Coronavirus Disease Predictor: An RNA-Seq based pipeline for dimension reduction and prediction of COVID-19

Naiyar Iqbal and Pradeep Kumar
Department of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India
Email: naiyariqbal.rs@manuu.edu.in

Abstract. SARS CoV-2, the novel coronavirus behind the COVID-19 infection, has caused destruction around the world with human life, detecting a range of complexity which has knocked medical care specialists to investigate new innovative solutions and diagnosis strategies. The soft computing-based approach has assumed a significant role in resolving complex issues, and numerous societies have been shifted to implement and convert these innovations in response to the encounters created by the COVID-19 pandemic. To perform genome-wide association studies using RNA-Seq of COVID-19 and identify gene biomarkers, classification, and prediction using soft computing techniques of Coronavirus disease studies to fight this emergency pandemic in the epidemiological domain, and disease prognosis. The RNA-Seq profiles of both healthy and COVID-19 positive patients’ samples were considered. We have proposed an integrated pipeline from bioinformatics in-silico phase for -omic profile data processing to dimension reduction using various prominent techniques such as formal concept analysis and principal component analysis followed by machine learning phase for prediction of the disease. In this experimental research, we have applied different eminent machine learning techniques to implement an effective integrated model using Classifier Subset Evaluator (CSE) followed by principal component analysis (PCA) for dimension reduction to select the highly significant features and then to do the classification and prediction of Coronavirus disease, different eminent classifiers have been applied on the selected features. In this analysis, the Hoeffding Tree model found the topmost performance classifier with a classification accuracy of 99.21% as well as sensitivity and specificity of 99% and 100% respectively.

Keyword: COVID-19, RNA-Seq, Gene Biomarker, Soft Computing, Machine Learning, CSE, PCA, Prediction

1. Introduction
The novel coronavirus outbreak (nCoV) also well known as COVID-19 has made a disturbing circumstance around the world. Coronavirus pandemic is currently pushing researchers and medical experts to make a continuous effort to research and development all over the globe to fight COVID-19 infection and trying to decrease and prevention of its spreading. Pathological examinations are utilized to detect whether the patient is corona positive or not, yet they are tedious and several test equipment and centers are limited [1]. The shocking thing is sixty percent of patients might not have any sign expression, but still, they are COVID-19 positive and might spread as a silent transporter cause of Corona disease infected. Clinical data related to a patient’s medical history; geographical location and symptoms can be analyzed using machine learning techniques to predict clinical outcomes. But as symptoms vary from person to person and are common in other diseases like flu, swine flu, and pneumonia, therefore, the correctness of experimental results is debatable [2]. To challenge this problematic situation, this research analysis proposed an integrated workflow approach to identify the existence of SARS CoV-2 in an individual with genome sequencing of the virus. This would be precisely assistance for detection of the existence of virus, for the development of target treatments, and to figure out how understanding will react to drugs [3]. This paper sums up an essential overview
of existing work and talks about an approach to distinguish. Along these lines, the proposed strategy can be utilized by specialists and experts as a tool for the decision-making system.

2. Integrated Workflow Pipeline
The complete pipeline for our work interconnected with Bioinformatics workflow is shown in Figure 1, that broadly classified into three phases, first biological -omic profile data processing target to differential gene identification and second dimension reduction phase, and third machine learning-based prediction phase. Biological -omic profile data processing phase that comprises different phases like -omic profile data gathering, data pre-processing (quality check, and quality improvement and adapter trimming, etc.), read map to the human genome, read count, and differential expression gene identification [4]; and a second phase, dimension reduction can be done by single or hybrid approach to select high significant gene as a feature; then prediction phase applies soft computing techniques for classification and prediction [5] of novel coronavirus disease.

