RESEARCH ARTICLE OPEN ACCESS

A Novel Machine Learning Based Screening Method for High-Risk Covid-19 Patients Based on Simple Blood Exams

Mrs R. Jhansi Rani MCA [1], B. Karunakar Reddy [2]

[1] Asst. Professor, Department Computer Application
[2] Student, Department of Computer Application
[1], [2] Chadalawada Ramanamma Engineering College (Autonomous)

ABSTRACT

This paper presents a predictive model to potentially identify high-risk COVID-19 infected patients based on easily analyzed circulatory blood markers. These findings can enable effective and efficient care programs for high-risk patients and periodic monitoring for the low-risk ones, thereby easing the hospital flow of patients and can further be utilized for hospital bed utilization assessment. The present machine learning-based SV-LAR model results in a high 87% f1 score, harmonic mean of 91% precision, and 83% recall to classify COVID-19, infected patients, as high-risk patients needing hospitalization.

Keywords: - The present machine learning-based SV-LAR model results in a high 87% f1 score, harmonic mean of 91% precision, and 83% recall to classify COVID-19, infected patients, as high-risk patients needing hospitalization.

I. INTRODUCTION

COVID-19 is a new disease for which effective treatment is still awaited. It was declared a pandemic by World Health Organization (WHO) on March 11, 2020. As of January 28, 2021, more than 100 million people have been affected by this infection causing more than 2 million fatalities. Global health care now faces unprecedented challenges with the widespread and rapid human-to-human transmission of SARS-CoV-2 and high morbidity and mortality with COVID-19 worldwide.

COVID-19 patients get worse quickly and aggressively. In addition to high transmissibility SARS-CoV-2 infection it is also characterized by fever, dry cough, weakness, headache, dyspnoea, and loss of smell and taste in the early stages, which are common symptom of cold and flu. The early onset of common symptoms can rapidly change to acute respiratory distress syndrome (ARDS), acute cardiac injury, cytokine storm, coagulation dysfunction, and multi-organ failure if the disease is not resolved, resulting in patient death. Early studies showed that COVID-19 patients with comorbidity may lead to poor prognosis, increasing the risk of severe illness from COVID-19. Among laboratory confirmed cases, patients with any comorbidity yielded poorer clinical outcomes than those without. Several studies have been conducted to find a correlation between pre-existing medical conditions and their impact on COVID 19 prognosis.

In a meta-analysis by Wang et al, Hypertension, diabetes, chronic obstructive pulmonary disease (COPD), cardiovascular disease, and cerebrovascular disease were found to be the major risk factors for patients with COVID-19. Several risk factors that led to the progression of COVID-19 pneumonia were identified, including age, history of smoking, maximum body temperature at admission, respiratory failure, albumin, and C-reactive protein. Given the virtually unstoppable global trend of SARS-CoV-2, together with the high prevalence of comorbidities worldwide, the combination of these two conditions poses greater clinical, societal, and economic burdens to healthcare systems.

Until now the source of the pathogenesis of the COVID-19 remains unclear, and no specific treatment has been recommended for coronavirus infection except for meticulous care. The world is ready to receive the vaccines as approved worldwide, but the threat continues with mutating strains of the virus. Therefore, the need for a better solution for providing care to those who absolutely need it and to predict the future requirements for better planning and management for better patient outcomes, continues. In

several articles, researches have indicated the need for better hospital management by early identification of patients requiring hospitalization and possible further triage.

In another study aimed to clarify high-risk factors for COVID-19, researchers used Multivariate Cox regression to identify risk factors associated with the progression of the disease. Univariate and multivariate analyses showed that comorbidity, older age, lower lymphocyte count, and higher lactate dehydrogenase at presentation were independent high-risk factors for progression. A novel scoring model, named CALL, with an area under the receiver operating characteristic curve (ROC) of 0.91 (95% CI, .86–.94) was established to help clinicians better choose a therapeutic strategy.

