Fuzzy Systems and Soft Computing
ISSN : 1819-4362

MACHINE LEARNING APPLIED IN SARS-COV-2 COVID 19 SCREENING USING CLINICAL ANALYSIS PARAMETERS

[1]Andluru Jawahar Reddy, [2]Gaddam Chakresh Reddy, [3]Vattikuti Sri Yagneswara Phani Hemanth, [4]Vootukuri Harshith,

[5]S. Dhanalakshmi,[6]Kumbala Pradeep Reddy

[1,2,3,4]UG Scholar, Department of CSE (AI&ML)

[5]Professor,[6]Associate Professor, Department of CSE (AI&ML)

CMR Institute of Technology, Hyderabad, Telangana, India-501401

PROJECT OBJECTIVES

- The main objective is to effectively handling a SARS-CoV2 covid-19 unbalanced dataset.
- To efficiently predict the severity from clinical data.
- Machine learning and Deep learning models are used to enhance the overall performance to predict the severity from the dataset.

PROBLEM STATEMENT

Handling unbalanced data the machine learning algorithm doesn't properly classify the dataset. By extracting the SARS-CoV2 features from the data for more number of patients is difficult one.

ABSTRACT

COVID-19 was considered a pandemic by the World Health Organization. Since then, world governments have coordinated information flows and issued guidelines to contain the overwhelming effects of this disease. At the same time, the scientific community is continually seeking information about transmission mechanisms, the clinical spectrum of the disease, new diagnoses, and strategies for prevention and treatment. At the same time, the scientific community is continually seeking information about transmission mechanisms, the clinical spectrum of the disease, new diagnoses, and strategies for prevention and treatment. One of the challenges is performing the tests for the diagnosis of the disease, whose technique adopted for the detection of the genetic material of COVID-19 requires equipment and specialized human resources, making it an expensive procedure. We hypothesize that machine learning techniques can be used to classify the test results for COVID-19 through the joint analysis of popular laboratory tests' clinical parameters. The Machine learning techniques, such as Random Forest, Multi-Layer Perceptron, Support Vector Machine and Deep learning technique like Artificial Neural Network algorithms are enable the creation of disease prediction models and to analyze SARS-Cov2. To generate the

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

result based on accuracy, precision, recall, F1-measure and Specificity.

OVERVIEW

A novel member of human coronavirus, newly identified in Wuhan, China, recently, now officially named as SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) by International Committee on Taxonomy of Viruses (ICTV) is a new strain of RNA viruses that has not been previously identified in humans. Studies have shown that the disease caused by SARS-CoV-2, recently named as COVID-19 (coronavirus disease 2019) by World Health Organization (WHO), could induce symptoms including fever, dry cough, dyspnea, fatigue, and lymphopenia in infected patients. In more severe cases, infections causing viral pneumonia may lead to severe acute respirator y syndrome (SARS) and even death.1-4 Since the first report of COVID-19 in December 2019 in Wuhan, China, the outbreak of the disease is currently continuously evolving. Until February 16, 2020, the locations with confirmed SARS-CoV-2 cases include 25 countries. Globally, 73 332 cases were confirmed, including 72 528 cases in China and 804 cases out-side of China.5 A total of 1870 patients have died from this viral infection.5It was established that the SARS-CoV-2 belongs to the be-ta-coronavirus 2b lineage in the phylogenetic tree. By examining the full-length genome of SARS-CoV-2, it was discovered that this novel virus shared 87.99% identity sequencing with the bat SARS-like coronavirus,6 and it shared ~80% identity nucleotide with the original SARS epidemic virus.7 Based on the preliminary information of this novel virus, it is considered that SARS-CoV-2 is the third zoonotic human coronavirus of the century.8 In addition, clinical evidences have suggested that this virus is transmissible from person to person.9,10 Currently, it is still unclear about the origins and possible intermediate animal vectors of SARS-CoV-2, as well as the mechanism of this virus, that is, spreading between persons.

The Machine Learning and Deep Leaning algorithm plays the major role to analyse the clinical data and classify the data to predict the severity from SARS-Cov2 clinical parameters.

