ISSN 2959-6157

The Comprehensive Investigation for Covid-19 Trend Prediction Through Machine Learning and Deep Learning

Jiayu Xun[1, *]

[1]Department of Mathematics, Hamilton College, New York, US
*Corresponding author: jxun@hamilton.edu

Abstract:

Corona Virus Disease (Covid-19) has surely been a challenging problem to solve for the past few years. Due to the diversity in the form of dataset, it is essential to obtain accurate predictive results of Covid-19 trends. This paper analyzes different artificial intelligence methods used in Covid-19 trend prediction, including several machine learning and deep learning methods. More specifically, this work investigates linear regression, random forest, and decision trees in terms of machine learning and delves into Artificial Neural Network (ANN) as well as Long Short-Term Memory (LSTM) for deep learning. By comparing various past works, the effectiveness of machine learning and deep learning methods is achieved by their hidden algorithms, such as the Multiple Linear Regression (MLR) model for linear regression analysis. Incorporation with other models or methods is applied in deep learning. For example, Ensemble Empirical Mode Decomposition (EEMD) is included in ANN structure to decrease the noises within the Covid-19 datasets. Furthermore, the paper also inquiries into potential improvement of some drawbacks in predictive results for Covid-19 trends by reviewing related works of expert system and transfer learning as well as domain adaptation. The machine learning and deep learning models could provide accurate predictive results as a reference for related organizations to consider or establish insightful policies.

Keywords: Machine learning; deep learning; Covid-19 prediction

1. Introduction

Covid-19 is an infectious disease that spreads over approximately 100 countries and has been a focal point for more than 3 years. Related medications, such as vaccines, were created by many organizations, and preventive actions, including wearing masks in public places and having social distance in busy environments, are administered [1]. Even though the number of cases could be controlled by such actions, the case number could not be controlled properly in the long-term because different countries may have various capabilities to test and report the number of patients. In addition, virus may have different transmission rates and enhanced levels of vaccines resistance, making it difficult for professionals to obtain updated predictions. However, to cease the disease effectively and immediately, more efficient methods to diagnose symptoms and forecast future trends are needed.

In recent times, Artificial Intelligence (AI) was applied in multiple fields, including disease ecology area. For instance, AI methods are applied to Interstitial Lung Diseases (ILDs) to make diagnosis by using data sources from ILDs in various settings [2]. Another example of AI application would be the improved treatment of Cardiovascular Disease (CVD) [3]. The application of AI mainly consists of two categories, Machine Learning (ML) and Deep Learning (DL), and is used to analyze medical images and electronic medical records to diagnosis Covid-19 symptoms [4]. In addition, machine learning has a capacity for early detection. For example, Goodman-Meza et al created seven machine learning algorithms to diagnosis symptoms in the inpatient setting, and the classification results achieved the receiver operator curve of 0.91 [5], indicating that the result is very close to the perfect classifier. Because of the large amount of imaging data in medical centers [6], deep learning is often applied in medical imaging, such as CT imaging, X-ray imaging, and ultrasound imaging, to make diagnosis. Predicting future trends is also a necessary step in preventing the disease from spreading out in a timely manner. For this aspect, regression model based on time series data is established to predict Covid-19 trends in a long term [7]. Through a comprehensive review, professionals could gain enhanced accuracies of forecasting with the assistance of AI, and governments or related organizations could acquire timely insights of potential outbreaks and provide public health strategies in a timely manner. Comprehensive understand-

ing of AI could also offer policy makers potential impacts of lockdown, mask intervention, and traffic restrictions, so that they could make better policies immediately.

The rest section of this work is organized as follows. First, this review will detail the specific machine learning algorithms and deep learning methods applied in predicting Covid-19 trends as well as their accuracies in foresting through different metrices. To achieve this goal, this paper will delve into specific methods in the machine learning and the deep learning category, including linear regression, random forest, artificial neural network and long short-term memory. Lastly, this paper will give conclusions about the efficiencies of those methods by reviewing related journals and provide suggestions in potential future development in artificial intelligence's effort to predict disease trends.

