Automatic COVID-19 Prediction with Optimized Machine Learning Classifiers Using Clinical Inpatient Data

Abbas Jafar[1], Myungho Lee[2]
[1]Dept. of Computer Engineering, Myongji University
[2] Dept. of Computer Engineering, Myongji University

jafarabbas1272@gmail.com, myunghol@mju.ac.kr

Abstract

COVID-19 is a viral pandemic disease that spreads widely all around the world. The only way to identify COVID-19 patients at an early stage is to stop the spread of the virus. Different approaches are used to diagnose, such as RT-PCR, Chest X-rays, and CT images. However, these are time-consuming and require a specialized lab. Therefore, there is a need to develop a time-efficient diagnosis method to detect COVID-19 patients. The proposed machine learning (ML) approach predicts the presence of coronavirus based on clinical symptoms. The clinical dataset is collected from the Israeli Ministry of Health. We used different ML classifiers (i.e., XGB, DT, RF, and NB) to diagnose COVID-19. Later, classifiers are optimized with the Bayesian hyperparameter optimization approach to improve the performance. The optimized RF outperformed the others and achieved an accuracy of 97.62% on the testing data that help the early diagnosis of COVID-19 patients.

1. Introduction

In late 2019, a severe disease named coronavirus (COVID-19) started in China, and within a short period disseminated worldwide. World Health Organization (WHO) announced COVID-19 as a pandemic disease in March 2020 [1]. The major reason for spreading the virus is the contaminated air with coronavirus droplets exhaled by COVID-19 patients. Early identification of COVID-19 can stop the spreading of the virus.

Currently, different COVID-19 detection methods are under examination. The widely used method is Polymerase Chain Reaction (PCR) to examine whether the actual COVID-19 virus exists or not. However, the drawbacks of this examination are the limited number of PCR laboratories, the lack of PCR kit supplies, and the long-awaited examination results [2]. In the last few years, ML approaches are diagnosing diseases artificially by learning from the training data and making correct predictions. For COVID-19, ML algorithms are employed to analyze the clinical features of the coronavirus and classify the accurate detection with their processing capabilities. Albahri S. [3] detected the COVID-19 cases with decision tree (DT) ML algorithms and achieved an overall 0.99% R_2 score. Khanday et al. [4] classify the coronavirus patients based on the clinical reports with the naïve Bayes (NB) classifier and obtained 96.2% accuracy.

In this paper, we build supervised ML models to detect COVID-19 accurately from clinical data. The dataset consists of symptom features such as fever, cough, headache, shortness of breath, and sore throat. We use Extreme Gradient Boosting (XGB), Decision Tree (DT), NB, and Random Forest (RF) ML classifiers. All the classifiers consist of multiple numbers of hyperparameters. We tune the hyperparameters of the models with the global Bayesian Optimization (BO) approach [5]. The evaluation metrics such as accuracy, sensitivity, specificity, and F1_score express the model's efficiency in detecting the coronavirus. The performance analysis of the ML classifiers with and without optimization is conducted (see Figure 1). The optimized RF classifier using selected features achieved the highest level of accuracy (97.62%) among all the classifiers.

The rest of the paper is explained; dataset collection and the pre-processing approaches to process the data are discussed in Section 2. Optimized ML classifiers with the BO approach are explained in Section 3. Section 4 presents the experimental results of different ML classifiers with and without optimization. Finally, we conclude the paper in Section 5.

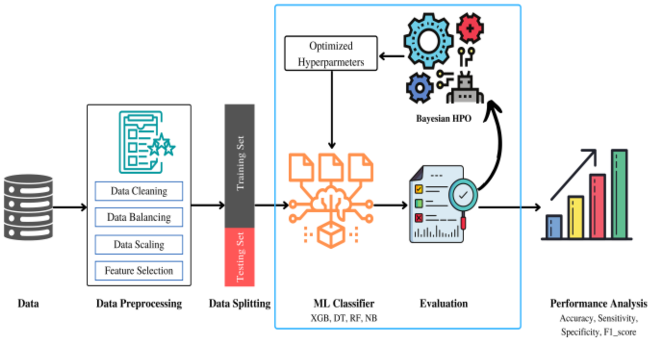

Figure 1. The overall optimized ML algorithms-based workflow of the classification of COVID-19.

