Evaluation of COVID-19 Cases based on Classification Algorithms in Machine Learning

Oqbah Salim Atiyah
College of Computer Science, University of Tikrit, Iraq.
E-mail: oqbah.s.atiyah35529@st.tu.edu.iq

Saadi Hamad Thalij
PHD. College of Computer Science, University of Tikrit, Iraq.
E-mail: saadi.s.alluhaibi@tu.edu.iq

Received September 29, 2021; Accepted December 21, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19326

Abstract

COVID-19 has appeared in china, spread rapidly the world wide and caused with many injuries, deaths between humans. It is possible to avoid the spread of the disease or reduce its spread with the machine learning and the diagnostic techniques, where the use classification algorithms are one of the fundamental issues for prediction and decision-making to help of the early detection, diagnose COVID-19 cases and identify dangerous cases that need admit Intensive Care Unit to provide treatment in a timely manner. In this paper, we use the machine learning algorithms to classify the COVID-19 cases, the dataset got from dataset search on google and used four algorithms, as (Logistic Regression, Naive Bayes, Random Forest, Stochastic Gradient Descent), the result of algorithms accuracy was 94.82%, 96.57%, 98.37%, 99.61% respectively and the execution time of each algorithm were 0.7s, 0.04s, 0.20s,0.02s respectively, and with the mislabeling Stochastic Gradient Descent algorithm was better.

Keywords

COVID-19, Novel Coronavirus, Diagnosis, Prediction, Machine Learning, Classification Algorithm.

Introduction

Covid-19 (Coronavirus) a new pandemic appeared at the end of 2019 in Wuhan city of the China, very quickly this pandemic spread to the world wide, it become concern for countries of the world due caused many injures, deaths (Abdulqader, Abdulazeez, & Zeebaree, 2020). World Health Organization (WHO) announced the status of emergency

after spreading the COVID-19 in most world, this need implementing strict procedures of governments to control or reduce the dangerous of pandemic. The virus disperses when infected person comes contact with healthy person or by the respiratory system (Mengist et al., 2021), the syndrome of appear in period of (2-14) days on injured persons based to the data WHO, the syndrome in state mild and moderate of patients: fatigue, dry cough, fever, and at the severe state suffocation, shortness of breath, fever, and tiredness (Chen et al., 2020). With the deployment of COVID-19 quickly there are require for use machine learning to assist in early detection about COVID-19 disease to avoid spread it. In the healthcare the machine learning very important by used for collected big data of COVID-19 patients and analyzed utilizing the sophisticated machine learning algorithms to better learn of the way that COVID-19 is spreading, and develop the speed, accuracy prognosis, and possibly find those most infection vulnerable based on a personal nature of physiological and genetic characteristics, and improve efficient treatment approaches (Alsharif et al., 2020; Siddique & Chow, 2021). Machine learning can develop and learn automatically depend on knowledge and experience without programmed obviously. algorithms basically based on features. A massive quantity and complex of data can be optimized by utilizing ML- techniques, it utilized for find out the epidemics and for prediction of diseases (Abdulkareem, Abdulazeez, Zeebaree, & Hasan, 2021). Machine learning helpful for classification, examining, forecasting and diagnose the diseases, ML algorithms has ability to forecast with number confirmed of infections possible of COVID-19 and number deaths possible in the future (Kumar, Gupta, & Srivastava, 2020). In this paper we use machine learning algorithms to classify COVID-19 cases need to Intensive Care Unit (ICU) or not, also measurement the algorithms performance of mislabeling for negative false and false positive, to determine the better ones in speed and accuracy of diagnostic, to help the doctors for recognize COVID-19 cases and prevent errors.

Literature Review

Now the world fear of the COVID-19 that It is related with fast growth of fundamental data and there is require to analyze the relations and hierarchy among data that lead to need to the machine learning techniques of the health system to diagnosis, prevention, and treatment of COVID-19 (Shi et al., 2020). This paper progress analyzes the recent studies of this area, determining the algorithm of most efficient with highest performance.

