Machine Learning to Predict COVID-19 and ICU Requirement

Prajoy Podder[1], M. Rubaiyat Hossain Mondal[2]
Institute of Information and Communication Technology,
Bangladesh University of Engineering and Technology (BUET), Dhaka-1205, Bangladesh.
[1]prajoypodder@gmail.com, [2]rubaiyat97@iict.buet.ac.bd

2020 11th International Conference on Electrical and Computer Engineering (ICECE) | 978-1-6654-2254-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICECE51571.2020.9393123

***Abstract*—This paper focuses on the application of machine learning (ML) algorithms to manage novel coronavirus disease (COVID-19). For this, different ML classifiers are used for two cases, one for the prediction of COVID-19 patients, and another for the prediction of the intensive care unit (ICU) requirement. A dataset of 5644 samples and 111 attributes collected at Hospital Israelita Albert Einstein, Brazil is considered in this paper. After necessary preprocessing 57 attributes are used for COVID-19 detection, while 67 attributes are considered for ICU requirement prediction. Using scikit-learn library of Python programming language, the most important features for both cases are found out. A number of base as well as ensemble classifiers are applied to the resultant datasets for the two cases. Results show that COVID-19 detection can be predicted with an accuracy of 94.39% and recall of 92% using stacking ensemble with random forest (RF), XGBoost (XGB) and logistic regression (LR). Results also show that ICU requirement can be predicted with an accuracy of 98.13% and recall of 99% using stacking ensemble with RF, extra trees and LR.**

Keywords—COVID-19, machine learning, recall, classification accuracy, feature selection.

I. Introduction

Novel coronavirus disease (COVID-19) has been first reported in China in December 2019 [1-6]. Since then this virus has spread in most parts of the world killing about one million people as of 10 October 2020. The virus can spread via respiratory droplets when people come in close contact. People are asked to use face masks and to maintain social distancing to reduce the spread of this virus. Because of the deadly nature of COVID-19, the World Health Organization (WHO) has declared this disease as a pandemic. There are several symptoms of COVID-19, including fever, dry cough, tiredness, loss of smell, sore throat, etc. However, some of the patients are asymptomatic, and some patients have only one or two symptoms [1-3]. Because of the huge number of patients, there is a burden on the hospitals and health care system in many countries. Hence, early diagnosis of this disease is important. A number of studies report the use of machine learning and deep learning for diagnosis of COVID-19 [7-12]. The study in [7] is shown to correctly classify COVID-19 patients in 85% cases using support vector machine (SVM) algorithm. The study in [7] uses a dataset [13] collected from Brazil and now available in Kaggle repository. A study in [8] uses a number of classifiers, including multilayer perceptron (MLP) and XGBoost to the same dataset [13] to diagnose COVID-19 patients with a classification accuracy of approximately 91%. A recent work [9] correctly classify COVID-19 patients in 91% cases for a dataset collected from UCLA Health System in Los Angeles, USA. The work in [11] and [12] predict COVID-19 with accuracies of 91% and 89%, respectively. Furthermore, a prediction is done for the requirement of ICU/semi-ICU successfully in 98% cases [10].

This paper focuses on improving the accuracy in predicting both COVID-19 and ICU/semi-ICU requirement. This paper applies a number of classifiers namely random forest (RF), MLP, light gradient boosting machine (LGBM), naïve Bayes (NB), as well as two ensemble algorithms namely stacking and majority voting. The main contributions of this work are to apply different ensemble algorithms to predict COVID-19 and ICU/semi-ICU requirement with an accuracy greater than that described in the literature. The rest of the paper is organized as follows. Section II discusses the dataset preprocessing and feature selection performed for both cases. Section III describes the performance results when different classifiers are applied to the dataset to predict COVID-19 as well as ICU/semi-ICU requirement. Finally, Section IV provides concluding remarks.

