A Comparative Study of Imputation Techniques on COVID-19 Infection Detection

Jiajun Gao, Yusuf Danisman
Queensborough Community College, CUNY

Abstract

There are two common methods for diagnosing COVID-19 infection, Molecular and Antigen Tests, which have been applied worldwide. These methods achieve high-accuracy, but at the same time, they have high false-negative rates. Moreover, these diagnostic methods require expensive medical equipment and professional medical personnel. As an alternative way to detect Covid-19 detection, machine learning methods trained with routine blood exam data can be used. Blood exams are one of the most commonly applicable fast exams in the medical area and only require cheap equipment and less medical personnel. It is common to identify missing values in a dataset and replace them with a numeric value which is called missing data imputation. In this study, the dataset that was used to train the machine learning methods, has a non-negligible amount of missing values that would affect the results. We propose machine learning models and imputation method pairs to improve accuracy of COVID-19 detection. As imputation methods mean, median, k-Nearest Neighbor, and iterative were applied to five different classification models separately. We have developed four robust machine learning classification models provided by San Raffaele Hospital. These models achieved average balanced accuracy above 90%, and have 90% to 96% accuracy in the 95% Confidence Interval.

Key words: COVID-19, Imputation, Machine Learning, Blood Exam

Introduction

More than a year has passed since the initial COVID-19 outbreak, but still there is no fast, reliable and low-cost way to detect the virus. The most common blood-related test is an antibody test, which identifies whether an individual has had COVID-19 at some point in the past. The positive test results indicate that you may have developed antibodies due to COVID-19 virus infection. However, a positive result may also be produced by other viruses of the same class. Seroprevalence estimates as of July 2020 indicate that the actual number of infections is much higher than the number of reported cases, which may reflect the fact that some people might have mild or no symptoms, or have not sought medical care or been tested, but are still likely to continue to spread the virus in the population [1]. Moreover, there is always limited medical equipment and sanitary environment for certain regions, which can affect the efficiency of detecting SARS-CoV-2 infections [2].

According to Brinati et al. [3], the dataset of blood exam has been separated into 80% training set and 20% validation set after being imputed by MICE method. [3] provided two outperformance models with optimal hyperparameters that have been selected as the best models,

respectively Logistic Regression (LR) and Random Forest (RF). Among them, Random Forest has higher performance. Therefore, RF has been trained on the same dataset but excludes "Gender" features that provide 86% accuracy, 95% sensitivity, 86% recall, and 75% specificity.

In Yan et al. [4], the dataset of 351 patients' blood exams was imputed by padding -1 on missing values. [4] developed an XGBoost machine learning model, which can predict the mortality of COVID-19 patients with more than 90% accuracy and 10 days in advance through blood testing.

Cabitza et al [2]. is the development research of the paper Brinati et al. [3]. There was a distinctive dataset which includes 1624 patients admitted at San Raphael Hospital (OSR). Five models were developed based on the OSR dataset. Among them, Random Forest achieved the best overall performance -- 88% accuracy, 86% sensitivity, and 91% specificity. Among COVID-specific and CBC dataset, KNN was the best overall performance model with 86% accuracy on both sets, 80% and 82% sensitivity, 92% and 82% specificity.

Due to the data fragmentation in the database used, in Brinati et al. [3], blood samples are used to detect the COVID-19 infection by using MICE imputation method. As a comparative method with MICE on [3], we applied various imputation methods such as mean, median, kNN, and iterative imputation.

The original dataset included 279 (samples), 15 imputation-ML algorithm pairs, and a label of reverse transcriptase Polymerase Chain Reaction (rt-PCR) performed in the laboratory by San Raffaele Hospital [3] for the detection of the virus. Before handling the unbalanced data, data was oversampled randomly and divided into two groups with equal size in terms of diagnosis results. Patients diagnosed with positivity were assigned to class 1 and those that were negative to class 0. With the same amount of data in the two classes, the multiple imputation methods mentioned above were used to perform data processing. In this project, we focus on five certain machine learning models to evaluate the result. The models considered in the paper are XGBoost (XGB), Logistic Regression (LG), Random Forest (RF), Decision Tree (DT), and LightGBM (LGB). Fivefold nested cross-validation [5] is used to get the average performance of each machine learning model and do hyperparameter tuning to further improve the performance. The information involved in this paper was followed up to June 2021.

The rest of this article is organized as follows. In the Data Description section, the characteristics and differences of all the data used in the study will be outlined. In Supporting Evidence, we gave examples of existing related experiments and their advantages and disadvantages. It provided a theoretical basis for this research. The implementation process is described in Method, including data cleaning, model training process, and model performance improvement. Next, our results would be discussed in the Results section. Discussion summarized this article and made inferences based on the results obtained.

