

Application of Machine Learning on Predicting the Risk of Death of Respiratory Infectious Diseases Patients — Using COVID-19 as Example

Wenhao Li

*Faculty of Science, McMaster University, Hamilton, Canada
li2306@mcmaster.ca*

Abstract: In modern society, with the accelerating pace of urbanization and the increasing frequency of global travel, densely populated areas have become hotspots for the outbreak of infectious diseases. These areas, such as large-scale urban centers and crowded public transportation hubs, facilitate the rapid spread of pathogens. The patients affected by infectious diseases span a wide spectrum of different groups, including the elderly with weakened immune systems, children whose immune systems are still developing, and individuals with pre-existing chronic health conditions. The impact of infectious diseases on patients varies significantly. Depending on a patient's underlying health condition, the course of the disease can range from a mild, self-limiting illness to a severe, life-threatening condition. For example, in the case of influenza, some healthy individuals may experience only mild symptoms like a runny nose and mild fever, while the elderly or those with respiratory diseases may develop severe pneumonia, which can be fatal. In the context of the ongoing global health challenges, such as the COVID - 19 pandemic, there is an urgent need to explore innovative methods for predicting the risk of patients infected by infectious diseases. To demonstrate whether machine learning techniques can be effectively used for this purpose, this study employs neural networks to develop a model specifically designed to predict the risk of death among COVID-19 patients. Through rigorous testing and validation, the model has been proven to have a high degree of accuracy in predicting these risks. This indicates that the method proposed in this study can be applied to identify which patients have a more urgent need for medical resources, thus providing valuable reference for healthcare providers to optimize the allocation of healthcare resources.

Keywords: Machine Learning, COVID-19, Python, data science, neural networks

1. Introduction

Infectious diseases affect patients with diverse health histories, including pre-existing chronic conditions like diabetes, heart disease, or autoimmune disorders. These conditions can complicate the course of the disease, making them more susceptible to infections. Lifestyle factors like smoking, alcohol consumption, and lack of exercise also influence a patient's susceptibility and symptoms. The healthcare needs of these patients are diverse, with mild infections requiring basic care and severe infections requiring intensive care. Efficient allocation of healthcare resources is crucial for cost containment and optimal use of limited resources. Prioritizing urgent needs can enhance treatment

effectiveness. To predict patients at higher risk, past patient data can be analyzed. This data can identify patients more likely to develop severe infections or require intensive medical care. This predictive ability can guide healthcare providers in resource allocation decisions. For example, if elderly patients with multiple comorbidities are at a higher risk of developing severe symptoms during an infectious disease outbreak, healthcare providers can allocate more resources, such as antiviral medications and specialized nursing care, to this high-risk group. Infectious diseases spread by direct or indirect contact between infected people and healthy people. During an infectious disease outbreak, the number of infected people is typically large, and these individuals possess diverse characteristics and medical histories. Analyzing the data of such a large and heterogeneous group of infected individuals is an arduous task for the human brain. Machine learning is a technique that leverages given data to construct computational models, aiming to assist in prediction and problem - solving [1]. This technique performed well in dealing with massive data, its advantage in this field is suitable for this kind of problems. In particular, during the COVID-19 pandemic, Machine Learning was widely applied to predict disease procession of patients [2], proving its application value in the field of infectious disease research.

The efficient allocation of medical resources is of paramount importance in the fight against infectious diseases. Scarce resources such as intensive care unit (ICU) beds, ventilators, and antiviral medications need to be distributed in a way that maximizes the survival rate of patients. By ensuring that these resources are directed towards patients who are most in need, the risk of patients succumbing to infectious diseases can be minimized. In recent years, the development of data-driven models has emerged as a promising approach to address this challenge. The creation of models that can translate implicit risks, such as a patient's underlying health status and the virulence of the infecting pathogen, into intuitive and actionable results is crucial. These models enable healthcare providers to more accurately evaluate which patients are more likely to require substantial medical resources, thereby promoting the effective allocation of resources based on evidence-based predictions. Since the COVID-19 pandemic is the closest case of a large-scale infectious disease outbreak and the relevant data are the most recent and detailed, this paper uses the health data of patients during the COVID-19 pandemic to train a neural network model, thus demonstrating the suitability of the trained model for predicting the risk of death of patients with a specific infectious disease [3]. Confirmation of the model will help healthcare providers to improve the realization of the allocation provided possible.