Iweendi et al. (Iwendi et al., 2020) proposed a Fine-Tuned of Random Forest algorithm, also to Adaboost algorithm. To forecast of the probable result, the model utilizes the COVID-19 patient spatial, demographic details, and health. The system has 0.86 F1 Score

and an accuracy rate 94%. The data review indicates a strong correlation among patient gender and cases of death, as well the patient's majority are between 20-70 ages.

Sarwar et al. (Sarwar, Ali, Manhas, & Sharma, 2020) prognosis diabetes utilizing the machine learning techniques, the result referred the ensemble technique 98.60% assured accuracy. These can be useful to prognosis and forecast COVID-19 exact and Firm prognosis of COVID-19 can salvage millions lives of human and can output a big of data for train the machine learning algorithms. ML may present beneficial input in this field, in particular of making prognosis depend on clinical text, Images and radiography etc.

Bayat et al. (Bayat et al., 2021) evolved a model to forecast COVID-19 depend on tests of standard laboratory. A big dataset including of 75,991 patients was acquired of the US department in the Veterans Affairs, used XGBoost to build the system, the results 86.8% specificity, 82.4% sensitivity and accuracy 86.4%. This study found the features of top 10 in descending of importance.

Tordjman et al. (Tordjman et al., 2020) utilized a dataset including 400 patients from three various hospitals in France: Ambroise Par´e Hospital, Cochin Hospital, and Raymond Poincar´e Hospital, this study used a binary Logistic Regression to construction the scoring model to forecast the likelihood in the positive COVID-19 prognosis. The system attained AUC of 88.9%, 80.3% in sensitivity, and 92.3% in positive predictive value.

Tschoellitsch et al. (Tschoellitsch, Dünser, Böck, Schwarzbauer, & Meier, 2021) evolved a model utilizing a Random Forest algorithm to forecast the prognosis of COVID-19 depend on routine tests of blood, dataset including 1528 patients was use to construction the model, which attained 81% in accuracy, 0.74 in AUC, 60% in sensitivity and 82% in specificity. This study found the features of most important in forecasting prognosis were: distribution of red blood cell, leukocyte count, serum calcium, and hemoglobin.

Zhou et al. (Zhou et al., 2020) evolved a model to forecast the severity of injured in COVID-19 patients. This study utilized a dataset including 377 patients (106 non-severe, 172 severe) from the hospital of Wuhan, the LR algorithm was used to build the forecasting model, which got 87.9% in AUC, 73.7% in specificity, and 88.6% in sensitivity. The results found three independent elements were related with severity with COVID-19 patients: age, C-reactive protein, and D-dimer.

Research Method

Figure 1 shows the main phases of the methodology used in this work, accuracy, sensitivity, specificity, precision, negative prevalence, positive prevalence,

ROC_AUC_Score, and execution time are utilized as performance measures. The python (notebook) was use for process the results to construct classification models (containing dataset preparation, preprocessing, data analysis, features selection, splitting data and classification) based on four classification algorithms, Logistic Regression (LR) Naive Bayes (BN), Random Forest (RF), Stochastic Gradient Descent (SGD).

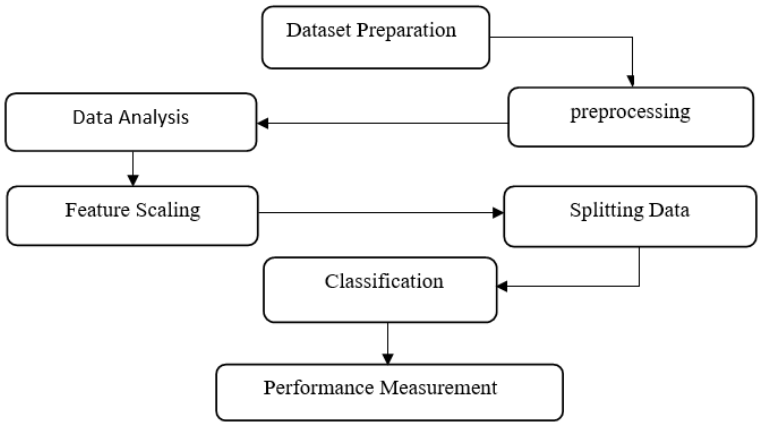


Figure 1 Show the Method Diagram

Dataset Preparation

To classify the COVID-19 cases, we have obtained on the data from dataset search engine on google is an open-source repository that includes most detailed and suitable information on COVID-19, which the dataset xlxl file format contain 1925 column and 231 row (team, 2020).