In another statistical analysis regarding the associations between increased cardiac injury markers and the risk of 28- day-all-cause death of COVID-19 patients in the Chinese population, the 5 myocardial biomarkers (high-sensitivity cardiac troponin I, creatine phosphokinase)-MB, N-terminal pro-B-type natriuretic peptide, creatine phosphokinase, and myoglobin) were found to be significantly prognostic of COVID-19 mortality

Despite several initiatives aimed at containing the spread of the disease, countries are faced with unmanageable increases in the demand for hospitalization and ICU beds. The health care system globally, has been stressed and stretched to its limit. In order to help in patient triage, several attempts have been made to discover early predictors of COVID-19 disease progression and spread. Identification of such factors that predict complications of COVID-19 is pivotal for guiding clinical care, improving patient outcomes, and allocating scarce resources effectively in a pandemic. Medical resource allocation assessments should be based on a risk/benefit approach considering the intensity of transmission, the health system's capacity to respond, other contextual considerations (such as upcoming events which may alter transmission or capacity) and the overall strategic approach to responding to COVID-19 in each specific setting.

We think that it requires agile decision-making based on ongoing situational assessments at the most local administrative level possible. We propose a predictive machine learning model that identifies a potential high-risk patient from the COVID-19 patient population based on blood based circulatory markers. These predictions would help the administrators to make provisions for the scarce 'hospital beds. Consequently, the model can help in providing better public health and social measures to alleviate patient care during the pandemic time thereby improving patient outcomes at large.

II. LITERATURE SURVEY

A. Shander et al., "Essential role of Patient Blood Management in a pandemic: A Call for Action: A call for action," Anesth. Analg., vol. 131, no. 1, pp. 74–85, 2020.

The World Health Organization (WHO) has declared coronavirus disease 2019 (COVID-19), the disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a pandemic. Global health care now faces unprecedented challenges with widespread and rapid human-to-human transmission of SARS-CoV-2 and high morbidity and mortality with COVID-19 worldwide. Across the world, medical care is hampered by a critical shortage of not only hand sanitizers, personal protective equipment, ventilators, and hospital beds, but also impediments to the blood supply. Blood donation centers in many areas around the globe have mostly closed. Donors, practicing social distancing, some either with illness or undergoing self-quarantine, are quickly diminishing. Drastic public health initiatives have focused on containment and "flattening the curve" while invaluable resources are being depleted. In some countries, the point has been reached at which the demand for such resources, including donor blood, outstrips the supply. Questions as to the safety of blood persist. Although it does not appear very likely that the virus can be transmitted through allogeneic blood transfusion, this still remains to be fully determined. As options dwindle, we must enact regional and national shortage plans worldwide and more vitally disseminate the knowledge of and immediately implement patient blood management (PBM). PBM is an evidence-based bundle of care to optimize medical and surgical patient outcomes by clinically managing and preserving a patient's own blood. This multinational and diverse group of authors

issue this "Call to Action" underscoring "The Essential Role of Patient Blood Management in the Management of Pandemics" and urging all stakeholders and providers to implement the practical and common-sense principles of PBM and its multiprofessional and multimodality approaches.

CDC, "COVID-19 and Your Health," Cdc.gov, 03-Feb-2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019

The Delta variant causes more infections and spreads faster than earlier forms of the virus that causes COVID-19. It might cause more severe illness than previous strains in unvaccinated people. Vaccines continue to reduce a person's risk of contracting the virus that cause COVID-19, including this variant. Vaccines continue to be highly effective at preventing hospitalization and death, including against this variant. Fully vaccinated people with breakthrough infections from this variant appear to be infectious for a shorter period. Get vaccinated and wear masks indoors in public spaces to reduce the spread of this variant.

Silva, and D. L. Guidoni, "Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data," bioRxiv, p. 2020.06.26.20140764, 2020.

