COVID-19 Early Diagnosis with the Use of Machine Learning

Anna Marciniak[1,2][0000−0002−9900−3951], Agata Giełczyk[1][0000−0002−5630−7461], Martyna Tarczewska[1], and Sylwester Kloska[2][0000−0002−5165−9302]

[1] University of Science and Technology, Bydgoszcz, Poland
agata.gielczyk@pbs.edu.pl
[2] Faculty of Medicine Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, Bydgoszcz, Poland

Abstract. Since December 2019 COVID-19 disease has spread all over the world, paralyzing human life and threatening our security. Thus, the need for a novel and fast approach to diagnosing COVID-19 infections became predominant. This work proposes a machine learning based method to classify the early symptoms of SARS-CoV-2 infection. Three classifiers have been selected: Random Forest, XGBoost, and LightGBM. Additionally, each classifier was tested on an unbalanced and balanced dataset and with the use of default and tuned hyper parameters. XGBoost model gave the best results after training at balanced dataset with Accuracy=79.94%, Precision=93.32%, Recall=64.50%, and F1-score=76.62%. The selected ML model was then linked to a mobile application that contained a questionnaire about symptoms. After completing the questionnaire, the result obtained from the ML model is returned to the user in the application. The obtained results are prognostic results, suggesting social isolation and/or performing additional tests, e.g., a PCR test, and are not a substitute for the professional medical diagnosis.

Keywords: SARS-CoV-2 · COVID-19 · Machine Learning · XGBoost · Random Forest · LightGBM · Classification

1 Introduction

Since December 2019 SARS-CoV-2 virus has spread all over the world from Wuhan, China, causing a disease known as coronavirus disease COVID-19. On 30 January 2020, the World Health Organization (WHO) announced a public health emergency, and the epidemic rapidly evolved into a pandemic by March 2020, with a high number of cases in the Europe, especially in Italy [5]. The healthcare systems were overstretched and, as the result, patients had serious obstacles in receiving needed medical help on time. Thus, a rapid tool for the diagnosis support were especially needed.

In this article, we evaluate three ML-based models. Additionally, we examine models with/without dataset balancing and with/without hyper parameters

tuning. We propose also a mobile application that can be used for COVID-19 self diagnosis. The proposed application cannot substitute the professional medical diagnosis, but it can be the first tool used in triaging system at medical centres. It uses the machine learning methods for binary classification (COVID positive vs COVID negative). The article is constructed as follows. In section 2, we describe the materials and methods, in section 3 we provide the obtained results. The last section contains the conclusions and future possibilities.

2 Materials and Methods

The schema of the proposed application and the machine learning-based pipeline are presented in Fig. 1. In the top part of the figure the machine learning process can be observed. We provided the data, processed them and finally, we used them in order to develop the ML models. Then, we performed experiments and the most promising model was implemented in the mobile application. In the bottom part of the figure the use of the application is presented. Firstly, the user fulfill the questionnaire. Then, the answers are analysed by the ML-based model and the application provides the result - COVID positive or negative.

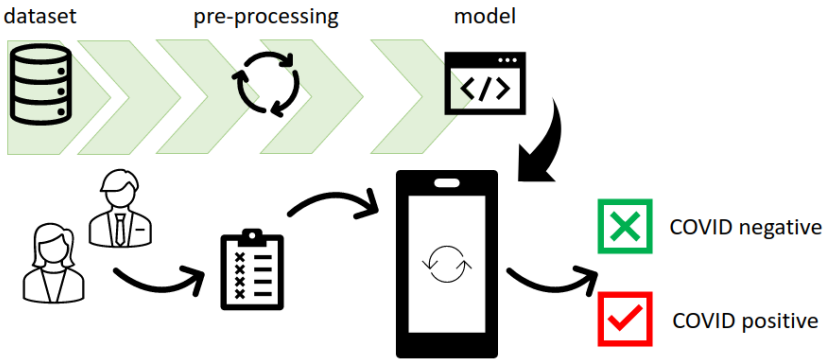

Fig. 1. The overview of the proposed method

2.1 Dataset and Data Pre-processing

In this research, we used the dataset provided by Yazeed Zaobi [7]. The dataset was created from the data gathered in Israel. It contains the following pieces of information: date of test, cough, fever, throat soreness, breath shortness, headache, age 60 and above, gender, and test indication. All negative and positive COVID-19 cases in this dataset were confirmed via RT-PCR assay.