Preprocessing

There are 1925 instances and 231 features of the dataset that used for this paper. Prior to implement the model, the data is enhanced with a better way to address data for consistency necessities. preprocessing has two main ways: handle missing values and categorical data encoding.

Data Analysis

Data analysis is modeling data and visualizing to find useful information to get conclusion for making decisions.

Feature Scaling

The widely dimensions of the dataset and the variation of inputs make of difficult to extract the data therefore the values dimensions of the dataset must be compatible to

obtained efficient model, and speeding up the computation in the algorithms, therefore, standardization used with standard deviation, mean becomes zero for features, the standard deviation is one for resulting distribution.

Dataset Splitting

Before implement the classification algorithms, we splitting the dataset to the 20% for testing set and 80% for training set.

Classification Algorithms

There are several classification algorithms of the ML. In this paper we used the following algorithm as, Naïve Bayes (NB), Logistic regression (LR), Random Forests (RF), Stochastic Gradient Descent (SGD).

Naïve Bayes

Naïve Bayes (NB) present a method to forecast the likelihood of different class depend on different features. This algorithm is usually utilized in text categorization and when handle to multiclass problems (Mahboob, Irfan, & Karamat, 2016).

Logistic Regression

Logistic regression (LR) is a supervised learning of classification algorithm that uses to forecast a probability of target variant. due to the character of goal or the dependent variant is dichotomous, this means will be two potential classes, the Logistic Regression is utilized as the interplay among various sets of a predictor variables and categorical result variant (Bhandari et al., 2020).

Random Forests

Random Forests (FR) is a supervised learning utilized in regression and classification, it's collecting big amounts of decision trees for training the data, and utilizes an instrument known packing to classification, every decision tree indicates to class prediction, this method gathering of votes at the decision trees, final class has more votes (Wang, Zhang, Lu, & Yu, 2020).

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) model is a very efficient way and simple and appropriate to the linear classifiers in accordance with convex loss functions like linear.

the SGD is applying in scattered and large-scale machine learning problem successfully to often in processing the natural language and classify the text (Bottou, 2012). The classifier implements regularized linear models in the SGD, the loss gradient is appreciated for every sample simultaneously and the module is updated through learning rate (Samuel, Ali, Rahman, Esawi, & Samuel, 2020).

Results and Discussion

This section offers the dataset, learning models, the algorithms used to classify COVID-19 cases, performance measures. Python (Notebook) is utilized to process the outcomes.

Describing Data Results

Describing data is important step in the collection phase of data, it enables of visualize the things by display the variants of the dataset that utilized. The table 1 presents a description for the variants of the dataset.

Table 1 Description of Dataset

Attribute	Type	Description
PATIENT_VISIT_IDENTIFIER	int64	The patient visit identifier of hospital
AGE_ABOVE65	int64	patients age who above 65 years
AGE_PERCENTIL	Object	percentile for patients age
GENDER	int64	gender the patient
DISEASE GROUPING	float64	six disease groups has available features with unknown information
	.....	
	.....	
RESPIRATORY_RATE_DIFF_REL	float64	Available features about respiratory rate relative diff
TEMPERATURE_DIFF_REL	float64	Available features about temperature relative diff
OXYGEN_SATURATION_DIFF_REL	float64	Available features about oxygen saturation relative diff
WINDOW	Object	WINDOW are five type of groups each one contain time hours of the admission
ICU	int64	Response attribute (0= not ICU admission and 1 = ICU admission)

Pre-Processing Results

Perform a statistical analysis and process the data. The output for any stage are input of the next stage, therefore the data must prepare with the same specifics. In this stage will handle missing values, if missing values are numeric will replaced with column rate mean, or if it is nominal will replaced with neighbor value, the dataset ready now for the next stage.

