www.ijiarec.com

ISSN:2348-2079

Volume-8 Issue-3

International Journal of Intellectual Advancements and Research in Engineering Computations

Comparison of Machine learning approaches to predict COVID-19 infection

Janani Prabu

ITM Business School, Student, Chennai, Data Scientist Intern in Dr.APJ Abdul Kalam International Research Foundation

ABSTRACT

In today's world, the SARS-CoV2 virus, which causes COVID-19 (coronavirus) has become a pandemic and has spread all over the world. Because of increasing number of cases day by day, it takes time to interpret the laboratory findings thus the limitations in terms of both treatment and findings are emerged. Due to such limitations, the need for clinical decisions making system with predictive algorithms has arisen. Predictive algorithms could potentially ease the strain on healthcare systems by identifying the diseases. In this study, we perform clinical predictive models that estimate, using machine learning. To evaluate the predictive performance of our models, precision, F1-score, recall, AUC, and accuracy scores calculated. The experimental results indicate that our predictive models identify the future confirmed cases (COVID- 19 disease) at accuracy of 94.60%, F1-score of 93.38%, precision of 89.48%, recall of 96.48%, and AUC of 78.50%. It is observed that predictive models trained on datasets could be used to predict COVID-19 infection, and can be helpful for medical experts to prioritize the resources correctly. The models (available at (https://github.com/JananiPrabu/COVID-19-Machine-learning-comparision))can be used to assist the medical experts and clinical prediction studies.

Keywords: SARS-CoV2, COVID-19, Coronavirus, Machine learning, Prediction

INTRODUCTION

Coronaviruses are a large family of viruses that can cause severe illness to the human being. The first known severe epidemic is Severe Acute Respiratory Syndrome (SARS) occurred in 2003, whereas the second outbreak of severe illness began in 2012 in Saudi Arabia with the Middle East Respiratory Syndrome (MERS). The current outbreak of illness due to coronavirus is reported in late December 2019. This new virus is very contagious and has quickly spread globally. On January 30, 2020, the World Health Organization (WHO) declared this outbreak a Public Health Emergency of International Concern (PHEIC) as it had spread to 18 countries. On Feb 11, 2020, WHO named this "COVID-19". On March 11, as the number of COVID-19 cases has increased thirteen times apart from China with more than 118,000 cases in 114 countries and over 4,000 deaths, WHO declared this a pandemic. As the outbreak of the COVID-19 has become a worldwide pandemic, the real-time analyses of epidemiological data are needed to prepare the society with better action plans against the disease. Since the birth of novel COVID19, the world is restlessly fighting with its cause. The main objective of this paper

- To predict and forecast COVID-19 cases, deaths, and recoveries through predictive modelling,
- Develop a model based on the machine learning to predict the virus spread in future.

Author for correspondence:
Janani Prabu
ITM Business School, Student, Chennai,
Data scientist intern in Dr.APJ Abdul Kalam International Research Foundation

In this study, we provide a prediction system for detection of COVID-19 infection by developing and applying various machine learning models.

RELATED WORK

It is important to predict clinical tasks for health base systems. Computer aided clinical predictive models have been used in various areas including risk of heart failure [29], mortality in pneumonia [30, 31], mortality risk in critical care [32–34]. With these systems medical experts are enable to comprehend and assess clinical findings better. In this study, we build on recent methodological advances to provide clinical predictive model for COVID-19. Similar studies about clinical prediction for COVID-19 are limited in the literature. In this study, Four clinical features were considered and two different - Linear regression, Facebook prophet were applied. The performance of the algorithms was evaluated with only accuracy values. Best accuracy was obtained with Facebook Prophet with 96.4%. In the another study [27], authors applied machine learning classifiers to predict COVID-19 diagnosis. Clinical data was obtained from Kaggle for COVID 19 spread in India. 18 clinical findings were considered in the study and classifiers were evaluated with AUC, sensitivity, specificity, F1-score, Brier score, positive predictive value, and negative predictive value. Only five different classifiers were applied including, SVM, random forests, neural networks, logistic regression, and gradient boosted trees. The best AUC scores were obtained with both SVM, and random forest classifiers with 0.847. In the study of [28] , clinical predictive model for COVID-19 was proposed. In the study, data was collected from Hospital Israelite Albert Einstein at Sao Paulo, Brazil like in this study and [27]. Authors applied various machine learning applications including RF, NN (Neural Network), LR, SVM, XFB (Gradient Boosting) and determined the performance of classifiers by calculating sensitivity, specificity, and AUC scores. The best performance was obtained with XGB with 66% AUC score.

