COVID-19 Severity Analysis Using Improved Machine Learning Algorithm

Balraj Preet Kaur[1], Harpreet Singh[2] ,Rahul Hans[3],Sanjeev Sharma[4]
[1]Research Scholar, [2]Assistant Professor, [3,4]Associate Professor
[2]Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology Patiala, India
[1,3,4] Department of Computer Science and Engineering, DAV University, Jalandhar,India
[1]*balrajpreetkaur2@gmail.com*,[2]*harpreet.singh1@thapar.edu*,,[3]rahulhans@gmail.com,[4]sanju3916@rediffmail.com

ABSTRACT

The new pandemic produced by the COVID-19 virus has resulted in an overflow of medical treatment in clinical centers all over the world. The fast and exponential growth in the number of COVID-19-infected individuals has necessitated an effective and timely prediction of probable infections and their effects in order to reduce health-care quality overload. As a result, intelligent models are being developed and used to assist medical workers in making more accurate diagnoses concerning the health condition of COVID-19-infected individuals. The purpose of this research is to present an alternative algorithmic approach for predicting the health status of COVID-19 patients in Mexico. Different prediction models were assessed and compared, including Adaboost, gradient boosting machine, random forests, and light gradient boosting machine. Additionally, Grid search hyperparameter optimization is used to improve the algorithm's success rate. The optimal model feature analysis procedure is being carried out. The purpose of this study is to analyses features in terms of feature importance as indicated by SHapely adaptive exPlanations (SHAP) values in order to identify relevant predictive factors that can identify patients at high risk of mortality.

Keywords—Machine learning; COVID-19; Hyperparameter tuning; SHAP analysis

INTRODUCTION

The COVID-19 pandemic has brought forth an urgent need for effective ways to battle the virus and reduce its effects on global health systems and economy. Machine learning techniques have emerged as a key tool in analyzing and tackling the pandemic's multiple difficulties [1]. A subset of artificial intelligence, machine learning allows computers to learn from and analyses enormous volumes of data in order to make predictions, detect patterns, and produce insights. Machine learning approaches have been used in numerous aspects of COVID-19, including epidemiological modeling, diagnosis, therapy development, vaccine development, and resource allocation. These applications have the potential to transform our approach to pandemic management and public health outcomes [3].

Machine learning is the use of algorithms and statistical approaches to allow computers to learn from data and make predictions or judgments without being explicitly programmed. Grid search, on the other hand, is a hyperparameter tuning strategy used to optimize machine learning models [1]. Grid search includes systematically exploring a preset range of hyperparameter values to discover the ideal combination that produces the best model performance [18]. Hyperparameters are configuration settings that affect the behavior of a model, and grid search [13] involves systematically exploring a predefined range of hyperparameter values to determine the optimal combination that yields the best model performance [6].

Furthermore, machine learning and grid search are important components of data-driven modeling, with machine learning allowing computers to learn from data and grid search being a helpful tool for optimizing model performance through hyperparameter tweaking [17]. They offer a strong foundation for constructing and optimizing machine learning models for a wide range of applications [7].Moreover, Identifying and analyzing the traits or qualities that may affect the severity of COVID-19 instances is what feature analysis for severity analysis in a COVID-19 dataset entails [15]. SHAP feature analysis was used in this study on high accuracy algorithms following grid search on each algorithm [16]. In this article, the relationship between COVID-19 and machine learning approaches, and how machine learning is being used to address the enormous problems of the current epidemic are discussed [9]. Furthermore, SHAP feature analysis can give useful insights into the value of various characteristics in predicting COVID-19 severity and aid in identifying the underlying causes that lead to COVID-19 case severity [11, 12].

LITERATURE OVERVIEW

Shekar et.al [14] discusses the challenges in analyzing microarray cancer data, which include the curse of dimensionality, small sample size, noisy data, and imbalance class problem. To address these challenges, the authors propose a method that uses grid search-based hyperparameter tuning for classification. The method is evaluated using several standard metrics such as classification accuracy, precision, recall, f1-score, misclassification error, Out-of-bag (OOB) error and confusion matrix. The results show promising performance across most of the test datasets.

