Machine Learning–Based Approaches for COVID-19 Transmission Modeling and Prediction: A Comprehensive Investigation

Yitong Zheng *

Ulink College of Shanghai, Shanghai, China

* Corresponding Author Email: yitong.zheng@ulink.cn

Abstract. Due to several outbreaks of large infectious diseases, it is crucial to make accurate predictions of dynamics of the pandemics. Among all the tools, machine learning proves to be an efficient way for predicting infectious diseases, such as Covid-19. This paper reviews the use of machine learning algorithms to forecast the transmission dynamics of COVID-19. It outlines the fundamental workflow of both traditional methods (Random Forest, Decision Tree and Linear Regression) and deep learning models (Bi-Gru, RNN, DNN). While these data-driven approaches demonstrate significant capability in capturing complex patterns and making predictions, there are still reliability issues of these methods in real-world health decision-making, such as interpretability, applicability and error accumulation. To address the limitations, the paper proposes three methods, including expert system, transfer learning and detailed data cleansing and collection process. The conclusion advocates for hybrid frameworks that merge professional epidemiological knowledge with machine learning to build more robust and reliable tools for public health decision-making.

Keywords: Machine-learning, Covid-19, deep learning models.

1. Introduction

The emergence and wide spread of respiratory pathogens, most recently exemplified by the Covid-19 pandemic, poses a long-lasting and profound threat to global health. The overloaded medical system along with economic disruptions caused by the epidemic have emphasized the critical need for robust statistical tools that can not only predict the trajectory of an outbreak but also set up a model for controlling the widespread of novel respiratory diseases in the future. The early development of an epidemic often preserves randomness and uncertainty which are hard to capture by simply using deterministic models. While traditional models such as the classic Susceptible-Infectious-Recovered (SIR) provides valuable insights in predicting the overall trend of the pandemics of large scales (Weiss, 2013), these deterministic models have limitations in dealing with random fluctuations and superspreading events during the very early stages of an epidemic (Lloyd-Smith et al., 2005).

Relevant researches have been done to evaluate the early pattern of international transmission of the epidemic (Riou and Althaus, 2020). However, traditional calculation of parameters often faces computational challenges. This problem can be solved by computer simulation and machine learning tools. Previous studies have shown traditional machine learning methods, such as random forests (Zhou, 2024), decision tree classification algorithms (Mohammadi-Pirouz et al., 2024), and deep learning models are efficient tools for forecasting infectious disease dynamics, such as Covid-19 (Qian, 2025).

Section 2 outlines traditional learning methods and deep learning methods. Section 3 outlines challenges of deep learning and future prospects of the application of machine learning in predicting COVID-19. Section 4 summarizes the previous review and concludes.

2. Machine learning-based modelling for spreading of covid 19 pandemic

2.1. Preliminaries of Machine Learning

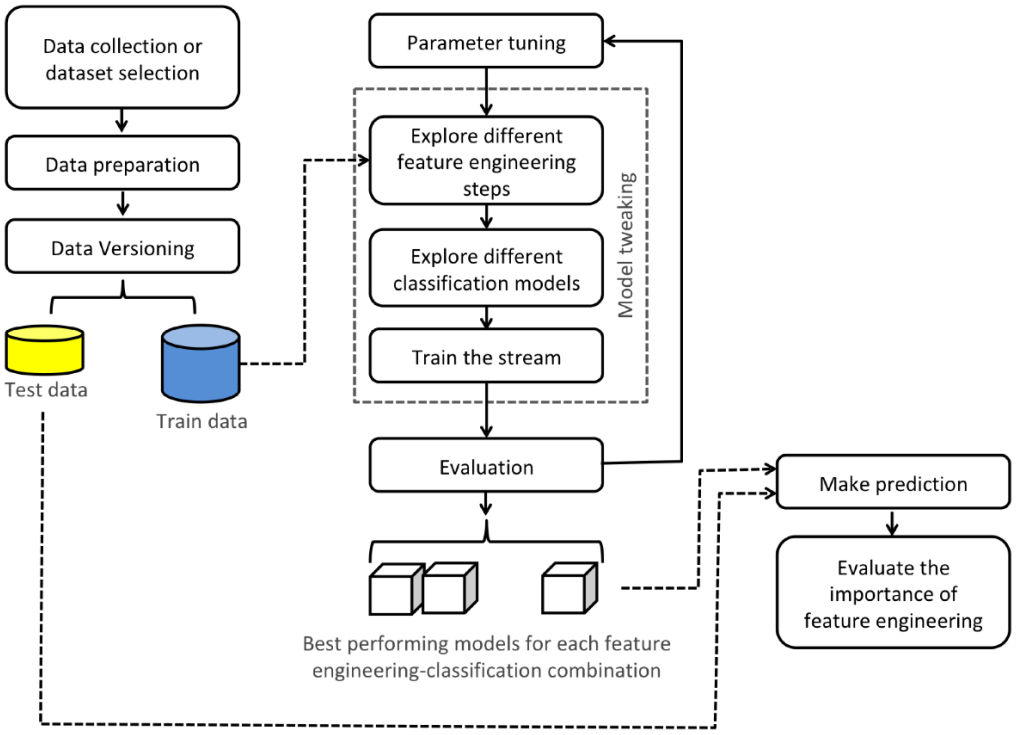


Figure 1. The schematic of machine learning framework (Vali et al., 2024)