II. Preprocessing and Feature Selection

This dataset [13] is created at the Hospital Israelita Albert Einstein, at São Paulo, Brazil. There are 5644 number of samples or suspected patients with 111 attributes. This dataset is used for two classification tasks. In the first classification, experiments are performed to classify the normal and COVID-19 patients from the suspected patients in the dataset. In the second classification, experiments are done to classify the need for general ward and ICU/semi-ICU. For simplicity, we term the ICU/semi-ICU requirement as only ICU requirement in the rest of the paper. Experiments are carried out using Scikit-learn library of Python programming language.

The case of COVID-19 prediction is considered first. The dataset is imbalanced as 90.10% samples are for negative cases representing people without COVID-19. There were many records with several missing values. Features that have null values in more than 99.80% positive cases are dropped. The resultant dataset has 1091 records and 57 attributes with 49.80% positive cases and 50.20% negative cases indicating a balanced dataset. The output or target attribute is converted to numerical value where 1 means positive case and 0 means negative case. Next feature selection is performed using univariate feature selection algorithm to find the most important attributes of COVID-19 detection. Fig. 1 shows the horizontal bar graph of the important features for COVID-19 prediction. It can be seen that the most important feature is Proteina C reativa MG/DL. This is followed by Leukocytes and Lymphocytes.

978-1-6654-2254-3/20/$31.00 ©2020 IEEE

Next, the case of ICU requirement prediction is considered. Since the dataset has a number of missing values, a number of columns that have null values of 99% or more are removed. After this process, the resultant dataset has 5644 records but only 67 attributes. The most important features are selected using feature importance approach with ExtraTrees classifier. Fig. 2 presents the horizontal bar graph of the important features for ICU prediction. It can be seen that the most important feature is Proteina C reativa mg/dl. This is followed by age_quantile and SARS-CoV-2-exam result.