EL-Kenawy et.al [4] discusses a study on novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. The study was conducted by a team of experts from different universities in Egypt, Australia, and South Korea. The paper provides details on the proposed algorithms and their effectiveness in classifying COVID-19 in CT images. The study also highlights the potential applications of these algorithms in biomedical imaging diagnoses. Statistical tests were carried out to ensure the quality of the proposed algorithms.

Kini et.al [7] represent a novel framework that uses deep learning and IoT technologies for automated COVID-19 diagnosis. The proposed model is designed to diagnose the type of infection and provide results, and if the patient is infected with COVID-19, their sample will be further processed for verification of actual COVID types such as delta variant and omicron variant. The framework can provide better performance and can be used for automated diagnosis of COVID-19 suspected cases.

Albataineh et.al [1] proposed system that uses machine learning algorithms to diagnose and assess the severity of COVID-19 through CT scans. The system categorizes the illness into three stages: mild, moderate, and severe, based on the segmentation method and features extracted from the CT images. The proposed system obtained excellent performance levels in segmentation, classification, and infection quantification.

Attallah et.al [2] presents an intelligent ECG-based tool for diagnosing COVID-19 using various AI methods. The proposed tool uses ensemble deep learning techniques and a hybrid feature selection approach to reduce the number of features used to train the classification models. The classification procedure of the proposed tool is performed on two levels: binary class level to classify ECG data into COVID-19 and normal cases, and multiclass level to distinguish COVID-19 cases from normal and other cardiac complications. Wang et.al [19] discusses the issue of class imbalance in classification and proposes an improved version of the Ada Boost algorithm to address this problem. The proposed solution involves adjusting the weighted vote parameters of weaker classifiers, including the global error rate and positive class accuracy rate, as well as considering the imbalanced index to improve classification performance.

Li et.al [8] proposes a new method for diagnosing transformer faults with higher accuracy using the Multi-class Ada Boost Algorithm. Traditional shallow machine learning algorithms have low fault diagnosis accuracy because they cannot effectively explore the relationship between the fault data of oil-immersed transformers. The paper shows that the proposed method has strong search ability and fast convergence speed and has a significant improvement in diagnostic accuracy compared with traditional methods. Gupta et.al [5] discusses the use of machine learning techniques for classifying diabetes disease. The paper is divided into seven sections, covering topics such as the objectives of the study, a literature review, working methodology and results, and a conclusion. The authors explore different classifiers and dataset models, as well as the impact of preprocessing on classification accuracy.

PROPOSED MODEL

The suggested design for the COVID-19 diagnosis system is depicted in Figure 1. The proposed framework has five phases. Data is preprocessed in the first stage by deleting irrelevant characteristics. In the second phase, the Machine learning algorithm is applied to the dataset to calculate the performance of four machine learning algorithms, including Adaboost, Gradient boosting machine, Random forest and light gradient boosting machine [10].

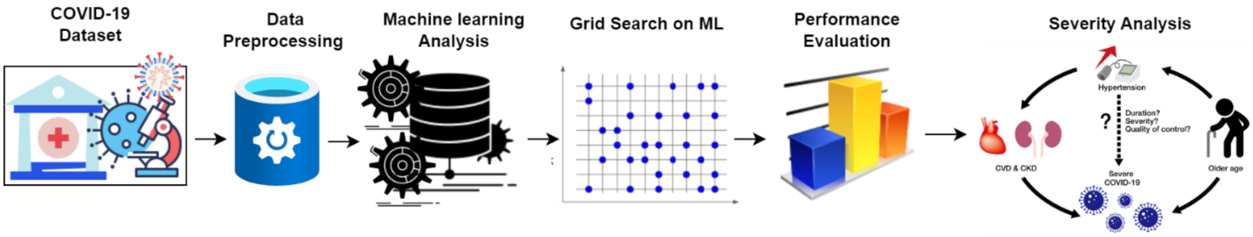

Figure - 1. Proposed Model architecture

In addition, the Grid search technique is used to discover the ideal hyperparameter to improve their performance. After selecting the optimal hyperparameter, accuracy is compared, and the SHAP analysis [12] is performed using the best hyperparameter-optimized method. This examination will offer strongly correlated characteristics and a severity rating for the COVID-19 virus. A high correlation value of attribute indicates a high risk of COVID-19 illness in the patient.