Data Description

In this paper, three distinctive datasets have been used [2, 3, 4]. The dataset from IRCCS Ospedale San Raffaele Hospital consists of 16 features, 279 patients' blood samples that were randomly extracted from the end of February 2020 to mid-March 2020 [3]. The correlation of features are shown in Fig. 1. Among 279 patients, 177 positive cases and 102 negative cases were marked respectively as 1 and 0 under the feature of "TARGET" which is the result of the RT-PCR test for COVID-19, and 14 other routine blood test indicators.

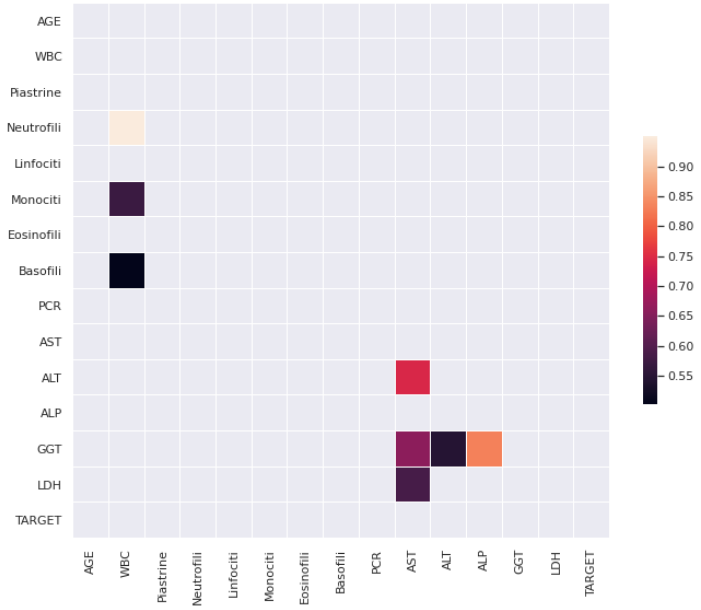

Fig. 1 Correlation coefficients, that are greater than 0.5, of 279 patients' blood samples from [3]

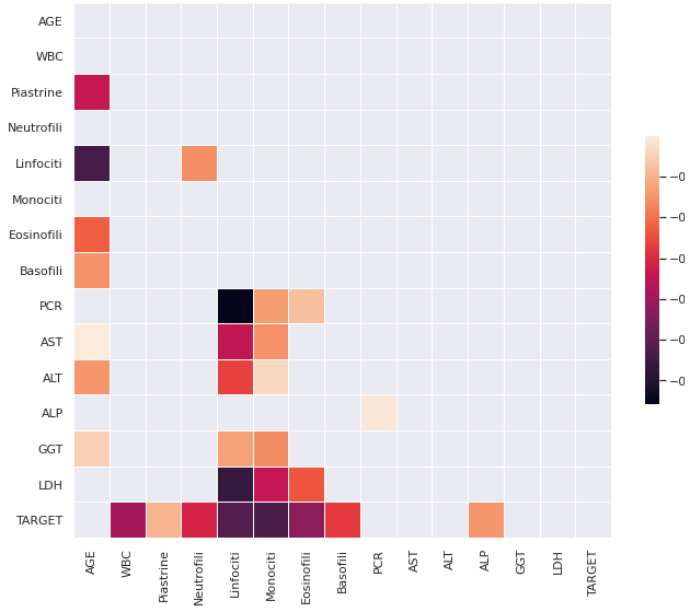

Fig. 2 Correlation coefficient that are less than 0, of 279 patients' blood samples from [3]

The dataset of mortality prediction [4] consists of 77 features for 375 patients who were tested positive from 10 January to 18 February 2020, are shown in Fig. 2. Each patient has multiple rows of routine blood exam data, totaling 6120 rows of available data in the dataset. Features include "age," "gender," "outcome" (in which 0 represents patient survived, 1 represents patient died, and other routine blood test indicators [4].

Supporting Evidence

By far, the most widely used and recognized virus detection in the world is to take samples of the respiratory tract based on reverse transcription polymerase chain reaction (rt-PCR). In addition to the previous method, the combination of chest CT scan [6], deep learning technology, and lung CT scan with machine learning detection methods [7] are all related to radiation dose, the number of related equipment and related operations. Due to the cost and other reasons, it is difficult to use this for screening tasks [3]. Also, 60% of patients, whose chest CT test is positive,

have a negative final test result [8]. Based on this phenomenon, the cost and uncertainty of using CT detection have increased again [8]. According to [9], the existing research reports have more or less high bias, over-fitting, or over-learning.