2. Data Source and Pre-processing

We collected COVID-19 data from the open-source GitHub repository, extracted from the Israeli Ministry of

Health website [6]. The real data consists of 10 features including test_date, gender, age_60_and_above, fever, headache, cough, sore_throat, shortness_of_breath, test_indication, and corona_result.

For the data pre-processing, we first cleaned the dataset and removed the missing values. We dropped those features which have no impact on the prediction such as gender, date, age, and test indications from the data. The data scaling approach is used to upscale the data so that the ML models would learn better. Figure 2 shows the correlation of the selected features contributing toward the detection of coronavirus. In the dataset, the target class contains 3.76% of COVID-19 positive and 96.34% of negative cases, so data balancing is performed. We applied the oversampling approach to the minority class (COVID_Positive) and downsampling to the majority class (COVID_Nagative) to generate synthetic data to balance the target class. Later, the dataset is split into training (80%) and testing sets (20%), respectively.

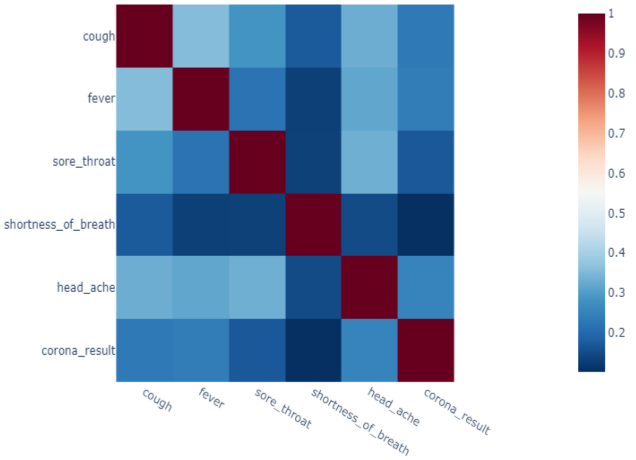

Figure 2. The correlation of selected features to predict COVID-19.

3. Optimized ML Classification Algorithms

For the identification of COVID-19-positive patients based on symptoms, different supervised ML algorithms such as XGB, DT, RF, and NB are used.

XGB is a supervised gradient-boosting (GB) algorithm that can effectively handle classification, regression, and ranking problems. XGB prediction is based on the residual, not the actual class labels. It uses the maximum size of trees similar to the GB to overcome overfitting. The new trees are generated based on the previous trees, and the final prediction is achieved based on all trees. The key hyperparameters used in the training of XGB are learning_rate, gamma, maximum_depth, and num_jobs.

DT is a non-parametric classifier to make human-like decisions based on decision rules. DT predicts the target class by learning the data. In tree models, the discrete values of the target class are classification trees, and the target class with continuous values is typically known as a regression tree. The hyperparameters of DT are maximum_feat, maximum_depth, and criterion.

Random forest is an ML algorithm that uses decision trees for classification. Each decision tree in RF selects the data samples and predicts the class with major voting. It is considered the most accurate algorithm and can be used for both classification and regression. It can handle the missing values in the data and helps extract the important features to overcome overfitting.

NB discriminates objects from large data features. NB is a Bayes theorem to classify the different objects from certain features using probability functionality. NB generates the conditional probability to allocate class labels properly so that one feature would not affect the other features and it predicts independently. Table 2 shows the hyperparameters of NB with a possible range of values.

Table 2. Machine learning classifiers with selected hyperparameters and their optimal configured values. (Hyp. = Hyperparameters)

HPO	ML Models	Selected Hyp.	Values Range	Optimal Values
BO	XGB	estimators_n	50, 100, 150	100
		depth_max	3, 5, 7, 9	5
		gamma	0, 0.1, 0.5	0.1
		learning_rate	0.001, 0.01, 0.1	0.03
	DT	feat_max,	sqrt, log2	sqrt
		depth_max,	3, 5, 7, 9	7
		criterion	gini, entropy, log_loss	gini
	RF	estimators_n	50, 100, 150, 200	150
		depth_max	3, 7, 9	7
		criterion	gini, entropy, log_loss	gini
	NB	alpha	0.001, 0.01, 1, 10	0.01
		fit_prior	true, false	true

Our methodology uses ML classifiers to classify COVID-19 patients based on their clinical symptoms. Each algorithm has its own hyperparameters to control the hyperplane. The performance of each classifier is observed with different evaluation matrices. The optimal choice of hyperparameters leads to better model performance which requires optimization. We used Bayesian Optimization (BO) approach to tune the hyperparameters of ML models and obtain the optimal search space. BO is a global optimization to find the optimal combination of hyperparameters by tracking past evaluations **[7]**. It computes the objective function iteratively and decides which search space is worthy of the current ML algorithms. Table 2 shows the obtained optimal search space of hyperparameters of each ML classifier after applying BO.

