Received: 12 March 2022 | *Revised: 18 April 2022* | *Accepted: 23 April 2022*
DOI: 10.37917/ijeee.18.1.15 *Vol. 18| Issue 1| June 2022*

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

A Comparison of COIVD-19 Cases Classification Based on Machine Learning Approaches

Oqbah Salim Atiyah*, Saadi Hamad Thalij
College of computer science, University of Tikrit, Iraq

Correspondence
* Oqbah Salim Atiyah
Computer Science College,
University of Tikrit, Tikrit, Iraq
Email: oqbah.s.atiyah35529@st.tu.edu.iq

Abstract
COVID-19 emerged in 2019 in china, the worldwide spread rapidly, and caused many injuries and deaths among humans. Accurate and early detection of COVID-19 can ensure the long-term survival of patients and help prohibit the spread of the epidemic. COVID-19 case classification techniques help health organizations quickly identify and treat severe cases. Algorithms of classification are one the essential matters for forecasting and making decisions to assist the diagnosis, early identification of COVID-19, and specify cases that require to intensive care unit to deliver the treatment at appropriate timing. This paper is intended to compare algorithms of classification of machine learning to diagnose COVID-19 cases and measure their performance with many metrics, and measure mislabeling (false-positive and false-negative) to specify the best algorithms for speed and accuracy diagnosis. In this paper, we focus onto classify the cases of COVID-19 using the algorithms of machine learning. we load the dataset and perform dataset preparation, pre-processing, analysis of data, selection of features, split of data, and use of classification algorithm. In the first using four classification algorithms, (Stochastic Gradient Descent, Logistic Regression, Random Forest, Naive Bayes), the outcome of algorithms accuracy respectively was 99.61%, 94.82% ,98.37%,96.57%, and the result of execution time for algorithms respectively were 0.01s, 0.7s, 0.20s, 0.04. The Stochastic Gradient Descent of mislabeling was better. Second, using four classification algorithms, (eXtreme-Gradient Boosting, Decision Tree, Support Vector Machines, K_Nearest Neighbors), the outcome of algorithms accuracy was 98.37%, 99%, 97%, 88.4%, and the result of execution time for algorithms respectively were 0.18s, 0.02s, 0.3s, 0.01s. The Decision Tree of mislabeling was better. Using machine learning helps improve allocate medical resources to maximize their utilization. Classification algorithm of clinical data for confirmed COVID-19 cases can help predict a patient's need to advance to the ICU or not need by using a global dataset of COVID-19 cases due to its accuracy and quality.

KEYWORDS: COVID-19, Classification algorithm, Prediction, Machine learning.

I. INTRODUCTION

Coronavirus is known as COVID-19, a new epidemic that appeared in 2019 in Wuhan city in China, this epidemic spread worldwide very rapidly, and it became a concern for all countries due caused to many injuries and deaths[1]. World-Health-Organization (WHO) proclaimed an emergency status after the COVID-19 spread in most countries, this requirement applies strict steps to control and decrease the danger of the epidemic. The virus is spread by the respiratory system or when an injured person comes into contact with an uninjured person [2], the symptom appears on the injured for a duration of (2 - 14) days depending on WHO information, the symptom in a moderate state of injures; fever, dry-cough, fatigue, and of the critical state: fever, asphyxia, tiredness, and breath distress[3]. With the spread of COVID-19 rapidly there is required to use Machine-Learning (ML) to help in the early disclosure of the disease to avert spreading it. In healthcare ML is important, it is utilized for collecting data of injures and analyzed using algorithms for best learn of the method that COVID-19 transition, and refines the velocity and accuracy of prognosis, it possible to exist those most infections of threatened depend on a personal genetic and physiologic, and ameliorate quality treatment ways [3],[4]. ML can learn and develop automatically based on experience and knowledge without being programmed explicitly. The algorithms really depend on attributes. A big and complex magnitude of the data can be optimal utilization using ML- algorithms. Machine learning is useful in the classification, diagnosis, and forecasting the diseases [1]. ML algorithms can predict of the number possible confirmed injuries of COVID-19 and the number of likely deaths in the future [5]. In this work. We compare the machine learning approaches in classifying COVID-19 cases into cases that require to Intensive-Care-

Unit (ICU) or wouldn't require one. also comparing the performance of algorithms in false-positive and false-negative mislabeling, to select the best ones in accuracy and speed of prognosis to aid the doctors in recognizing the COVID-19 and avoid the mistake.