DATASET DESCRIPTION

The dataset comprises 29 columns of clinical data as well as an RT-PCR test. There are statistics on 263007 patients accessible [20].

Table (1) – Dataset Description

S. No.	Attribute Name	Description
1.	Entidad_um	Region where hospital performed admission
2.	Entidad_Res	Residence of the patient at which region
3.	Delay	Lag in the process of lab report
4.	Entidad_Registro	The actual region from where case assigned
5.	Origen	surveillance of patient (1 = yes, 2= no)
6.	Sector	identify the institute of national health system
7.	Sexo	gender of patient 1 = female, 2 = male and 99 for undisclosed
8.	Entidad_nac	patient birth state or region
9.	Tipo_paciente	type of care patient received (1 = outpatient, 2= impatient)
10.	Neumonia	Identifies whether the patient was diagnosed with pneumonia
11.	Edad	Age of the patient
12.	Nacionalidad	check whether patient is Mexican(1) or foreign(2)
13.	Embrazo	Identifies patient is pregnant or not
14.	Habla_lengua_Indig	Patient speaks an indigenous language
15.	Diabetes	Identifies whether the patient was diagnosed with diabetes

16.	EPOC	Classify whether the patient detect with pulmonary disease
17.	Asma	Classify whether the patient diagnosed with asthma or not
18.	Immusupr	Identifies if the patient is immune suppressed
19.	Hipertension	Classify whether the patient diagnosed with hypertension
20.	Otra_Com	Identifies if the patient presents another disease
21.	Cardiovascular	Classify whether the patient diagnosed with cardiovascular disease or not
22.	Obesidad	Classify whether the patient diagnosed with obesity or not
23.	Renal_Cronica	Identifies if the patient presents chronic renal insufficiency
24.	Tabaquismo	Identifies if the patient has tobacco addiction
25.	Otro_Caso	Classify whether the patient diagnosed with any other case diagnosed with SARS COV-2
26.	Migrante	Identifies if the patient is migrant
27.	UCI	Identifies if the patient was admitted to ICU
28.	Intubado	patient need intubation or not(1= yes, 2=no,97=not applicable)
29.	Resultado	The RT-PCR test(1 = positive , 2 = negative)

EXPERIMENTAL RESULTS AND DISCUSSION

This section represents and analyze the results obtained after performing an experiment with 70:30 dataset splitting. The proposed approaches have been executed with python using sklearn libraries [20].The efficiency outcomes of the algorithms with default hyper-parameters are presented in figure 2. The four algorithms, with assessment metrics recall, precision and f-measure compared in figure 2.The result of the success rate of four machine learning algorithms with and without grid search is computed in figure 3. Furthermore, the Adaboost success rate is 0.9445 which is higher than other algorithm. The components of binary classification in order to create the confusion matrix for Adaboost algorithm classification and ROC are shown in figure 4.