INTRODUCTION

SARS-CoV-2 has caused the current pandemic of COVID-19, a disease that first emerged as an outbreak in December 2019 in the Chinese province of Hubei [1]. The management of patients with COVID-19 remains problematic and controversial,

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

although this is to be expected in such a recently emerged disease. The first symptoms of COVID-19 resemble those of many other infections and inflammatory conditions that affect the respiratory system; they include fever, sneezing and rhinitis, persistent cough, and fatigue with body ache [2]. However, an infected patient can rapidly develop additional and more severe symptoms that can be life-threatening and require intensive care intervention; these include pneumonia, severe shortness of breath, diarrhea, dispersed thrombosis, and vascular inflammation [3,4]. An additional issue in caring for patients with COVID-19 is the presence of comorbidities that interact with COVID-19, particularly pulmonary and vascular conditions, which can greatly worsen the patient's prognosis [5]. This is an important consideration given the current lack of effective therapy for COVID-19. However, there have been notable advances in treating patients with advanced disease; therefore, the ability to predict that a patient will have poor outcomes, indicating a need for more aggressive treatment, has the potential to save lives and enable more effective allocation of resources.Intensive care units (ICUs) are key to increasing the survival of patients with severe COVID-19; they provide oxygen, 24-hour monitoring and care, and assisted ventilation when needed. Therefore, ICU beds are a precious resource in locations where COVID-19 case numbers are high [6-8]. Allocating hospital wards or ICU beds for infected patients thus requires rapid decision-making processes, both to use resources efficiently and reduce patient suffering and mortality. In many parts of the world, stressed care systems face significant difficulty in deciding on ICU bed allocation; therefore, a smart, automated system could be useful to improve care and resource allocation. The World Health Organization has recommended that all suspected patients with COVID-19 be tested by reverse transcription–polymerase chain reaction (RT-PCR)–based diagnosis methods that directly detect viral RNA [9]. Testing by approaches other than RT-PCR does not yet show acceptable accuracy. However, RT-PCR tests can take many hours or days to finalize the test outcomes, by which time the health condition and infectious status of confirmed patients may deteriorate. Rather than seeking a new single rapid test that improves on RT-PCR, an alternative approach could be to use results from many different profiling tests that are already available and can be performed quickly using existing equipment [10,11]. The best way to use the resulting multidimensional data is currently controversial.Rapid blood and serology testing of clinical samples by

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

current equipment enables monitoring of many peripheral blood parameters of interest, some of which indicate changes in organ functions and are used to diagnose a range of conditions and diseases [7,12]. This raises the possibility that such profiling of blood samples could provide predictive information about the disease trajectory and risk of comorbidities for patients with COVID-19. Some data is already used in physician deliberations; however, the many available test parameters suggest that an agnostic statistical or machine learning (ML) approach would improve the quality of those decisions. Therefore, we undertook a comprehensive assessment that examined the utility of a range of statistical and ML approaches. Indeed, we identified algorithms that showed significantly improved outcome estimates. Therefore, this work has the potential to optimize decision processes regarding patient care by clinicians who are under significant time and resource pressure during the current COVID-19 pandemic.

SYSTEM PROPOSAL

Existing System

The covid 19 SARS-CoV-2 laboratory test clinical data was implemented and predict the severity of the patient. The existing system applied only the machine learning algorithms to predict the severity from the data. So it doesn't effectively classify and predict the COVID-19 severity.

DISADVANTAGES

- Data inconsistency Problems
- Theoretical Limits.
- Loss of Information.
- Incorrect Prediction Results.

PROPOSED SYSTEM

The proposed model is introduced to overcome all the disadvantages that arise in the existing system. By applying the machine learning and deep learning algorithms to effectively predict the results. It enhances the performance of the overall classification results.

ADVANTAGES

- High performance.
- Provide accurate prediction results.
- It avoids data inconsistency.

METHODS

DATA SETS AND ANALYSES

We used two different data sets in this study; the first included data from 89 patients, and the second included data from 1945 patients with confirmed positive COVID-19 tests identified by RT-PCR. For the first data set [13], we use statistical methods such as the Student t test, chi-square test, and Pearson

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

correlation to identify the most significant and associative blood parameters that can strongly distinguish between patients with COVID-19 and healthy people. Moreover, to compare the blood parameter values of patients with COVID-19 with those of healthy patients, we considered the standard value ranges as reference values for each parameter. For the second data set [14], in addition to statistical methods, we used several ML models to further identify blood parameters that can discriminate between COVID-19–positive patients who are at risk of serious illness and those who are not. Figure 1 depicts a schematic of the ML analysis workflow of our approach.

