International Journal of Multidisciplinary Research and Growth Evaluation.

Investigation study on classification based COVID disease prediction with patient data

A Mary Theresa [1*], **Dr. V Saravanan** [2]

[1] Assistant Professor, Department of Information Technology, Nirmala College for Women, Coimbatore, Tamil Nadu, India

[2] Professor & HEAD, Department of Information Technology, Hindusthan College of Arts and Science, Coimbatore, Tamil Nadu, India

* Corresponding Author: **A Mary Theresa**

Article Info

ISSN (online): 2582-7138
Volume: 03
Issue: 02
March-April 2022
Received: 15-02-2022;
Accepted: 02-02-2022
Page No: 92-97

Abstract

Data mining is the process of discovering essential patterns and extraction of large data. The evolution of medical field technology creates the rapid effect in every field. With the development in medical field, predictive analysis has become an essential component for future disease prediction. Early diagnosis is an essential for accurate treatment to minimize the pressure in the health care system. Coronavirus disease (COVID-19) has created the restriction to the public health and considered as the greatest pandemic in the world history. Coronavirus disease in human is an extremely hard task because of their symptoms. Many researchers introduced feature engineering and classification techniques for medical data analytics. However, the true negative rate and time consumption was not focused through existing techniques. In order to address the problems, medical data analytics via feature engineering and classification is performed using machine learning and deep learning techniques.

Keywords: Data mining, disease prediction, pandemic, feature engineering, classification, medical data analytics

Introduction

World has gained fast progression in technology and it shows an essential part in developed countries. The healthcare center is an essential field that strongly needs to use new technologies from defining the symptoms for accurate diagnosis and digital patient triage. Coronavirus-2 (SARSCoV-2) cause severe respiratory infection and respiratory disorders in Wuhan city of China in December 2019. COVID-19 has become global pandemic because of their rapid spread. It is demanding one to identify the exposed persons as they show disease symptom instantly. It is an essential technique for estimating the number of potentially infected persons on the regular basis to adopt the suitable measures. COVID-19 pandemic increases rapidly around the world. It is the most leading cause of the death and heath disaster in world. COVID-19 detection of severe patients at earlier stage is the great significance to reduce the disease course and mortality.

This paper is organized as follows: Section 2 reviews the corona virus prediction methods. Section 3 explains the existing corona virus prediction methods. Section 4 describes the experimental settings with possible comparison between them. Section 5 discusses the limitation of existing corona virus prediction techniques. Section 6 concludes the paper.

2. Literature Review

Coronavirus disease 2019 (COVID-19) threatens the nation of the world irrespective of health infrastructure conditions. An optimized machine learning framework termed ADAptive SYNthetic (ADASYN) method was introduced in [1] with an inpatient facility data to provide the user-friendly, cost efficient and timely solution. ADASYN method employed Bayesian optimization with hyper parameters of classifier that balanced COVID and non-COVID classes in an efficient manner. Though the accuracy was improved, the classification performance was reduced.

A novel *K*NN variant (*K*NNV) was introduced in [2] with better result through addressing the issues related to incompleteness and heterogeneity. *K*NN variant algorithm was employed for managing the incomplete data by imputation towards heterogeneity by making the conversion of categorical data. But, the accuracy level was not improved by *K*NNV algorithm.

Machine Learning Scheme was designed in [3] for identifying the COVID-19 cases in India. The data analysis was carried out with occurrence of cases in many states of India chronologically by multi-class classification. However, the complexity level of prediction was not minimized by Machine Learning Scheme.

A hybrid method was introduced in [4] for attribute selection through classification process. An optimal attribute subset was generated through genetic algorithm. The classification model increased the prediction accuracy of COVID-19 patients, precision and recall to large extent. The hospitalization need prediction of COVID-19 patients was not reduced by designed method.

A novel ensemble-based classifier method was designed in [5] for predicting the COVID-19 cases at early stage to take the necessary action taken by patients and doctors. A convex hull-based method was introduced to the data for enhancing the accuracy and speed. But, the time complexity was not reduced by designed method.

A machine learning pipeline involving feature selection over sampling and supervised classification was introduced in [6] for predicting the different levels of hospitals infected by the corona virus disease. The hyper parameter optimization grid search algorithm was employed to predict the hospitals consistent with the severity of pandemic disease. However, the computational cost was not reduced by classification method.