![Integrated Workflow Pipeline](image)

**Figure 1:** An integrated pipeline for data processing, dimension reduction, and prediction of COVID-19

2.1 Phase-I: Dataset Collection
We considered the RNA-Seq expression profile of the frontal cortex from healthy patients (Control) and COVID-19 positive patients (Case). A dataset of 126 samples was considered, including 26 healthy patients and 100 COVID-19 patient samples [6] accessible at Gene Expression Omnibus (GEO) with Accession No. GSE157103.
We get the transcript per base million (tpm) dataset from the NCBI-GEO repository as a normalized read count mentioned in the first phase of the pipeline. After that, we have applied transpose to convert genes as feature sets. There is a total number of 2728 genes are available in the selected datasets. Therefore, the curse of dimensionality problem occurs in the dataset. So that, we have applied PCA followed by CSE that are described in the subsequent section. This dataset collection has been done from the Gene Expression Omnibus repository of the National Centre for Biotechnology Information (NCBI).

3. Phase-II: Dimension Reduction

Dimension reduction is a process to reduce the number of feature vector from a dataset which segregates feature sets between higher and lower significant attributes [7]. It is used to resolve the curse of dimensionality problem of the feature vector [8].

For our target analysis, we directly perform machine learning analysis using various algorithms on raw datasets with 2728 gene feature sets without dimension reduction. After that, we have taken an integrated approach for this task to select the most relevant feature sets from the dataset. In the first step, we use the CSE attribute evaluator to the Best First search method of the full training dataset and got 110 significant genes out of 2728 feature sets. When we apply machine learning algorithms, we get an accuracy rate increased up to 7% approximately shown in Table 1. Again, we used PCA attribute evaluator to Ranker search method on 110 significant genes, then we get 57 most relevant gene feature sets which increase the accuracy rate up to 2% approximately shown in Table 1. In conclusion, we found that the integrated approach works better than the individual technique. The dimension reduction and machine learning classification phases have been done in an eminent Waikato Environment for Knowledge Analysis (WEKA) machine learning tool [9].

| ML Algorithm            | Feature Dimension | Accuracy (%) | Accuracy increased by (%) | Sensitivity (%) | Specificity (%) |
|-------------------------|-------------------|--------------|---------------------------|----------------|-----------------|
| Hoeffding Tree          | 2728              | 91.27        | +6.35                     | 95.00           | 76.92           |
|                         | 110               | 97.62        |                           | 98.00           | 96.15           |
|                         | 57                | 99.21        | +1.59                     | 99.00           | 100.00          |
| Naive Bayes             | 2728              | 92.06        | +5.56                     | 95.00           | 80.77           |
|                         | 110               | 97.62        |                           | 98.00           | 96.15           |
|                         | 57                | 98.41        | +0.79                     | 98.00           | 100.00          |
| Sequential Minimal Optimization | 2728       | 94.44        | +2.39                     | 95.00           | 92.31           |
|                         | 110               | 96.83        |                           | 96.00           | 100.00          |
|                         | 57                | 98.41        | +1.58                     | 98.00           | 100.00          |
| AdaBoost                | 2728              | 89.68        | +5.56                     | 95.00           | 69.23           |
|                         | 110               | 95.24        |                           | 97.00           | 88.46           |
|                         | 57                | 95.24        | 0.00                      | 95.00           | 96.00           |
| Multilayer Perceptron  | 2728              | 79.37        | +17.46                    | 97.00           | 11.54           |
|                         | 110               | 96.83        |                           | 97.00           | 96.15           |
|                         | 57                | 94.44        | -2.39                     | 96.00           | 88.46           |

Table 1: Comparison of machine learning performance based on different feature dimension
Classifier Subset Evaluator

Classifier subset evaluator (CSE) assesses the feature subsets on a training dataset or a distinct grip out testing dataset [10]. CSE applies a classification technique for the evaluation of the importance of a feature set. Classifier subset evaluator comprises of three main elements as the attributes or features evaluator, search method, and attribute selection mode. Attributes or features classifier assesses feature subsets on the training dataset and applies classification algorithms, for example, decision tree, Naïve Bayes, Random Forest, and Support Vector Machine to calculate the importance of features set. Whereas, search method explores the entire features set to identify the ideal node. In this analysis, the best first search method has been applied. This technique observes the scope of feature subsets by applying a greedy hill-climbing approach as enhanced with a backtracking capability [11]. In the context of the attribute selection mode, the features, applying the whole training dataset have been chosen.