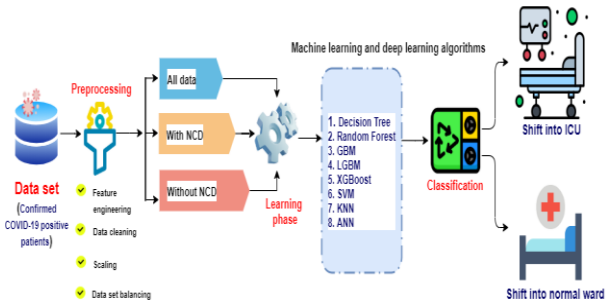

Figure 1. Proposed methodology and workflow of the machine learning analysis in this study. ANN: artificial neural network; GBM: gradient boosting machine; ICU: intensive care unit; LGBM: light gradient boosting machine; NCD: noncommunicable disease; SVM: support vector machine; KNN: k-nearest neighbor; XGBoost: extreme gradient boosting

We formulated the task of identifying patients with severe COVID-19 to enable selection of the appropriate hospital ward for their care as a classification problem by training ML models with features of clinical data collected from blood samples of patients with COVID-19. Raw data of interest collected from the data sets underwent a data-wrangling pipeline, including denoising, missing value imputation, transformation, normalization, and partition. Next, several statistical comparisons and correlation methods were adopted for feature engineering, including the Student t test, chi-square test, and Pearson correlation. After this, each data set was further split into three categories based on the criteria of existing noncommunicable disease (NCD): with NCD, without NCD, and all data. In our study, “NCD” refers to patients with pre-existing noncommunicable diseases or conditions. Finally, a range of state-of-the-art ML methods were trained and evaluated. The algorithms used included decision tree (DT), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), support vector machine (SVM), light gradient boosting machine (LGBM), k-nearest neighbor (KNN), and artificial neural network (ANN)–based deep learning sequential models. Each of these steps is discussed in the following subsections.

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

Data Collection

We obtained two different data sets of patients with COVID-19. The first data set was produced by Zenodo [13], and it contains demographic information and blood sample information from 89 COVID-19–positive patients. In this data set, 31 patients were alive at the point of data collection, while 58 patients had died. The second, larger data set was obtained from the Kaggle web-based resource [14], which contains grouped information regarding previous diseases, blood sample results, and vital sign data of 1945 COVID-19–positive patients. The primary sources of the data in this set are Brazilian hospitals, including Sirio Libanes, São Paulo, and Brasilia. The parameters of the data set included patient age percentile, gender, and demographic information. Some patients had pre-existing NCDs, including hypertension and immunocompromised status. The blood parameters examined included lactate, respiratory rate, diastolic blood pressure, hemoglobin, hematocrit, venous base excess, leukocytes, neutrophils, albumin, arterial base excess, urea, platelets, potassium, systolic blood pressure, venous PO2, arterial O2 saturation, partial thromboplastin time, temperature, gamma-glutamyl transferase, venous O2 saturation, creatinine, international normalized ratio (INR), venous PCO2, venous pH, arterial bicarbonate, labels of free fatty acids, venous bicarbonate, calcium, lymphocytes, alanine aminotransferase, aspartate aminotransferase, arterial PCO2, dimerized plasmin fragment D (D-dimer), oxygen saturation, bilirubin, arterial PO2, arterial pH, heart rate, blast, and glucose. During the feature-engineering phase in our study, all these blood parameters were considered as features.

Data Processing

For the Zenodo data set [13], which consists of 89 COVID-19–positive patients, we first removed any unwanted parameters (eg, ethnicity, BMI, drinking or smoking habits). We then eliminated all the missing values, resulting in a data set of 70 patients. In the Sirio Libanes data set [14] from Kaggle, there were 1945 individual patients with 54 types of tests. The primary data set contained a large number of missing values. This data set was prepared from information received from local hospitals and some of this information was not well prepared, which is a significant reason why most of the data have missing entries. The rationale behind the removal of entries with missing parameter values is that when we conducted a pilot study with the imputation of missing values with mean,

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

median, or regression values, poor predictive performance was observed. In the raw data set, the dimensions were 1925 × 205, and almost 57% of the data units (cell values) were missing; after eliminating unwanted attributes, the amount of missing data increased above 70%. If we considered all the data and imputed the missing values, most of the values would be inferred, and the analysis results would be unreliable. Therefore, we eliminated entries that contained at least one missing value. This elimination resulted in 545 sets of patient data entries in the second data set that contained no missing values. Among the patients in this data set, 264 had sufficiently severe symptoms to be admitted to the ICU. Both data sets underwent a denoising step, in which we removed unwanted strings. Standard scaling techniques were performed, such as feature scaling, in which the variance values of the data are scaled between 0 and 1; this is calculated by subtracting the mean value of a feature from the original value and then dividing by the standard deviation. After preprocessing, we considered data from 545 patients for the analysis. For a precise study, we then divided this data set according to whether a patient had a coexisting NCD (NCD) or not (no NCD). We found 264 patients with NCDs and 281 patients without NCDs; in the NCD and no NCD groups, 156 and 108 patients were respectively classed as displaying severe conditions. After this data preparation and preprocessing, we considered all these data for the statistical analysis. Due to the possibility of data leakage in ML analysis if we separated the test set and train sets after preprocessing, we first separated a randomly selected 80% of the grouped patient data for model training and used the rest for model validation testing, then performed the preprocessing steps.