A machine learning technique depending on feature selection and classification phase termed Non-dominated Sorting Genetic Algorithm (NSGA-II) was introduced in [7] for forecasting the COVID- 19 patient infection. AdaBoost classifier was introduced for attaining higher sensitivity and specificity. Though the sensitivity was improved, the time complexity was not reduced.

Nature inspired optimization method was introduced in [8] for examining the clinical data with 26 distinct features. An artificial bee colony optimization was carried out with the features for attaining the optimum feature set. But, the accuracy was not improved by designed method. LSTM Recurrent Neural Network was introduced in [9] for extracting the discussion related to the COVID-19 in an automatic manner. The natural language process (NLP) method was employed with topic modeling to address COVID- 19 issues acquired from public opinions. However, the computational cost was not reduced by LSTM Recurrent Neural Network.

A random forest classifier was introduced in [10] for categorizing the medical data in accurate manner. A feature ranking based technique was introduced with the features were ranked employing ranker algorithm. Random Forest classifier was applied with the highly ranked features for enhancing prediction accuracy rate. But, the error rate was not minimized by random forest classifier.

Machine learning classifier with SVM for classification of COVID-19 diseases was introduced in [11]. A modified cuckoo search algorithm was introduced with the hyper parameter optimization technique for enhancing the classification accuracy. A hybrid feature selection model using Minimum Redundancy Maximum Relevance algorithm was introduced to identify the unrelated and misleading features. However, the accuracy level was not reduced by Machine learning classifier.

COVID-19 pandemic has widened throughout the world with foremost mainsprings of death and heath calamity globally. A predictive model was introduced in [12] with an artificial neural network (ANN). With ANN, the deaths due to COVID-19 were analyzed with higher accuracy. But, the time consumption was not minimized by designed model.

3. Coronavirus Disease Prediction

Coronavirus Disease 2019 (COVID-19) is a deadly infection that affects respiratory organs in humans and animals. The disease is turned as pandemic one that affects millions of individuals across the globe. Many tests are conducted for large number of suspects preventing the virus spread. Coronaviruses are the large family of viruses that cause illness in animals or humans. Coronaviruses causes the respiratory infections varying from common cold to the severe diseases like MERS and SARS. COVID-19 pandemic initiated from the Wuhan city has affected health, socio-economic and financial matters of different countries. COVID-19 symptoms are fever, cough, breath shortness, loss of sense and fatigue.

3.1 Novel Bayesian Optimization-based Machine Learning Framework for COVID-19 Detection from Inpatient Facility Data

A machine learning-based framework was introduced with inpatient facility data to provide user-friendly, cost-effective and time-efficient solution. Bayesian optimization framework was introduced to optimize hyper-parameters of classifier and Adaptive Synthetic (ADASYN) algorithm to balance the COVID and non-COVID classes. A fast and user-friendly model was employed to identify the COVID-19 patients with help of machine learning methods. A large quantity of COVID-19 data was collected from different laboratories and test centers. The dataset included the additional features to design the automatic COVID-19 detection model.

Linear Discriminant Analysis (LDA) was the productive classification technique to reform the n-dimensional space structure into 2-dimensional space structure that separated by hyper-plane. LDA was employed to trace the mean function for every class and the function was positioned on directions that optimized between-group variance and minimized within-group variance. The majority of the class was downsampled to the amount minority class. The process minimized the amount of data that cause data inadequacy and loses the COVID information. In designed framework, the first phase was pre-processing the large amount of gathered raw data where raw data was assigned and scaled. The imbalanced data was balanced with help of ADASYN algorithm. The pre-processed data were divided into training and testing set employed through different classifiers to compute the classification performance. The optimized classification model was tested with different performance metrics for the evaluation.

3.2 Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm

Coronavirus disease 2019 (COVID-19) is wreak havoc around the world. COVID-19 caused main threat to the human life with severe economic consequences. A new KNN variant (KNNV) algorithm was introduced for attaining accurate classification results. The rough set theoretic methods were employed to handle incompleteness and heterogeneity for identifying the ideal value for 'K'. KNNV algorithm considered an incomplete and heterogeneous data with medical records of people. KNNV identified the cases with COVID-19 disease. The distance metrics like Euclidean and Mahalanobis were employed to enhance the operational scope. KNN variant (KNNV) algorithm increased COVID-19 case identification within IHC datasets in accurate manner. KNNV classification algorithm selected the parameter 'K' adaptively for every unknown patient.