Data Analysis Results

In data analysis are summarized and interpreted the data assembled through analytical and logical reasoning to determine models, relationships, trends and prescribe preprocessed data to realize the characteristics. Table 2 show total patients. Table 3 show the total patients of age distribution. Table 4 show the age distribution of ICU admitted patients.

Table 2 Total patients

Total Patients after preprocessing	**293**
not admitted to ICU	188
admitted to ICU	105

Table 3 Total Patients Age Distribution

Age Distribution	
Patients below age 65	172
Patients above age 65	121

Table 4 Age Distribution of ICU Admitted patients

Age Distribution	
Patients below age 65:	45
Patients above age 65:	60

Classification Results

The classification algorithms performance of LR, GNB, RF, SGD on COVID-19 is then evaluated. The accuracy, sensitivity, specificity, precision, negative prevalence, positive prevalence, ROC_AUC_Score, and execution time are used to measures the performance. Figure 2 offer the execution time to all algorithms that used. Figure 3 offers the classification results for the dataset. Figure 4 offers the mislabeling for all algorithm used.

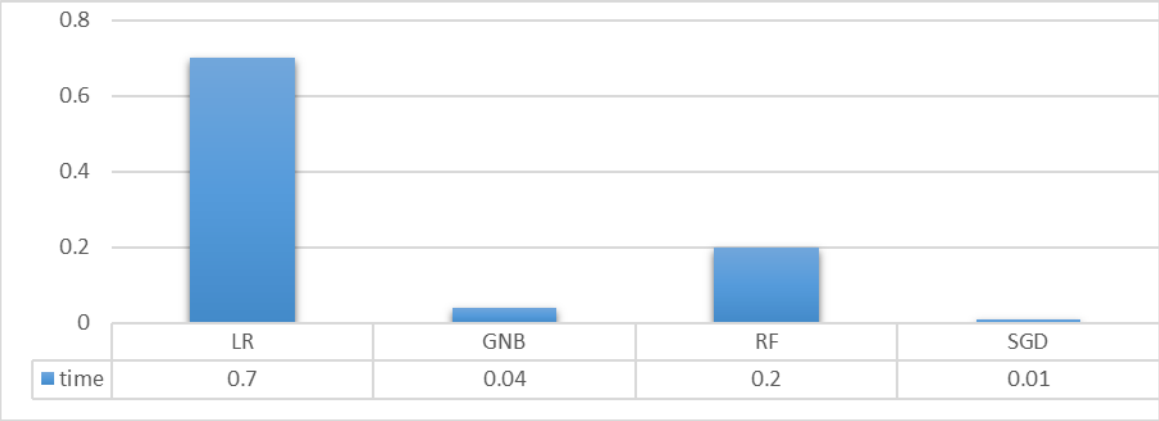
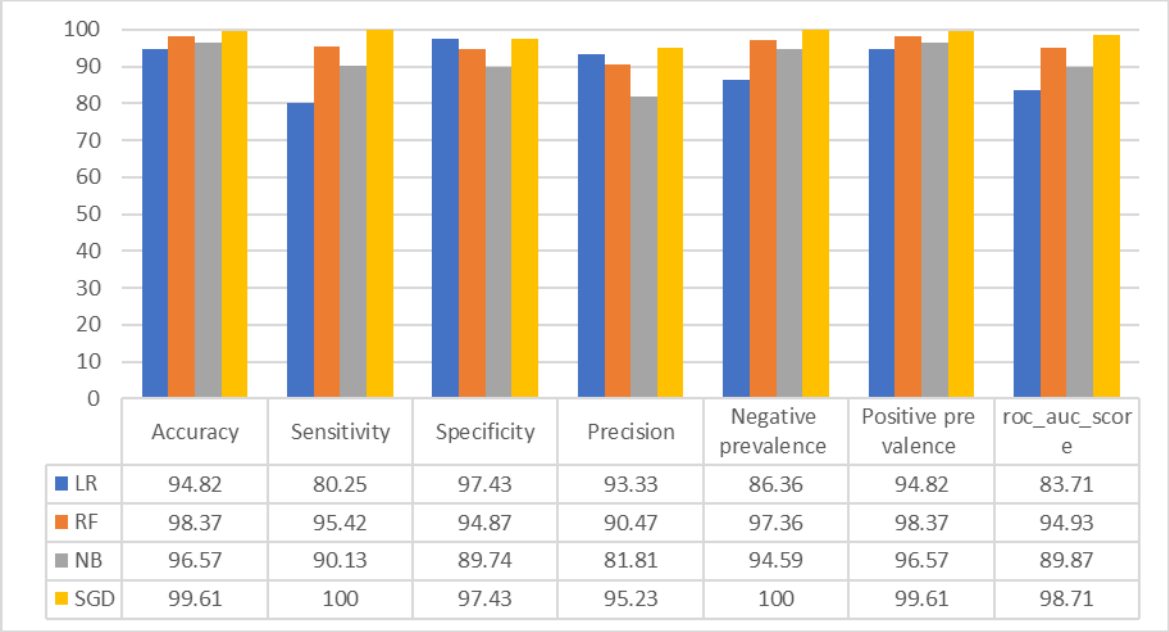
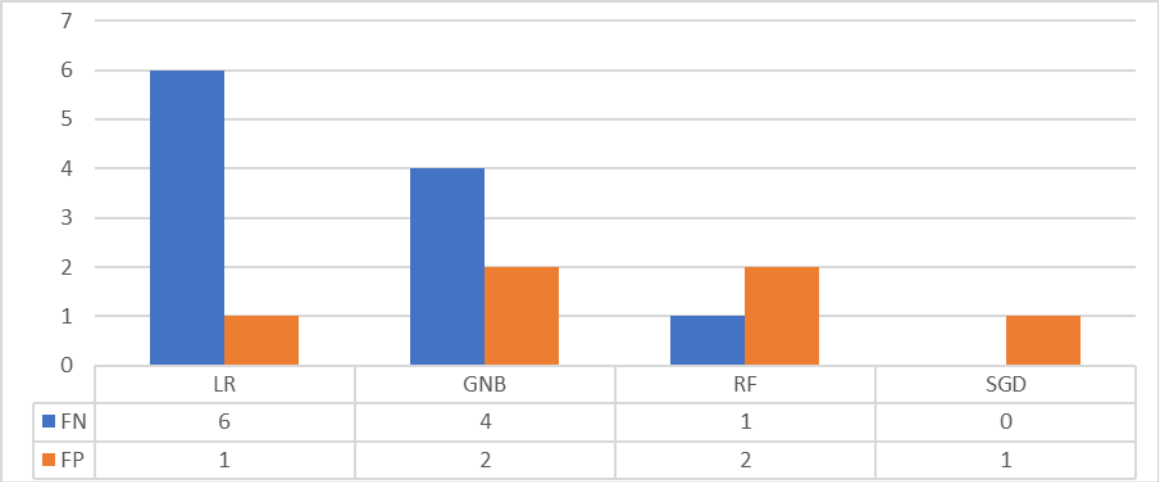