Statistical Methods to Identify the Most Significant and Associative Blood Parameters

In the statistical analysis, we used chi-square tests for categorical variables, Student t tests for continuous variables, and Pearson correlations among various blood sample counts. The null hypothesis was that the data from the patients with COVID-19 and the healthy population were independent. Significant blood parameters were chosen based on a P value <.05, while in some cases, the selection criteria were a false discovery rate–adjusted P value <.05 and an absolute value log 2 fold change (LFC) <1. To understand the changes (positive or negative) of the parameters and the number of changes, we have calculated the LFC. LFC=1 indicates a fold change of value 2. Furthermore, hierarchical

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

clustering was conducted on the Pearson correlation coefficients for grouping significant parameters [15-17].

ML Models to Classify COVID-19 Disease Severity

To identify a set of important blood samples as a feature selection step, we employed a set of ML algorithms using COVID-19 data sets that included data from severely and nonseverely affected patients. We chose ML algorithms that are known to perform classification tasks with superior performance and fast execution [18,19]. For this purpose, we considered a basic ensemble learning approach based on max-voting, averaging, and weighted averaging for some classifiers, as well as advanced ensemble learning algorithms that function by stacking, blending, bagging, and boosting. Ensemble learning algorithms are combinations of one or more basic algorithms that are high-performing, efficient, effective, and easy to debug [20,21].We next address the parameters of the ML algorithms that were considered when they were run. In the DT algorithm, we used a random state of 42, a criterion of Gini, and a minimum sample split of 2. Similarly, in the RF algorithm, the minimum sample split was 2 and the number of estimators was 100. Degree and kernel cache size are parameters of the SVM algorithm; the algorithm sets a polynomial kernel with a degree of 3, and we set the kernel cache size at 200 MB for fast execution. In the GBM algorithm, the learning rate was 0.1, the criterion was friedman_mse, and the number of estimators was 100. The learning rate in the LGBM algorithm was 0.05, the feature fraction was 0.9, the bagging fraction was 0.8, and the bagging frequency was 5. In the XGB algorithm, we used a tree-based booster with a maximum depth of 6, a learning rate of 0.1, and 1000 estimators. For the KNN algorithms, we used Minkowski matrices; the weights were uniform, and the number of neighbors was 3 (k=3).We also experimented with a sequential deep learning model, namely, a feed-forward 1D ANN. This model consists of an input layer, three hidden layers, and an output layer [22]. Each layer contains a collection of parallel processing nodes, called neurons, that take input from the nodes of the previous layer. All the hidden layers are activated by rectified linear units, and the output layer is activated by a softmax function, providing the class probability of the input sample. The network was trained in 1000 epochs using the stochastic gradient descent optimization algorithm with categorical cross-entropy loss as a convergence indicator and a learning rate of 0.0001.

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

Shapley Additive Explanation Value Calculations

To measure the feature importance, we calculated the Shapley Additive Explanation (SHAP) values from all the models to estimate the degree of contribution of each of the features in the samples of the training data set to the overall decision-making of the model [23]. SHAP uses game theory rules to determine the contributions of particular features to the decision-making of the model. We used the TreeExplainer [24] for tree-based models and the KernelExplainer [23] for kernel-based models to calculate the feature importance. After finding the SHAP values for all the models, we normalized the values in a fixed range and considered the average values

CONCLUSION

The results of our analysis indicated that there is a strong relationship between particular abnormal blood parameters and disease severity status in hospitalized patients with COVID-19. The primary utility of our findings is that the subset of routine blood parameters linked to disease severity could be used in a predictive algorithm that would better enable appropriate care to be given before the onset of severe symptoms. This is of particular importance in developing countries, where ICU beds in hospitals are a limited resource. This can be achieved using a relatively small number of currently available blood-based hospital tests to properly use ICU resources and identify patients who need to be monitored closely.