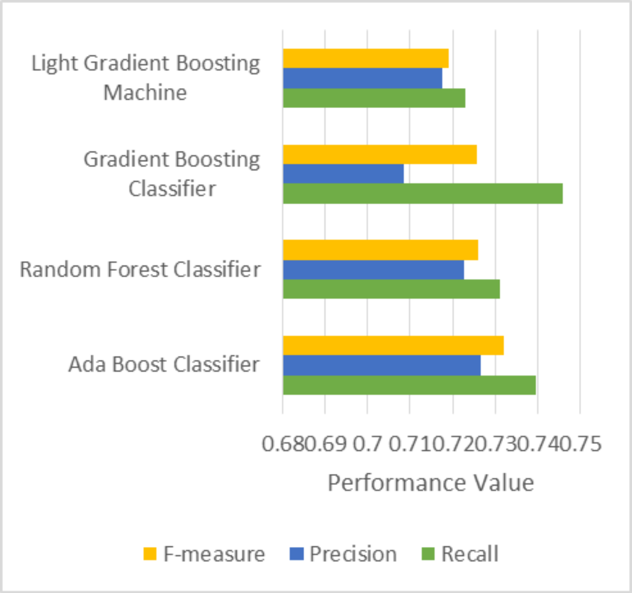

Figure – 2 Comparison of Machine learning Algorithm with default hyper parameter

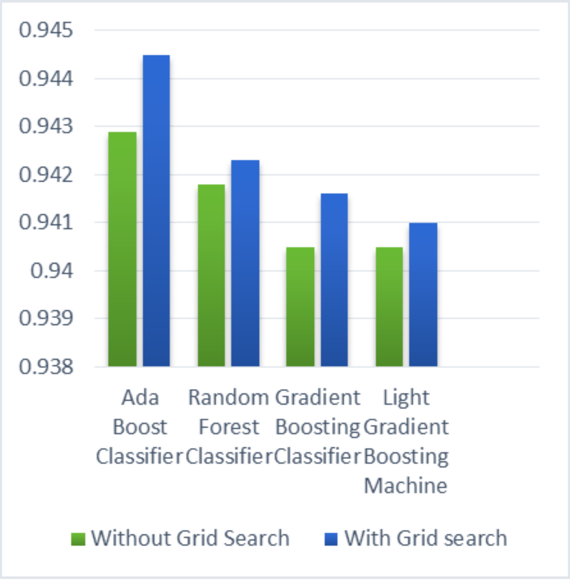

Figure – 3 Comparison of Machine learning Algorithm with and without grid search

Adaboost outperforms than other machine learning algorithms after grid search hyperparameter optimization on Adaboost, Gradient boosting machine, Random forest, and Light gradient boosting machine.

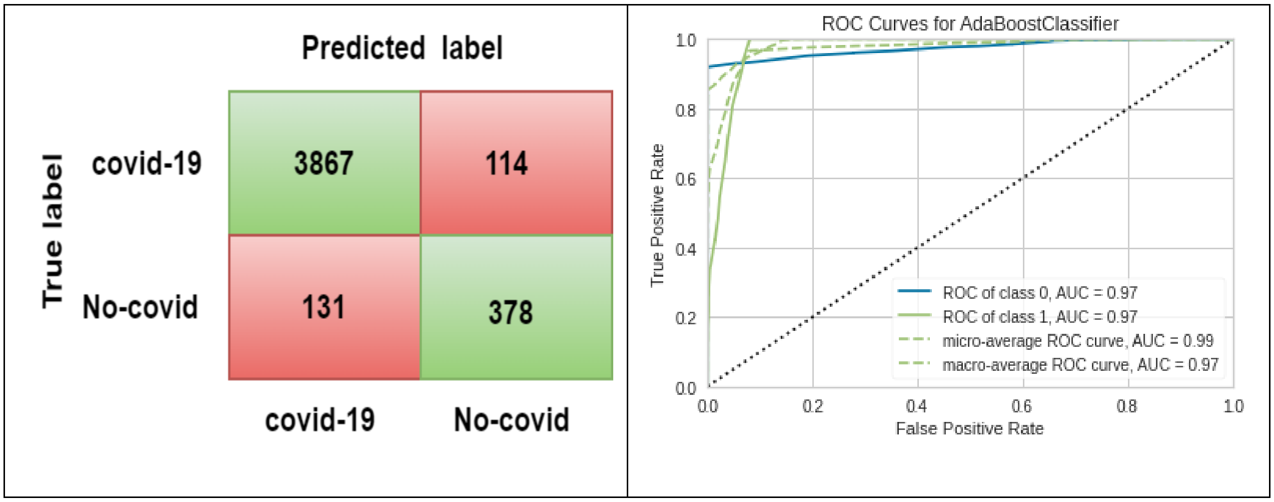

Figure - 4 Confusion matrix of Adaboost and ROC of Adaboost after Grid search algorithm

The table 2 show the hyperparameter of machine learning algorithm after applying grid search hyperparameter optimization technique. These parameters help in increase the success rate of each machine learning algorithm. The SHAP analysis is then run using the Adaboost method to uncover highly correlated characteristics with class covid-19. The feature importance is shown in figure 5. High-value characteristics investigate risk variables linked to COVID-19 illness severity.

