Open Access
Full Length Article

ORIGINAL ARTICLE

Machine Learning Techniques for Early Classification of Covid-19 Disease in Patients

Aremu TB[1], Oyelakin AM[2], Akanbi MB[3], Sulaimon HF[4], Odeyale KM[5]

[1]Postgraduate Student, Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria
[2]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria
[3]Department of Computer Science, Kwara State Polytechnic, Ilorin, Nigeria
[4]ICT Unit, Al-Hikmah University, Ilorin, Nigeria

ABSTRACT

Coronavirus (COVID-19) is a disease that is caused by the SARS-CoV-2 virus. Patients that are infected by this disease have common symptoms that range from mild to moderate respiratory sickness. Machine learning (ML) techniques have been identified to be very promising for the identification of COVID-19 evidence in patients. Some of the past ML-based studies focused more on the use of clinical images for COVID-19 disease classification. Thus, this work used identified medical symptoms in the chosen dataset for the classification of the disease. The study specifically seeks to investigate the performances of two ensemble learning models when the dataset is pre-processed and promising features are used to train the models. Exploratory analysis was carried out on the dataset with a view to understanding the patterns better. Then, a categorical variable in the dataset was encoded and promising features were selected with the aid of feature importance method. Thereafter, Random forest and AdaBoost algorithms were used to build COVID-19 classification models from the dataset. The results showed that the two ensemble models performed better when a filter-based feature selection technique was used on the dataset compared to when all the features were used for building the COVID-19 classification models. For instance, the RF-based model record an accuracy of 0.89 and 0.96 without and with filter based feature selection respectively. Similarly, the Adaboost-based model recorded an accuracy of 0.90 and 0.97 without and with feature sub-set selection, respectively. Thus, the study established that the feature selection method used was able to achieve improved predictive accuracy.

Keywords: Medical diagnosis, Machine Learning techniques, Disease Classification, Detection Accuracy

Author`s Contribution
[1,2,3] Data analysis, interpretation and manuscript writing, Active participation in data collection,[4,5] Conception, synthesis, planning of research, Interpretation, and discussion

Address of Correspondence
Oyelakin AM
Email: amoyelakin@alhikmah.edu.ng

Article info.
Received: March 26, 2022
Accepted: November 21, 2022
Published: December 30,2022

Cite this article: *Aremu TB, Oyelakin AM, Akanbi MB, Sulaimon HF, Odeyale KM. Machine Learning Techniques for Early Classification of COVID-19 Diseases in Patients. J. inf. commun. technol. robot. appl.2022; 13(1):8-14.*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

COVID-19 disease has been ravaging countries of the world for more than two years now. The disease is caused by the SARS-CoV-2 virus.[1] The symptoms that are exhibited by patients carrying the virus include mild and moderate respiratory illnesses[1,2;3] World Health Organization has claimed that as at November 11 2022, there are 630,832,131 cases of the disease and 6,584,104 deaths have been reported.[3]

As a way to explain the various ways in which technology can be used for curbing the disease, Authors in[4] described about ten technologies that can help in the fight against coronavirus fight disease. On the list, Artificial Intelligence

(AI) techniques took the lead. As a sub-filed of AI, machine learning (ML) techniques are very promising. The paper argued that as the COVID-19 disease keep evolving, the various technological applications and initiatives are coming up with a view to stopping the spread of the disease among global people. This study is in support of the fact that AI and machine learning techniques can be used to support clinical decisions early enough. Thus, it is argued herein that the use of a dataset that has symptoms for the identification of the corona virus disease is a step in the right direction.

Many past studies have pointed out that the use of ML methods can promote accurate and early detection of COVID-19 disease.[5,6,7,8] Apart from this, a study by9 argued that ML approaches are very promising for identifying COVID-19 patterns in the evolving coronavirus disease datasets. The study specifically pointed out that with the growing of the disease globally, machine learning techniques can aid in early diagnosis of the disease in infected patients.