KNNV determined the distance between patients accurately to recognize the nearest neighbors. It mitigated the incompleteness and the heterogeneity issues through new rigorous rules spelled out in sequel. KNNV treated the distinct features in a different way from numerical features. KNNV employed the rough set theory (RST) methods to handle the categorical features and conventional distance metrics to manage the numerical features. KNNV not converted the categorical feature values into numbers for making all numerical features. The distance metrics were used to identify the nearest neighbors in standard KNN. By using the RST methods to categorical features, KNNV addressed incompleteness of features and vagueness of K value. KNNV was tested on available IHC dataset from Italian Society of Medical and Interventional Radiology (SIRM).

3.3 Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model

A new coronavirus termed SARS-CoV-2 was introduced with an unusual viral pneumonia in patients. It identified late in December 2019 and declared as a pandemic through World Health Organizations due to their fatal effects on public health. The COVID-19 pandemic cases were increasing day by day slowly in world. The COVID-19 cases were confirmed, death and cured cases in India only. The analysis was carried out with the cases occurring in diverse states of India in chronological dates. The designed machine learning scheme was introduced depending on data-driven approach. The approach provided the prediction about number of infected people with COVID-19 in upcoming days with available data. A designed model was employed to predict the count of fresh COVID-19 cases in order that management makes preparation to manage the cases. During the model building process, feature selection was employed to select the relevant features outside all features. Feature selection minimized the complexity of prediction model. The feature selection was carried out with random forest importance algorithm in R programming language. The classification model was determined with input parameters of COVID-19 cases in India. The features were employed for multi-class classification model with random forest importance algorithm. The features were discarded as they impact only at beginning of COVID-19 infection.

4. Performance analysis of coronavirus disease prediction methods

Experimental evaluation of existing coronavirus disease prediction methods is implemented using Java language. The experiment of existing coronavirus disease prediction methods is conducted using Novel Corona Virus 2019 Dataset taken from the Kaggle. The URL of the mentioned dataset is given as https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset. The dataset includes the daily level information on the number of affected cases, deaths and recovery from the 2019 novel corona virus. It is a time series data and number of cases on specified day is the cumulative number. The dataset includes the eight files. Among the eight files, we considered the COVID19_open_line_list file for experimental consideration. The dataset comprises 44 features and 13174 instances. The features are ID, age, sex, city, country, province, etc. Among the features, relevant features are chosen to perform classification for COVID prediction. Result analysis are carried out with existing methods with parameters are,

- Prediction accuracy
- Prediction time and
- Error rate

4.1 Prediction accuracy

Prediction accuracy is described as the ratio of number of patient data that are correctly predicted the risk to the total number of patient data taken. The prediction accuracy (PA) is formulated as,

$$PA = \left(\frac{Number\ of\ patient\ data\ that\ correctly\ predicted\ the\ risk}{Number\ of\ patient\ data}\right) * 100 \quad (1)$$

From (11), the prediction accuracy is determined. The prediction accuracy is computed in terms of percentage (%).

Table 1: Tabulation for Prediction Accuracy

Number of patient data (Number)	Prediction accuracy (%)		
	ADASYN algorithm	KNNV	Machine Learning Scheme
100	89	84	86
200	93	86	88
300	90	83	85
400	88	81	83
500	86	79	81
600	89	82	84
700	91	85	87
800	93	87	89
900	94	90	92
1000	96	93	94

Table 1 explains the COVID disease prediction accuracy with respect to number of data points ranging from 100 to 1000. Prediction accuracy comparison takes place on the existing ADAptive SYNthetic (ADASYN) algorithm, KNN variant (KNNV) algorithm and Machine Learning Scheme. The graphical representation of prediction accuracy is explained in figure 1.

