**COVID-19 FUTURE FORECASTING USING SUPERVISED MACHINE LEARNING MODELS**

**ABSTRACT**

Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.

INTRODUCTION

MACHINE learning (ML) has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle (AV), business applications, natural language processing (NLP), intelligent robots, gaming, climate modeling, voice, and image processing. ML algorithms’ learning is typically based on trial and error method quite opposite of conventional algorithms, which follows the programming instructions based on decision statements like if-else [1]. One of the most significant areas of ML is forecasting [2], numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis. Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease [3]. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease [4], cardiovascular disease prediction [5], and breast cancer prediction [6]. In particular, the study [7] is focused on live forecasting of COVID-19 confirmed cases and study [8] is also focused on the forecast of COVID19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively. This study aims to provide an early forecast model for the VOLUME 4, 2016 1 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/. This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997311, IEEE Access Author et al.: Preparation of Papers for IEEE TRANSACTIONS and JOURNALS spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO) [9]. COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia [10]. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration [11]. Hundreds of thousands of people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close person to person physical contacts, by respiratory droplets, or by touching the contaminated surfaces. The most challenging aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared either partial or strict lockdowns throughout the affected regions and cities. Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity. To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. The forecasting is done for the three important variables of the disease for the coming 10 days: 1) the number 0f New confirmed cases. 2) the number of death cases 3) the number of recoveries. This problem of forecasting has been considered as a regression problem in this study, so the study is based on some state-of-art supervised ML regression models such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). The learning models have been trained using the COVID-19 patient stats dataset provided by Johns Hopkins. The dataset has been preprocessed and divided into two subsets: training set (85% records) and testing set (15% records). The performance evaluation has been done in terms of important measures including R-squared score (R2 score), Adjusted R-squared Score (R2 adjusted), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). This study has some key findings which are listed below: • ES performs best when the time-series dataset has very limited entries. • Different ML algorithms seem to perform better in different class predictions. • Most of the ML algorithms require an ample amount of data to predict the future, as the size of the dataset increases the model performances improve. • ML model based forecasting can be very useful for decision-makers to contain pandemics like COVID-19. The rest of the paper consists of six sections. Section I presents the introduction, section II contains the description of the dataset and methods used in this study. Section III presents the methodology, section IV presents the results, and section V summarizes the paper and presents the conclusion.

**Existing System:**

The existing system for forecasting COVID-19 cases typically relies on traditional statistical models and basic predictive techniques. Many countries and health organizations use epidemic models such as SEIR (Susceptible-Exposed-Infectious-Recovered) or simple curve-fitting methods to forecast trends. However, these models often fail to capture the complex, multifactorial nature of pandemic progression, as they might not account for various dynamic factors such as regional lockdowns, population behavior, and vaccination rates. Therefore, they lack the flexibility and accuracy required for precise forecasting in rapidly changing scenarios like COVID-19.

**Proposed System:**

In this study, the proposed system leverages machine learning models to improve the forecasting of COVID-19 trends. The system uses four standard ML-based forecasting models—Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), and Exponential Smoothing (ES)—to predict key factors such as the number of newly infected cases, deaths, and recoveries for the next 10 days. This system aims to harness the strength of machine learning to adapt to more complex patterns in the data, offering better predictive accuracy than traditional models. The use of these multiple models allows for a comparative analysis to determine which models perform best in different scenarios.

**Conclusion:**

The study concludes that machine learning models, especially Exponential Smoothing (ES), can be highly effective in forecasting COVID-19 cases. The results show that ES outperforms the other models in predicting the new confirmed cases, death rates, and recovery rates, proving to be the most promising method. LR and LASSO also perform well, providing reliable forecasts for new infections and death rates, whereas SVM performs poorly across all predictions due to the limitations posed by the available dataset. Overall, the study demonstrates the potential of machine learning-based forecasting mechanisms in improving decision-making during the ongoing COVID-19 pandemic.

REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” PloS one, vol. 13, no. 3, 2018.

[2] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, “Machine learning strategies for time series forecasting,” in European business intelligence summer school. Springer, 2012, pp. 62–77.

[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, “Regression models for prognostic prediction: advantages, problems, and suggested solutions.” Cancer treatment reports, vol. 69, no. 10, pp. 1071–1077, 1985.