On the use of various ML methods for early detection of other diseases, there have been several studies published. For instance, in recent times researchers have built ML-based approaches for the early detection of diseases such as lung cancer,[10-12], cervical cancer by[13], COVID-19 disease detection by [5,6,7,8,10] and many others. Unlike some of the past studies that used clinical images (X-rays and CT scans) for COVID-19 disease classification, this work employed a dataset that contains medical symptoms for the ML-based coronavirus disease classification. The study first of all carried out exploratory data analysis so as to understand the patterns in the dataset better. Then, focus was shifted to building two ensemble-based models that can be used for the early detection of coronavirus disease in patients based on the identified symptoms in the dataset. Emphasis is on investigating how a filter-based feature sub-set selection will impact on the performances of the proposed ML models in the detection of the disease. The two algorithms chosen for the disease classification are Random Forest (RF) and AdaBoost algorithms. The features in the pre-processed dataset were fed into the algorithms to build the disease classification models.

LITERATURE REVIEW

Authors in[15] built a ML model that is based on probabilistic distribution. The work involves classification and prediction with the use of most important CT images features that are peculiar to Coronavirus. The authors argued that a combination of statistical and machine learning techniques were applied for feature extraction from the CT images and selected features were classified using hybridised stack classification technique. Writers in[16] proposed a ML-based technique for the prediction of criticality in patients that have severe Covid-19 infection using three clinical features. The study is a machine learning-based prognostic model with clinical data in Wuhan. The study was interested in predicting the mortality risk in patients before they become seriously sick. The authors argued that their model claimed that the approach is of great clinical significance.

Moreso, authors in[5] proposed a model for the classification of corona virus disease by making use of X-ray images and Deep Convolutional Neural Networks. Authors used three different convolutional neural network approaches namely: ResNet50, InceptionV3 and Inception for the detection of corona virus pneumonia infected patient using chest X-ray radiographs. It was reported that ResNet50 model provided the highest classification performance with 98% accuracy. Researchers in[8] carried out a study focused on the prediction of the epidemic peak for COVID-19 in Japan. The study used the real-time data from 15 January to 29 February 2020. They took the uncertainty due to the incomplete identification of infective population, into consideration by applying the well-known SEIR compartmental model for the prediction. The study was able to predict the reproduction level of the epidemic in the country from nearest time to come using the model.

Authors in[17] analysed, modeled and forecasted COVID-19 outbreak in China. The authors provided estimates of the main epidemiological parameter of corona virus. These parameters include: such as the basic reproduction number (R0) and the infection, recovery and mortality rate. The study then predicted the growth of corona virus epidemic. The study further used data analysis to know the rate of reduction of mortality rate in Hubei, China. The measures are quarantine and

hospitalization of infected individuals. 90% confidence intervals were computed for the three chosen parameters.

Researchers in[5] developed a ML-based approach for predicting Corona virus. The authors focused on algorithmically identifying the combinations of clinical characteristics of COVID-19 that predict COVID-19 outcomes. They then come up with AI-driven model that is capable of carrying out the prediction in patients who may have risk for more severe illness. The emphasis of the work was on the promises that machine learning techniques has for corona virus prediction.

Similarly, writers in[6] predicted the COVID-19 epidemic trend in China and across the world using the machine learning approach. The authors built the models for predicting the daily numbers of cumulative confirmed cases (CCCs), new cases (NCs), and death cases (DCs) of COVID-19 in China based on the data from Jan 20, 2020, to Mar 1, 2020. However, with the growing number of cases of corona virus in the world, the number has surpassed the predicted values. Thus, identifying the patients with the symptoms earlier is a promising research approach.

RESEARCH METHODOLOGY

The basic approaches used in the study include: collection of relevant COVID-19 dataset, carrying out detailed exploratory analysis of the chosen COVID-19 dataset with a view to understanding it better. The EDA revealed the actual patterns in the dataset which further enable the researcher to settle for three linear machine learning algorithms. Two different ensemble learning algorithms, namely: Random Forest (RF) and AdaBoost algorithms were chosen for the classification of the COVID-19 disease. Then, the impact of a filter-based feature selection technique in the classification was investigated. All the experimentations were carried out in Anaconda Python Environment. Split test ratio was used for validating the results of the ML-based COVID-19 classification models.

Dataset Collection

The study first of all collected COID-19 dataset. The dataset used in this study contains the possible symptoms. It was generated based on different guidelines from WHO. The dataset is available at https://www.kaggle.com/datasets/martuza/early-stage-symptoms-of-covid19-patients?resource=download. The author of the dataset termed the features in it as early stage symptoms.

Exploratory Data Analysis

The exploratory Data analysis (EDA) carried out on the dataset revealed the values in table 1, figure 1, table 2 and figure 2.