The disadvantages described above greatly threaten the lives of patients and increase the risk of outbreak or recurrence of the epidemics within the surrounding area. In addition, blood sample data has missing values is a common case. Nevertheless, there won't be precise results without high-quality data. It is crucial to find an efficient imputation method to deal with missing values. In [7], chest CT-based diagnosis, clinical symptoms, exposure history and laboratory testing are used as a reference to rapidly diagnose potential COVID-19 patients, but the testing only used white blood cell count. In [3], only one imputation method was used to train the model -- multivariate imputation by chained equation (MICE). The scarcity of such similar studies were also mentioned in [3]. There is no study that compared the potential impact of different imputation methods on blood sample data for covid-19 detection, nor mortality risk evaluation. Therefore, our project aims to develop a universal clinically diagnostic machine learning model by selecting reliable imputation methods based on patients' results of blood exams. Specifically, we used a 95% confidence interval to compare various imputation and machine learning models. Then, four best combinations among them were chosen for further improvement. These models only demand simple blood sample results which increases the efficiency of the testing.

Methods

We were focusing on applying the suitable imputation methods, including K-Nearest Neighbors (Knn), iterative, median, mean imputer, and non-imputation for XGB and LGB. Specifically, K-Nearest Neighbors imputes missing values by using values from a fixed number of nearest neighbors that have value for the feature. Iterative Imputer is a strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion which is similar to multivariate imputation by chained equation [10]. Mean Imputer uses the average value of the dataset to fill in missing data. Similarly, Median Imputer uses the median of the dataset.

During the development of machine learning algorithms, four imputation methods have been used with five machine learning models. In particular, we considered the following models. Support Vector Classifier (SVC), finds a hyperplane in an N-dimensional space that distinctly classifies the data points at once. Decision Tree (DT) learns a hierarchy of if/else questions, then leads to a decision [5]. Random Forest (RF) is essentially a collection of decision trees and it reduces the amount of overfitting by averaging their results [5]. XGBoost Classifier (XGB), is an implementation of gradient boosted decision trees designed for speed and performance. LightGBM Classifier (LGB), is also a gradient boosting framework that uses tree based learning algorithms.

For a total of 22 combined models, shown in Fig. 3, we used bootstrap's algorithm to calculate their prediction accuracy. In the bootstrap inner loop, the data group was divided into two groups of 0 and 1 based on the target. Each imputation-model combination was oversampled 20 times to avoid the impact of sample size imbalance, and knowing that the number of data is 177

from each class. The oversampled dataset would be separated into two group corresponding training and test sets with a ratio of 8 to 2, as the dataset that would go through the nested CV. Inside the nested CV, GridSearchCV with 5-fold was applied. The grid search tervase all the combination hyperparameters and would return the best one to cross validate. As a result, 100 results of model accuracy obtained for each imputation and model combination. The mean of sample and standard deviation were used to obtain the model accuracy range of each combination with a 95% confidence interval. We use the same method to obtain the 95% confidence interval of precision, recall and specificity to comprehensively evaluate the model.

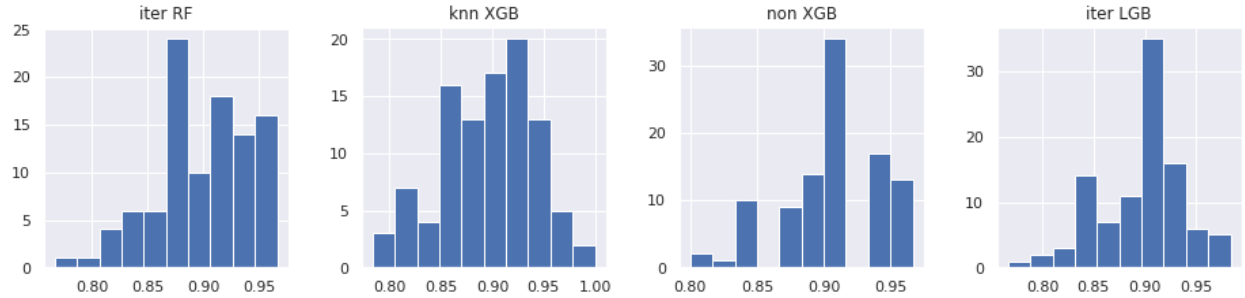

Fig. 3 Histogram of accuracy distribution for selected imputation-model combinations

In order to make the detection model generalize and adapt to new data, the four models use the hyperparameters optimized by GridSearchCV to refit and test the original data set and the resampled data set. After we got the best estimators, we fitted the estimators with the training set of imputed dataset of corresponding pairs. Then we transformed the datasets from [4] and [2] respectively to the same scale as the fitted training set. Ultimately, we evaluated the results by score function.

Results

All the processing was finished on Google Colaboratory with Python. The majority library includes Pandas, NumPy, and Scikit-learn. Data visualization and exploratory data analysis were done by using matplotlib and seaborn libraries. Certain machine learning algorithms were imported from XGBoost [11] and LightGBM [12].