The first officially registered case of COVID-19 in Brazil was on February 26, 2020. Since then, the situation has worsened with more than 672, 000 confirmed cases and at least 36, 000 reported deaths by June 2020. Accurate diagnosis of patients with COVID-19 is extremely important to offer adequate treatment, and avoid overloading the healthcare system. Characteristics of patients such as age, comorbidities and varied clinical symptoms can help in classifying the level of infection severity, predict the disease outcome and the need for hospitalization. Here, we present a study to predict a poor prognosis in positive COVID-19 patients and possible outcomes using machine learning. The study dataset comprises information of 8, 443 patients concerning closed cases due to cure or death. Our experimental results show the disease outcome can be predicted with a Receiver Operating Characteristic AUC of 0.92, Sensitivity of 0.88 and Specificity of 0.82 for the best prediction model. This is a preliminary retrospective study which can be improved with the inclusion of further data. Conclusion: Machine learning techniques fed with demographic and clinical data along with comorbidities of the patients can assist in the prognostic prediction and physician decision-making, allowing a faster response and contributing to the non-overload of healthcare systems.

W.-J. Guan et al., "Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis," Eur. Respir. J., vol. 55, no. 5, p. 2000547, 2020

We analysed data from 1590 laboratory confirmed hospitalised patients from 575 hospitals in 31 provinces/autonomous regions/provincial municipalities across mainland China between 11 December 2019 and 31 January 2020. We analysed the composite end-points, which consisted of admission to an intensive care unit, invasive ventilation or death. The risk of reaching the composite end-points was compared according to the presence and number of comorbidities. The mean age was 48.9 years and 686 (42.7%) patients were female. Severe cases accounted for 16.0% of the study population. 131 (8.2%) patients reached the composite end-points. 399 (25.1%) reported having at least one comorbidity. The most prevalent comorbidity was hypertension (16.9%), followed by diabetes (8.2%). 130 (8.2%) patients reported having two or more comorbidities. After adjusting for age and smoking status, COPD (HR (95% CI) 2.681 (1.424–5.048)), diabetes (1.59 (1.03–2.45)), hypertension (1.58 (1.07–2.32)) and malignancy (3.50 (1.60–7.64)) were risk factors of reaching the composite end-points. The hazard ratio (95% CI) was 1.79 (1.16–2.77) among patients with at least one comorbidity and 2.59 (1.61–4.17) among patients with two or more comorbidities.

III. SYSTEM ANALYSIS & FEASIBILITY STUDY

Existing Method:

The increasing growth of machine learning, computer techniques divided into traditional methods and machine learning methods. This section describes the related works of how covid-19 testing and are classified how machine learning methods are better than traditional methods. The existing method in this

project have a certain flow and also normal sentimental analysis is used for development. But it requires large memory and result is not accurate.

Disadvantages:

- Low Accuracy
- High complexity.
- Highly inefficient.
- Requires skilled persons

Proposed System:

We propose this application that can be considered a useful system since it helps to reduce the limitations obtained from traditional and other existing methods. The objective of this study to develop fast and reliable method which detects the sentiment accurately. To design this system is we used a powerful algorithm svm in a based Python environment.