Description of the COVID-19 Dataset

Table 1: Features and Feature Type

S/N	Column (Attribute)	Number of Non-Null Count	Data type
0	gender	6512	object
1	age_year	6512	int64
2	fever	6512	int64
3	cough	6512	int64
4	runny nose	6512	int64
5	muscle soreness	6512	int64
6	pneumonia	6512	int64
7	diarrhea	6512	int64
8	lung infection	6512	int64
9	travel_history	6512	int64
10	isolation treatment	6512	int64
11	SARS-CoV-2	6512	int64

From the exploratory analysis in this study, the summary of number of instances' and data types in the dataset is as shown in table 1. As shown in table 1, there are numerical and categorical data types in the dataset used in this study.

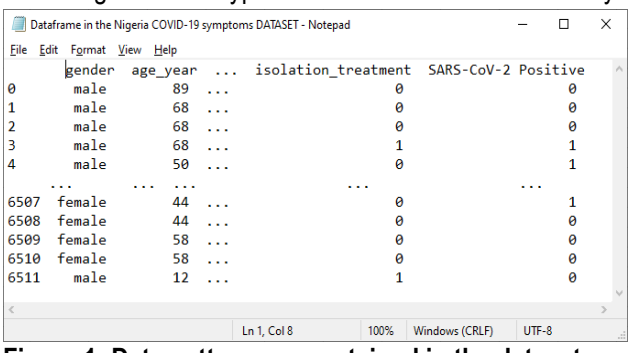

Figure 1: Data patterns as contained in the dataset

Based on the values in table 1 and figure 1, it is also evident that the input features in the dataset include: gender, age year, fever, cough, runny nose, muscle soreness, pneumonia, diarrhea, lung infection, travel history. Isolation treatment, while the target class is SARS-

CoV-2. Furthermore, judging from the data frame obtained in analysis as shown in figure 2, the summary of the description in the chosen dataset is captured in table 1.

Table 2: Dataset Summary	
No of input features	11
No of samples	6512
Missing values	NO
DataTypes(for Input and Output Features)	Mixed

Data Pre-processing

The categorical data (gender) in the dataset was pre-processed using label encoding. This is needed with a view to generating a complete dataset that has been transformed to the usable format for the selected learning classifiers.

Statistical Summary

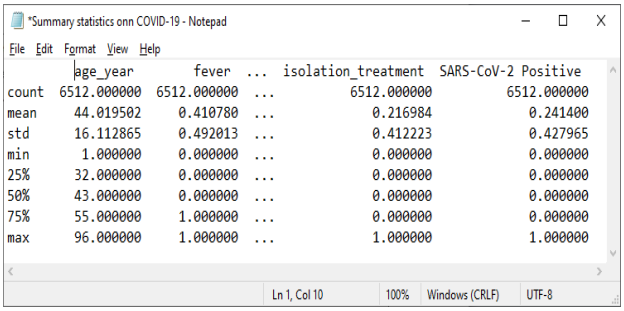

Figure 2. Statistical Summary of the Dataset

The statistical summary provided the information about the individual count, mean, standard deviation, minimum value, 25th percentile, 50 percentile, 75th percentile and maximum value.

Chosen Ensemble Machine Learning Algorithms

The ensemble machine learning algorithms used include Random Forest and AdaBoost Algorithms. Both algorithms are tree-based and were used to learn from the dataset features fed into them.

Feature Selection Approaches

Authors in[18] have argued the importance of selecting promising attributes in ML-based classification models. Feature selection methods are of different categories. In this study, a filter-based feature sub-section method called feature importance is chosen. In all, scoring of the attribute is derived from the feature importance provided by the base classifiers (Decision Trees) in both ensembles. That is, the feature importance in both RF and Adaboost-based models is determined by the average feature importance provided by each decision tree used as base classifier. The feature selection techniques selected a subset of the features that were used for building the ensemble COVID-19 classification models. The feature selection approaches arrived at seven attributes out of the eleven input features in the dataset.

Model Building

Based on the patterns of the target class in the dataset, it is evident that the problem is a binary classification one. In each of the machine-learning based COVID-19 identification approaches, the dataset was cleaned and selected subset features were fed into the two algorithms for the detection of COVID-19 evidence. The rate at which the two models can detect COVID-19 symptoms was evaluated using accuracy, precision, recall and f1-score as metrics. Decision trees are the weak classifiers that are used to build the ensembles. Also, the test split ratios of 85% for the training set and 15% for the testing set were settled for.