METHODS AND DATA

Data description

Dataset includes the Age group details, Hospital beds in India, ICMR Testing Labs, Individual Details. Population of India details and State wise Testing details. In this our objective is to predict the spread of corona virus in future. So only the necessary fields like confirmed cases, daily updated cases and date pf the patient admitted in the hospital are considered. Samples were collected from patients to detect SARS-CoV2 in the early months of 2020. Dataset contains 111 laboratory findings from 5644 various patients. In the dataset, the rate of positive patients was around 10% of which around 6.5% and 2.5% required hospitalization and critical care. In the dataset, there is no gender information. According to the study of [26– 28], 18 laboratory findings have a vital role on COVID-19 disease. Thus, we wiped away remaining laboratory features to balance the dataset and to perform COVID-19 detection. After the balancing process, dataset includes 18 laboratory findings from 600 patients,

AI based algorithms learn from the historical data to provide predictions for the future outcomes. Machine learning (ML) and deep learning (DL) algorithms can be considered as a subset of the AI. It is an area that is based on learning and improving on its own by analyzing computer algorithms. There are certain differences between machine learning and deep learning. To assess the predictive performance of each of the developed predictive models, we calculated their performance in terms of accuracy, f1-score, precision, recall, and area under roc curve (AUC). To validate the data, we both used 10-fold cross validation and 80– 20 train-test split approach

DEEP ANALYSIS INTO ALGORITHMS

Linear regression

Linear regression is perhaps one of the most well-known and well understood algorithms in statistics and machine learning. The reason is because linear regression has been around for so long (more than 200 years). It has been studied

from every possible angle and often each angle has a new and different name. Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression. Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression. Linear regression is an attractive model because the representation is so simple.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric. The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example). It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model. When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available. There are many more techniques because the model is so well studied. Here we are going to implement the normal equation method through the linear regression

Facebook Prophet

When a forecasting model doesn't run as planned, we want to be able tune the parameters of the method with regards to the specific problem at hand. Tuning these methods requires a thorough understanding of how the underlying time series models work. The first input parameters to automated ARIMA, for instance, are the maximum orders of the differencing, the auto-regressive components, and the moving average components. A typical analyst will not know how to adjust these orders to avoid the behavior and this is the type of expertise that is hard to acquire and scale. The Prophet package provides intuitive parameters which are easy to tune. Even someone who lacks expertise in forecasting models can use this to make meaningful predictions for a variety of problems in a business scenario. We use a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- g(t): piecewise linear or logistic growth curve for modelling non-periodic changes in time series
- s(t): periodic changes (e.g. weekly/yearly seasonality)
- h(t): effects of holidays (user provided) with irregular schedules
- ε_t: error term accounts for any unusual changes not accommodated by the model

Using time as a regressor, Prophet is trying to fit several linear and non linear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in Holt-Winters technique. We are, in effect, framing the

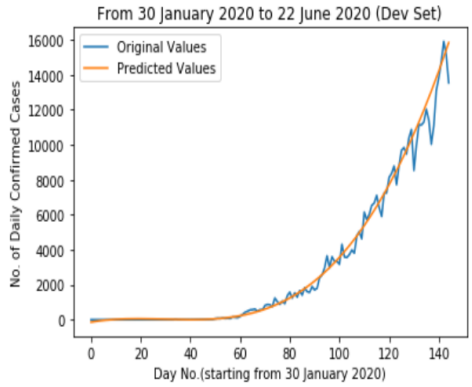
forecasting problem as a curve-fitting exercise rather than looking explicitly at the time-based dependence of each observation within a time series.