	Combinations	Low	High
0	mean SVC	0.629000	0.860000
1	median SVC	0.640957	0.857043
2	knn SVC	0.610760	0.876573
3	iter SVC	0.635549	0.874451
4	mean DT	0.696617	0.932716
5	median DT	0.699828	0.933172
6	knn DT	0.720237	0.930096
7	iter DT	0.740775	0.945892
8	mean RF	0.793234	0.974099
9	median RF	0.799311	0.972022
10	knn RF	0.805058	0.966609
11	iter RF	0.812552	0.984114
12	mean XGB	0.792878	0.984789
13	median XGB	0.791263	0.974404
14	knn XGB	0.803874	0.988126
15	iter XGB	0.789905	0.979762
16	non XGB	0.827213	0.976121
17	mean LGB	0.805249	0.969417
18	median LGB	0.789316	0.985017
19	knn LGB	0.783257	0.975410
20	iter LGB	0.812700	0.980966
21	non LGB	0.792829	0.976838

Table 1. The table results of 95% confidence interval for 5-fold nested CV model accuracy

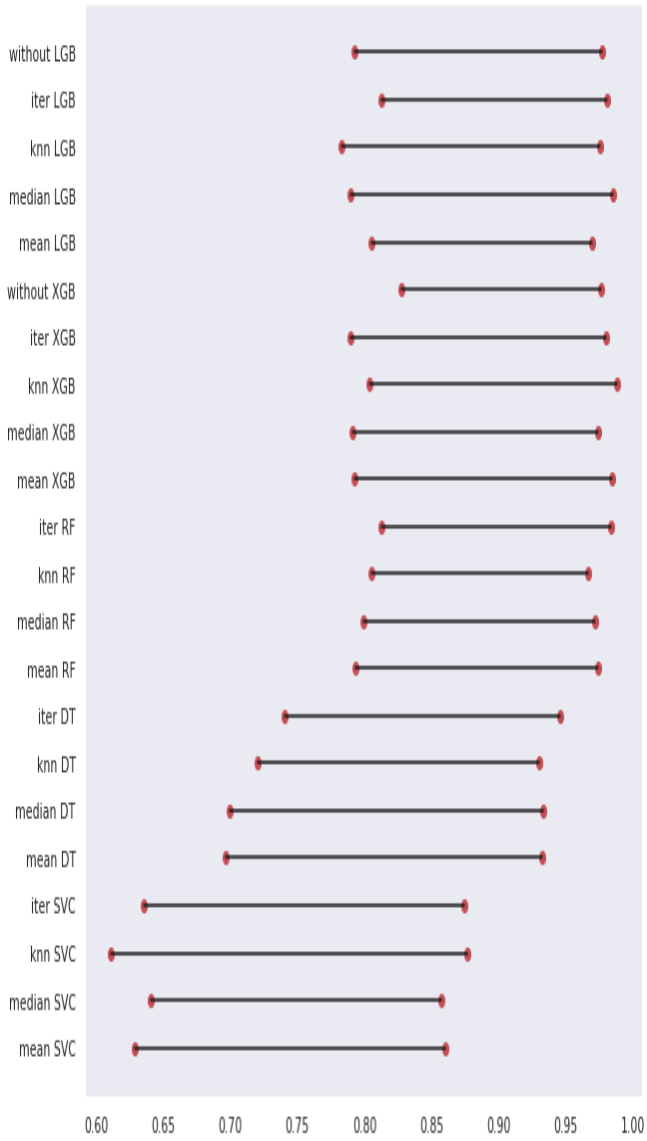

Fig. 4. The line plot results of 95% confidence interval for 5-fold nested CV model accuracy

Table 1 and Fig. 4 show the 95% confidence intervals of model accuracy based on 20 times recursion with 5-fold nested cross-validation and the dataset from IRCCS Ospedale San Raffaele Hospital with 279 patients. The dataset was separated into 80% training and 20% testing sets. Table 2 shows the result of model performance measurements, respectively, Precision, Recall (Sensitivity), Specificity, and F1.