Table (2) – Grid search hyperparameter

S.NO.	**Machine learning algorithm**	**Grid search parameter**
1.	Adaboost	{'algorithm': 'SAMME.R', 'learning_rate': 1.01, 'n_estimators': 4}
2.	Gradient Boosting machine	{'learning_rate': 0.1, 'max_depth': 3,'n_estimators':250,'random_state': 1, 'subsample': 0.7}
3.	Random Forest	{'criterion': 'gini', 'max_depth': 15, 'n_estimators': 20}
4.	Light gradient boosting machine	(boosting_type='dart', colsample_bytree=0.6, learning_rate=1, max_depth=5,n_estimators=20, num_leaves=5, reg_lambda=1)

Patients who have a lack of oxygen in their bodies are at risk of developing COVID-19. The severity of COVID-19 is also affected by the person's age. Age, pneumonia, pregnancy, diabetes, and a variety of other factors are also essential.

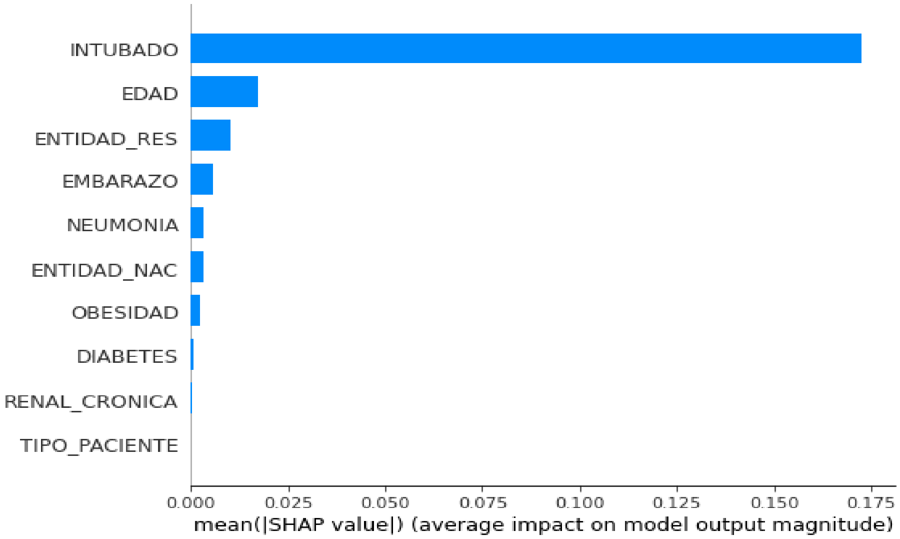

Figure – 5. SHAP analysis on Adaboost algorithm

In descending order, Fig. 5 demonstrates the significance of each characteristic in predicting COVID-19 mortality. The most critical criterion in this study is intubation, followed by age, followed by contact and pneumonia. Pregnant patients are especially susceptible. Additionally, Diabetes, obesity, hypertension, cardiovascular disease, asthma, and renal disease are all risk factors.

CONCLUSION AND FUTURE WORK

The present use of data analysis techniques around the world is becoming more important in order to provide models to analyses patient data in order to establish efficient treatment plans. In this study, machine learning algorithms were compared with the goal of assessing the likelihood that a person infected with COVID-19 will recover. We presented many prediction models, with the Adaboost method being picked as the most accurate framework since it not only allows us to receive the patient's categorization but also allows us to obtain it in terms of probability. The benefits of this research include the discovery and evaluation of novel characteristics for health status prediction models. And, based on the results, it is conceivable to see good indicators that it is possible to build alternative models that are similarly efficient based on the fast diagnosis. Our prediction model varies from others in various ways. In future, for starters, it will allows users to forecast a patient's health state based on a simple and faster medical diagnosis (by using the mobile app, for instance). By analyzing data like vital signs, patient region, blood oxygen, and so on, this method enables speedier medical evaluations to identify patients who are more likely to die. In the future, more hyperparameter tuning and feature selection analysis can be used to increase efficiency.

REFERENCES

[1] Albataineh, Zaid, Fatima Aldrweesh, and Mohammad A. Alzubaidi. 2023. “COVID-19 CT-Images Diagnosis and Severity Assessment Using Machine Learning Algorithm.” *Cluster Computing* 5(May 2022).

[2] Attallah, Omneya. 2022. “An Intelligent ECG-Based Tool for Diagnosing COVID-19 via Ensemble Deep Learning Techniques.” *Biosensors* 12(5).