4. Experimental Results

It is important to measure the performance of different classifiers for a specific task in ML. The performance can be measured using various evaluation matrices. To calculate these matrices, we need to evaluate the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The mathematical equations to calculate the metrics are

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{all predictions}$$
$$\text{Specificity} = \text{TN}/(\text{TP} + \text{FP})$$
$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

We used Sickit learn tool with compulsory libraries for the experiments. All the experiments are conducted on GeForce GTX 1060. Initially, the models are trained on the COVID-19 dataset, and results are evaluated without optimization. Later, evaluation results are improved with BO hyperparameter optimization (HPO). The list of selected hyperparameters with the optimal configured values for each classifier using BO is shown in Table 2.

Table 3. Classification performance of ML classifiers on the testing set with and without HPO. (Acc.=Accuracy)

Without HPO				
ML Models	**Acc.**	**Sensitivity**	**Specificity**	**F1_score**
XGB	**94.45**	93.78	93.18	94.74
DT	93.76	93.01	92.85	93.47
RF	94.23	93.96	93.62	94.52
NB	92.94	93.75	92.86	93.01
With Bayesian Hyperparameter Optimization				
XGB	97.46	96.82	96.68	97.06
DT	96.37	95.73	96.08	96.27
RF	**97.62**	97.04	96.87	97.23
NB	95.83	96.11	95.45	95.94

Table 3 shows the performance of each classifier with and without HPO. The upper portion of Table 3 presents a performance evaluation of classifiers for the COVID-19 dataset, without the use of HPO. It can be analyzed that XGB outperformed all the other classifiers and achieved the highest classification accuracy of 94.74%. However, DT and RF perform well, and the accuracy is close to the XGB. For the second part of Table 3, we applied the BO, obtained the optimal configured value of hyperparameters for each classifier, and later trained each classifier to improve the results. RF performed exceptionally well, achieving the highest scores in terms of accuracy, sensitivity, specificity, and F1_score of 97.62%, 97.04%, 96.87%, and 97.23%, respectively. The results highlight that our optimized methodology improved the accuracy to predict COVID-19 patients by a valuable margin.

5. Conclusion

Due to the COVID-19 pandemic, the diagnosis of coronavirus became challenging. This paper proposed a fast-track solution to detect the coronavirus with clinical symptoms using machine learning. Different ML classifiers are used to predict the presence of coronavirus based on the severe symptoms. The performance of each classifier is assessed with varying matrices of evaluation. Moreover, the evaluation results are improved with the Bayesian hyperparameter optimization. Experimental results show that optimized RF model achieved a best classification accuracy of 97.62%. This approach can be readily integrated into mobile devices, which would benefit the clinical staff and the end-users to predict the COVID-19 patients.

Acknowledgements

This work was supported by the Supercomputer Development Leading Program of the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2020M3H6A1084984).

References

[1] N. Jebril, “World Health Organization Declared a Pandemic Public Health Menace: A Systematic Review of the Coronavirus Disease 2019 ‘COVID-19,’” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3566298, Apr. 2020.

[2] H. X. Bai *et al.*, “Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT,” *Radiology*, vol. 296, no. 2, pp. E46–E54, Aug. 2020.

[3] A. S. Albahri *et al.*, “Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review,” *J. Med. Syst.*, vol. 44, no. 7, p. 122, 2020.

[4] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data.,” *Int J Inf Technol*, pp. 1–9, 2020.

[5] M. Nour, Z. Cömert, and K. Polat, “A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization,” *Appl. Soft Comput.*, vol. 97, p. 106580, Dec. 2020.

[6] “https://github.com/nshomron/covidpred.” (Accessed on Mar.22, 2023).

[7] E. C. Garrido-Merchán and D. Hernández-Lobato, “Dealing with Categorical and Integer-valued Variables in Bayesian Optimization with Gaussian Processes,” *Neurocomputing*, 2020.