APPLICATION RESULT

The below table shows the accuracy and other criteria when the model is developed in those two algorithms.

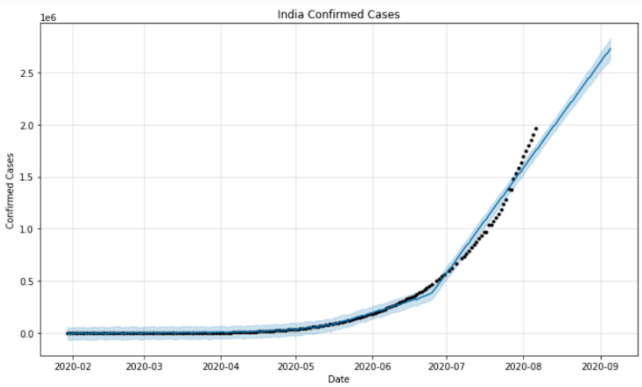
	Accuracy	F1-Score	Precision	Recall	AUC
Linear Regression	0.9032	0.9134	0.8855	0.9578	0.7115
Facebook Prophet	0.94600	0.9338	0.8948	0.9648	0.7849

In addition to these, we tested the performance of the algorithms using 80–20 train-test split approach. As can be seen in above Table, the accuracy results of all machine learning models were reached at least 90.00% and above.

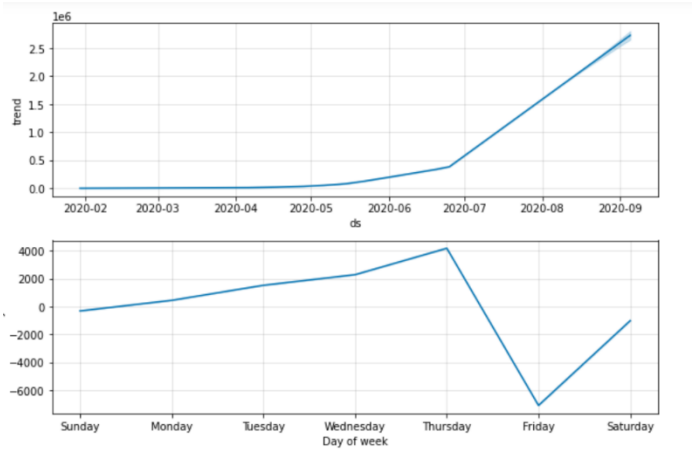


5.1 Prediction using linear regression

The major limitation in this study is the size of the data. Data up to August 3rd 2020 were used, and some laboratory findings could not be measured for some patients. In addition to these, the data was imbalanced, thus we balanced the data by deleting some materials. The performance of these models can be enhanced with a larger data set.



5.2 Prediction using Facebook Prophet (Date wise)



5.3 Prediction using Facebook Prophet (Day wise)

CONCLUSION

In conclusion, we found evidence to suggest that machine learning application models can be applied to predict COVID-19 infection with laboratory findings. Our experimental results indicate that government can take some effective measures to reduce the confirmed cases in future. Based on our study's results, we conclude that health- care systems should explore the use of predictive models that assess individual COVID-19 risk in order to improve healthcare re- source prioritization and inform patient care.

REFERENCES

[1]. World Health Organization, Report of the WHO-China joint mission on coronavirus disease (COVID-19). 2020 https://www.who.int/docs/default-source/ coronaviruses/who- china- joint- mission- on- Covid- 19- final-report.pdf .

[2]. World Health Organization, Health topics, coronavirus. 2020 https://www.who.int/health-topics/coronavirus#tab=tab _ 3.