One area of artificial intelligence is machine learning shown in Figure 1. In order to make predictions or choices, it seeks to create algorithms that can automatically identify patterns and principles in data. Supervised, unsupervised, semi-supervised, and reinforcement learning are the three primary areas of machine learning. A methodical workflow is used in most machine learning initiatives. In supervised machine learning, for instance, pertinent data is labeled and cleaned before being fed into the algorithms. According to the goal of the tasks, appropriate models are chosen and trained to optimize the model parameters on the training set. After the model can make more accurate predictions on the training set, an independent test set is used to evaluate the performance and ensure its reliability before finally deploying the application. Common assessment indicators include accuracy and F1 score, which are used in dealing with classification tasks. Indicators, such as Mean Absolute Error (MAE) and R2can be applied to regression tasks.

2.2. Traditional Machine Learning Methods

2.2.1. Linear Regression

As a fundamental learning process in supervised learning, linear regression builds a model by minimizing the prediction error and assuming a linear relationship between data and continuous target values.

The trend of SARS-CoV-2 is predicted in this study using a variety of machine learning approaches, including Logistic Regression, SVM, Random Forest, and Linear Regression. With the lowest MSE and the highest accuracy, Linear Regression outperforms the other approaches in terms of prediction (Kaur et al., 2023).

2.2.2. Decision tree

Zahra suggested using decision tree classification methods to forecast Covid-19 patient mortality. Decision tree is a method used in classification which sorts a population into branch-like segments, including internal nodes, leaf nodes and a root node. The algorithm is efficient in dealing with large, complicated datasets without using a complicated parametric structure (Song and Lu, 2015). The

main idea of the research is to develop a clinical risk assessment model based on the decision tree algorithm to identify high-risk Covid-19 patients at an early stage. A variety of decision trees such as CART have been trained with the data and compared to find the optimal tree structure. After evaluation, the findings show that the model managed to identify specific laboratory indicators, such as age, admission to ICU, which are the crucial factors for predicting mortality (Mohammadi-Pirouz et al., 2024).

2.2.3. Random forest

Random Forest aims to improve the prediction performance by building decision trees and conducting group decision-making. It introduces the Bootstrap sampling, making it more effective to handle complex nonlinear relationships between high-dimensional features (Zhou, 2024).

Sihan Zhou proposed the Random Forest model to predict the number of Covid-19 cases in four European nations in April and May of 2022: Germany, Italy, Malta, and Sweden. First, the main data are collected were processed by using Interquartile Range (IQR) method and logarithmic transformation. In the training of Random Forests, lagged infection case values are used as inputs for the model and the hyperparameters are optimized using cross-validation. The final forecast is created by averaging the forecasts from several decision trees.

The results show that the Seasonal Autoregressive Integrated Moving Average (SARIMA) has a significantly higher prediction than the random forecast model by using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and other indicators to evaluate and compare the performance of Random Forest and SARIMA. This suggests that Random Forests can more accurately capture complex nonlinear patterns and interactions in the data, surpassing the traditional time series method.

2.3. Deep Learning Methods

2.3.1. DNN and RNN

Models that capture the dynamics of sequences across the loop in the network of nodes are called recurrent neural networks, or RNNs. Any lengthy context window can be represented by the state it maintains (Lipton, 2015).

Deep Neural Network (DNN) are models which contain multiple hidden layers. Its core is a multi-layer nonlinear transformation. These networks can provide more abstract data representation by adding number of layers.

In Jung’ s study, the main goal is to try an exploratory data analysis to determine how lifestyle data affects the number of confirmed cases of COVID-19. By incorporating lifestyle information into previously verified cases, a "lifestyle addition" set was created. The analysis makes use of both recurrent neural networks (RNNs) and deep neural networks (DNNs). This paper examined and contrasted how different data sets and deep learning models performed. Results showed that the forecast curve with lifestyle data in Busan almost identical with the predicted graph (Jung, 2025).

2.3.2. Bi-Gru

Bi-GRU consists of. a two-way gate-controlled loop unit (GRU) networks, which can effectively capture the timing relationship in the time series data.