II. Related work

COVID-19 is related to the fast evolution of data basics, so there is a need to analyze the relations and hierarchy of the data utilizing the ML algorithms to aid the health system in the diagnosis of COVID-19, [6].This paper provides recent studies analysis of this area:

Sarwar et al. [7], diagnose diabetes using the algorithms of machine learning, the outcome referred to assured accuracy of 98.60%. These can be helpful to forecast COVID-19, the exact identification of COVID-19 can rescue many the people, the output enormous of data to train algorithms. ML is possible to offer beneficial entries in this area, of performing prognoses based on Images, radiography, clinical text, etc.

Iweendi et al. [8], presented a Fine-Tuned RF model, as well as the adaboost model. To predict the possible result, the system uses the spatial, demographic, and health details of COVID-19 patients, the results were an F1-Score of 0.86, an accuracy rate is 94%. **A** review of data refers to a strong relationship between the state of death and patient gender, patient majority are among 20-70 ages.

Bayat et al. [9], presented the system to anticipate COVID-19 depending on testing in a standardized lab. A massive dataset containing 75,991 infections was gained from US Veterans-Affairs, utilized XGB to create the model, the outcomes were 86.4% of accuracy, 86.8% of specificity, and 82.4% of sensitivity. This work found the privileges of the top (10) are of downward significance.

Zhou et al. [10], presented a system to anticipate the disease seriousness of COVID-19 infections. they used a dataset containing 377 infections (172 are seriousness, 106 are non-seriousness) from one of the china's-hospitals, the Logistic Regression was utilized to create the forecasting system, the results were 87.9% of AUC, and 88.6% of sensitivity, 73.7% of specificity The outcomes were existed three separate elements linked strongly to COVID-19 infections: C_reactive proteins, age, and d-dime.

III. Methodology

The main stage of the methodology utilized in this study is displayed in figure 1. Specificity, sensitivity, accuracy, precision, ROC_AUC_Score, the positive and negative prevalence, mislabeling and execution time are used to measure performance. we used python to process the outcomes to create a classification system (including preparation of dataset, performing the pre-processing, analysis of data, scaling of features, split of data, and algorithm of classification), in the first model [11], we use algorithms such as Stochastic Gradient Descent (SGD), Naïve-Bayes (BN), Logistic-Regression (LR), and Random Forest (RF). Second model [12], we use algorithms like Support-Vector Machines (SVM), Decision-Tree (DT), eXtreme-Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN).

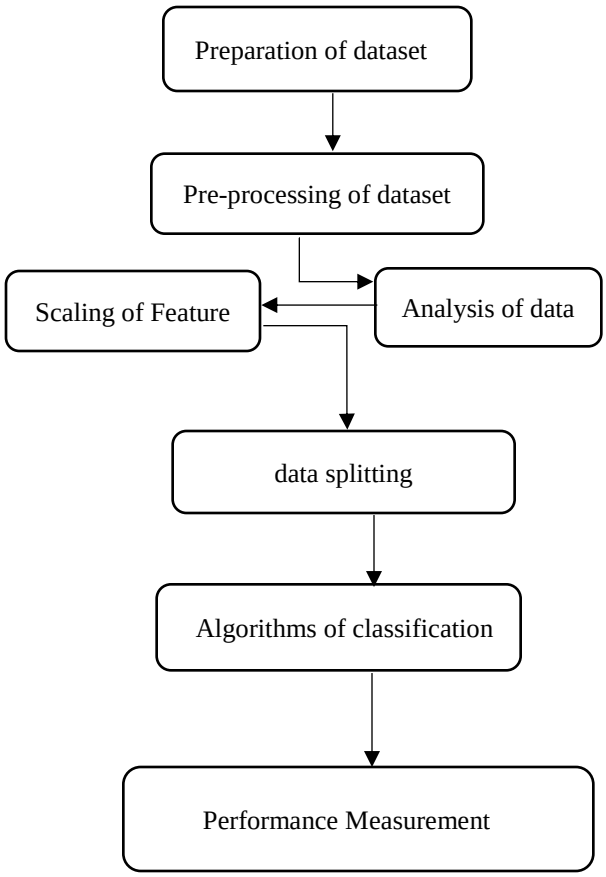

Fig.1. Show the Method Diagram

A. *Preparation of Dataset*

We have gotten the dataset from a search of the dataset in the google engine is a repository open_source that contains the most suitable information and details of COVID-19, the file format is the dataset of xlxl which involve 1925 columns and 231 rows[13].