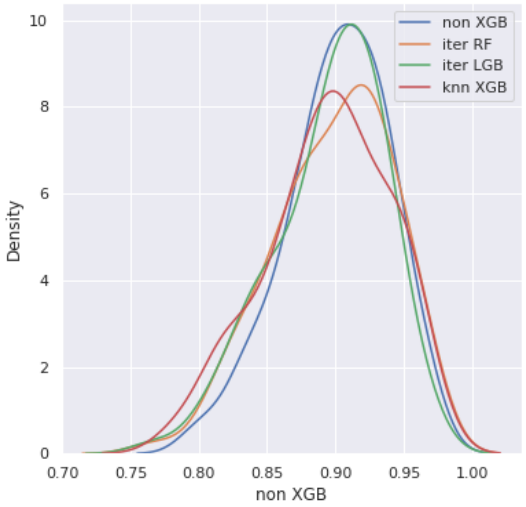

Fig. 5 The model accuracy distribution of four selected models

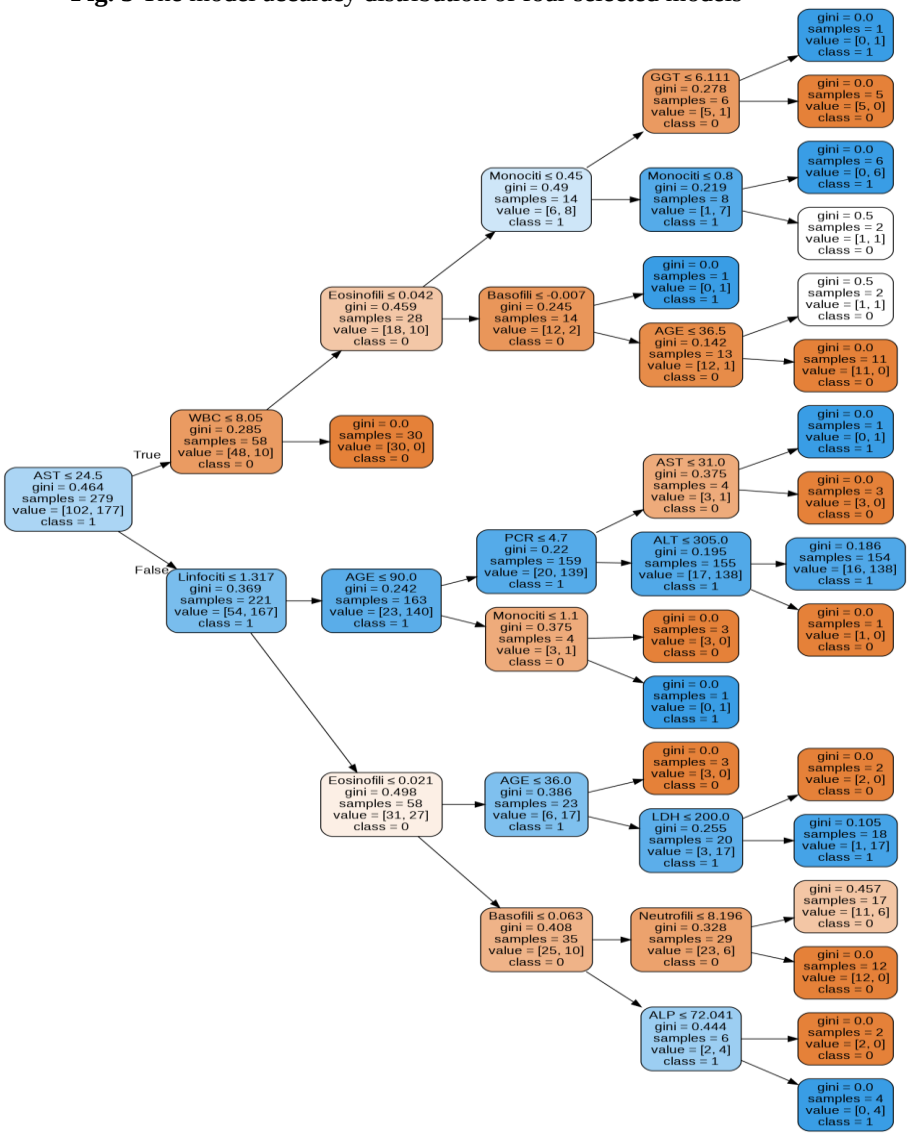

Fig. 6 The decision-making process of Decision Tree (max_depth=5)

Finally, with all 22 imputation-model combinations, we have selected four of the best models by comparing the comprehensive results from both indicators. They are iterative imputer with Random Forest (iter-RF), Knn with XGBoost (knn-XGB), iterative imputer with LightGBM (iter-LGB), and XGBoost itself (non-XGB, no imputation). The cross-validation scores are similar for each model, which is around 90%. non-XGB reported 84% recall and 94% specificity and sensitivity. In contrast, knn-XGB had a higher overall performance of 94% recall and 92% sensitivity. Similarly, iter-RF had a 92% recall and 85% sensitivity and specificity on resampled dataset. iter-LGB reported a 92% recall and 85% sensitivity.

The accuracy distribution of selected models are shown in Fig. 5. As we observed, non-XGB and iter-LGB's data distribution mainly concentrated between 87% and 93%. Fig 4. shows a more intuitive visualized confidence interval line plot distribution. The higher the minimum limit of the interval, the better the performance of the model. Therefore, we determine the four models by comparing the minimum limit of the interval, but at the same time ensure that the maximum limit is as large as possible. Then, we selected four models to visualize the data distribution. As shown in Fig. 5, the more the data is concentrated in the right area of the picture, the more accurate the model will be. As we observed, the data of non-XGB and iter-LGB are the most concentrated and tend to 1.