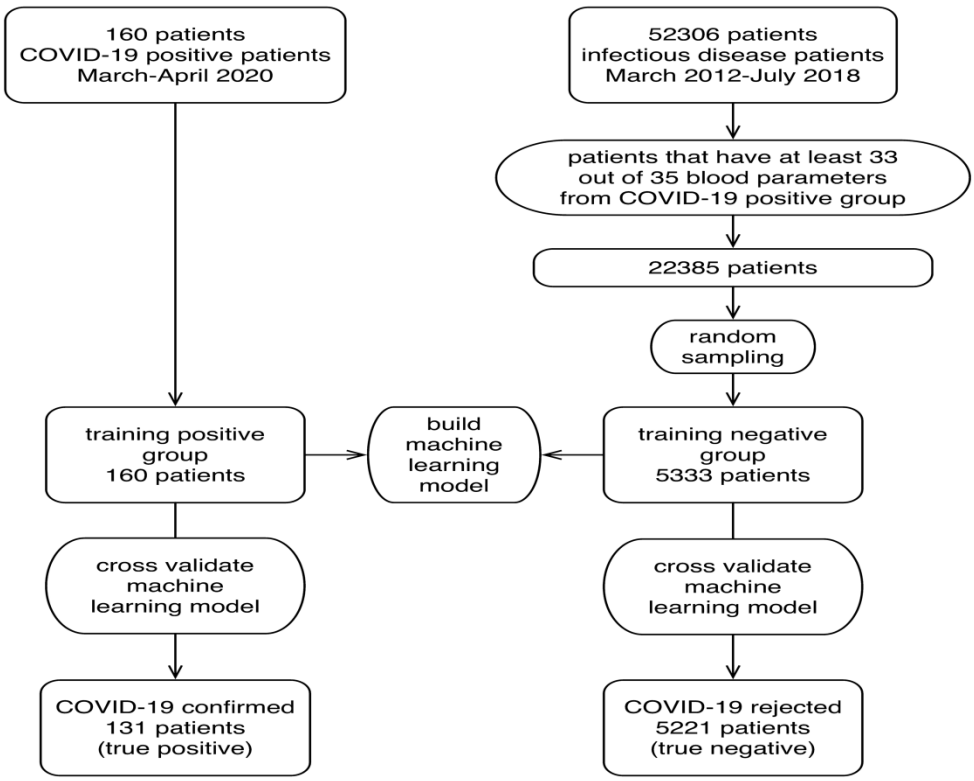
Advantages:

- Accuracy is good.
- Low complexity.
- Highly efficient.
- No need of skilled persons

IV. MATERIALS AND METHODS

- The Dataset

  We have used data published on a public forum , that of Hospital Israelita Albert Einstein, at São Paulo, Brazil [19]. The dataset contains records of patients that were tested for COVID-19 using SARS-CoV-2 Reverse transcription polymerase chain reaction (RT-PCR) and additional blood tests between the 28th of March 2020 and 3rd of April 2020. All data were anonymized following the best international practices and recommendations. The full dataset released included 5,644 individual patients' clinical test results that were standardized to have a mean of zero and a unit standard deviation. It provided information of patient hospitalization into three types of wards in the hospital, such as regular ward, semi-intensive care unit, and intensive care unit as depicted in Fig 1. The information of patients admission to various wards in the hospital was used to create the target variable for the current problem statement. Hospitalization is needed by patients needing extra care and monitoring due to health who despite being infected do not need hospital admission constitute the low-risk patient population. This formed the basis of the binary classification and the target label for classification of patients {needing hospitalization in any of the hospital wards = 1, no hospitalization needed = 0} for the current objective. As the current hypothesis is set around blood analysis, we have carefully selected features of routine blood analysis only. Parameter related to patient age was not considered in order to avoid any age related bias in the present analysis. Tests pertaining to viral or bacterial infections other than SARS-CoV-2 were also dropped. It is our objective to find blood based markers to identify high-risk patients and therefore features related to routine urine analysis were also dropped from the current model. Blood gas analysis either on venous blood or arterial blood is also not included in the current analysis. Largely because the blood samples are required to be tested in a 30 minute window or need a cold supply chain [19]. It is our intent to find markers that eases the hospital workload during the pandemic and therefore it is counterintuitive to include tests that need immediate attention and hospital setting to give good results. It is for this reason that the working dataset for the model building exercise includes test parameters form a simple blood workup, keeping in mind that the sample could be collected form patients' home environment and not necessarily in the hospital setting. Fig 3 presents the frequency of each test performed amongst the blood analysis related parameters considered for the model building, in the select dataset of 558 patient records tested positive for SARS-CoV-2.



