JOURNAL OF MACHINE LEARNING AND APPLICATIONS
Jan-Jun 2025, VOL. 1, ISSUE 1, pp. 12-14
https://doi.org/10.61577/jmla.2025.100003

RESEAPRO JOURNALS

COMMENTARY ARTICLE

OPEN ACCESS

Analyzing and Categorizing COVID-19 Symptom Severity by Integrating Machine Learning and Statistical Techniques

Monalisha Biswal

Department of Humanities and Basic Sciences, G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India

ABSTRACT

The COVID-19 pandemic, originating in late 2019, has significantly affected worldwide health systems, economies, and everyday life. One of the major challenges faced by healthcare providers has been the varying severity of symptoms among individuals infected with the virus. While some patients experience mild or even asymptomatic cases, others suffer from severe symptoms that require intensive medical intervention. Early detection and accurate classification of symptom severity are crucial for effective treatment, resource allocation, and monitoring disease progression. In recent years, the integration of statistical methods and machine learning (ML) techniques has shown great promise in analyzing and classifying the severity of COVID-19 symptoms. These advanced approaches help health professionals identify high-risk patients, predict outcomes, and tailor treatment plans more effectively. In this article, we will explore how statistical techniques and machine learning models can be used to analyze COVID-19 symptom severity and discuss the benefits, challenges, and future directions of these approaches.

KEYWORDS

Artificial neural networks; Big data; Clinical decision making; Decision support; Machine learning

ARTICLE HISTORY

Received 15 January 2025; Revised 31 January 2025; Accepted 07 February 2025

Introduction

Symptoms of COVID-19 can vary from mild, like fever, cough, and fatigue, to severe, such as difficulty breathing, chest pain, and organ failure. The intensity of symptoms may differ due to various factors, such as the patient's age, existing health issues (e.g., diabetes, hypertension), and genetic influences. In some cases, patients with seemingly mild symptoms may suddenly deteriorate, while others may remain stable despite having high viral loads [1].

Given this variability, healthcare systems need reliable methods to assess symptom severity early in the infection to optimize treatment plans, allocate medical resources, and improve patient outcomes. Accurate classification of symptom severity involves analyzing a wide array of data, including clinical symptoms, demographic information, medical history, and test results [2].

Statistical Methods for COVID-19 Symptom Analysis

Statistical methods have long been used in epidemiological studies to analyze disease patterns, identify risk factors, and predict outcomes. In the context of COVID-19 symptom severity, statistical techniques can help identify the most significant factors associated with severe outcomes, as well as model the distribution of symptoms across different populations [3].

Descriptive statistics

Descriptive statistics provide a basic understanding of the distribution and central tendency of symptoms in a population. Measures such as the mean, median, standard deviation, and percentiles can be used to summarize key aspects of the dataset, including age, gender, comorbidities, and symptom severity [4]. These statistics provide a general overview of trends and allow healthcare providers to better understand the general characteristics of COVID-19 patients.

Inferential statistics

Inferential statistics, such as hypothesis testing and regression analysis, are essential for identifying factors that influence the severity of COVID-19 symptoms. For example, logistic regression can be used to assess the relationship between demographic and clinical variables (e.g., age, comorbidities) and the likelihood of severe outcome [5]. Chi-square tests can help identify associations between categorical variables, such as the presence or absence of certain symptoms, and the severity of the disease [6].

Survival analysis

Survival analysis methods, such as the Cox proportional hazards model, can be applied to model the time to a particular event (e.g., hospitalization or death) based on various covariates. This is particularly useful for understanding the progression of COVID-19 and predicting outcomes for individual patients based on their symptom severity and other risk factors.

Machine Learning Techniques for Classifying COVID-19 Symptom Severity

While statistical methods are valuable for understanding relationships between variables and predicting outcomes, machine learning (ML) techniques offer more powerful tools for building predictive models that can automatically learn from data and classify the severity of COVID-19 symptoms. ML algorithms can handle large, complex datasets and

*Correspondence: Ms. Monalisha Biswal, Department of Humanities and Basic Sciences, G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India, 518007. e-mail: monalisha.b123@gmail.com

identify intricate patterns that may not be immediately apparent through traditional statistical methods [7].