Among the association between blood parameters that can give predictive information regarding the severity of COVID-19 symptoms, the levels of lactate and immature granulocytes (absolute) appeared to have the strongest predictive value. Levels of hemoglobin, procalcitonin, erythrocyte sedimentation rate, brain natriuretic peptide, ferritin, D-dimer, and platelets likewise showed significant deviation from the normal control group for prediction of disease severity. Other parameters, namely respiratory rate, lactate, blood pressure (systolic and diastolic), hematocrit, venous and arterial base excess, neutrophils, albumin, and urea, showed less obvious deviations but clearly had predictive value. Our work suggests that links exist between these parameters and COVID-19, and similar proinflammatory infectious diseases may merit more detailed physiological investigations.There were a few limitations to our study. First, the small sample size may restrict the

http://doi.org/10.36893/FSCC.2024.V09I01.088-098

precision of the identification of severity. Second, the absence of more detailed clinical information in the data sets that were used (such as patient age, sex, and comorbidities) may hinder better classification, although this suggests that in future studies, we could use new data sets to address this and improve on our work. Finally, the disease severity and mortality of COVID-19 varies significantly from country to country; the reasons for this are very poorly understood, but it is suggested that this type of predictive analysis should be conducted on data from other parts of the world to improve the performance of the algorithm. Nevertheless, we hope our study can be used by practitioners and help policy makers to improve resource allocation and outcomes for patients with COVID-19.

REFERENCES

[1] Machine Learning Applied in SARS-CoV-2 COVID 19 Screening using Clinical Analysis Parameters, R. F. A. P. Oliveira, C. J. A. Bastos-Filho, A. C. A. M. V. F. Medeiros, P. Buarque, D. L. Freire, IEEE LATIN AMERICA TRANSACTIONS, VOL. 19, NO. 6, JUNE 2021.

[2] Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey, Yassine Meraihi1 · Asma Benmessaoud Gabis, SeyedaliMirjalili,Amar Ramdane-Cherif, Fawaz E. Alsaadi, SN Computer Science (2022) 3:286 https://doi.org/10.1007/s42979-022-01184-z, Springer journal.

[3] Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): The epidemic and the challenges. Int J Antimicrob Agents. 2020;55(3): 105924.

[4] Mehta N, Shukla S. Pandemic analytics: How countries are leveraging big data analytics and artifcial intelligence to fght covid19? SN Computer Science. 2022;3(1):1–20.

[5] Haruna Chiroma, Absalom E Ezugwu, Fatsuma Jauro, Mohammed A Al-Garadi, Idris N Abdullahi, and Liyana Shuib. Early survey with bibliometric analysis on machine learning approaches in controlling covid-19 outbreaks. PeerJ Computer Science, 6:e313, 2020.

[6] Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE transactions on Knowledge and Data Engineering, 25(4):734–750, 2012.

[7] Iqbal Muhammad and Zhu Yan. Supervised machine learning approaches: A survey. ICTACT Journal on Soft Computing, 5(3), 2015.

[8] S.-Y. Xiao, Y. Wu, and H. Liu, “Evolving status of the 2019 novel coronavirus infection: Proposal of conventional serologic assays for disease diagnosis and infection monitoring,” Journal of medical virology, vol. 92, no. 5, pp. 464–467, 2020.

1. D.Vijayasekar, S.Dhivya, and S.Dhanalakshmi, “Wiener Filter Operation on Blurred Images”, International Journal of Applied Engineering Research Technology(IJAER), ISSN 0973-4562 Vol. 10 No.85 (2015), October 2015,PP 197-200

2. Shiva Prasanth.A, and S.Dhanalakshmi,, “Automated Testing Tools for Different Coverage Metrics”, International Journal of Advanced Innovative Research (IJAIR), Volume 5, Issue 10, ISSN: 2278-7844, October 2016, PP 120-123

3. K.Sindhu, and S.Dhanalakshmi,, “Disclosure of Malevolent in MANET: A Survey", International Journal of Research in Technological Studies(IJRTS),Volume 4,Issue 1,ISSN:2348-1439,December 2016,PP 31-34

S. M. Babu, P. P. Kumar, B. S. Devi, K. P. Reddy, M. Satish and A. Prakash, "Enhancing Efficiency and Productivity: IoT in Industrial Manufacturing," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023, pp. 693-697, doi: 10.1109/ICCCMLA58983.2023.10346807 .

19. Prakash, S. M. Babu, P. P. Kumar, S. Devi, K. P. Reddy and M. Satish, "Predicting Consumer Behaviour with Artificial Intelligence," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg, Germany, 2023, pp. 698-703, doi: 10.1109/ICCCMLA58983.2023.10346660 .

http://doi.org/10.36893/FSCC.2024.V09I01.088-098