- Model building
  We started model building after data processing. The working data had missing values and columns with all null values. Features with more than 95% missing values were dropped and remaining missing values were imputed with the mean. We started with simple linear algorithms such as logistic regression, ridge classification and elastic net, and moved on to non-parametric algorithm such as Nearest Neighbours Classifier and Gaussian Naive Bayes classification. We also used tree based algorithms like decision tree classifier and extra tree classifier. Multiple ensemble techniques like random forest classifier, bagging classifier, adaboost classifier were also used to model the target label with select features of the prepared dataset. We used scikit learn library of machine learning algorithms [20]. Based on our findings, we propose the SV-LAR model for our 2-class (SARS-CoV-2 positive induced hospitalization or not) classification. The proposed model uses voting classifier ensemble based on logistic regression, random forest and adaboost classifier. The working dataset also has class imbalance. Only 10 % labels of the working dataset are positive class of hospitalised COVID positive patients, in the three available hospital wards. We have used SMOTE (Synthetic Minority Oversampling Technique) on the training data to deal with the class imbalance by upsampling the positive class [21].
- Model performance measures
  The performance of the model is expressed in terms F1 score, precision and recall. As we attempt imbalanced classification problem F1 score metric becomes more relevant. It is a measurement that considers both precision and recall to compute the score that can be interpreted as a weighted average of the precision and recall values. High F1 score (closer to 1.0) is desirable in our model. Precision is determined by the number of correctly labeled annotations divided by the total number of annotations added by the machine-learning annotator. It indicates how accurately the model has labelled the two classes. Another metric, recall specifies how many mentions that should have been annotated by a given label were actually annotated with that label. A recall score of 1.0 means that every mention that should

have been labeled as entity type A was labeled correctly. In this imbalanced healthcare dataset high scores for precision and recall are desirable for an ultimate high f1 score [23, 24].

V. CONCLUSION

By the simple intervention of machine learning model (SVLAR) with f1-score of 87%, the identification of a potential high-risk patient can be performed easily. The differential costs of tests required to the prediction is also underwhelming. SV-LAR model can be used to benefit healthcare workers by identifying about 10% high-risk COVID-19 patients from the increasing COVID-19 patient population. The precision of the model is a high 91% and 83% recall for the positive class. Simply put, 83 out of 100 high-risk patients can be identified correctly using this model and can be taken into hospital care for further treatment. SV-LAR model is potentially fastest way to triage COVID19 patients into high-risk and low-risk groups. Not only that, it enables to monitor patients via simple, non-expensive, quick, robust and minimally-invasive blood analysis. The identified high-risk patient population can then be put through more tests and procedures and can be treated accordingly. The low-risk patients, on the other hand, can be remotely monitored for any change in the patients' prognosis via the same model. Our proposed model can be utilised globally. It relies on basic blood analysis, which is the most simple and established diagnostic service readily available in the healthcare system of any nation. This enables improved management of pandemic agnostic of the socio-economic standing of the nation. An additional positive impact is related to hospital and patient flow management. It allows patient journey to be monitored from a distance providing better isolation of COVID-19 patients. Given that blood samples could be drawn periodically and analysed away from hospital emergency rooms, it allows ERs to work more efficiently despite the pandemic. A limitation of our study however, is that not every patient hospitalized needs ICU. Due to the lack of data we could not segment the ICU needing patients from hospitalization requirements. As more data is collected and made available we can further refine the model. We believe that as more data will be incorporated in the model its performance and reliability will increase.

REFERENCES

[1] "Archived: WHO Timeline - COVID-19," Who.int. [Online]. Available: https://www.who.int/news/item/27-04-2020-who-timeline- --covid-19. [Accessed: 14-Feb-2021].

[2] "COVID-19 map - johns Hopkins Coronavirus resource Center," Jhu.edu. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 14-Feb-2021].

[3] A. Shander et al., "Essential role of Patient Blood Management in a pandemic: A Call for Action: A call for action," Anesth. Analg., vol. 131, no. 1, pp. 74–85, 2020.

[4] CDC, "COVID-19 and Your Health," Cdc.gov, 03-Feb-2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019- ncov/need-extra-precautions/people-with-medical-conditions.html. [Accessed: 14-Feb-2021].