[3]. National Institute of Infection Diseases, Field briefing: diamond princess COVID-19 cases. 2020 https://www.niid.go.jp/niid/en/2019-ncov-e/ 9407- covid- dp- fe- 01.html.

[4]. Del Rio C, Malani PN. Novel coronavirus – important information for clinicians. J Am Med Assoc 323(11), 2019, 2020. doi: 10.1001/jama.2020.1490 .

[5]. Wang D, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. J Am Med Assoc 323(11), 2020. DOI: 1061-1069. doi: 10.1001/jama.2020.1585.

[6]. Jiehao C, et al. A case series of children with 2019 novel coronavirus infection: clinical and epidemiological features. Clin Infect Dis 2020; ciaa198. doi: 10.1093/ cid/ciaa198 .

[7]. Karm KQ, et al. A well infant with coronavirus diseases 2019 (COVID-19) with high viral load. Clin Infect Dis 2020; ciaa201. doi: 10.1093/cid/ciaa201 .

[8]. Bai Y, Yao L, Wei T, et al. Presumed asymptomatic carrier transmission of COVID-19. J Am Med Assoc 323(14), 2020, 1406–7. doi: 10.1001/jama.2020.2565 .

[9]. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4), 2017. doi: 10.1136/svn- 2017- 0 0 0101 .

[10]. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare J 6(2), 2019, 92–8. doi: 10.7861/futurehosp.6- 2- 94 .

[11]. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 112(1), 2019, 22–8. doi: 10.1177/014107681881551 .

[12]. Alakus TB, Turkoglu I . Detection of pre-epileptic seizure by using wavelet packet decomposition and artificial neural networks. In: 10th International Conference on Electrical and Electronic Engineering; 2017, 511–15.
[13]. Memarian N, Kim S, Dewar S, Engel J, Staba RJ. Multimodal data and ma- chine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. Comput Biol Med; 64(1), 2015, 67–78. doi: 10.1016/j. compbiomed. 06.008.
[14]. Yousefi J, Hamilton-Wright A. Characterizing EMG data using machine-learning tools. Comput Biol Med 51, 2014, 1–13. doi: 10.1016/j.compbiomed.2014.04.018 .
[15]. Karthick PA, Ghosh DM, Ramakrishnan S. Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms. Comput Methods Programs Biomed 154, 2018, 45–56. doi: 10.1016/j.cmpb.2017.10.024 .
[16]. Alfaras M, Soriano MC, Ortin S. A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection. Front Phys 2019. doi: 10. 3389/fphy.2019.00103 .
[17]. Ledezma CA, Zhou X, Rodriguez B, Tan PJ, Diaz-Zuccarini V. A modeling and machine learning approach to ECG feature engineering for the detection of ischemia using pseudo-ECG. PLoS ONE 14(8), 2019. PMC6690680. doi: 10.1371/ journal.pone.0220294 .
[18]. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer diagnosis using deep learn- ing: a bibliographic review. Cancers (Basel) 11(9), 2019, E1235. doi: 10.3390/ cancers11091235 .
[19]. Andriasyan, V., Yakimovich, Georgi, F. et al., Deep learning of virus infec- tions reveals mechanics of lytic cells, bioRxiv, 2019. doi: https://doi.org/10. 1101/798074.
[20]. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 2020, 706–10. doi: 10.1038/ s41586- 019- 1923- 7 .
[21]. Bosco G, Gangi MA. Deep learning architectures for DNA sequence classifica- tion. Lect Notes Comput Sci 2017, 162–71. doi: 10.1007/978- 3- 319- 52962- 2 _ 14 .
[22]. Krishna MM, Neelima M, Harshali M, Rao MVG. Image classification using deep learning. Int J Eng Technol 7(2.7), 2018, 614–17. doi: 10.14419/ijet.v7i2.7.10892 .
[23]. Nassif AB, Shahin I, Attilli I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: a systematic review. IEEE Access 7, 2019, 19413-19165. doi: 10. 1109/ACCESS.2019.2896880.
[24]. Li, Y, Huang, C, Ding, L, Li, Z, Pan, Y, Gao, X. Deep learning in bioinformatics: introduction, application, and perspective in big data era, *arXiv*, 2019.
[25]. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assess- ment. J Thoracic Oncol 5(9), 2010, 1315–16. doi: 10.1097/JTO.0b013e3181ec173d .
[26]. Jiang X, Coffee M, Bari A, Wang J, Jiang X, et al. Towards an artificial intel- ligence framework for data-driven prediction of coronavirus clinical severity. Compu Mater Continua 63(1), 2020, 537–51. doi: 10.32604/cmc.2020.010691 .
[27]. Batista, A.F., Miraglia, J.L., Donato, T.H.R., and Filho, A.D.P.C., COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, *medRxiv*, 2020. doi: 10.1101/2020.04.04.20052092 .
[28]. Schwab, P., Schütte, A.D., Dietz, B., and Bauer, S. "predCOVID-19: a systematic study of clinical predictive models for coronavirus disease 2019, arXiv: 08302, 2005, 2020.
[29]. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 48(6), 2010, 106–13. doi: 10.1097/MLR.0b013e3181de9e17 .
[30]. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 9(2), 1997, 107–38. doi: 10.1016/S0933-3657(96)00367- 3 .
[31]. Wu C, Rosenfeld R, Clermont G. Using data-driven rules to predict mortality in severe community acquired pneumonia. Plos ONE" 9(4), 2014, e89053. doi: 10. 1371/journal.pone.0089053.
[32]. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predict- ing hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 29(2), 2001, 291–6.

