ANALYSIS OF COVID-19 DISEASE USING MACHINE LEARNING - PERSONALIZED MODEL

Anđela Blagojević[1,2], Tijana Šušteršič[1,2], Ivan Lorencin[3], Sandi Baressi Šegota[3], Nikola Anđelić[3], Dragan Milovanović[4,5], Dejan Baskić[5,6], Zlatan Car[3], Nenad Filipović[1,2]

[1] University of Kragujevac, Faculty of Engineering, Sestre Janjić 6, 34000 Kragujevac, Serbia

[2] Bioengineering Research and Development Center (BioIRC), Prvoslava Stojanovića 6, 34000 Kragujevac, Serbia

[3] University of Rijeka, Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia

[4] Clinical Centre Kragujevac, Zmaj Jovina 30, 34000 Kragujevac, Serbia

[5] University of Kragujevac, Faculty of Medical Sciences, Svetozara Markovića 69, 34000 Kragujevac, Serbia

[6] Institute of Public Health Kragujevac, Nikole Pašića 1, 34000 Kragujevac, Serbia

e-mails: andjela.blagojevic@kg.ac.rs, tijanas@kg.ac.rs, ilorencin@riteh.hr, sbaressisegota@riteh.hr, nandelic@riteh.hr, piki@medf.kg.ac.rs, dejan.baskic@gmail.com, car@riteh.hr, fica@kg.ac.rs

Abstract:

Coronavirus disease (COVID-19), since its appearance, has put a large burden on the global health system which have strived to mitigate the pandemic, but mortality of COVID-19 continues to increase. Many authors have employed machine learning (ML) algorithms in the investigation of COVID-19 in order to identify infected individuals, predict their condition in time, predict the outbreaks and forecast certain numbers. Although there are many studies that examine the application of ML in the diagnosis of prognostic biomarkers and survival prediction several days in advance, there is a limited literature dealing with evidence to label patients in more categories (mild, moderate, severe, etc.) that would help not only to respond in a timely manner to prevent lethal results, but also to minimize the number of patients in hospitals where this is not the case. In this paper we present a methodology for classification of patients into 3 distinct classes of clinical condition (mild, moderate and severe) of COVID-19 disease and prediction of the outcome (change of severity of clinical condition) in advance. The results show that XGBoost classifier achieved average accuracy of 88%. The main advantage of our system is that it is a rule-based algorithm which is easier to implement in a real clinical practice, instead of the use of black box models, which are not appealing for real clinical use.

Key words: COVID-19, personalized model**,** clinical condition assessment, ensemble model, rule-based machine learning

1. Introduction

The development of COVID-19 disease caused by novel coronavirus SARS-Cov-2 has been investigated for two years and researchers are still making the effort to find the adequate solution

that will repress the virus and reduce the burden of the healthcare system [1]. Machine learning (ML) proved to be adequate methodology for many purposes including epidemiological and personalized predictions [2, 3]. There are many studies related with clinical issues such as fast diagnosis, prediction and classification of blood test analyses, as well as prediction of final clinical outcomes [4, 5, 6]. The main drawback of the most studies is reflected in binary outcomes – survival/death [7]. Binary classification may not be the best type of classification in situations where the healthcare systems are overloaded. On the other hand, knowledge extracted from multiclass models could help doctors to determine which patients will develop critical condition and therefore should stay at the hospital, and which patients could be treated at home, having only mild condition. Multiclass-based research could be also helpful for hospital managers in decision-making process to improve other important secondary treatment endpoints and institutional performances, which are deteriorated by inappropriate measures such as unnecessary prescription of adjunctive drugs and inappropriate allocation of intensive care beds. Therefore, in this paper we propose a methodology based on ML to classify patients into several classes and predict the outcome in advance (change of severity of clinical condition).

2. Materials and Methods

The dataset used in this paper consists of blood test analyses from 105 patients. Patients' data were collected in two hospitals – Clinical Center of Kragujevac, Serbia and Clinical Center of Rijeka, Croatia. The dataset consisted of 44 female and 61 male patients, with age distribution given in the form mean ± standard deviation – 52.77 ± 16.63. We divided the clinical data into three subgroups: demographic data (gender and age), symptoms (fever, cough, fatigue, chest pain, muscle pain, headache, dyspnea, loss of taste or smell) and blood analysis (complete blood count, coagulation, kidney function, hepatic function, enzymes, electrolytes, oxygenation and acid-base balance, inflammation indices, carbohydrate metabolism).
From the perspective of machine learning methodology, there are two main tasks, prediction of the blood analysis in advance and determination of the severity of clinical condition (mild, moderate and severe) based on the prediction from the first task. The proposed methodology is presented in Figure 1.