Principal Component Analysis

Principal component analysis (PCA) is a prominent dimension reduction method based on recognizing associations and patterns from a dataset intended to be converted into a dataset of expressively lower dimension without damage of any significant data. It is a standard approach to decrease the dimension of the biological expression dataset, increases the interpretability to conceive the relationships among the biological instances, and clean the noise. It is frequently applied as a preprocessing task of succeeding analysis [12]. This happens so by forming new uncorrelated features which make sequentially maximize variance. It builds linear groupings of the gene expression profile datasets, known as principal components (PCs). The PCs are statistically independent of one another, can efficiently describe the deviation of gene expression profile data, and possibly will have a much lesser dimension [13]. PCA is exceptionally valuable to accelerate the computation by decreasing the dimension of the feature set. Besides, when there are high dimensionalities with higher interrelated features of each other, the PCA can increase the accuracy of classification and prediction models.

Phase-III: Soft Computing Algorithms

In our experiment, we have applied various eminent simple binary soft computing classifiers such as Naïve Bayes, Multilayer Perceptron-ANN, SMO-SVM, IBk, Hoeffding Tree-Decision Tree, J48-DT, to ensemble-based classifiers like AdaBoost, Bagging, LogitBoost, and Random Forest. During the analysis, all the mentioned classifiers come with a better performance rate of above 90% classification accuracy. We have picked above 95% of accuracy rate to depict in this research article.

Hoeffding Tree

The Hoeffding tree is a progressive decision tree machine learning algorithm for the huge dataset, which expects that the distribution of the dataset is not fluctuating over the long run. It increments progressively based on a decision tree of the logical assurances of the Hoeffding bound [14]. A node is extended when there is adequate factual proof that an optimum splitting feature occurs, a result based on the distribution independent Hoeffding bound. The system trained through the Hoeffding tree is asymptotically similar to the other developed through a non-incremental learner if the quantity of training samples is sufficiently huge.

Naïve Bayes

Naïve Bayes applies a comparable technique for the prediction of the probability of diverse classes based on several features. This algorithm is widely applied in text classification and with the solution of multi-class problems. Naïve Bayes is a powerful based on statistics classification technique and it has been also played an effective role in the domain of biomedical analysis [15]. The fundamental concept of Naïve Bayes is analogous to the Covariance Determinant. However, in Naïve Bayes, there
is an assumption of the feature variables to be autonomous from one another on the given result. This presumption incredibly improves on the computation of contingent probabilities and defeats the different issues when applying the Covariance Determinant prediction model to manage with the procedures in which the parts of constituent feature vectors are regularized.

### 4.3 Sequential Minimum Optimization

Sequential minimal optimization (SMO) is a technique for solving the quadratic programming (QP) problem that emerges throughout the training of support vector machines. To train a support vector machine, it needs the solution of an enormous quadratic programming optimization problem. The research on the sequential minimal optimization technique has produced lots of interest in the SVM group, as earlier existing approaches for SVM modeling were considerable with more complexity and needed costly mediator, quadratic programming solver. The volume of storage needed for sequential minimal optimization is linearly fashioned in the training dataset size, which permits it to deal with exceptionally enormous training datasets [15]. It is because the matrix calculation is escaped, SMO measures somewhere close to linear and quadratic in the training dataset, volume for several analysis issues, whereas the regular type of SVM technique measures somewhere close to linear and cubic. The computation time is overwhelmed by support vector machine calculation; subsequently, SMO is the quickest technique for linear SVMs and sparse datasets.