(i) Random Forest Algorithm for COVID-19 Classification

Random Forest algorithm is built based on base classifiers. It uses bagging to select different variations of the training data with a view to ensuring that the decision trees are different.

Algorithm 1: Algorithm for Random Forest-based COVID-19 Classification model

(a) Input: COVID-19 dataset containing basic symptoms
(b) Output: COVID-19 classification based on Random Forest Algorithm runs
1. Load the disease dataset in csv format
2. Select samples randomly from the given dataset.
3. Designt a decision tree for every sample.
4. Compute the prediction result from every decision tree.
5. obtain voting for every predicted result
6. Select the most voted prediction result as the final prediction result
7. Stop

(ii) AdaBoost Algorithm for COVID-19 Classification

Adaptive Boosting algorithm is popularly called AdaBoost. It is also a learning algorithm that is built from base learners.

Algorithm 2: Algorithm for AdaBoost-based COVID-19 Classification model

I. Input: COVID-19 dataset containing basic symptoms

II. Output: COVID-19 classification based on AdaBoost Algorithm runs

1. load the dataset
2. initialize weight Di(i)=1/m
3. For t= 1...T

   Refer to a weak learn which returns a weak classifier (DT as base estimator) ht: X ∈ {-1,1} with minimum error w.r.t Di

   Select α ∈ R
4. obtain the classifier from the DT-based weak estimators
5. End For
6. Stop

Algorithms 1 and 2 are used for building the coronavirus disease classification models.

Experimental Results

Results of Data Exploratory Analysis

Summary Statistics of Features in the Dataset

*Summary statistics onn COVID-19 - Notepad

File Edit Format View Help

	age_year	fever	...	isolation_treatment	SARS-CoV-2 Positive
count	6512.000000	6512.000000	...	6512.000000	6512.000000
mean	44.019502	0.410780	...	0.216984	0.241400
std	16.112865	0.492013	...	0.412223	0.427965
min	1.000000	0.000000	...	0.000000	0.000000
25%	32.000000	0.000000	...	0.000000	0.000000
50%	43.000000	0.000000	...	0.000000	0.000000
75%	55.000000	1.000000	...	0.000000	0.000000
max	96.000000	1.000000	...	1.000000	1.000000

Ln 1, Col 10 100% Windows (CRLF) UTF-8

Figure 3. Summary Statistics of the COVID-19 Dataset

Visualizing the features in the Dataset

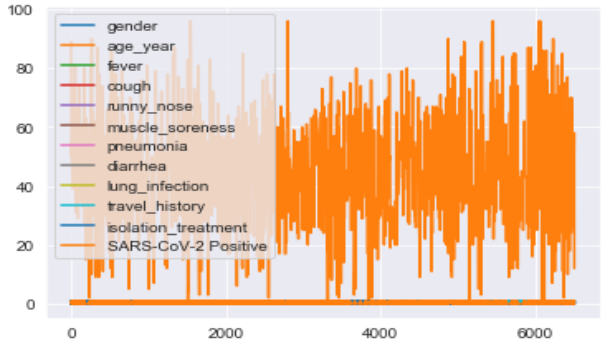

Figure 4: Showing patterns of features in the dataset

The patterns in the dataset are as captured in the figures Was part of the exploratory data analysis provides some insights for a better understanding of the dataset.

Results of Data Pre-processing

Encoding

Aside the target class, the dataset contains a categorical feature. Thus, the categorical attribute in in the dataset was converted to numerical equivalent before being fed in the chosen classification algorithms used for building the model. Label encoding technique was used for the encoding.

Dataset Patterns after pre-processing

New Data frameCOVID - Notepad

File Edit Format View Help

	gender	age_year	...	isolation_treatment	SARS-CoV-2 Positive
0	1	89	...	0	0
1	1	68	...	0	0
2	1	68	...	0	0
3	1	68	...	1	1
4	1	50	...	0	1
...	...	...	...	...	...
6507	0	44	...	0	1
6508	0	44	...	0	0
6509	0	58	...	0	0
6510	0	58	...	0	0
6511	1	12	...	1	0