In this study, the main idea is to improve long-term prediction by integrating the features of traditional epidemiological model (SEIR), government prevention and etc. and analyzing and weighing each stage of transmission with self-attention mechanism and breakpoint detection. The model is based on the bi-directional gated recurrent unit network (Bi-GRU). The results indicate the model has robust performance with average absolute error of the control model during the 210-day long-term forecast (Chen et al., 2024).

3. DISCUSSIONS

3.1. Challenges

3.1.1. Interpretability

After training the datasets, deep learning models are ideal for accurately predicting trends. However, epidemiologists and public health decision makers frequently lack understanding of the internal decision-making process, such as how to determine whether an intervention program is beneficial. Algorithms have a difficult time explaining the cause of the anticipated trend and phenomenon. To trust the output projections, transparent thinking is necessary when making decisions about health, stock, and disease transmission.

3.1.2. Applicability

Deep learning models require large scale, high-quality training sets. The datasets chosen in training the models often have limitations. In Jung's study (Jung, 2025), the data collected in Busan and Seoul was used, which showed different results in training. Models trained on data from one region or phase of a pandemic may suffer from poor generalizability when applied to another region with different distribution of population, lifestyles and policies, etc.

3.1.3. Error Accumulation

Accuracy of prediction relied greatly on the data input to the deep-learning models. If there are errors in the input data, the prediction will have errors too, which will affect the following predictions as the errors accumulate. It becomes extraordinarily difficult to distinguish whether the predicted result comes from the original input or the errors.

3.2. Future Prospects

To overcome these challenges, future research should focus more on the following directions.

3.2.1. Expert System

To address the challenge in interpretability, embedding expert system which includes previous pandemics or branching process is a good way to combine the accuracy of data-driven models with traditional biological models with concrete epidemiological theories, which enhance the model's interpretability as part of the output is based on understandable rules.

3.2.2. Transfer Learning

Models can use transfer learning approaches to increase their generalizability by training on large amounts of data from a well-documented domain. This enables the models to be successfully applied to other domains with smaller amounts of data. Additionally, domain adaptation reduces the disparities in distributions between domains, allowing the models to function well in various target domains.

3.2.3. Detailed data cleansing process and statistic models

More careful data collection and cleansing process is required, including manual processing or preprocessing algorithms for better detection of errors. In addition, the proposed model should have better robustness against noise or erroneous data to improve the accuracy in prediction. Researchers can also adopt branching process to produce output with different circumstances and conditions.

4. Conclusion

This work introduced the framework and workflow of traditional machine learning and deep learning models. Although these methods demonstrate high accuracy in predicting the trajectories of the epidemic, they have limitations in generalizability and interpretability. This paper then proposed some potential methods, such as expert systems and transfer learning, in dealing with these problems.

It can be concluded that a hybrid framework, integrating branching process with machine learning, is a useful to approach to improve scientific interpretability and reduce influence of public behavior.

References

[1] Chen, W., Luo, H., Li, J., & Chi, J. 2024. Long-term trend prediction of pandemic combining the compartmental and deep learning models. *Scientific Reports*, 14 (1), 21068.

[2] Jung, S. 2025. Can the number of confirmed COVID-19 cases be predicted more accurately by including lifestyle data? An exploratory study for data-driven prediction of COVID-19 cases in metropolitan cities using deep learning models. *Digital Health*, 11, 20552076251314528.

[3] Kaur, G., et al. 2023. Forecasting prediction of COVID-19 outbreak using linear regression. In I. J. Jacob, S. Kolandapalayam Shanmugam, & I. Izonin (Eds.), *Data Intelligence and Cognitive Informatics*. *Algorithms for Intelligent Systems*. Springer, Singapore.

[4] Lipton, Z. C. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv* (abs/1506.00019).

[5] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438 (7066), 355 – 359.

[6] Mohammadi-Pirouz, H., et al. 2024. Development of decision tree classification algorithms in predicting mortality of COVID-19 patients. *International Journal of Emergency Medicine*, 17 (1), 126.

[7] Qian, Y., et al. 2025. Physics-informed deep learning for infectious disease forecasting.

[8] Riou, J., & Althaus, C. L. 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, 25 (4), 2000058.

[9] Song, Y., & Lu, Y. 2015. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27 (2), 130 – 135.

[10] Vali, A., Comai, S., & Matteucci, M. 2024. An automated machine learning framework for adaptive and optimized hyperspectral-based land cover and land-use segmentation. *Remote Sensing*, 16 (14), 2561.

[11] Weiss, H. M. 2013. The SIR model and the foundations of public health.

[12] Zhou, S. 2024. Using SARIMA method and random forest to predict the COVID-19 infection cases. *Advances in Economics, Management and Political Sciences*, 140 (1), 1 – 13.

[13] Zhang, L., Shen, L., Ding, L., Tao, D., & Duan, L. Y. 2022. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10174 - 10183).