Supervised learning

Supervised learning requires training a model with a labeled dataset where the result (i.e., symptom severity) is known. The model learns to predict severity based on input features such as age, comorbidities, and clinical symptoms [8]. Common supervised learning techniques used in COVID-19 severity classification include:

Logistic regression

This is a simple and interpretable method for binary classification tasks (e.g., severe vs. mild cases). It can be used to predict the likelihood of a patient experiencing severe symptoms based on various input features [9].

Random forest

A random forest classifier, an ensemble technique, can identify intricate relationships among features by constructing numerous decision trees and combining their outputs. Random forests are especially effective for handling high-dimensional data [9].

Support vector machines (SVM)

SVMs are applicable for both binary and multi-class classification tasks. By discovering an ideal hyperplane that distinguishes various classes (e.g., mild, moderate, and severe), SVMs can provide accurate classification even in complex datasets [9].

Neural networks

Deep learning techniques, such as artificial neural networks (ANNs), are well-suited for handling large datasets with multiple variables. ANNs are capable of learning non-linear relationships and are often used when there is a large amount of data and a need for high predictive accuracy [10].

Unsupervised learning

Unsupervised learning methods can be applied in situations where labeled data is absent, or when the aim is to uncover hidden patterns in the data without any predetermined labels. Clustering methods like K-means or hierarchical clustering can categorize patients with comparable symptom profiles, potentially aiding in recognizing new severity categories or uncovering previously overlooked patterns in patient data [11].

Natural language processing (NLP)

In addition to structured data such as vital signs and laboratory test results, patient reports, and clinical notes often contain valuable unstructured data. Natural language processing (NLP) techniques can be used to extract relevant information from textual data, such as symptom descriptions in electronic health records or patient surveys. By applying text mining and sentiment analysis, NLP tools can extract critical symptom severity information and enhance the prediction models [12].

Integrating statistical methods and machine learning

The integration of statistical methods and machine learning techniques provides a comprehensive approach to analysing and classifying COVID-19 symptom severity [13]. Here's how these two methodologies can complement each other:

Feature selection

Statistical techniques like correlation analysis and principal component analysis (PCA) can help identify the most important features (e.g., age, underlying conditions) that contribute to symptom severity. These features can then be used to train machine learning models more efficiently.

Model interpretation

Statistical methods can help interpret the results of machine learning models. For example, logistic regression coefficients or decision tree rules can be analysed to understand how various factors influence symptom severity.

Model evaluation

While machine learning models can offer high accuracy, statistical methods can provide insights into their robustness. Cross-validation and confidence intervals from statistical methods can be used to assess the uncertainty of the model's predictions and ensure they are reliable.

Risk stratification

By combining the power of machine learning with statistical analysis, healthcare providers can create more accurate risk stratification systems. These systems can identify patients who are at high risk of severe outcomes and prioritize them for early intervention [14].

Challenges and future directions

While integrating statistical methods and machine learning for COVID-19 symptom severity analysis holds great promise, there are several challenges:

Quality and accessibility of data

Quality labeled datasets are crucial for developing precise machine learning models. Data that is incomplete or biased may result in incorrect predictions.

Model interpretability

Machine learning models, particularly deep learning methods, can occasionally function as "black boxes," complicating the interpretation of their decision-making processes. Combining these models with statistical methods can improve interpretability.

Generalization

Models trained on data from one region or population may not generalize well to others. Ensuring that models are adaptable to diverse populations is crucial for their widespread implementation.

Real-time applications

To make an immediate impact in healthcare settings, these models must be able to provide real-time predictions with minimal computational overhead. In the future, integrating more advanced techniques such as reinforcement learning and transfer learning could further improve predictive accuracy. Additionally, real-time data collection through wearable devices and mobile applications may allow for continuous symptom monitoring, enabling even more accurate and dynamic risk assessments [15].