[33]. Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, Szolovits P. Unfolding physiological state: mortality modelling in intensive care units. KDD 2014, 75–84. doi: 10.1145/2623330.2623742 .

[34]. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. Proc Mach Learn Res 68, 2017, 361–76.

[35]. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guaddarrama S, Saenko K, et al. Long-term recurrent convolutional networks for visual recogni- tion and description. IEEE Trans. Patt. Analy. Mach. Intelli. 39(4), 2017, 677–91. doi: 10.1109/TPAMI.2016.2599174 .

[36]. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. doi: 10.1109/CVPR.2015.7298935 .

[37]. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 18(4), 2018. doi: 10.1186/ s12911- 018- 0677- 8 .

[38]. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for med- ical diagnostic test evaluation. Caspian J Intern Med 4(2), 2013, 627–35.

[39]. Kamarudin AN, Cox T, Kolamunnaage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 17(53), 2017. doi: 10.1186/s12874-017-0332-6 .

[40]. Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction model for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369, 2020. doi: 10.1136/bmj.m1328 .

[41]. Pierce R. Evaluating information: validity, reliability, accuracy, triangulation. Res Methods Polit 2008:79–99. doi: 10.4135/9780857024589 .

[42]. Li H, Li C, Liu HG. Clinical characteristics of novel coronavirus cases in ter- tiary hospitals in Hubei Province. Chin Med J 133(9), 2020, 1025–31. doi: 10. 1097/CM9.0000000000000744.

[43]. Huang C, Wang Y, Li X, Ren L, Zhao J, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223), 2020, 497– 506. doi: 10.1016/S0140-6736(20)30183-5.

[44]. World Health Organization. Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance. https://www.who.int/publications-detail/laboratory-testing-for-2019-novel-coronavirus-in-susp ected- human-cases- 20200117 . (Updated on 2020).

[45]. Wölfel R, Corman VM, Guggemos W, et al. Virological assessment of hos- pitalized patients with COVID-2019. Nature 581, 2020, 465–9. doi: 10.1038/ s41586- 020- 2196- x .

[46]. Wang, W., Xu, Y., Gao, R., and et al. "Detection of SARS-CoV-2 in different types of clinical specimens," JAMA, 323(18), 1843–4, 220. doi: 10.1001/jama. 2020.3786.

[47]. Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020. doi: 10.1148/radiol.2020200432.