[3] Ciotti, Marco et al. 2020. “The COVID-19 Pandemic.” *Critical Reviews in Clinical Laboratory Sciences* 0(0): 365–88. https://doi.org/10.1080/10408363.2020.1783198.

[4] El-Kenawy, El Sayed M. et al. 2020. “Novel Feature Selection and Voting Classifier Algorithms for COVID-19 Classification in CT Images.” *IEEE Access* 8.

[5] Gupta, Subhash Chandra, and Noopur Goel. 2023. “Predictive Modeling and Analytics for Diabetes Using Hyperparameter Tuned Machine Learning Techniques.” *Procedia Computer Science* 218(2022): 1257–69. https://doi.org/10.1016/j.procs.2023.01.104.

[6] Kassania, Sara Hosseinzadeh et al. 2021. “Automatic Detection of Coronavirus Disease (COVID-19) in X-Ray and CT Images: A Machine Learning Based Approach.” *Biocybernetics and Biomedical Engineering* 41(3): 867–79.

[7] Kini, Anita S. et al. 2022. “Ensemble Deep Learning and Internet of Things-Based Automated COVID-19 Diagnosis Framework.” *Contrast Media and Molecular Imaging* 2022.

[8] Li, Jifang, Genxu Li, Chen Hai, and Mengbo Guo. 2022. “Transformer Fault Diagnosis Based on Multi-Class AdaBoost Algorithm.” *IEEE Access* 10: 1522–32.

[9] Madoery, Pablo G, Ramiro Detke, Lucas Blanco, and Sandro Comerci. 2020. “Since January 2020 Elsevier Has Created a COVID-19 Resource Centre with Free Information in English and Mandarin on the Novel Coronavirus COVID- 19 . The COVID-19

Resource Centre Is Hosted on Elsevier Connect , the Company ' s Public News and Information .” (January).

[10] Mansbridge, Nicola et al. 2018. “Feature Selection and Comparison of Machine Learning Algorithms in Classification of Grazing and Rumination Behaviour in Sheep.” *Sensors (Switzerland)* 18(10): 1–16.

[11] Ndwandwe, Duduzile, and Charles S. Wiysonge. 2021. “COVID-19 Vaccines.” *Current Opinion in Immunology* 71(Figure 1): 111–16. https://doi.org/10.1016/j.coi.2021.07.003.

[12] Patel, Dhruv et al. 2021. “Machine Learning Based Predictors for COVID-19 Disease Severity.” *Scientific Reports* 11(1): 1–7. https://doi.org/10.1038/s41598-021-83967-7.

[13] Rostami, Mehrdad, and Mourad Oussalah. 2022. “A Novel Explainable COVID-19 Diagnosis Method by Integration of Feature Selection with Random Forest.” *Informatics in Medicine Unlocked* 30(January): 100941. https://doi.org/10.1016/j.imu.2022.100941.

[14] Shekar, B. H., and Guesh Dagnew. 2019. “Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data.” *2019 2nd International Conference on Advanced Computational and Communication Paradigms, ICACCP 2019* (November): 1–8.

[15] Siji George, C. G., and B. Sumathi. 2020. “Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction.” *International Journal of Advanced Computer Science and Applications* 11(9): 173–78.

[16] Sreedharan, Radhika, and Archana Praveen Kumar. 2020. “Analysis and Prediction of Smart Data Using Machine Learning.” *AIP Conference Proceedings* 2240(Ml): 15–21.

[17] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. “Comparing Different Supervised Machine Learning Algorithms for Disease Prediction.” *BMC Medical Informatics and Decision Making* 19(1): 1–16.

[18] Velavan, Thirumalaisamy P., and Christian G. Meyer. 2020. “The COVID-19 Epidemic.” *Tropical Medicine and International Health* 25(3): 278–80.

[19] Wang, Wenyang, and Dongchu Sun. 2021. “The Improved AdaBoost Algorithms for Imbalanced Data Classification.” *Information Sciences* 563: 358–74. https://doi.org/10.1016/j.ins.2021.03.042.

[20] https://www.kaggle.com/marianarfranklin/mexico-covid19-clinical-data/