Conclusions

The integration of statistical methods and machine learning techniques offers significant potential for analyzing and

https://doi.org/10.61577/jmla.2025.100003

classifying COVID-19 symptom severity. By combining the interpretability of traditional statistics with the predictive power of machine learning, healthcare providers can develop more accurate and efficient tools for managing the COVID-19 pandemic. These approaches not only assist in early detection and risk stratification but also help optimize treatment plans, allocate resources, and ultimately improve patient outcomes. As data availability and model sophistication continue to improve, the role of these integrated techniques in managing COVID-19 and other future pandemics will only grow.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

1. Liu W, Liu D, Yan J, Liu X, Wang Q. Research progress on genetic susceptibility of patients with severe novel coronavirus pneumonia. Heredity. 2022;44(8):672-681. https://doi.org/10.16288/j.yczz.22-058
2. Kalapuge V, Kasthurirathna D. Predictive Modeling for Early Identification of Disease Severity in Acute Respiratory Infections: A Case Study with COVID-19. In2023 5th International Conference on Advancements in Computing. 2023. 203-208p. https://doi.org/10.1109/ICAC60630.2023.10417268
3. Cugnata F, Scarale MG, De Lorenzo R, Simonini M, Citterio L, Querini PR, et al. Profiling Covid-19 patients with respect to level of severity: an integrated statistical approach. Sci Rep. 2023;13(1): 5498. https://doi.org/10.1038/s41598-023-32089-3
4. Bulanov NM, Suvorov AY, Blyuss OB, Munblit DB, Butnaru DV, Nadinskaia MY, et al. Basic principles of descriptive statistics in medical research. Сеченовский вестник. 2021;12(3):4-16. https://doi.org/10.47093/2218-7332.2021.12.3.4-16
5. Nachev A. Analysis of Factors Influencing the Severity of Coronavirus Symptoms Using Predictive Modeling. In2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE) 2023. 157-162p. https://doi.org/10.1109/CSCE60160.2023.00030
6. Hazra A, Gogtay N. Biostatistics series module 4: comparing groups–categorical variables. Indian J Dermatol. 2016;61(4):385-392. https://doi.org/10.4103/0019-5154.185700
7. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad AK, Uddin S, et al. Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. JMIR Med Inform. 2021;9(4):25884. https://doi.org/10.2196/25884
8. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, Uddin S, Liò P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl. 2020;160:113-661. https://doi.org/10.1016/j.eswa.2020.113661
9. Xiong Y, Ma Y, Ruan L, Li D, Lu C, Huang L. National Traditional Chinese Medicine Medical Team. Comparing different machine learning techniques for predicting COVID-19 severity. Infect Dis Poverty. 2022;11(1):19. https://doi.org/10.1186/s40249-022-00946-4
10. Sanchez T, Cury J, Charpiat G, Jay F. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Mol Ecol Resour. 2021;21(8): 2645-2660. https://doi.org/10.1111/1755-0998.13224
11. Alsayat A, El-Sayed H. Efficient genetic K-means clustering for health care knowledge discovery. In2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications. 2016. 45-52p. https://doi.org/10.1109/SERA.2016.7516127
12. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform. 2019;125:37-46. https://doi.org/10.1016/j.ijmedinf.2019.02.008
13. Raddad Y, Hasasneh A, Abdallah O, Rishmawi C, Qutob N. Integrating Statistical Methods and Machine Learning Techniques to Analyze and Classify COVID-19 Symptom Severity. Big Data Cogn Comput. 2024;8(12):192. https://doi.org/10.3390/bdcc8120192
14. Wiens J, Guttag J, Horvitz E. Patient risk stratification with time-varying parameters: a multitask learning approach. J Mach Learn Res. 2016;17(79):1-23. http://doi.org/10.1109/5.771073
15. Adeniran IA, Efunniyi CP, Osundare OS, Abhulimen AO. Data-driven decision-making in healthcare: improving patient outcomes through predictive modeling. Eng Sci Technol Int J. 2024;5(8). https://doi.org/10.56781/ijsrms.2024.5.1.0040

https://doi.org/10.61577/jmla.2025.100003