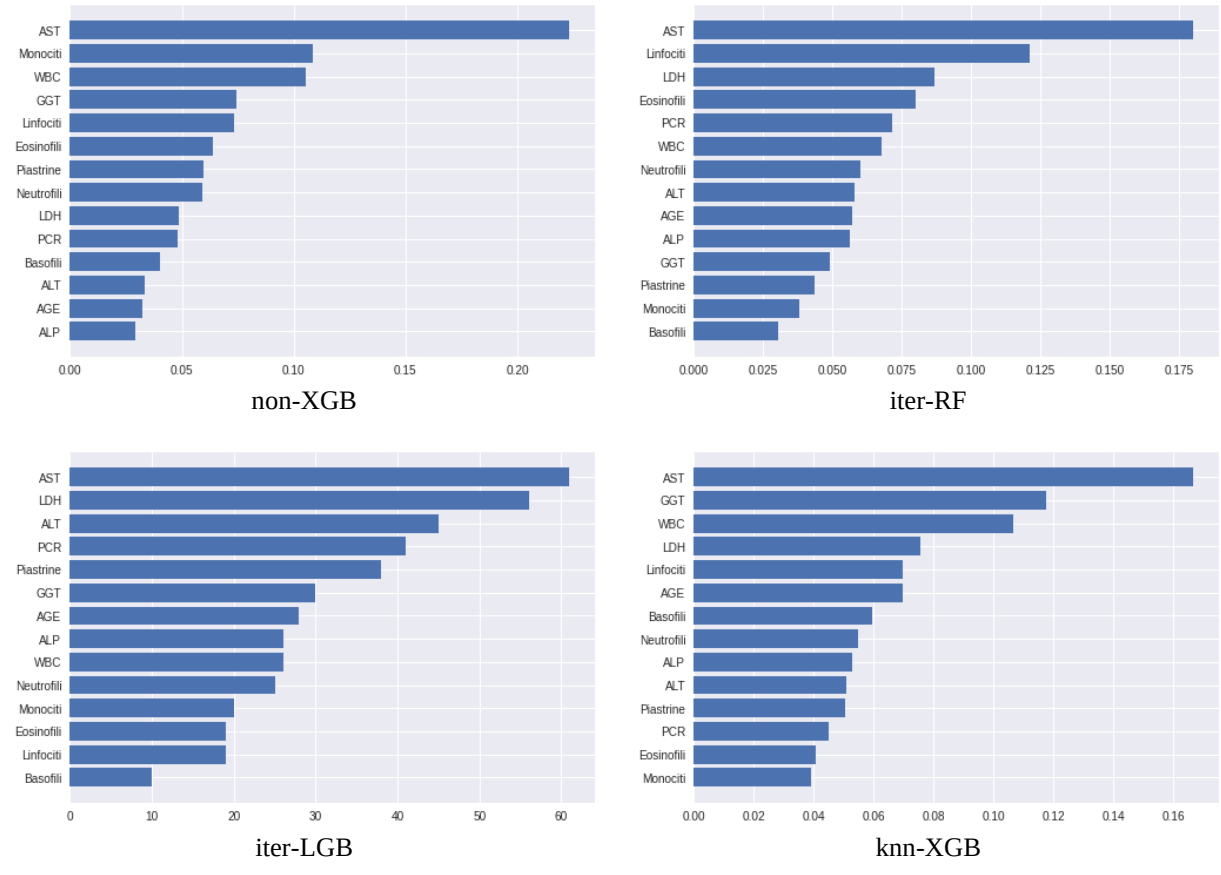

Fig. 7 Feature importance scores for the selected models

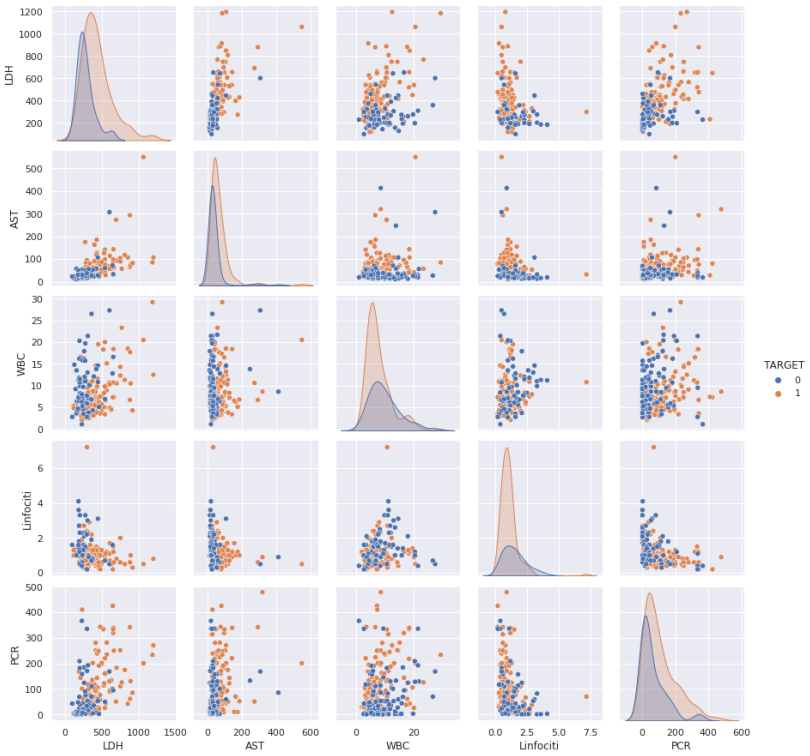

Fig. 8 The distribution plots and pairwise bivariate distributions of five selected features. Blue points represent the negative cases (0) and orange points represent the positive cases (1).

Among the four models, we first selected iter-RF as our benchmark model. Fig. 7 shows the feature importances of the iter-RF model. According to the previous procedure from the recursive feature elimination, the model is in the best performance by considering the four to five features. Therefore, Fig. 8 indicates the results of cross-comparing the feature importance of the remaining three groups of models with the first group, on the most influential five features.