B. *Pre-processing of Dataset*

The dataset used in this work contains about 1925 instances and 231 features, the dataset should be improved with a better form to process the data to the consistency requirements, before implementing the model. preprocessing has two main stages: processing the missing values and encoding the data for classification.

C. *Analysis of Data*

It is an operation modeling the data, examining, and imagining to extract helpful information and knowledge to make conclusions of performing an important role in decision-making.

D. *Scaling of Feature*

The large-scale dimensions and the discrepancy of entries in the dataset make it challenging to find the data. So the dimensions of values should be compatible in the dataset to get an efficient model and computation speeding up in the

models. Therefore, the standard deviation is utilized with standardization, making the mean zero to the features, and the standard deviation becomes one for the dissemination of the results.

E. Dataset Splitting

The dataset must be segmented into a set of trains and a set of tests before applying the algorithms of classification. the dataset dividing for the set of training is 80% and the set of testing is 20%.

F. Classification Algorithms

There are many classification algorithms in machine learning. In the first classification model, we use the algorithms like Stochastic Gradient Descent (SGD), Naive-Bayes (BN), Logistic-Regression (LR), and Random-Forest (RF). In the second model, we use the algorithms like Support-Vector Machines (SVM), Decision-Tree (DT), eXtreme-Gradient Boosting (XGBoost), and K_Nearest Neighbors (KNN).

1) Naïve Bayes

NB provides a way to predict different class potentials relying on the various features. it is mostly used in classifying texts and processing of multi-class problems [14].

2) Logistic Regression

LR is a classification algorithm of learning supervised that is utilized to predict a target variant possibility, due to the dichotomous dependent various or the goal nature, two potential classes will be made, LR is used as the reaction between different sets from prediction variants and the categorical result variables [15].

3) Stochastic Gradient Descent

It is very qualified method and proper for the linear classifiers based on functions of convex lack like linear. The SGD is implementing sporadic and large-scale ML problems successfully in classifying the text and processing normal language [16]. In the SGD the classifier applies regularized linear models, for every sample, the loss gradient is estimated concurrently and the model is upgraded during the learning ratio [17].

4) Random Forests

It is an algorithm learning of supervised used in classification problems, it assembling amounts of data training of decision trees, utilize in classification a means known packing. Each decision tree refers to forecasting of class, this way collecting the volumes in the decision trees, the final layer is who has further volumes [18].

5) Support Vector Machines

It is an algorithm learning of supervised based on the decision planes idea, it works to isolate the data through creating the hyperplane, where utilized the hyperplanes of classification the specific classes set [19]. It works to discover the boundaries and lines for classification the training dataset correctly and determine the line of data points nearest [20].

6) eXtreme Gradient Boosting

XGBoost is an algorithm of supervised learning used of regression and classification. It is an application open-source common and effective for gradient boosted in the tree model. It attempts to forecast the accurate target variables by combining an estimates group for the set of weaker models and simpler ones. The idea of the algorithm is to add continuously trees and apply the fragmentation of features to grow a tree [21].

7) K_Nearest Neighbors

It is a supervised learning algorithm utilized in problems of classification. It utilized the 'feature similarity' to create the coordinates of new data values. These assigned coordinates values will be based on how identical to the training set coordinates values. The training phase keeps the dataset only. while the test phase classifies the new data that a much identical to the class of the dataset [22].

8) Decision Tree

A Decision Tree is an algorithm supervised utilized for regression problems and classification. It works for both continuous and categorical output variables. A decision tree contains two stages at the classification learning stages and predicting. The system is training to use the training data granted in the learning stage. It is used to predict the outcome that indicated to test the data in the predicting stage [23].

IV. THE RESULTS

This part shows detail of the dataset, the learning system, the algorithms utilized in classifying the COVID-19, and the metrics of performance. and display the results of comparing the performance of the algorithms.

A. Characterizing Data

Characterizing data is a significant procedure in the data preparation stage, it gives visualizes things by showing the variables of the dataset used. Table I display a description of the variables in the dataset.

B. Pre-Processing Results

Pre-Processing is an operation of processing data and working on statistical analysis. The results of any phase are the entry of the next phase, so the data need to be prepared in similar details. In this phase, the miss of values is processed, if the miss of values is numeric should be a substitution with the mean of the value in the column, or if it was nominal must be a substitution with a value in the neighbor, the dataset is ready to the next stage.

C. Data Analysis Results

In this stage, the data assembled is summarized and interpreted during logical reasoning and analysis to specify modules, trends links, and describe the pre-processed data to perceive features. Table (2) displays the patients' total. Table (3) displays the distribution of ages for the total patients. Table (4) display the distribution of age for patients who need to the ICU.