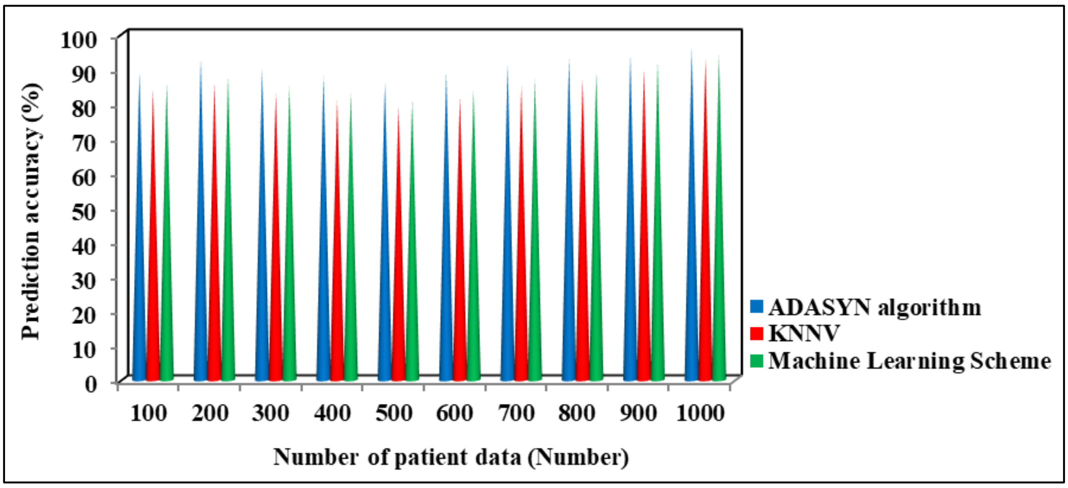

Fig 1: Measurement on Prediction Accuracy

From figure 1, the prediction accuracy for different number of data points is explained. The blue color line denotes the prediction accuracy of ADASYN algorithm. The red colour line and green colour line symbolizes the prediction accuracy of KNN variant (KNNV) algorithm and Machine Learning Scheme correspondingly. It is clear that the prediction accuracy using ADASYN algorithm is higher when compared to the KNN variant (KNNV) algorithm and Machine Learning Scheme. This is due to the application of LDA with productive classification technique to reform the n-dimensional space structure into 2-dimensional space structure divided by the hyperplane. Consequently, prediction accuracy of ADASYN algorithm is increased by 7% when compared to the KNN variant (KNNV) algorithm and 5% when compared to the Machine Learning Scheme.

4.2 Prediction time

Prediction time is described as the amount of time consumed for predicting the risk level of patient data. Prediction time is the multiplication of number of patient data and amount of time consumed for predicting one patient data. Consequently, the prediction time is determined as,

$$PT = Number\ of\ patient\ data * time\ consumed\ for\ predicting\ one\ data \quad (12)$$

From (12), the prediction time is computed. The prediction time is computed in terms of milliseconds (ms).

Table 2: Tabulation for Prediction Time

Number of patient data (Number)	Prediction time (ms)		
	ADASYN algorithm	*K*NNV	Machine Learning Scheme
100	28	21	32
200	31	25	36
300	33	27	39
400	36	30	42
500	39	32	45
600	41	35	48
700	43	38	51
800	46	40	53
900	49	42	56
1000	51	45	59

Table 2 describes the COVID disease prediction time with respect to number of data points varying from 100 to 1000. Prediction time comparison takes place on existing ADAptive SYNthetic (ADASYN) algorithm, KNN variant (KNNV) algorithm and Machine Learning Scheme. The graphical representation of prediction time is described in figure 2.

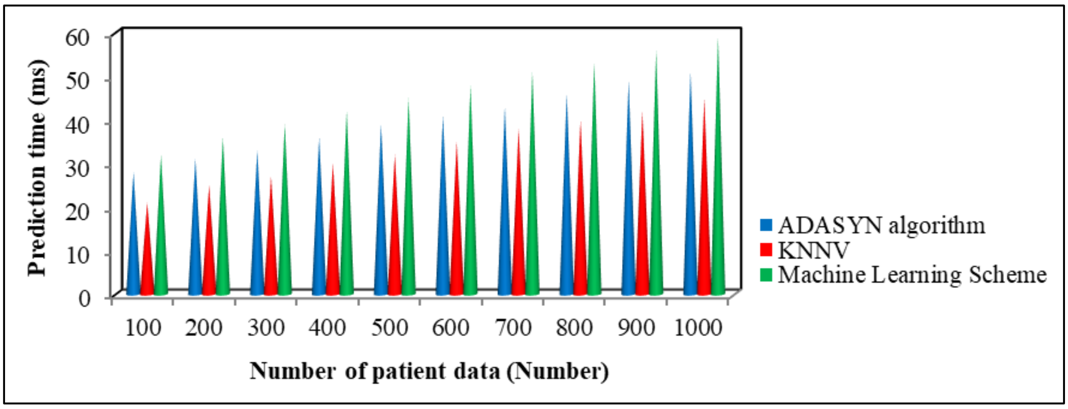

Fig 2: Measurement on Prediction Time

From the figure 2, the prediction time for different number of data points are described. The blue colour line represents the prediction time of ADASYN algorithm. The red colour line and green colour line represents the prediction time of KNN variant (KNNV) algorithm and Machine Learning Scheme correspondingly. It is observed that the prediction time using