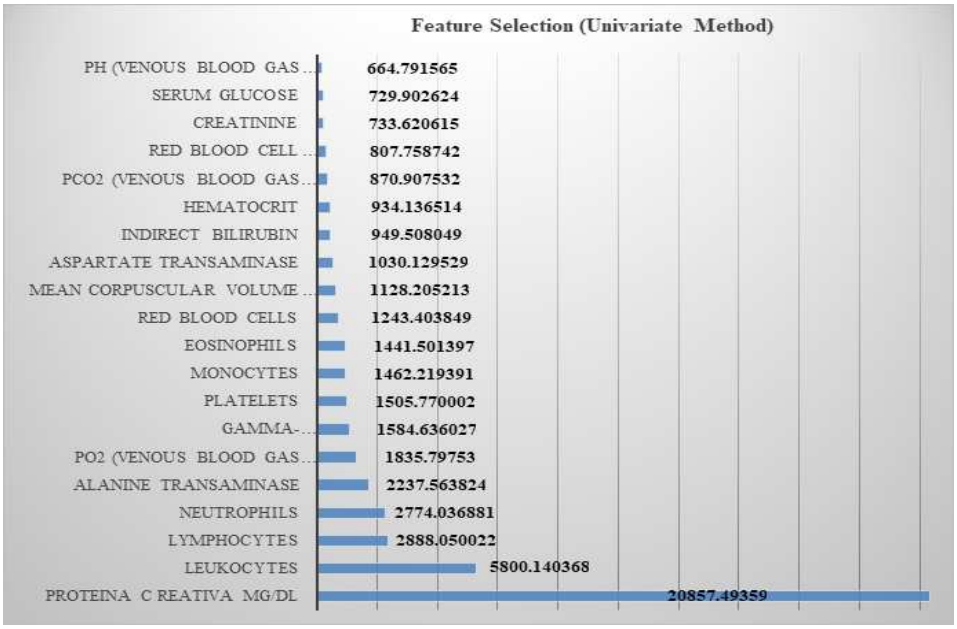

Fig. 1: Best 20 features for COVID-19 prediction

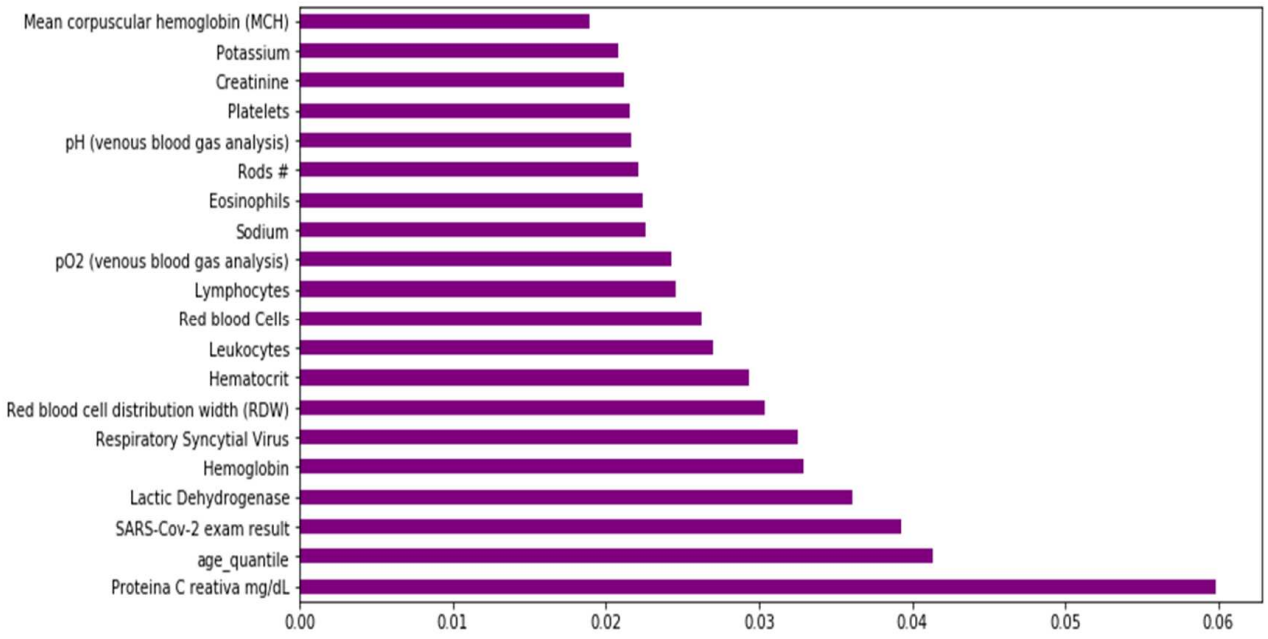

Fig. 2. Best 20 features for ICU requirement prediction

III. Results and Discussion

This section provides the results of the application of different classification algorithm to the dataset. In the experiments, 10-fold cross-validation method is used to split the data into training and testing portions. The case of COVID-19 prediction is considered first. Table 1 shows the elements of confusion matrix for the case of COVID-19 prediction. In this case, TP, FP, TN and FN indicate true positive, false positive, true negative and false negative, respectively [14]. Table 2 presents the performance results of different classifiers in terms of precision, recall, specificity, F1 score, accuracy, and area under the receiver operating characteristics curve (AUC). Results show that COVID-19 detection can be predicted with an accuracy of 94.39% and recall of 92% using stacking ensemble with

random forest (RF), XGBoost (XGB) and logistic regression (LR). In this case, a precision and F1 score of 91% and an AUC value of 90% are achieved. The second highest accuracy of 92.78% is achieved when stacking is applied with NB, LGBM, and LR. Majority voting (hard) achieves the third-highest accuracy of 92.47% using base classifiers of extra trees, RF and LGBM. Now the case of ICU prediction is considered. Table 3 shows the performance results of different classifiers for ICU prediction. Results also show that ICU requirement can be predicted with an accuracy of 98.13% and recall of 99% using stacking ensemble with RF, extra trees and LR. In this case, a precision of 99%, F1 score of 98% and an AUC value of 97% are achieved. Table 3 also shows that a high AUC value of 99%, an accuracy of 97.71% and a recall of 98% can be achieved for majority voting algorithm when used with Extra Trees, RF and LGBM.