Discussion/Conclusion

Four reliable machine learning models were developed to detect COVID-19 virus through the results of blood tests. The three sets of data [2, 3, 4] involved are all routine blood exam results from different regions and time periods. The results of confirmed and undiagnosed patients are all obtained by RT-PCR. In order to reduce the error of the experiment, we screened the number of features of the three sets of data. The number of features in [2, 4] are 77 and 35, respectively. After comparison and screening, we use the features in the data in [3] as the standard ones, and reduce the remaining two sets of data accordingly.

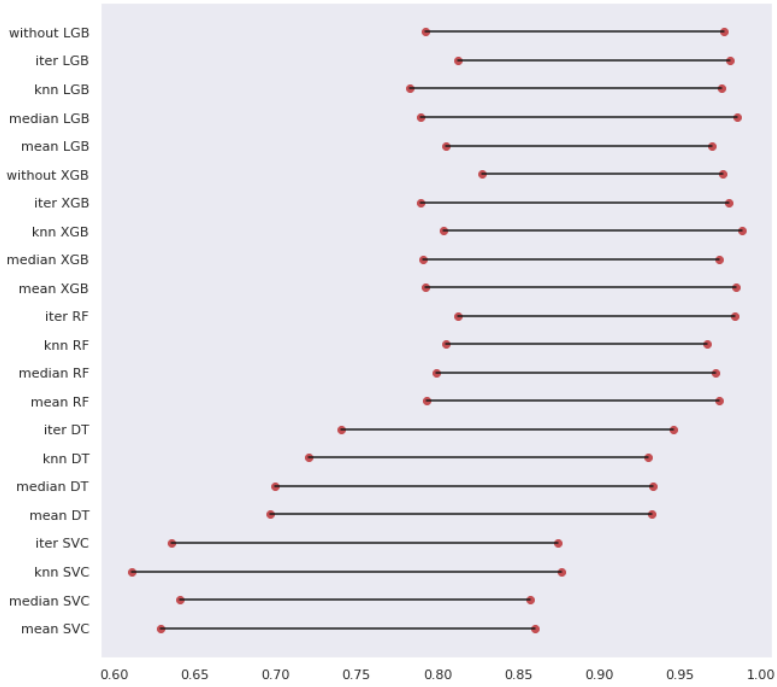

Fig. 9 The line plot results of 95% confidence interval 5-fold nested CV model accuracy