TABLE I
DESCRIPTION OF DATASET

Feature	Kind	Prescribing
PATIENT_VISIT_ IDENTIFIER	int_64	identifier of patient that visited the hospital
GENDER	int_64	gender of patients
AGE_PERCENTIL	Object	percentile of ages the patients.
AGE_ABOVE65	int_64	the ages of patients that above 65 years.
DISEASE_ GROUPING	float_64	six sets of the diseases that have available attributes of the nameless information
	.....	
RESPIRATORY _RATE_ DIFF_REL	float_64	the available attributes around the respiratory average relative -diff
TEMPERATURE _DIFF -REL	float_64	the available attributes around the temperature relative_diff
OXYGEN SATURATION DIFF -REL	float_64	the accessible attributes around the oxygen sated relative _diff
WINDOWS	object	The windows has five kinds of sets everyone containing hours for acceptance
ICU	int_64	reply features (0 represent not need to ICU and 1 represent need to ICU)

TABLE II
THE PATIENTS' TOTAL.

Patients Total after pre-processing	293
need to ICU	105
not need to ICU	188

TABLE III
DISSEMINATION OF AGE FOR TOTAL infections.

Age Dissemination	
infections under of 65age	172
infections over of 65 age	121

TABLE IV
DISSEMINATION OF AGE FOR INFECTIONS IN ICU

Age Dissemination	
Infections over of 65 age:	60
Infections under of 65 age:	45

D. Results of Classification

The performance of ML-algorithms that used in classifying COVID-19 disease is evaluated. Accuracy, specificity, accuracy, sensitivity, negative-prevalence and positive-prevalence, ROC-AUC-Score, mislabeling and execution time measures were used for these algorithms. A first model comprising four algorithms was conducted, and then a second classification model was made, which also included four algorithms, on the same global dataset to test the largest number of algorithms to choose the best. Table 5 shows the performance of the algorithms used on the test set, while Tables 6 show the performance of the algorithms in terms of execution time and misclassification. After making a comparison of the algorithms in Tables 5 and 6, SGD the best performance among the classification algorithms used when we chose classification is based on the labeling of the ICU. While KNN was the worst performance among the algorithms in those works.

TABLE V
COMPARISN OF ALGORITHMS PERFORMANCE

measurements unit	SGD	DT	RF	XGB	SVM	NB	LR	KNN
Accuracy	99.61	99	98.4	98.4	97	96.6	94.8	88.4
Sensitivity	100	100	95.4	90	85	90.1	80.3	45
Specificity	97.43	94.8	94.9	97.4	94.9	89.7	97.4	100
Precision	95.23	90.9	90.5	94.7	89.5	81.8	93.3	100
Negative prevalence	100	100	97.4	95	92.5	94.6	86.4	78
Positive prevalence	99.61	96.6	98.4	94.9	91.5	96.6	94.8	81.4
Roc-auc-score	98.71	97.4	94.9	93.7	89.9	89.9	83.7	72.5

TABLE VI
COMPARISN OF EXECUTION TIME AND MISLABELING

Measurements unit	SGD	DT	RF	XGB	SVM	NB	LR	KNN
Execution Time	0.01	0.02	0.2	0.18	0.03	0.04	0.7	0.01
mislabeling	0.39	1	1.63	1.63	3	3.43	5.18	11.6

V. CONCLUSIONS

The COVID-19 epidemic a become a major disquieted for countries worldwide and has affected millions of people, has become represented a major economic and social challenge, and poses a serious threat to public health, so there is required to classify COVID-19 dataset to forecast the number of injuries when disease outbreak to avoid a collapse the system healthy. Moreover, in cases of COVID-19 with varying severity grades, some require ICU. Therefore, it is necessary to forecast of injuries number that require ICU to

present accurate information to health care institutions and hospitals to accommodate as many COVID-19 patients as possible. This work aims to help professionals understand the spread of COVID-19 around the world. The dataset was obtained from the dataset website at Google. We performed a preprocessing to compensate for the missing values and encoded the categorical data to convert it to numeric, as well as analyzed the data to provide a visualization of it and a feature scaling was made to match the dimensions of the values in the data set to obtain an effective model to speed up the computation process in the modules. The dataset was segmented into a set of trains of 80% and 20% for a set of tests and eight algorithms (LR, GNB, RF, SGD, KNN, SVM, XGBoost, and DT) were used, in the form of two models. The algorithm's performance was evaluated, and the outcome appeared that SGD is the best algorithm for classification with 99% of accuracy, and the execution time was 0.01 sec. SGD was the least.