Ln 1, Col 26 100% Windows (CRLF) UTF-8

Figure 5. Datasets distribution after encoding

Results of Selected Features

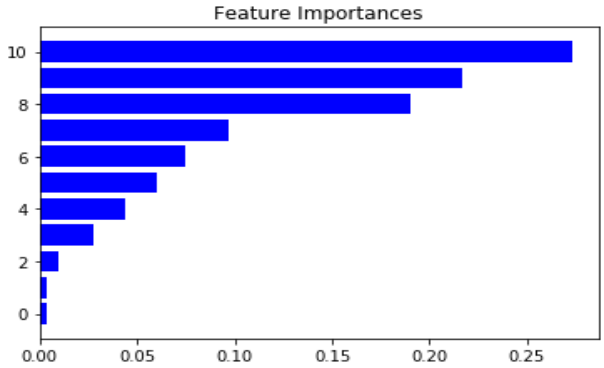

Figure 6. Pictorial representation of scores of features in the dataset.

Sub-set features were selected from the dataset based on the feature importance technique. Seven features were settled based on their scores that are better than the chosen threshold of 0.05.

Results of COVID-19 ML-based Classification

The results are based on the chosen metrics for the evaluation of the ensemble-based COVID-19 classification models. First, results of the models built without considering feature selection were reported in table 2. Thereafter, the results of the models built with feature selection being considered are captured n table 3.

Model Performances

Case One: Ensemble-based COVID-19 Classification Model without Feature Selection

Table 3. RF-based Model performance

Algorithm used	Performance Metrics	Value
Random Forest	Accuracy	0.89
	Precision	0.89
	Recall	0.88
	F-Measure	0.86

Table 4. Adaboost-based Model performance

Algorithm used	Performance Metrics	Value
AdaBoost	Accuracy	0.90
	Precision	0.89
	Recall	0.88
	F-Measure	0.88

Case Two: Ensemble-based COVID-19 Classification Model with Feature Selection

Table 5. RF-based Model performance

Algorithm used	Performance Metrics	Value
Random Forest	Accuracy	0.96
	Precision	0.95
	Recall	0.94
	F-Measure	0.95

Table 6. AdaBoost-based Model performance

Algorithm used	Performance Metrics	Value
AdaBoost	Accuracy	0.97
	Precision	0.96
	Recall	0.96
	F-Measure	0.95

RESULT AND DISCUSSION

The exploratory data analysis carried out in the study revealed the different patterns and features in the dataset. The study specifically seeks to investigate the performances of two ensemble learning models under two cases (scenarios). The first scenario is when all features were used while the second case is when selected features were employed. Since one of the input features in the dataset is a categorical type, the feature was handled through encoding. Thereafter, the study obtained sub-set features from the pre-processed dataset using feature importance. In each of the experimentations, promising results were obtained for the two ensemble algorithms in the classification of COVID-19 evidence in patients. Eighty-five percent (85%) and fifteen percent (15%) of the dataset were used as training and test tests respectively. The results shown in tables 3, 4, 5 and 6 revealed that the two models performed better when the filter-based feature selection technique called feature importance was used compared to when all the features were used for building the classification models for the disease. For instance, the RF-based model records an accuracy of 0.89 and 0.96 without and with filter-based feature selection. Similarly, the Adaboost-based model recorded an accuracy of 0.90 and 0.97 without and with feature sub-set selection. Detailed resources of the disease classification are shown in tables 3, 4, 5 and 6 respectively. Thus, this study further confirmed the importance of feature selection in any machine learning related researches as argued by authors.[18]

CONCLUSION

This study focused on the investigation of how ensemble-based models can be used for the early detection of coronavirus disease in patients based on the identified symptoms in the chosen dataset. The study first of all introduced how ML techniques can be very promising for COVID-19 identification. Then the work used the pre-processed features in the dataset to build two ensembles for the disease identification and then reports their performances. The dataset used was obtained from Kaggle repository and was found relevant for COVID-19 detection studies. The chosen dataset contains symptoms of patients that have suffered from COVID-19 infections. The dataset was pre-processed and the features in two different cases were used to train the models. The two ensemble learning algorithms used for building the coronavirus classification models are: Random Forest and AdaBoost. The study specifically investigated the

performances of the two models when all features in the dataset were used and when selected features were employed. The results revealed that the two models performed better when a filter-based feature selection technique called feature importance was used compared to when all the features were used for building the classification models for the disease.