Next, the results of this work are compared to that of the literature. For the case of COVID-19 detection, we reliably classify COVID-19 patients and normal people with an accuracy of 94.39% and a recall value of 92% which is better than the highest 91% accuracy reported in the literature [7-12]. The appropriate choice of stacking classifier enables us to outperform existing results in COVID-19 detection. We also classify the need for general ward or ICU admission at an accuracy of 98.13%, an AUC of 97% and a recall of 99% using stacking algorithm, and we obtain an AUC of 99% and a recall of 98% using majority voting classifier. These results are better than the work in [10] having an AUC of 98% and recall value of only 82%.

Table 1: Elements of confusion matrix for COVID-19 prediction

Classifier	**TP**	**FP**	**TN**	**FN**
RF	124	5	119	25
MLP	138	17	107	11
LGBM	131	8	116	18
Naïve Bayes	126	5	119	23
Extra Trees Classifier	123	6	119	25
Stacking 1 (RF, XGB, LR)	129	6	118	20
Stacking 2 (NB, LGBM, LR)	130	6	118	19
Voting 1 (Extra Trees, RF, LGBM) Hard	125	5	119	24
Voting 1 (Extra Trees, RF, LGBM) Soft	126	7	117	23
Voting 2 (MLP, NB, LGBM) Hard	129	6	118	20
Voting 2 (MLP, NB, LGBM) Soft	128	7	117	21

Table 2: Classification results for COVID-19 prediction

Classifier	**Precision**	**Recall**	**Specificity**	**F1 score**	**Accuracy**	**AUC**
RF	90%	89%	95.9677%	89%	91.10%	89%
MLP	90%	91%	86.2903%	91%	90.55%	90%
LGBM	91%	90%	93.5484%	90%	91.29%	90%
Naïve Bayes	91%	90%	95.9677%	90%	90.95%	90%
Extra Tress Classifier	90%	89%	95.9677%	89%	91.56%	89%
Stacking 1 (RF, XGB, LR)	**91%**	**92%**	**95.1613%**	**91%**	**94.39%**	**90%**
Stacking 2 (NB, LGBM, LR)	91%	91%	95.1613%	91%	92.78%	91%
Voting 1 (Extra Trees, RF, LGBM) Hard	90%	90%	95.9677%	89%	92.47%	90%
Voting 1 (RF, LR, SVM) Soft	90%	89%	94.3548%	89%	91.56%	89%
Voting 2 (MLP, NB, LGBM) Hard	91%	90%	95.1613%	90%	91.95%	91%
Voting 2 (MLP, NB, LGBM) Soft	90%	90%	94.3548%	90%	91.65%	90%

Table 3: Classification results for ICU prediction

Classifier	Precision	Recall	F1 score	Accuracy	AUC
RF	98%	98%	98%	97.66%	99%
MLP	97%	98%	97%	97.59%	95%
LGBM	98%	98%	98%	97.18%	88%
Naïve Bayes	98%	94%	96%	93.16%	64%
Extra Tress Classifier	98%	97%	98%	97.94%	91%
Stacking 1 (RF, Extra Trees, LR)	**99%**	**99%**	**98%**	**98.13%**	**97%**
Stacking 2 (NB, LGBM, LR)	98%	99%	98%	97.95%	88%
Voting 1 (Extra Trees, RF, LGBM) Hard	98%	98%	98%	97.71%	99%
Voting 1 (RF, LR, SVM) Soft	98%	97%	96%	97.41%	99%
Voting 2 (MLP, NB, LGBM) Hard	98%	98%	98%	97.39%	85%
Voting 2 (MLP, NB, LGBM) Soft	97%	98%	97%	97.34%	84%