[5] Silva, and D. L. Guidoni, "Predicting the disease outcome in COVID19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data," bioRxiv, p. 2020.06.26.20140764, 2020.

[6] W.-J. Guan et al., "Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis," Eur. Respir. J., vol. 55, no. 5, p. 2000547, 2020

[7] B. Wang, R. Li, Z. Lu, and Y. Huang, "Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis," Aging (Albany NY), vol. 12, no. 7, pp. 6049–6057, 2020.

[8] W. Liu et al., "Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease," Chin. Med. J. (Engl.), vol. 133, no. 9, pp. 1032–1038, 2020.

[9] Y. Zhou et al., "Comorbidities and the risk of severe or fatal outcomes associated with coronavirus disease 2019: A systematic review and meta-analysis," Int. J. Infect. Dis., vol. 99, pp. 47–56, 2020.

[10] Swiss Society Of Intensive Care Medicine, "Recommendations for the admission of patients with COVID-19 to intensive care and intermediate care units (ICUs and IMCUs)," Swiss Med. Wkly, vol. 150, no. 1314, p. w20227, 2020.

[11] Z. Zhao et al., "Prediction model and risk scores of ICU admission and mortality in COVID-19," PLoS One, vol. 15, no. 7, p. e0236618, 2020.
[12] D. Ji et al., "Prediction for progression risk in patients with COVID19 pneumonia: The CALL score," Clin. Infect. Dis., vol. 71, no. 6, pp. 1393–1399, 2020.
[13] E. Grifoni et al., "The CALL score for predicting outcomes in patients with COVID-19," Clin. Infect. Dis., vol. 72, no. 1, pp. 182–183, 2021.
[14] J.-J. Qin et al., "Redefining cardiac biomarkers in predicting mortality of inpatients with COVID-19," Hypertension, vol. 76, no. 4, pp. 1104– 1112, 2020.
[15] S. Schalekamp et al., "Model-based prediction of critical illness in hospitalized patients with COVID-19," Radiology, vol. 298, no. 1, pp. E46–E54, 2021.
[16] Y. Zhou et al., "Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: a multi-center study in China," Scand. J. Trauma Resusc. Emerg. Med., vol. 28, no. 1, p. 106, 2020
[17] G. Wang et al., "C-reactive protein level may predict the risk of COVID-19 aggravation," Open Forum Infect. Dis., vol. 7, no. 5, p. ofaa153, 2020.
[18] S. Chikode, N. Hindlekar, P. Padhye, N. Darapaneni, and A. R. Paduri, "COVID-19: Prediction of Confirmed cases, active cases and health infrastructure requirements for India," International Journal of Future Generation Communication and Networking, vol. 13, no. 4, pp. 2479– 2488–2479–2488, 2020Allen Institute For AI, "COVID-19 Open Research Dataset Challenge (CORD-19)."
[19] Allen Institute For AI, "COVID-19 Open Research Dataset Challenge (CORD-19)." .
[20] P. K. Nigam, "Correct blood sampling for blood gas analysis," J. Clin. Diagn. Res., vol. 10, no. 10, pp. BL01–BL02, 2016.
[21] 1. Supervised learning — scikit-learn 0.24.1 documentation," Scikitlearn.org. [Online]. Available: https://scikitlearn.org/stable/supervised_learning.html . [Accessed: 14-Feb-2021].
[22] "Welcome to imbalanced-learn documentation! — imbalanced-learn 0.7.0 documentation," Imbalanced-learn.org. [Online]. Available: https://imbalanced-learn.org/stable/index.html. [Accessed: 14-Feb2021].
[23] K. P. Shung, "Accuracy, precision, recall or F1? - towards data science," Towards Data Science, 15-Mar-2018. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1- 331fb37c5cb9. [Accessed: 14-Feb-2021].
[24] "API Reference — scikit-learn 0.24.1 documentation," Scikitlearn.org. [Online]. Available: https://scikitlearn.org/stable/modules/classes.html. [Accessed: 14-Feb-2021].