2. Methods

2.1 Introduction of Machine Learning Workflow

Based on Fig. 1, machine learning includes multiple procedures, including data collection, preprocessing, model building, model training and testing. After identifying the target problem, professionals should apply data collection, which involves gathering data from various resources. Data preprocessing, such as cleaning out missing values or removing duplicates, is needed. Next, Exploratory Data Analysis (EDA) uses graphs to understand the distribution of data and outliers. After selecting the appropriate model based on the problem, professionals would train the model to fit the training data and evaluate the performance through different metrics, such as precision and recall. If the model does not perform well, redefining the model by trying different algorithms and changing hyperparameters is required. Then, professionals would test the model on test set to see the generalization on a new dataset. If the model works, professionals document the whole procedures and report necessary results to related organizations.

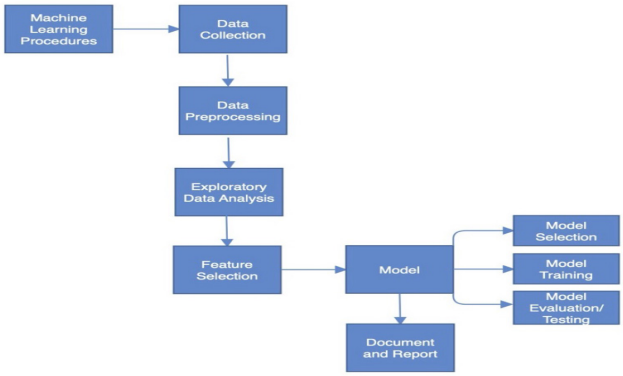

Fig. 1 Machine learning workflow (Photo/ Picture credit : Original).

2.2 Machine Learning-based Prediction

2.2.1 Linear regression

This method can form a linear relationship between the dependent and independent variables. From the mathematical aspect, by getting a linear function, the value of dependent variable can be obtained based on the value of independent variables. For instance, Gothai et al present that they have classified the original dataset into four classifications, including confirmed case, active case, death case and closed case, and obtain the number of confirmed cases through linear regression [1].

However, linear regression may not show a comprehensive understanding of variables' relations, thus Multiple Linear Regression (MLR) model may be further considered, which include one dependent variable and multiple independent variables. For instance, Rath et al consider that the three independent variables should be daily positive cases, recovered cases, and deceased cases, and the dependent variable would be the daily active cases [8].

2.2.2 Random forest

Random forests are often applied to find out the optimal predictors by combining multiple decision trees. In an example, random forest can be applied to forecast the prognoses of patients and recognize prime parameters for clinical prognoses. For instance, Wang et al show that by establishing the random forest based on the dataset collected from Wuhan Fourth Hospital and using the Gini index, two clinical risk predictors are chosen, including lactate dehydrogenase (LDH) and Myoglobin (Myo) [9].

2.2.3 Decision trees

Decision Trees can establish models for multiple predictors and the sequence of the predictors can be varied to find out the optimal result. Hence, decision trees show possible outcomes by collecting series of related decisions. Unlike random forests which consist of many decision trees that need to be selected or extracted from the bootstrap sampling, decision trees are simpler because they only explore nodes along with sub-nodes (leaf-nodes) by [10]. For instance, Hamdi et al present that the mean value along with the standard deviation may vary by selecting various nodes, and the predictive results also vary based on the node selection and related values [11].

2.3 Deep Learning-based Prediction

2.3.1 Artificial Neural Network (ANN)

Artificial Neural Network is considered to be a subset of neural network, which consists of an input, outputs, and hidden layers. Based on the Covid-19 data, the datasets could be separated into training set and testing set, and

the training set can fit into ANN model to predict future positive cases and be compared with the results of testing set. For instance, Niazkar et al show that the ANN structure can gain accurate predictions by considering the data of the previous fourteen days [12]. The ANN model can also be incorporated with other methods to generate more decent and accurate forecasts of Covid-19. According to Hasan, ANN architecture is incorporated with Ensemble Empirical Mode Decomposition (EEMD), which could denoise real-time Covid-19 data and trains the denoised datasets [13].