### 4.4 AdaBoost: An ensemble approach

Adaptive Boosting, also abbreviated as AdaBoost is one of the most widely used ensemble models based on boosting approach which mixed several weak classifiers into a single strong classifier. In AdaBoost, the weight vectors are reassigned to every sample and it assigned higher weights to incorrectly classified samples [16]. This algorithm performs boost better in decision trees with the binary classification problem solve. Further currently, it can be mentioned as discrete AdaBoost because it is applied for classification rather than regression.

### 4.5 Multilayer Perceptron

Multilayer perceptron (MLP) is a category of feed-forward artificial neural network (ANN). In this architecture, there are an input layer, a hidden layer, and an output layer composition as a minimum of three layers of nodes. Apart from the input nodes, every node is like a neuron that applies a non-linear activation function. A multilayer perceptron is also known as a deep artificial neural network. The input layer collects the signals, the one or more arbitrary hidden layers in the middle are the real computing machine of the MLP and an output layer that produce a decision class or prediction onto the input signals [17]. The multilayer perceptron algorithm is appropriate to classification problems where the input signals are given a class label, as well as it is also working properly for regression problems associated with a real value measure that has to be predicted.

### 5. Performance Metrics

In this experiment, ten supervised machine learning techniques have been used to classify and predict the Coronavirus disease and show the outcome and performance of the top five techniques based on accuracy rate. The efficiency of the classification approaches has been validated based on tenfold cross validation. The eight gold standard factors have been observed in the interpretation of the model classifiers [15]. Instances with the negative of Coronavirus disease have been treated as a control class (-ive), and instances with the positive of Coronavirus disease have been treated as a case class (+ive). Standard terms related to confusion matrix as defined subsequently in Table 2.
### 6. Results & Discussions

The performance estimation of Coronavirus disease prediction by five machine learning techniques is calculated with associated gold standard eight metrics as listed in Table 2. There were a sum of 126 instances considered with 100 cases (+ive) and 26 controls (-ive) of Coronavirus disease. COVID-19 normalized read count datasets in the form of transcript per base million (.tpm) has been collected through the Bioinformatics in-silico phases using RNA-Seq gene expression profile dataset mentioned in the proposed integrated workflow in Figure 1.

All the instances have been distributed tenfold and every fold is used in testing and the rest folds are utilized to train during cross validation. The confusion matrix of classification outcome has presented in Table 3 on behalf of the Hoeffding Tree machine learning technique, and along with other classifiers such as Naïve Bayes, Sequential minimal optimization, AdaBoost and Multilayer perceptron techniques have been displayed in Figure 2. It portrays the prediction models based on all above mentioned machine learning techniques.

This can be stated onto the outcomes that prediction of (Figure 2) Hoeffding Tree is the uppermost amount of true positive (The number of instances predicted as case and it does have Coronavirus disease) and also Hoeffding Tree and other two Naïve Bayes and SMO predict the highest amount of true negative (The number of instances predicted as control and it does not have Coronavirus disease).
Table 3: Confusion Matrix for Hoeffding Tree

| HT-DT      | Predicted Class | Total Actual |
|------------|-----------------|--------------|
| Actual Class | CASE | CONTROL |                  |
| CASE       | 99    | 1       | 100             |
| CONTROL    | 0     | 26      | 26              |
| Total Predicted | 99   | 27      | 126             |

The second uppermost true positive has been detected by Naïve Bayes and SMO confusion matrix and AdaBoost techniques predict the second uppermost true negative. Whereas the third uppermost true positive has been detected by AdaBoost matrix and MLP techniques predict the third uppermost true negative. The third uppermost true positive has been detected by MLP (Figure 2).

All the standard performance metrics based on confusion matrix for classification and prediction exclusively classification accuracy, true positive rate, true negative rate, Positive Predicted Value, False Positive Rate, Negative predictive value, error rate, and F1 Measure are mentioned in Table 4.

Table 4 acknowledged that Hoeffding Tree performs outstanding over alternative machine learning models based on the uppermost classification performance of 99.21%, however, the second uppermost classification performance rate has been obtained by Naïve Bayes and SMO with 98.41% for both, whereas AdaBoost and MLP stated at third and fourth position with 95.24% and 94.44% respectively.