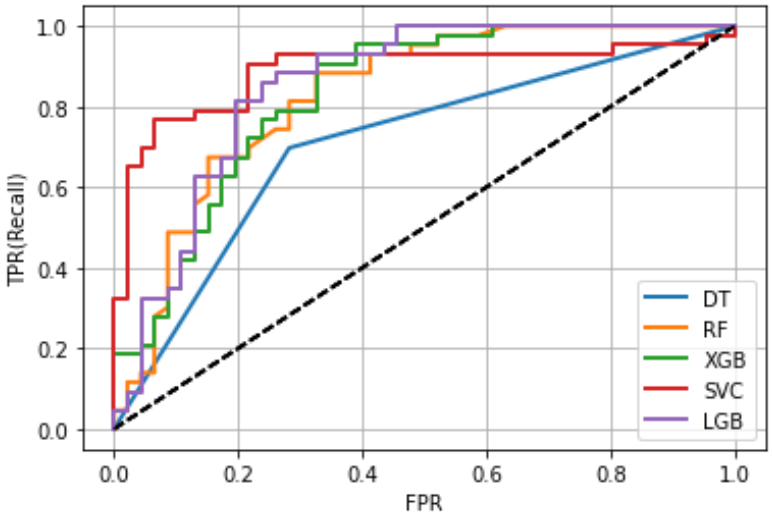

Fig. 10 ROC Curve

In this study, the priority considerations are recall and specificity. Recall, also referred to true positive rate, is used to measure the percentage of actual positives which are correctly identified. We wouldn't want to miss any of them. Therefore, according to the result of the recall among four best models, iter-RF has the best performance with 78% to 100% in 95% confidence interval. The confidence intervals for the specificity of non-XGB and iter-LGB are 80% to 100% and 77% to 100%, respectively. Non-XGB represents the highest specificity among the four groups.. Moreover, knn-XGB has the highest recall, which is 95%, and iter-LGB comes out with 93%. These are the two models that could maximize the true positive rate. By observing the precision, iter-RF provides a 97% precision ML model, which is the highest. Second best is the non-XGB model, with a score of 94%. These models could maximize the positive predictive rate

to make sure the positive result is solid. 97% specificity was provided by iter-RF and 94% provided by non-XGB, which maximizes the true negative rate.

In addition, feature importances must be considered when applying to clinical practice. As shown in Fig. 7, the top five features for selected models are similar. According to the iter-RF, the most impact biomarkers are aspartate aminotransferase (AST), lymphocyte (Linfociti), C-reactive protein (PCR), lactate dehydrogenase (LDH), and decreased lymphocyte count (WBC). For iter-LGB, those are Piastrine, AST, Linfociti, Gamma-glutamyltransferase (GGT), and AGE. In knn-XGB, those are AST, LDH, AGE, WBC, and Eosinophil (Eosinofili). For non-XGB, the most important are AST, Linfociti, WBC, LDH, and AGE. After we calculated the weighted average of each of them, the most predictive features are AST, LDH, Linfociti, PCR, and WBC. As a result, recent studies have reported up to 20% false negative results for the rRt-PCR test according to [14]. In contrast, our research reduced the false negative rates by applying different imputation methods. However, the clinical application in reality requires multiple times of testing and verifying its reliability. The study still has its limitations, but it does not prevent it from becoming an motivation to keep optimizing our method to diagnose COVID-19 infection more efficiently and accurately.

References

1. "About Serology Surveillance." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, cdc.gov/coronavirus/2019-ncov/cases-updates/about-serology-surveillance.html.

2. Cabitza, Federico, Campagner, Andrea, Ferrari, Davide, Di Resta, Chiara, Ceriotti, Daniele, Sabetta, Eleonora, Colombini, Alessandra, De Vecchi, Elena, Banfi, Giuseppe, Locatelli, Massimo and Carobene, Anna. "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests" Clinical Chemistry and Laboratory Medicine (CCLM), vol. 59, no. 2, 2021, pp. 421-431. https://doi.org/10.1515/cclm-2020-1294

3. Brinati, D., Campagner, A., Ferrari, D. et al. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst 44, 135 (2020). https://doi.org/10.1007/s10916-020-01597-4

4. Yan, L., Zhang, H., Goncalves, J. et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2, 283–288 (2020). https://doi.org/10.1038/s42256-020-0180-7

5. Introduction to Machine Learning with Python by Muller & Guido, 1st Edition, 2016

6. Gozes, O., Frid-Adar, M., Sagie, N., Zhang, H., Ji, W., and Greenspan, H., Coronavirus detection and analysis on chest ct with deep learning. arXiv:200402640, 2020

7. Mei, X., Lee, H. C., Ky, D., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P. M., Chung, M. et al., Artificial intelligence–enabled rapid diagnosis of patients with covid-19.Nature Medicine pp 1–55, 2020.

8. Mea, W., Chest x-ray findings in 636 ambulatory patients with covid-19 presenting to an urgent care center: a normal chest xray is no guarantee. The Journal of Urgent Care Medicin (2):1–9, 2020.

9. Wynants, L., Van Calster, B., Bonten, M. M., Collins, G. S., Debray, T. P., De Vos, M., Haller, M. C., Heinze, G.,Moons, K. G., Riley, R. D. et al., Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj 369, 2020.

10. sklearn.impute.IterativeImputer — scikit-learn 0.24.2 documentation

11. XGBoost Documentation — xgboost 1.5.0-SNAPSHOT documentation

12. Welcome to LightGBM's documentation! — LightGBM 3.2.1.99 documentation

13. Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

14. Kim, S., Kim, D. M., and Lee, B., Insufficient sensitivity of rna dependent rna polymerase gene of sars-cov-2 viral genome as confirmatory test using korean covid-19 cases, 2020.