2.3.2 Long Short-Term Memory (LSTM)

LSTM can be well-applied for predicting Covid-19 trends because long-term dependencies can be captured in time series data. However, LSTM could be highly dependent on the data source, and for LSTM model, the prediction error would increase due to limited data. For instance, Ma et al come up with a way to reverse this situation by combining LSTM model with Markov model and getting the prediction results by calculating the training errors and structuring those errors with the probability transfer matrix of Markov model [14]. With this attempt, the LSTM-Markov model could achieve a higher accuracy in predicting Covid-19 trends than only using LSTM model.

3. Discussion

3.1 Advantages and Disadvantages of ML and DL

The prediction of Covid-19 trends immerses researchers in machine learning and deep learning. For machine learning, not only can it provide sufficient mathematical models for both categorical and numerical data, but also it can recognize patterns automatically without being specifically designed. Besides, the interpretation behind traditional machine learning algorithms, including regression analysis and random forest, is easy to understand so that it becomes widely used in different trend forecasts. However, machine learning also has its drawbacks. For instance, building a large tree can be complicated to analyze when using decision trees for a large dataset. Apart from machine learning, deep learning emersed and improved with more powerful algorithms to represent layered structures that can be interpreted as human brains [15]. However, deep learning requires large number of labeled data which can be difficult to achieve, and generalization to new data may not be effective due to overfitting.

3.2 Challenges in Covid-19 Trend Prediction

For complicated infectious diseases, like Covid-19, trend prediction is critical in controlling such diseases. However, there exists challenges for predicting trends of Covid-19 due to issues such as lacking interpretability and applicability of artificial intelligence models. The absence of interpretability and understandability would lead to insufficient accountability and potential reduction in the quality of predictive results for Covid-19 trends [16]. For users who involved in developing vaccines or establishing public health policies, understanding the learning process or how models make decisions would be a difficult task due to limited training. Because of regional and time differences, the Covid-19 trend prediction for one place may not be applicable for another place. For instance, when giving the daily confirmed number of Covid-19 deaths in specific regions [17], there exists plentiful fluctuations in a specific time slot for each country, hence it is difficult to apply one country's predictive results to study another country's case. Thus, the lack of applicability makes Covid-19 a difficult case to gain accurate predictive results.

3.3 Future Developments

3.3.1 Expert system for improving interpretability

In terms of future perspectives, the expert system could be widely implemented. For instance, González-Pérez et al apply a non-linear regression model to implement the expert system which can be accessed by users to manipulate through graphs because its machine learning-based algorithms allow parallel running and real-time data [18]. Besides, optimization of the expert system is also needed with the enhancement of AI methods. As data mining develops, models can be trained to create input-output mappings, and many big-data driven expert systems have largely appeared [19]. Hence, applying the expert system may resolve the challenges in Covid-19 trend prediction efficiently.

3.3.2 Transfer learning and domain adaptation for improving applicability

As countries may enter different stages of Covid-19 based on its population size, city distribution, and other related factors, transfer learning would be a proper method for generalization of each country's case [20]. Li et al proposed ALeRT-COVID which is a transfer learning model by applying attention-based Recurrent Neural Network (RNN) to train a model on pre-defined countries and transfer it to other countries [21]. By using transfer learning, less data and computational requirements are needed, and in emergency cases, such as Covid-19, transfer learning could provide decent predictive results in a timely manner for many countries. Being a subset of transfer learning, domain adaptation has been used in Covid-19 diagnosis through Multi-attention Representation Network Partial Domain Adaptation (MARPDA) [22]. Even though

domain adaptation has shortcomings, including mixing noises and emphasizing local information, it still has the potential to be applied in Covid-19 trend prediction.

4. Conclusion

This paper summarized several related ML and DL models or methods used in Covid-19 trend prediction by analyzing how the methods are used and their effectiveness. For the methods, this work mainly investigates linear regression, random forest, and decision trees for ML approaches, and ANN and LSTM for DL approaches. To gain a more comprehensive review, this paper studied similar past works of linear regression, decision tree, and ANN to compare their efficiencies in Covid-19 trend forecasts. By comparing different experiments and related works, it can be found out that both machine learning and deep learning methods could give decent predictive results by improving existed models and combining pre-existed algorithms, such as the LSTM-Markov model. However, due to the lack of understandability and applicability resulted from Covid-19 stage differences among countries, future developments of expert system and transfer learning is needed to solve more complicated infectious diseases trend prediction.