ADASYN algorithm is lesser when compared to the KNN variant (KNNV) algorithm and Machine Learning Scheme. This is because of applying the KNNV classification algorithm to select the parameter 'K' adaptively for every unknown patient. KNNV determined the distance between patients accurately to identify the nearest neighbors. Consequently, prediction time of KNNV algorithm is reduced by 16% when compared to the ADASYN algorithm and 28% when compared to the Machine Learning Scheme.

4.3 Error rate

Error rate is defined as the ratio of number of data points that are incorrectly predicted to the total number of data points considered. Consequently, the error rate 'ER' is determined as,

$$ER = \left(\frac{Number\ of\ data\ points\ that\ are\ incorrectly\ predicted}{Number\ of\ data\ points}\right) * 100 \qquad (3)$$

From (3), the error rate is computed. The error rate is determined in terms of percentage (%).

Table 3: Tabulation for Error rate

Number of patient data (Number)	Error rate (%)		
	ADASYN algorithm	KNNV	Machine Learning Scheme
100	24	20	11
200	26	23	14
300	23	21	12
400	20	18	10
500	24	21	13
600	27	23	15
700	29	25	17
800	32	28	19
900	30	26	21
1000	33	29	24

Table 3 describes the error rate with respect to number of data points varying from 100 to 1000. Error rate comparison takes place on existing ADAptive SYNthetic (ADASYN) algorithm, KNN variant (KNNV) algorithm and Machine Learning Scheme. The graphical representation of error rate is described in figure 3.

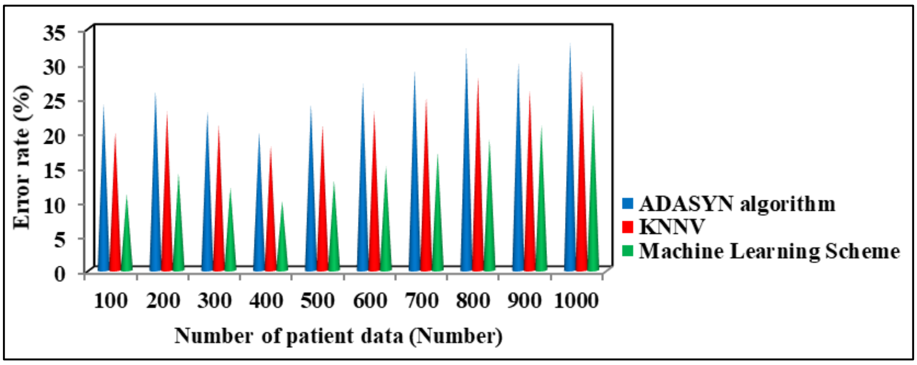

Fig 3: Measurement on Error Rate

From figure 3, the error rate for different number of data points is described. The blue color line represents the error rate of ADASYN algorithm. The red color line and green color line denotes the error rate of KNN variant (KNNV) algorithm and Machine Learning Scheme correspondingly. It is observed that the error rate using ADASYN algorithm is lesser when compared to the KNN variant (KNNV) algorithm and Machine Learning Scheme. This is because of applying the random forest model to forecast the count of fresh COVID-19 cases with the management makes the preparation to manage the cases. The feature selection had chosen relevant features outside all the features. Feature selection reduced the prediction complexity and error rate. As a result, the error rate of ADASYN algorithm is reduced by 43% when compared to the KNN variant (KNNV) algorithm and 34% when compared to the Machine Learning Scheme.

5. Discussion and limitation on existing marine weather forecasting methods

Adaptive synthetic (ADASYN) algorithm was introduced to balance COVID and non-COVID classes. A machine learning-based framework was employed with inpatient facility data for user-friendly, cost-effectiveand time-efficient solution. The designed framework employed Bayesian optimization to optimize the hyperparameters of classifier. Though the accuracy was improved, the classification performance was reduced.