2. Methods

In this study Python was used to process the data and Recurrent Neural Networks (RNN) were used to train the model, as the problem of interest falls under supervised learning, which uses labeled data to train the computational model [4].

Since the focus of the study is on the risk of patient death in infectious diseases using a binary outcome label. This is chosen to provide a clear measure for evaluating the impact of infectious disease factors on patient survival. The binary label distinguishes between two distinct states of patient outcome, allowing for accurate analysis of complex relationships between variables like disease type, pre-existing health conditions, medical intervention timeliness, and death risk. This label serves as the foundation for building a predictive model, identifying key determinants contributing to a patient's death. Researchers use this label to analyze vast data on patients with infectious diseases, separating those who survived from those who did not. This helps healthcare providers make informed decisions about patient treatment and resource allocation, ultimately reducing the mortality rate associated with infectious diseases.

2.1. Dataset and preprocessing

The dataset used in this study is COVID-19 patient data provided by Mexican government, which contains data of 1,048,576 patients with 21 recorded features [5]. The features included: sex, age, classification, patient type, pneumonia, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, cardiovascular, renal chronic, other disease, obesity, tobacco, usmr, medical unit, intubed, icu and date died.

In order to improve computing speed, we randomly select the data of 1,000 patients from the original dataset. Date of death was manipulated and reshaped to 1 if the patient died and 0 if they did not die [6].

The dataset have been split into training set and validation set. 75% of the dataset forms the training set. Data in the training set will be used to train a computational model, and others will be used to verify the accuracy [7].

2.2. Neural networks

Neural network is a method to find the relationship between a dependent variable and independent variables [8]. This method assumes the variables have a functional relationship, although this relationship would not be easy to describe. In this study, the date of death was selected as a label from the 21 features of the dataset as the independent variable. Other features are viewed as the independent variable [9].

3. Result

This study embarked on a meticulous data-handling process. Initially, the collected patient data underwent comprehensive pre-processing follow Table 1. This involved tasks such as data cleaning to remove outliers, imputing missing values, and standardizing numerical features. After pre - processing, a strategic decision was made to utilize 75% of the data to form a training set. This proportion was carefully determined through a series of pilot analyses and industry - standard best practices, as it provided a sufficient amount of data for the model to learn the underlying patterns while also reserving an adequate portion for validation.

The training set was then employed to develop a sophisticated computational model follow Table 2, specifically designed to predict whether a patient would succumb to a COVID - 19 infection. In this binary classification task, for the sake of computational simplicity and clarity, patients were labeled “1” if they unfortunately died due to COVID - 19 and “0” if they survived. This labeling system allowed the model to distinguish between the two critical outcomes and learn the relationships between various input variables, such as the patient's age, pre - existing comorbidities like diabetes or heart disease, the severity of initial symptoms, and the time elapsed from the onset of symptoms to hospitalization.

The computational model, which was based on advanced machine - learning algorithms such as neural networks or decision trees, underwent an iterative training process. During each iteration, the model adjusted its internal parameters to minimize the difference between its predicted outcomes and the actual labels in the training set. After the model was trained, a crucial step was to evaluate its performance. One way to do this was to record a list of error values after each training step. These error values, which could be calculated using metrics like mean squared error or cross - entropy loss, provided valuable insights into how well the model was converging. The error values were then plotted as shown in Figure 1 [10]. This graphical representation enabled researchers to visually assess the model's training progress. A steadily decreasing error curve over time indicated that the model was learning effectively, while erratic or increasing error values might suggest issues such as overfitting or inappropriate model selection. By analyzing these error plots, the research team could

fine - tune the model, for example, by adjusting hyperparameters like the learning rate or the number of hidden layers in a neural network, to improve its predictive accuracy and generalization ability, ultimately enhancing its utility in predicting the risk of death among COVID-19 patients.