IV. CONCLUSION

This work uses ML classifiers to predict COVID-19 and ICU requirement. A dataset of 5644 samples and 111 attributes collected at a hospital in Brazil is considered in this paper. After necessary preprocessing 57 attributes are used for COVID-19 detection, while 67 attributes are considered for ICU requirement prediction. For the case of COVID-19 prediction, the three most important attributes are found to be Proteina C reativa MG/DL, Leukocytes and Lymphocytes. On the other hand, the three most important features for ICU requirement prediction are Proteina C reativa MG/DL, age_quantile and SARS-CoV-2-exam result. A number of base as well as ensemble classifiers are applied to the resultant datasets for the two cases. Stacking ensemble classifier and majority voting classifier are appropriately formed using different combinations of base classifiers. Results show that COVID-19 detection can be predicted with an accuracy of 94.39% and recall of 92% using stacking ensemble with random forest (RF), XGBoost (XGB) and logistic regression (LR). Results also show that ICU requirement can be predicted with an accuracy of 98.13% and recall of 99% using stacking ensemble with RF, extra trees and LR. The results indicate that using appropriate stacking algorithms, we can classify COVID-19 patients and normal patients, as well as we can classify whether suspected patients require general ward or ICU admission. Note that the results of the classifiers depend on the datasets and the attributes of the dataset. In future, the effectiveness of these classifiers should be validated for other datasets.

REFERENCES

[1] F.-Y. Lan, C.-F. Wei, Y.-T. Hsu, D.C. Christiani, S.N. Kales, "Work-related COVID-19 transmission in six Asian countries/areas: a follow-up study", *Plos One*, 15 (5) (2020), Article e0233588.

[2] D. Fisher, A. Wilder-Smith, "The global community needs to swiftly ramp up the response to contain COVID-19". *The Lancet* 2020; S0140-6736(20)30679-6.

[3] S. Sanche, Y.T. Lin, C. Xu, E. Romero-Severson, N.W. Hengartner, R. Ke, "The novel Coronavirus 2019-nCoV is highly contagious and more infectious than initially estimated", arXiv.

[4] M. Gilbert, G. Pullano, F. Pinotti, et al. "Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study". *The Lancet* 2020, 395(10227):871-877.

[5] H. Nishiura, "The extent of transmission of novel Coronavirus in Wuhan, China", *J Clin Med*, vol. 9, no. 2, 2020, p. 330.

[6] F. Khanam, I. Nowrin, and M. R. H. Mondal, "Data Visualization and Analyzation of COVID-19", *Journal of Scientific Research and Reports*, vol. 26, no. 3, pp. 42-52, Apr. 2020.

[7] A. F. M. Batista, J. L. Miraglia, T. H. R. Donato, A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach", medRxiv, 2020.

[8] M. R. H. Mondal, S. Bharati, P. Podder, P. Podder, "Data analytics for novel coronavirus disease", *Informatics in Medicine Unlocked*, Elsevier, vol. 20, 2020, 100374.

[9] D. Goodman-Meza, A. Rudas, J. N. Chiang, P. C. Adamson, J. Ebinger, et al., "A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity", PLOS ONE 15(9), 2020: e0239474.

[10] P. Schwab, A. D. Schütte, B. Dietz, S. Bauer, "Clinical predictive models for COVID-19: systematic study", *J Med Internet Res* 2020; 22(10):e21439, DOI: 10.2196/21439.

[11] Y. Sun, V. Koh, K. Marimuthu, O. T. Ng, B. Young, S. Vasoo, M. Chan, et al., "Epidemiological and clinical predictors of COVID-19", *Clin Infect Dis.* 2020 Jul 28; 71(15): 786-792.

[12] Z. Meng, M. Wang, H. Song, S. Guo, Y. Zhou, W. Li, et al., "Development and utilization of an intelligent application for aiding COVID-19 diagnosis". medRxiv. 2020..

[13] https://www.kaggle.com/einsteindata4u/covid19, last accessed on 15 October 2020

[14] M. Raihan-Al-Masud, M. R. H. Mondal. "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms". PLOS ONE. 2020; 15(2): e0228422.