Figure 2: Classification outcome of Machine Learning Techniques
Furthermore, Hoeffding Tree has got a maximum sensitivity of 99%, and Naïve Bayes and SMO both have acquired the second maximum sensitivity of 98%. However, Hoeffding Tree, Naïve Bayes and SMO all three acquire uppermost specificity and precision value of 100%. Based on this experiment, it found that all the machine learning performs better, but overall the Hoeffding Tree model got the most suitable to classify and predict Coronavirus infected (case) or non-infected (control) patients. Table 4 expressed that Hoeffding Tree has the maximum negative predictive value of 96.30%. Also, the Hoeffding Tree got an F1 score of 99.50%. Hoeffding Tree, Naïve Bayes, and SMO achieve the lowest False Positive rate of 0%, whereas Hoeffding Tree performs the best with the lowermost Rate of misclassification with 0.79%. The ROC area value was recorded at 99% for four models such as Hoeffding Tree, Naïve Bayes, SMO, and AdaBoost except that MLP has got ROC value of 96%.

|                | HT-DT  | Naive Bayes | SMO_SVM | AdaBoost | MLP_ANN |
|----------------|--------|-------------|---------|----------|---------|
| Classification Accuracy | 99.21  | 98.41       | 98.41   | 95.24    | 94.44   |
| Sensitivity     | 99.00  | 98.00       | 98.00   | 95.00    | 96.00   |
| Specificity     | 100.00 | 100.00      | 100.00  | 96.15    | 88.46   |
| Precision       | 100.00 | 100.00      | 100.00  | 98.96    | 96.97   |
| FPR             | 0.00   | 0.00        | 0.00    | 3.85     | 11.50   |
| NPV             | 96.30  | 92.86       | 92.86   | 83.33    | 85.19   |
| RMC             | 0.79   | 1.59        | 1.59    | 4.76     | 5.56    |
| F1              | 99.50  | 98.99       | 98.99   | 96.94    | 96.48   |
| ROC             | 99.00  | 99.00       | 99.00   | 96.00    | 99.00   |

Table 4: Classification performance metrics of machine learning algorithms

7. Limitations and Future Research Directions
In this experimental analysis, we applied the integrated dimension reduction technique and selected 57 RNA-Seq based gene expression feature set out of 2728 gene expression features with 126 instances in the form of normalized transcript per base million (TPM). Out of 126 dataset instances, there are 100 Coronavirus case (+ive) and 26 corona control (-ive) instances and accomplishes classification activities of machine learning. Then, we used ten eminent machine learning techniques in which the Hoeffding Tree, one of the decision tree models outperform.

The limitations of this experiment are as follow i) it has done on the imbalance of case and control class instances, ii) requires to analyze the DEGs identification iii) need to perform enrichment analysis step for finding the association of identified gene in the Coronavirus disease as well as to other associated diseases mentioned in the proposed section of Bioinformatics phase. Moreover, we are hopeful to enrich the accuracy of the model with additional associated expressive and responsive gene expression features using a massive quantity of datasets in the forthcoming plan, and along with we are furthermore attentive to the development of an online toolkit that would be helpful for medical practitioners to yield an assessment with the highest possible accuracy of the Coronavirus disease prediction as well as prognosis and diagnosis. Furthermore, we have pointed out some of the important challenges which should be tackled in future research work.

- Insufficient availability of control (-ive) class instances
- Identification of severity level of Coronavirus disease
- Analysis of Coronavirus mutation from different countries
- Web-based implementation of this whole model.