Figure 2 Show Execution Time



	Accuracy	Sensitivity	Specificity	Precision	Negative prevalence	Positive pre valence	roc_auc_scor e
LR	94.82	80.25	97.43	93.33	86.36	94.82	83.71
RF	98.37	95.42	94.87	90.47	97.36	98.37	94.93
NB	96.57	90.13	89.74	81.81	94.59	96.57	89.87
SGD	99.61	100	97.43	95.23	100	99.61	98.71

Figure 3 Show Algorithms Perfomance



	LR	GNB	RF	SGD
FN	6	4	1	0
FP	1	2	2	1

Figure 4 Show The Mislabeling

Conclusion

COVID-19 pandemic affects millions of human on the world wide as a major concern of public health to predict on the COVID-19 pandemic, we'll need an efficient to classify the COVID-19 cases. Therefore, we retrieved the COVID-19 cases datasets processed this data using four algorithms (Logistic Regression, Naive Bayes, Random Forest, Stochastic Gradient Descent). The performance evaluation is important; it helps to select the suitable algorithm for prediction activities. The results after apply algorithms on the dataset and compute the accuracy for all classification algorithms appears the Stochastic Gradient

Descent that best algorithm with accuracy of 99% and execution time is 0.01 seconds, and with mislabeling SGD was better.