Table 1: Training set accuracy

	Precision	Recall	F1-score	Support
0	0.98	0.99	0.99	701
1	0.83	0.71	0.77	49
Accuracy			0.97	750
Macro Avg	0.91	0.85	0.88	750
Weighted Avg	0.97	0.97	0.97	750

Table 2: Validation set accuracy

	Precision	Recall	F1-score	Support
0	0.96	0.96	0.96	239
1	0.23	0.25	0.24	12
Accuracy			0.92	251
Macro Avg	0.60	0.60	0.60	251
Weighted Avg	0.93	0.92	0.93	251

Training set and validation set are used to show the accuracy of the model as well:

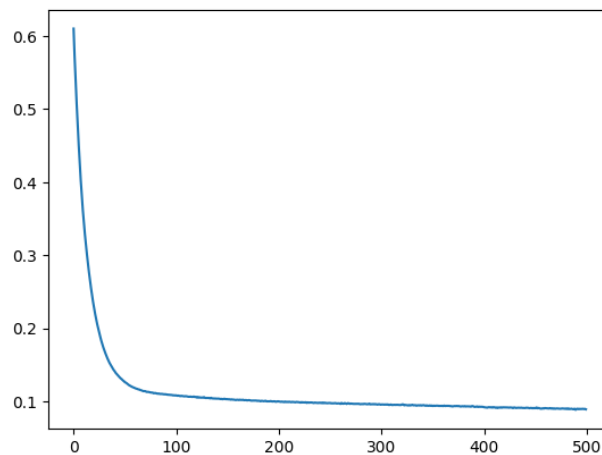


Figure 1: Plot of the errors

4. Conclusion and discussion

The model trained using the COVID-19 patient dataset has shown high accuracy in predicting the risk of death for COVID-19 patients, suggesting the potential for other infectious diseases using similar methods. However, the study has several limitations, including a potential lack of data representativeness, which may not fully capture the wide-ranging characteristics of other infectious diseases, such as symptoms, incubation periods, and treatment responses. The data collection methods and variables considered in this study might not be sufficient or appropriate for other diseases, leading to inaccurate predictions.

The model's complexity and flexibility may be insufficient, as it was optimized for COVID-19 and may struggle to adapt to the unique patterns and factors associated with different infectious agents. It

may also struggle to handle complex interactions between multiple co-infections effectively, such as when a patient is infected with both influenza and pneumonia-causing bacteria simultaneously.

The study did not account for potential changes in the healthcare environment and patient demographics across different regions and over time, such as healthcare infrastructure, access to medical resources, and the genetic makeup of patient populations. In resource-poor regions, the availability of diagnostic tools and treatment options may be limited, impacting disease progression and mortality. Moreover, the method may not work well in researching infectious pathogens due to their various characteristics, and some patients may be infected by multiple diseases during an outbreak, affecting the predicting accuracy of the model.

References

- [1] Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.
- [2] Syeda, H. B., Syed, M., Sexton, K. W., Syed, S., Begum, S., Syed, F., ... & Yu Jr, F. (2021). Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. *JMIR medical informatics*, 9(1), e23811.
- [3] Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 11(4), 193.
- [4] Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.
- [5] Meir Nizri (2022). COVID-19 Dataset. Kaggle [<https://www.kaggle.com/datasets/meirnazri/covid19-dataset>].
- [6] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, 139, 110059.
- [7] Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics*, 8(6), 890.
- [8] Sujath, R. A. A., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, 34, 959-972.
- [9] Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd international conference on machine learning and soft computing* (pp. 5-9).
- [10] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj digital medicine*, 4(1), 3.