A new KNN variant (KNNV) algorithm was designed to attain improved results. The rough set theoretic technique to manage incompleteness and heterogeneity problems to identify an ideal value for K. KNNV algorithm considered heterogeneous dataset with medical records of people and recognized the cases with COVID-19. But, the accuracy level was not improved by KNNV algorithm.

The feature selection method was carried out to predict all classes using random forest, linear model, support vector machine, decision tree, and neural network. The random forest was employed for prediction and analysis. K-fold cross-validation was carried out to determine the consistency. But, the complexity level of prediction was not minimized by Machine Learning Scheme.

5.1 Future Direction

The future direction of work can be carried out using deep learning techniques for increasing the COVID disease prediction performance with improved accuracy and lesser time consumption.

6. Conclusion

A comparison of different existing COVID disease prediction methods is illustrated. From the study, it is examined that the complexity level of prediction was not minimized by machine learning scheme was not reduced. The survival review shows that the accuracy level was not improved by

KNNV algorithm. In addition, the classification performance was reduced by ADASYN algorithm. The wide range of experiments on many existing COVID disease prediction determines the performance with its limitations. Finally, the research work can be carried out using deep learning and machine learning methods for increasing the COVID disease prediction performance.

References

1. Md. Abdul Awal, Mehedi Masud, Md. Shahadat Hossain, Abdullah Al-Mamun Bulbul, S. M. Hasan Mahmud, Anupam Kumar Bairagi, A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection From Inpatient Facility Data, IEEE Access. 2021; 9:10263-10281.
2. Ahmed Hamed, Ahmed Sobhy and Hamed Nassar, Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a *K*NN Variant Algorithm, Arabian Journal for Science and Engineering, Springer, 2021, 1-12
3. Vishan Kumar Gupta, Avdhesh Gupta, Dinesh Kumar, and Anjali Sardana, Prediction of Covid-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model, Big Data Mining and Analytics. 2021; 4(2):116-123
4. Miriam Pizzatto Colpo, Bruno Cascaes Alves, Kevin Soares Pereira, Anna Flávia Zimmermann Brandão, Marilton Sanchotene de Aguiar, Tiago Thompsen Primo. Attribute selection based on genetic and classification algorithms in the prediction of hospitalization need of COVID-19 patients, Association for Computing Machinery. 2021; 2:1-8
5. Prabh Deep Singh, Rajbir Kaur, Kiran Deep Singh, Gaurav Dhiman, A Novel Ensemble-based Classifier for Detecting the COVID-19 Disease for Infected Patients, Information Systems Frontiers, Springer. 2021, 1-17.
6. Elena Hernandez-Pereira, Oscar Fontenla-Romero, Veronica Bol on-Canedo, Brais Cancela-Barizo, Bertha Guijarro-Berdi nas, Amparo Alonso-Betanzos. Machine learning techniques to predict different levels of hospital care of CoVid-19, Applied Intelligence, Springer, 2021, 1-15
7. Makram Soui, Nesrine Mansouri, Raed Alhamad, Marouane Kessentini, Khaled Ghedira. NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient’s symptoms, Nonlinear Dynamics, Springer. 2021; 106:1453-1475
8. Suma LS, Anand HS, Vinod chandra SS. Nature inspired optimization model for classification and severity prediction in COVID-19 clinical dataset, Journal of Ambient Intelligence and Humanized Computing, Springer, 2021, 1-13.
9. Hamed Jelodar, Yongli Wang, Rita Orji, Shucheng Huang. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach”, IEEE Journal of Biomedial and Health Informatics. 2020; 24(10):1-12.
10. Md. Zahangir Alam M, Saifur Rahman M Sohel Rahman. A Random Forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked, Elsevier. 2019; 15:1-18.
11. Dilip Kumar Sharma, Muthukumar Subramanian, Pacha Malyadri, Bojja Suryanarayana Reddy, Mukta Sharma, Madiha Tahreem. Classification of COVID-19 by using supervised optimized machine learning technique, Materials Today: Proceedings, Elsevier, 2021, 1-16.
12. Yusuf Kuvvetli, Muhammet Deveci, Turan Paksoy, Harish Garg. A predictive analytics model for COVID-19 pandemic using artificial neural networks, Decision Analytics Journal, Elsevier. 2021; 1:1-15.