References

Abdulkareem, N.M., Abdulazeez, A.M., Zeebaree, D.Q., & Hasan, D.A. (2021). COVID-19 world vaccination progress using machine learning classification algorithms. *Qubahan Academic Journal, 1*(2), 100-105.

Abdulqader, D.M., Abdulazeez, A.M., & Zeebaree, D Q. (2020). Machine learning supervised algorithms of gene selection: A review. *Machine learning, 62*(03).

Alsharif, M., Alsharif, Y., Chaudhry, S., Albreem, M., Jahid, A., & Hwang, E. (2020). Artificial intelligence technology for diagnosing COVID-19 cases: a review of substantial issues. *European Review for Medical and Pharmacological Sciences, 24*(17), 9226-9233.

Bayat, V., Phelps, S., Ryono, R., Lee, C., Parekh, H., Mewton, J., & Holodniy, M. (2021). A severe acute respiratory syndrome coronavirus 2 (sars-cov-2) prediction model from standard laboratory tests. *Clinical Infectious Diseases, 73*(9), e2901-e2907.

Bhandari, S., Shaktawat, A.S., Tak, A., Patel, B., Shukla, J., Singhal, S., & Dube, A. (2020). Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. *Ibnosina Journal of Medicine and Biomedical Sciences, 12*(2).

Bottou, L. (2012). Stochastic gradient descent tricks. *In Neural networks: Tricks of the trade*, 421-436.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., & Wei, Y. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet, 395*(10223), 507-513.

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., & Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in public health, 8*.

Kumar, A., Gupta, P.K., & Srivastava, A. (2020). A review of modern technologies for tackling COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14*(4), 569-573.

Mahboob, T., Irfan, S., & Karamat, A. (2016). A machine learning approach for student assessment in E-learning using Quinlan's C4. 5, Naive Bayes and Random Forest algorithms. *In 19th International Multi-Topic Conference (INMIC),* 1-8.

Mengist, H.M., Kombe, A.J.K., Mekonnen, D., Abebaw, A., Getachew, M., & Jin, T. (2021). Mutations of SARS-CoV-2 spike protein: Implications on immune evasion and vaccine-induced immunity. *In Seminars in immunology, 55.*

Samuel, J., Ali, G., Rahman, M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information, 11*(6).

Sarwar, A., Ali, M., Manhas, J., & Sharma, V. (2020). Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *International Journal of Information Technology, 12*(2), 419-428.

Shi, L., Lu, Z.A., Que, J.Y., Huang, X.L., Liu, L., Ran, M.-S., & Sun, Y.K. (2020). Prevalence of and risk factors associated with mental health symptoms among the general population in China during the coronavirus disease 2019 pandemic. *JAMA network open, 3*(7), e2014053-e2014053.

Siddique, S., & Chow, J.C. (2021). Machine learning in healthcare communication. *Encyclopedia, 1*(1), 220-239.

Team, D.I. (Producer). (2020). *COVID-19 - Clinical Data to assess diagnosis*. https://datasetsearch.research.google.com/search?query=S%C3%ADrio-Liban%C3%AAs

Tordjman, M., Mekki, A., Mali, R. D., Saab, I., Chassagnon, G., Guillo, E., & Madelin, G. (2020). Pre-test probability for SARS-Cov-2-related infection score: The PARIS score. *PloS one, 15*(12).

Tschoellitsch, T., Dünser, M., Böck, C., Schwarzbauer, K., & Meier, J. (2021). Machine learning prediction of sars-cov-2 polymerase chain reaction results with routine blood tests. *Laboratory medicine, 52*(2), 146-149.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. *Procedia Computer Science, 174*, 141-149.

Zhou, Y., Yang, Z., Guo, Y., Geng, S., Gao, S., Ye, S., & Wang, Y. (2020). A new predictor of disease severity in patients with COVID-19 in Wuhan, China. *medRxiv*.