REFERENCES

1. WHOa, Laboratory testing strategy recommendations for COVID-19 [illegible] 21st March, 2020 retrieved https://apps.who.int/iris/bitstream/handle/10665/331509/WHO-COVID-19-lab_testing-2020.1-eng.pdf
2. WHOb, Coronavirus disease 2019 (COVID-19) Situation Report, published on 11th April, 2020, retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200411-sitrep-82-covid-19.pdf?sfvrsn=74a5d15_2
3. WHO (2022).Overview of Coronavirus (COVID-19) disease, a publication of WHO retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_1|
4. M, Kritikos. Ten Technologies to fight Coronavirus, European Parliamentary Research Service. Scientific Foresight Unit (STOA),2020, 1-20, doi:10.2861/58070
5. Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Analysis and Applications, 24, 1207-1220.
6. Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J & Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 63(1), 537-551. https://doi.org/10.32604/cmc.2020.010691
7. Li, M., Zhang, Z., Jiang, S., Liu, Q., Chen, C., Zhang, Y., & Wang, X. (2020). Predicting the epidemic trend of COVID-19 in China and across the world using the machine learning approach. medRxiv, 2020-03. https://doi.org/10.1101/2020.03.18.20038117
8. Choi S, Ki M. Estimating the reproductive number and the outbreak size of COVID-19 in Korea. Epidemiology and health. 2020 Mar 12;42:e2020011. https://doi.org/10.3390/jcm9030789
9. A. M. Oyelakin., T. T. Salau-Ibrahim., B. S. Ogidan., R. D. Azeez, & I. K. Ajiboye. On the Use of Machine Learning Techniques for Predicting Covid-19 Cases-An Overview of Evolving Datasets and their Promising Features. 2020, Anale. Seria Informatică. Vol. XVIII fasc. 2 - 2020
10. Cai, Z., Yu, Z., Zhou, H., & Gu, Z. (2018, November). The early stage lung cancer prognosis prediction model based on support vector machine. In 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP) (pp. 1-4). IEEE., https://doi.org/10.1109/ICDSP.2018.8631657
11. A. M. Oyelakin., B. Ogidan, H. I. Olufadi. , A. Raheem-Giwa, O. H. Buraimah, D. M. Rilwan.A Study on Lung Cancer Identification Using Extra Trees-Based Model, Journal of Computer Science and Control Systems,2022, 15(1), 9-13, a publication of Oredia University Oredia, Romania, available at https://electroinf.uoradea.ro/images/articles/CERCETARE/Revista/JCSCS/JCSCS_V15_N1_MAY_2022/JCSCS%20VOL%2015%20NO%201%20MAY%202022%20Oyelakin_A_Study.pdf
12. Salau-Ibrahim, T. T., & Rilwan, M. D. Performance Analysis of Selected Machine Learning Algorithms for the Detection of Cervical Cancer Based on Behavioral Risk Dataset. available at: https://library.ncs.org.ng/journal-of-information-security-privacy-and-digital-forensic-volume-5-no-1-june-2021/
13. Olatunde, O. S., Mofiyinfoluwa, O., Akande, O. N., Misra, S., Ahuja, R., Agrawal, A., & Oluranti, J. (2022). Comparison of Selected Algorithms on Breast Cancer Classification. In Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2021 (pp. 161-171). Singapore: Springer Nature Singapore.https://doi.org/10.1007/978-981-19-1111-8_14
14. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y & Zhang, L. (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. https://doi.org/10.1016/S0140-6736(20)30211-7
15. Farid, A. A., Selim, G. I., & Khater, H. A. A. (2020). A novel approach of CT images feature analysis and prediction to screen for corona virus disease (COVID-19). https://doi.org/10.20944/preprints202003.0284.v1
16. Yan, L., Zhang, H. T., Xiao, Y., Wang, M., Guo, Y., Sun, C., ... & Yuan, Y. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv, 2020-02.02.https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2.full.pdf
17. Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PloS one, 15(3), e0230405.https://doi.org/10.1371/journal.pone.0230405
18. Oyelakin, A. M. (2021). A survey of feature extraction and feature selection techniques used in machine learning-based botnet detection schemes. available at https://vfast.org/journals/index.php/VTCS/article/view/604/658
19. Moreton R. A process model for software maintenance. Journal of information technology. 1990 Jun;5(2):100-4.
20. Musa JD. Software Reliability Engineering: More Reliable Software Faster and Cheaper 2nd. Edition.
21. Yang G. Life cycle reliability engineering. John Wiley & Sons; 2007 Feb 2.
22. Dhillon, B. S. (2007). Applied reliability and quality: fundamentals, methods and procedures. Springer Science & Business Media.