Firstly, the column 'date of test' was removed due to its irrelevance. Secondly, all the uncertain COVID results were removed. Then, all the data were

converted into a numerical format. After the pre-processing dataset contained 2 701 378 observations. In the research we used two separable subsets with random elements: 80% of data for training and 20% for testing. Due to the significant lack of balance in the dataset (90% of observations was COVID negative), we decided to use the dataset with and without balancing. The balancing was ensured by the SMOTE algorithm [1].

2.2 Classification

In the classification step of the method we implemented and evaluated three ML-based models: Random Forest (used also in [2]), XGBoost [3] and LightGBM [4]. All models were evaluated in four different experiments - on unbalanced and balanced dataset, and with different hyper parameter values - default and tuned using the Optuna framework [6].

3 Results

The results of the conducted research are presented in Table 1. It presents the evaluation of three models by Accuracy, Precision, Recall, and F1-score. In the table, we additionally provide two time related parameters: time of learning and time of predicting. The results in time domains were evaluated using the Google Colaboratory environment. Individual classifiers results do not differ significantly for different research variants. However, for XGBoost classifier (marked in bold in Table 1) trained on default hyper parameters and balanced dataset the evaluation measures and time are the most promising, namely Accuracy=79.94%, Precision=93.32%, Recall=64.50%, and F1-score=76.62%.

Table 1. The obtained results: Accuracy, Precision, Recall, F1-score, Training time, and Prediction time for Random Forest, XGBoost, and LightGBM, both with and without balancing and hyper parameters tuning

Model	Hyper parameters	Data balance	Acc.	Prec.	Recall	F1-score	Training time[s]	Pred. time[s]
Random Forest	Default	None	0.9395	0.6489	0.5593	0.6008	109	4.68
		SMOTE	0.7993	0.9332	0.6447	0.7626	279	0.925
	Optuna	None	0.9395	0.6489	0.5594	0.6008	604	2.58
		SMOTE	0.7993	0.9332	0.6447	0.7626	844	2.61
XGBoost	Default	None	0.9386	0.6517	0.5274	0.5830	98	0.942
		SMOTE	**0.7994**	**0.9332**	**0.6450**	**0.7628**	**199**	**0.197**
	Optuna	None	0.9395	0.6489	0.5593	0.6009	857	3.37
		SMOTE	0.7994	0.9332	0.6450	0.7628	954	0.338
LightGBM	Default	None	0.9396	0.6492	0.5594	0.6010	20.3	1.79
		SMOTE	0.7994	0.9332	0.6450	0.7628	52.1	0.323
	Optuna	None	0.9396	0.6493	0.5593	0.6009	34.1	2.21
		SMOTE	0.7995	0.9325	0.6458	0.7631	55.6	0.295

4 Conclusions

Since the start of the pandemic in 2020, many new mutations of the SARS-CoV-2 virus have emerged, each with a different set of disease symptoms. This may be one of the ways of developing the prepared application - distinguishing not only sick/healthy, but also typing the virus variant. Applications of this type can potentially be used as a tool in the healthcare sector. The result obtained by analysing the survey responses by the machine learning model may help healthcare professionals in the procedure of triaging patients. However, the diagnosis obtained from a machine learning model is not a substitute for a visit to the doctor, but an indication for further diagnosis or its omission.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
2. Gupta, V.K., Gupta, A., Kumar, D., Sardana, A.: Prediction of covid-19 confirmed, death, and cured cases in india using random forest model. Big Data Mining and Analytics **4**(2), 116–123 (2021)
3. Irawati, M.E., Zakaria, H.: Classification model for covid-19 detection through recording of cough using xgboost classifier algorithm. In: 2021 International Symposium on Electronics and Smart Devices (ISESD). pp. 1–5. IEEE (2021)
4. Kelter, D., Ghiassi, K., Patel, S., Connors, C., Bonk, M., Gray, E., Zarbiv, S., Menon, A., Juneja, P.: Use of feature engineering to predict covid-19 mortality. In: TP51. TP051 COVID: LUNG INFECTION, MULTIORGAN FAILURE, AND CARDIOVASCULAR, pp. A2630–A2630. American Thoracic Society (2021)
5. Macera, M., De Angelis, G., Sagnelli, C., Coppola, N., COVID, V., et al.: Clinical presentation of covid-19: case series and review of the literature. International journal of environmental research and public health **17**(14), 5062 (2020)
6. Srinivas, P., Katarya, R.: hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost. Biomedical Signal Processing and Control **73**, 103456 (2022)
7. Zoabi, Y., Deri-Rozov, S., Shomron, N.: Machine learning-based prediction of covid-19 diagnosis based on symptoms. npj digital medicine **4**(1), 1–5 (2021)