8. Conclusion
Coronavirus disease or COVID-19 is still currently a continuing pandemic overtaking the earlier histories of all transmittable outbreaks in the context of infection and death rate with different variants and phases of the wave [18, 19]. Scientists and medical practitioners are continuously investigating all
the possible actions to combat the COVID-19 pandemic [20]. The soft computing approach has been playing an important role in the identification of Coronavirus disease at an initial phase for the appropriate diagnosis [21, 22]. Hopefully, this machine learning-based model would be supportive and contribute to the implementation of suitable solutions to combat the existing pandemic situation. Each activity has a series of Bioinformatics methods and tools, including FastQC for read quality check, Trimmomatic for adapter trimming, Bowtie2 for read mapping, HT-Seq for read count, and normalization, edgeR for identification of differential expression genes [23]. After that, we have applied various eminent soft computing approaches for the classification and prediction of COVID-19 on the significant genes as identified feature vectors. In machine learning, the solutions to various biological problems are the trending research domain. The consciousness of the most significant features, especially at the genetic level of Coronavirus or other associated diseases would be aided to the researchers with an emphasis on how to enhance the classification accuracy [24]. This experimental research has effectively hybrid the different techniques such as CSE followed by PCA for dimension reduction to select the highly significant features and then to do the classification and prediction of Coronavirus disease, different eminent machine learning techniques have been applied to the selected features [25].

The main emphasis of the experiment is intended to classify and predict Coronavirus disease. Throughout this analysis work, it is found that when we compare the machine learning models with dimension reduction and without dimension reduction, it gives a better difference in the performance of the model as shown in Table 1. Furthermore, to do this experiment analysis, ten eminent machine learning approaches have been used with eight associated factors are applied to performance measurement based on confusion matrix. Out of ten, we have depicted five topmost techniques based on their accuracy rate followed by sensitivity and specificity. Based on this experimental analysis, it is concluded that the Hoeffding Tree model is the uppermost classification model which has achieved classification accuracy by 99.21%, followed by sensitivity and specificity of 90% and 100% respectively, also Hoeffding Tree performs the best with lowermost Rate of misclassification with 0.79%.

References

[1] Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, Prior F, Yu Jr F. Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review. JMIR medical informatics. 2021 Jan 11;9(1):e23811. https://doi.org/10.2196/23811

[2] Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics. 2020 Jun;8(6):890. https://doi.org/10.3390/math8060890

[3] Raza K, Qazi S. An introduction to computational intelligence in COVID-19: surveillance, prevention, prediction, and diagnosis. InComputational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis 2021 (pp. 3-18). Springer, Singapore. https://doi.org/10.1007/978-981-15-8534-0_1

[4] Iqbal N, Kumar P. A framework for the RNA-Seq based classification and prediction of disease. InICDSMLA 2019 2020 (pp. 74-81). Springer, Singapore. https://doi.org/10.1007/978-981-15-1420-3_8

[5] Iqbal N, Kumar P. I-NFG: An integrated neuro-fuzzy-genetic based soft computing techniques for feature selection and disease prediction using gene expression. Journal of Applied Computing. 2019;4(1):1-8.

[6] Overnnyer KA, Shishkova E, Miller JJ, Bahnis J, Bernstein MN, Peters-Clarke TM, Meyer JG, Quan Q, Muehlbauer LK, Trujillo EA, He Y. Large-scale multi-omic analysis of COVID-19 severity. Cell systems. 2021 Jan 20;12(1):23-40. https://doi.org/10.1016/j.cels.2020.10.003

[7] Ali RH, Abdulsalam WH. The Prediction of COVID 19 Disease Using Feature Selection Techniques. InJournal of Physics: Conference Series 2021 May 1 (Vol. 1879, No. 2, p. 022083). IOP Publishing. https://doi.org/10.1088/1742-6596/1879/2/022083

[8] Enríquez M, Naranjo S, Amaro I, Camacho F. Dimensionality Reduction Using PCA and CUR Algorithm for Data on COVID-19 Tests. Artificial Intelligence, Computer and Software Engineering Advances. 2021;1326:121. https://doi.org/10.1007/978-3-030-68080-9_9

[9] Awadh WA, Alasady AS, Mustafa HI. Predictions of COVID-19 Spread by Using Supervised Data Mining Techniques. InJournal of Physics: Conference Series 2021 May 1 (Vol. 1879, No. 2, p. 022081). IOP Publishing. https://doi.org/10.1088/1742-6596/1879/2/022081