[4] P. Lapuerta, S. P. Azen, and L. LaBree, “Use of neural networks in predicting the risk of coronary artery disease,” Computers and Biomedical Research, vol. 28, no. 1, pp. 38–52, 1995.

[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, “Cardiovascular disease risk profiles,” American heart journal, vol. 121, no. 1, pp. 293–298, 1991.

[6] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[7] F. Petropoulos and S. Makridakis, “Forecasting the novel coronavirus covid-19,” Plos one, vol. 15, no. 3, p. e0231236, 2020.

[8] G. Grasselli, A. Pesenti, and M. Cecconi, “Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response,” Jama, 2020.

[9] WHO. Naming the coronavirus disease (covid-19) and the virus that causes it. [Online]. Available: https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease- (covid-2019)-and-the-virus-that-causes-it

[10] C. P. E. R. E. Novel et al., “The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china,” Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi, vol. 41, no. 2, p. 145, 2020.

[11] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, “Identification of a new human coronavirus,” Nature medicine, vol. 10, no. 4, pp. 368–373, 2004.

[12] J. H. U. data repository. Cssegisanddata. [Online]. Available: https://github.com/CSSEGISandData [13] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, “Chapter 1 - analytics defined,” in Information Security Analytics, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston: Syngress, 2015, pp. 1 – 12. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/B9780128002070000010>

[14] H.-L. Hwa, W.-H. Kuo, L.-Y. Chang, M.-Y. Wang, T.-H. Tung, K.-J. Chang, and F.-J. Hsieh, “Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models,” Journal of evaluation in clinical practice, vol. 14, no. 2, pp. 275–280, 2008.

[15] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

[16] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[17] X. F. Du, S. C. Leung, J. L. Zhang, and K. K. Lai, “Demand forecasting of perishable farm products using support vector machine,” International journal of systems Science, vol. 44, no. 3, pp. 556–567, 2013.

[18] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, “Tweets classification on the base of sentiments for us airline companies,” Entropy, vol. 21, no. 11, p. 1078, 2019. 10 VOLUME 4, 2016 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/. This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997311, IEEE Access Author et al.: Preparation of Papers for IEEE TRANSACTIONS and JOURNALS

[19] E. Cadenas, O. A. Jaramillo, and W. Rivera, “Analysis and forecasting of wind velocity in chetumal, quintana roo, using the single exponential smoothing method,” Renewable Energy, vol. 35, no. 5, pp. 925–930, 2010.

[20] J. Lupón, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galán, E. Zamora, J. Vila, J. Peñafiel, A. Urrutia, E. Ferrer et al., “Biomarkerassist score for reverse remodeling prediction in heart failure: the st2-r2 score,” International journal of cardiology, vol. 184, pp. 337–343, 2015.

[21] J.-H. Han and S.-Y. Chi, “Consideration of manufacturing data to apply machine learning methods for predictive manufacturing,” in 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, 2016, pp. 109–113.

[22] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance,” Climate research, vol. 30, no. 1, pp. 79–82, 2005.

[23] R. Kaundal, A. S. Kapoor, and G. P. Raghava, “Machine learning techniques in disease forecasting: a case study on rice blast prediction,” BMC bioinformatics, vol. 7, no. 1, p. 485, 2006.

[24] S. Baran and D. Nemoda, “Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting,” Environmetrics, vol. 27, no. 5, pp. 280–292, 2016.

[25] Y. Grushka-Cockayne and V. R. R. Jose, “Combining prediction intervals in the m4 competition,” International Journal of Forecasting, vol. 36, no. 1, pp. 178–185, 2020.

[26] N. C. Mediaite. Harvard professor sounds alarm on ‘likely’ coronavirus pandemic: 40% to 70% of world could be infected this year. Accessed on 2020.02.18. [Online]. Available: <https://www.mediaite.com/news/harvardprofessor-sounds-alarm-on-likely-coronavirus-pandemic-40-to-70-ofworld-could-be-infected-this-year/>

[27] BBC. Coronavirus: Up to 70% of germany could become infected - merkel. Accessed on 2020.03.15. [Online]. Available: https://www.bbc.com/news/world-us-canada-51835856