y. Shi et al., "An overview of COVID-19," *Journal of Zhejiang University Science B*, vol. 21, no. 5, pp. 343–360, 2020. doi: 10.1631/jzus.B2000083
4. I. W. Suryasa, M. Rodríguez-Gámez, and T. Koldoris, "The COVID-19 Pandemic," *International Journal of Health Sciences (Qassim)*, vol. 5, no. 2, pp. VI–IX, 2021. doi: 10.53730/IJHS.V5N2.2937
5. S. J. Daniel, "Education and the COVID-19 pandemic," *Prospects*, vol. 49, no. 1–2, pp. 91–96, 2020. doi: 10.1007/s11125-020-09464-3
6. K. Yuki, M. Fujiogi, and S. Koutsogiannaki, "COVID-19 pathophysiology: A review," *Clinical Immunology*, vol. 215, 2020. doi: 10.1016/j.clim.2020.108427
7. J. Singh and J. Singh, "COVID Impact on Society," *Electronic Research Journal of Social Sciences and Humanities*, vol. 2, no. 1, pp. 168–172, 2020.
8. M. Sepandi, M. Taghdir, Y. Alimohamadi, S. Afrashteh, and H. Hosamirudsari, "Factors associated with mortality in COVID-19 patients: A systematic review and meta-analysis," *Iranian Journal of Public Health*, vol. 49, no. 7, pp. 1211–1221, 2020. doi: 10.18502/ijph.v49i7.3574
9. A. Nalbandian et al., "Post-acute COVID-19 syndrome," *Nature Medicine*, vol. 27, no. 4, pp. 601–615, 2021. doi: 10.1038/s41591-021-01283-z
10. N. D. Yanez, N. S. Weiss, J. A. Romand, and M. M. Treggiari, "COVID-19 mortality risk for older men and women," *BMC Public Health*, vol. 20, no. 1, pp. 1–7, 2020. doi: 10.1186/s12889-020-09826-8
11. Almulhim, M., Islam, N., & Zaman, N. (2019). A lightweight and secure authentication scheme for IoT based e-health applications. *International Journal of Computer Science and Network Security*, 19(1), 107-120.
12. Zaman, N., Low, T. J., & Alghamdi, T. (2014, February). Energy efficient routing protocol for wireless sensor network. In *16th international conference on advanced communication technology* (pp. 808-814). IEEE.

13. Azeem, M., Ullah, A., Ashraf, H., Jhanjhi, N. Z., Humayun, M., Aljahdali, S., & Tabbakh, T. A. (2021). Fog-oriented secure and lightweight data aggregation in iomt. *IEEE Access*, 9, 111072-111082.
14. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. *Electronics*, 11(13), 2071.
15. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. *CMC-Computers Materials & Continua*, 70(1), 601-618.
16. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In *Employing Recent Technologies for Improved Digital Governance* (pp. 151-163). IGI Global Scientific Publishing.
17. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. *Sensors*, 23(11), 5055.
18. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications* (pp. 49-64). IGI Global.
19. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. *Electronics*, 11(15), 2324.
20. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In *Cyber Security Applications for Industry 4.0* (pp. 97-109). Chapman and Hall/CRC.
21. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. *Electronics*, 11(4), 510.
22. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In *2020 International Conference on Computational Intelligence (ICCI)* (pp. 305-310). IEEE.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.