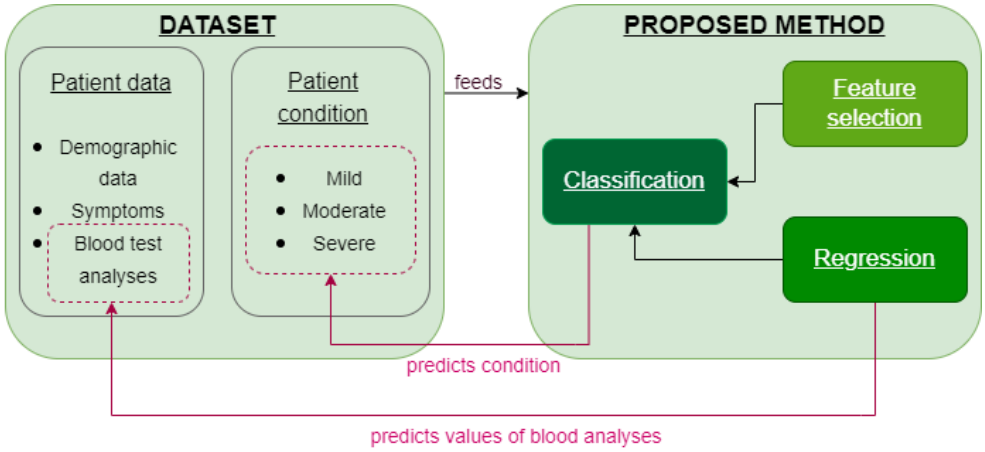

Figure 1. Proposed methodology.

Firstly, in order to evaluate which blood tests (biomarkers) have the highest impact on the disease development, five features have been chosen through a feature selection process. Following previous findings on the importance of biomarkers the next step is to predict the change of biomarkers values in time. The aim is to predict the patient's clinical condition two weeks after hospital admission based on blood test analyses from previous days. The dataset related to blood

test analyses is divided in training and test sets (training set consists of data from days 2,5,7,9 and 11, therefore, test set consists of data from the 14th day). For the prediction of each blood test analysis, Gradient boosting regressor (GBR) was used. As it was mentioned before, the main task of this paper is assessment of clinical condition of patients in advance. Firstly, it is necessary to assess the value of blood analyses, then classify them into one of 3 classes (mild, moderate and severe). For the described classification task, the aim was to construct a simplified, rule-based decision model, and for that purpose, the model of extreme gradient boosting (XGBoost) was used. XGBoost as well as GBR, was trained with optimal hyperparameters obtained by grid search method.

3. Results and discussion

Blood test analyses WBC, CRP, creatinine, urea, and LDH were selected as the features that had contributed the most to the classification task. This feature selection method reduced the amount of blood test analyses that need to be assessed. Values of all five selected blood biomarkers are assessed for 34 patients on the 14th day in hospital by gradient boost regressor model. Due to the small number of patients' data available in time, we decided to select these 34 patients with a full blood analysis for all days, according to the described methodology in the previous section. After evaluation of the results of patients' data, it is possible to predict the patient's clinical condition in advance by the XGBoost classification algorithm. Model was tested on 34 patients and achieved an accuracy of 88% in predicting the patient's condition on the 14th day. For the mentioned test set, the confusion matrix was computed, and it is presented in Figure 2.

Figure 2. Confusion matrix with regular values (left) and normalized values (right).

In addition to the accuracy, we have computed precision, recall, F1-score, AUC and PR values. The average value of precision is 0.87, average value of the recall is 0.9 and average F1-score is 0.88, value of AUC is 0.95 and average PR score is 0.9.

4. Conclusion

The aim of this paper is to create automatic ML methods for classification of patients with COVID-19 into classes (mild, moderate and severe). This methodology is also capable to predict the change of category in advance, which means that the model predicts the development of patient's clinical condition during hospital stay until discharge or death. This research represents a proof of concept that a ML model is an efficient and informative method to gain insight into the COVID-19 disease process. Further research would be focused on collecting larger database of

patients as well as investigation in terms of upgrading the existing ML models in order to achieve higher accuracy.

Acknowledgement

The research was funded by Serbian Ministry of Education, Science, and Technological Development, grant [451-03-68/2022-14/200107 (Faculty of Engineering, University of Kragujevac)]. This research is also supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 952603 (SGABU project). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

Bibliography

[1] S. Cabaro, V. D'Esposito, T. Di Matola, S. Sale, M. Cennamo, D. Terracciano, V. Parisi, F. Oriente, G. Portella, F. Beguinot, L. Atripaldi, M. Sansone and P. Formisano, "Cytokine signature and COVID-19 prediction models in the two waves of pandemics," *Scientific reports*, vol. 11, no. 1, pp. 1-11, 2021.

[2] K. Ikemura, E. Bellin, Y. Yagi, H. Billett, M. Saada, K. Simone, L. Stahl, J. Szymanski, D. Goldstein and M. Reyes Gil, "Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study," *Journal of medical Internet research*, vol. 23, no. 2, p. e23458, 2021.

[3] S. Sen, S. Saha, S. Chatterjee, S. Mirjalili and R. Sarkar, "A bi-stage feature selection approach for COVID-19 prediction using chest CT images," *Applied Intelligence*, vol. 51, no. 12, pp. 8985-9000, 2021.

[4] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo and Y. Yuan, "An interpretable mortality prediction model for COVID-19 patients," *Nature machine intelligence*, vol. 2, no. 5, pp. 283-288, 2020.

[5] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie and Y. Yang, "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature medicine*, vol. 26, no. 8, pp. 1224-1228, 2020.

[6] G. Wu, P. Yang, Y. Xie, H. C. Woodruff, X. Rao, J. Guiot, A.-N. Frix and P. Lambin, "Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study," *European Respiratory Journal*, vol. 56, no. 2, 2020.

[7] L. Yan, H. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang and Y. Yuan, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," *medRxiv*, 2020.