[10] Almustafa KM. Prediction of heart disease and classifiers’ sensitivity analysis. BMC bioinformatics. 2020 Dec;21(1):1-8. https://doi.org/10.1186/s12859-020-03626-y
[11] Chidambaram S, Srinivasagan KG. Performance evaluation of support vector machine classification approaches in data mining. Cluster Computing. 2019 Jan;22(1):189-96. https://doi.org/10.1007/s10586-018-2036-z
[12] Mahmoudi MR, Heydari MH, Qasem SN, Mosavi A, Band SS. Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. Alexandria Engineering Journal. 2021 Feb 1;60(1):457-64. https://doi.org/10.1016/j.aej.2020.09.013
[13] Hasan BM, Abdelazeez AM. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining. 2021 Apr 15;2(1):20-30. https://doi.org/10.30880/jscdm.2021.02.01.003
[14] Thaiarparnt S, Kritsanasun S, Chumuang N. A classification for patients with heart disease based on hoeffding tree. In2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) 2019 Jul 10 (pp. 352-357). IEEE. https://doi.org/10.1109/JCSSE.2019.8864158
[15] Iqbal N, Islam M. Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica. 2019 Sep 30;43(3). https://doi.org/10.31449/inf.v43i3.1548
[16] Soui M, Mansouri N, Alhamad R, Kessentini M, Ghedira K. NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient’s symptoms. Nonlinear Dynamics. 2021 May 18;1:23. https://doi.org/10.1016/s1170-7103(19)32588-3
[17] Qazi S, Raza K, Iqbal N. 5 Artificial intelligence in medicine (AIM): machine learning in cancer diagnosis, prognosis and therapy. InArtificial Intelligence for Data-Driven Medical Diagnosis 2021 Feb 8 (pp. 103-126). De Gruyter. https://doi.org/10.1515/9783110668322-005
[18] Overmyer KA, Shishkova E, Miller JJ, Balnis J, Bernstein MN, Peters-Clarke TM, Meyer JG, Quan Q, Muehlbauer LK, Trujillo EA, He Y. Large-scale multi-omic analysis of COVID-19 severity. Cell systems. 2021 Jan 20;12(1):23-40. https://doi.org/10.1016/j.cels.2020.10.003
[19] Majumder AB, Gupta S, Singh D, Majumder S. An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. InJournal of Physics: Conference Series 2021 Feb 1 (Vol. 1797, No. 1, p. 012011). IOP Publishing. https://doi.org/10.1088/1742-6596/1797/1/012011
[20] Bachtiger P, Peters NS, Walsh SL. Machine learning for COVID-19—asking the right questions. The Lancet Digital Health. 2020 Aug 1;2(8):e391-2. https://doi.org/10.1016/S2589-7500(20)30162-X
[21] Rakshit P, Debnath S, Mistri J, Kumar S. Assessment of Prediction Models of Confirmed, Recovered and Deceased cases due to COVID-19. InJournal of Physics: Conference Series 2021 Feb 1 (Vol. 1797, No. 1, p. 012004). IOP Publishing. https://doi.org/10.1088/1742-6596/1797/1/012004
[22] Ononiyi TA, Alabere HO, Sede EE. Diagnosis of covid-19 using artificial intelligence based model. InJournal of Physics: Conference Series 2021 (Vol. 1734, No. 1, p. 012007). IOP Publishing. https://doi.org/10.1088/1742-6596/1734/1/012007
[23] Auwul MR, Rahman MR, Gov E, Shahjaman M, Moni MA. Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19. Briefings in Bioinformatics. 2021 Apr 12. https://doi.org/10.1093/bib/bbaa0120
[24] Loey M, Manogaran G, Taha MH, Khalifa NE. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. Measurement. 2021 Jan 1;167:108288. https://doi.org/10.1016/j.measurement.2020.108288
[25] Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. https://doi.org/10.1152/physiolgenomics.00029.2020