References

[1] Gothai E, Thamilselvan R, Rajalaxmi RR, Sadana RM, Ragavi A, Sakthivel R. Prediction of COVID-19 growth and trend using machine learning approach. Materials Today, 2023.

[2] Exarchos KP, Gkrepi G, Kostikas K, Gogali A. Recent advances of artificial intelligence applications in interstitial lung diseases. Diagnostics, 2023, 13(13): 2303.

[3] Chen Z, Xiao C, Qiu H, Tan X, Jin L, He Y, Guo Y, He N. Recent advances of artificial intelligence in cardiovascular disease. Journal of Biomedical Nanotechnology, 2020, 16(7): 1065–1081.

[4] Huang S, Yang J, Fong S, Zhao Q. Artificial Intelligence in the diagnosis of COVID-19: Challenges and perspectives. International Journal of Biological Sciences, 2021, 17(6): 1581–1587.

[5] Goodman-Meza D, et al. A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. PLOS ONE, 2020, 15(9).

[6] Desai SB, Pareek A, Lungren MP. Deep learning and its role in COVID-19 medical imaging. Intelligence-Based Medicine, 2020, 3–4: 100013.

[7] Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in covid-19 with logistic model and Machine Learning Technics. Chaos, Solitons & Fractals, 2020, 139: 110058.

[8] Rath S, Tripathy A, Tripathy AR. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2020, 14(5): 1467-1474.

[9] Wang J, Yu H, Hua Q, Jing S, Liu Z, Peng X, Cao C, Luo Y. A descriptive study of random forest algorithm for predicting COVID-19 patients outcome. PeerJ, 2020, 8.

[10] Zaidi SA, Tariq S, Belhaouari SB. Future prediction of COVID-19 vaccine trends using a voting classifier. Data, 2021, 6(11): 112.

[11] Hamdi M, Hilali-Jaghdam I, Elnaim BE, Elhag AA. Forecasting and classification of new cases of COVID-19 before vaccination using decision trees and Gaussian mixture model. Alexandria Engineering Journal, 2023, 62: 327–333.

[12] Niazkar HR, Niazkar M. Application of artificial neural networks to predict the COVID-19 outbreak. Global Health Research and Policy, 2020, 5(1).

[13] Hasan N. A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. Internet of Things, 2020, 11: 100228.

[14] Ma R, Zheng X, Wang P, Liu H, Zhang C. The prediction and analysis of covid-19 epidemic trend by combining LSTM and Markov method. Scientific Reports, 2021, 11(1).

[15] Möller DPF. Machine Learning and Deep Learning. Guide to Cybersecurity in Digital Transformation. Advances in Information Security, vol 103. Springer, Cham, 2023.

[16] Ennab M, Mcheick H. Designing an interpretability-based model to explain the artificial intelligence algorithms in Healthcare. Diagnostics, 2022, 12(7): 1557.

[17] Jamshidi M (Behdad), et al. A review on potentials of artificial intelligence approaches to forecasting COVID-19 spreading. AI, 2022, 3(2): 493–511.

[18] González-Pérez B, et al. Expert system to model and forecast time series of epidemiological counts with applications to COVID-19. Mathematics, 2021, 9(13): 1485.

[19] Ho C-T, Wang C-Y. A robust design-based expert system for feature selection and covid-19 pandemic prediction in Japan. Healthcare, 2022, 10(9): 1759.

[20] Qiu Y, Hui Y, Zhao P, Wang M, Guo S, Dai B, et al. The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process. Fuel, 2024, 372: 132162.

[21] Li Y, et al. Alert-covid: Attentive lockdown-aware transfer learning for predicting COVID-19 pandemics in different countries. Journal of Healthcare Informatics Research, 2021, 5(1): 98–113.

[22] He C, Zheng L, Tan T, Fan X, Ye Z. Multi-attention representation network partial domain adaptation for covid-19 diagnosis. Applied Soft Computing, 2022, 125: 109205.