ACKNOWLEDGMENT

I would like to extend my sincere appreciation and gratitude to team that giving us a dataset of COVID-19 with quality as well as my family and teachers for encouraging and supporting every stage of life.

CONFLICT OF INTEREST

The authors have no conflict of relevant interest to this article.

REFERENCES

[1] Abdulkareem, N.M., et al., "COVID-19 world vaccination progress using machine learning classification algorithms", Qubahan Academic Journal, vol. 1, no. 2, pp.100-105, 2021.

[2] Abdulqader, D.M., A.M. Abdulazeez, and D.Q. Zeebaree, "Machine learning supervised algorithms of gene selection: A review", Machine Learning, vol. 62, no. 3, 2020.

[3] Alsharif, M., et al., "Artificial intelligence technology for diagnosing COVID-19 cases: A review of substantial issues", Eur. Rev. Med. Pharmacol. Sci, vol. 24, pp.9226-9233, 2021.

[4] Siddique, S. and J.C. Chow, "Machine learning in healthcare communication. Encyclopedia", vol. 1, no. 1, pp.220-239, 2021.

[5] Kumar, A., P.K. Gupta, and A. Srivastava, "A review of modern technologies for tackling COVID-19 pandemic", Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 4, pp.569-573, 2020.

[6] Bao, Y., et al., "2019-nCoV epidemic: address mental health care to empower society", The Lancet, vol. 395, no. 10224, pp.e37-e38, 2020.

[7] Sarwar, A., et al., "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model", International Journal of Information Technology, vol. 12, no. 2, pp.419-428, 2020.

[8] Du, L., et al., "The spike protein of SARS-CoV—a target for vaccine and therapeutic development", Nature Reviews Microbiology, vol. 7, no. 3, pp.226-236, 2009.

[9] Bayat, V., et al., "A severe acute respiratory syndrome coronavirus 2 (sars-cov-2) prediction model from standard laboratory tests", Clinical Infectious Diseases, vol. 73, no. 9, pp.e2901-e2907, 2021.

[10] Zhou, Y., et al., "A new predictor of disease severity in patients with COVID-19 in Wuhan", China. MedRxiv, 2020.

[11] Atiyah, O.S. and S.H. Thalij, "Evaluation of COVID-19 Cases based on Classification Algorithms in Machine Learning", Webology, vol. 19, no. 1, 2022.

[12] O. S. Atiyah, S.H.T., "Using Classification Algorithms in Machine Learning for COVID-19 Cases Diagnosis", International Journal of Mechanical Engineering, vol. 7, no. 1, pp.6472-6478, 2022.

[13] https://www.kaggle.com / S % C3 % ADrio-Libanes/covid19.

[14] Mahboob, T., S. Irfan, and A. Karamat, "A machine learning approach for student assessment in E-learning using Quinlan's C4. 5", Naive Bayes and Random Forest algorithms. in 2016 19th International Multi-Topic Conference (INMIC). 2016.

[15] Bhandari, S., et al., "Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters", Ibnosina Journal of Medicine and Biomedical Sciences, vol. 12, no. 2, 2020.

[16] LeCun, Y.A., et al., Efficient backprop, in Neural networks: Tricks of the trade. 2012, Springer. p. 9-48.

[17] Samuel, J., et al., "Covid-19 public sentiment insights and machine learning for tweets classification", Information, vol. 11, no. 6, 2020.

[18] Wang, Y., et al., "A Comparative Assessment of Credit Risk Model Based on Machine Learning a case study of bank loan data", Procedia Computer Science, vol. 174, pp.141-149, 2020.

[19] Noori, N.A. and A.A. Yassin, "Towards for Designing Intelligent Health Care System Based on Machine Learning", Iraqi Journal for Electrical & Electronic Engineering, vol. 17, no. 2, 2021.

[20] Lodhi, H., et al., "Text classification using string kernels", in NIPS. 2000.

[21] Xu, Y., et al., "Research on a mixed gas classification algorithm based on extreme random tree", Applied Sciences, vol. 9, no. 9, 2019.

[22] Piryonesi, S.M. and T.E. El-Diraby, "Role of data analytics in infrastructure asset management: Overcoming data size and quality problems", Journal of Transportation Engineering, Part B: Pavements, vol. 146, no. 2, 2020.

[23] Rokach, L. and O. Maimon, "Top-down induction of decision trees classifiers-a survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 35, no. 4, pp.476-487, 2005.