INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Hybrid Feature Selection for COVID-19 Severity Prediction Using Cuckoo Search with SVM Framework

N Hari Priya[1]*, **S Rajeswari**[2]

1 Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India
2 Associate Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India

Received: 13-10-2023
Accepted: 28-12-2023
Published: 20–01-2024

Citation: Priya NH, Rajeswari S (2024) Hybrid Feature Selection for COVID-19 Severity Prediction Using Cuckoo Search with SVM Framework. Indian Journal of Science and Technology 17(4): 317-324. https://doi.org/10.17485/IJST/v17i4.2601

* **Corresponding author**.

haripriyanarasimma@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: The main objective of this study is to determine the most important blood test markers that may indicate the presence of COVID-19 in a patient. To utilize the Cuckoo Search algorithm with SVM to explore the feature space efficiently and select features that contribute significantly to the model's performance. **Methods:** A novel hybrid method for feature selection has been proposed with the goal of improving the predictive capabilities of Support Vector Machines (SVM) for determining COVID-19 severity. Blood test datasets are used in the implementation of this study. The dataset has been split into two parts: 80% for training and 20% for testing. First, we use two statistical measures, chi-squared and mutual information, from the filter approach to minimize the feature dimensions. As a wrapper for SVM, we then use a modified Cuckoo Search algorithm. To measure how well the proposed approach works, we used evaluation metrics such as accuracy, precision, recall, and F1 score. **Findings:** The SVM classifier achieved the best performance with the features obtained from the proposed hybrid method, and the SVM classifier obtained an accuracy of 92% using the blood test dataset. The outcomes demonstrate that our hybrid approach effectively picks a subset of features that makes the model simpler while also making it more accurate and faster to compute. **Novelty:** This research work proposes a new hybrid feature selection technique by combining filter and wrapper methods to find the best feature set. This combination is introduced for the first time in this type of work related to COVID-19 prediction in which the results of Chi-Square and Mutual Information are used by the modified Cuckoo-Search algorithm to find the top features pertaining to COVID-19 severity and also to improve the performance of SVM model.

Keywords: Feature selection; Cuckoo search; Machine learning; Support Vector Machine (SVM); Severity prediction; Healthcare

1 Introduction

The emergence of the Coronavirus Disease 2019 (COVID-19) in December 2019 in Wuhan, China, caused unprecedented challenges for healthcare systems, researchers, and politicians worldwide(1). As the virus rapidly spreads, causing millions of infections and deaths, there is an immediate need for accurate predictive models that can aid in early detection, disease progression analysis, and strategy for intervention(2). Machine learning has become a strong tool in the fight against this pandemic because it can handle very large and complicated datasets(3). The Support Vector Machine (SVM) is a widely used supervised machine learning technique that has demonstrated promising results in diverse healthcare domains. Within the context of the COVID-19 pandemic, SVM can be utilized to forecast the probability of contracting the virus, the extent of disease advancement, and various associated effects(4). However, the efficacy of SVM is heavily contingent upon the careful identification and selection of relevant characteristics from the given dataset. The presence of unnecessary or irrelevant features has the potential to negatively affect the performance of a machine learning model.

Several authors have already employed machine learning to diagnose and forecast this fatal disease. Several studies employ standard evaluation techniques, such as antigen tests, CT scans, and X-rays, to detect COVID-19 infection.

Two hospitals in China, Kunshan People's Hospital and Wuhan Union Hospital, provided 294 blood samples, which were analysed using Random Forest and SVM. Experimentally, SVM outperforms random forest classifier with an accuracy of 84%, based on the fifteen features chosen for analysis(5). Authors in(6) suggested the diagnosis of COVID-19 using five distinct ML algorithms, including neural networks, random forest, gradient boost trees, logistic regression, and SVM. The Albert Einstein Hospital in Brazil provided a dataset consisting of 235 blood samples and 102 confirmed cases of COVID-19. From this dataset, 15 relevant characteristics were selected for analysis, with an AUC score of 85%, sensitivity of 68%, and specificity of 85%, respectively.

Based on nine simple survey questions, Shoer et al.(7) developed a prediction model. Over two million people participated in a national survey of symptoms in Israel, and that data was used in the study. There was a total of 43,752 adults, and 498 of them self-reported a positive status for COVID-19. The study included age, gender, medical history, smoking practices, and self-reported symptoms such as fever, sore throat, cough, shortness of breath, and loss of taste or smell. After training with Logistic regression, the model achieved an AUC of 74%.

Nemati et al.(8) developed a model to predict the duration of hospital stay for COVID-19 patients based on the clinical data of the patients. An open-access data set with information on 1182 hospitalized patients was put together by a group of researchers from different institutions and study labs. Various statistical analysis techniques and ML strategies were utilized to implement multiple survival analysis models. The most precise estimate of discharge time was found to be generated by the stage wise gradient-boosting survival model with C-index of 71.47. Male and older age categories were found to have lower discharge probabilities.

The previous works on predicting COVID-19 using machine learning techniques did not incorporate feature selection approaches to identify the most relevant features. Feature selection is a critical stage in machine learning as it entails determining which subset of the original set of features (variables or attributes) are most pertinent. The objective is to enhance model interpretability, minimize overfitting, and optimize model performance. The application of feature selection guarantees that the chosen features make a substantial contribution to the predictive model by identifying the most pertinent blood test indicators for predicting COVID-19. This process streamlines the diagnostic procedure by identifying a limited number of informative features that are crucial for predicting COVID-19. This improves the practicality of the method for real-world applications. Despite the limited number of features, feature selection plays a significant role in facilitating model generalization. By focusing on the most relevant blood test indicators, the predictive model is more likely to generalize well to new and previously unseen COVID-19 instances.

The process of selecting a subset of pertinent features is known as feature selection, and it is essential for the creation of effective and precise predictive models(9). Feature Selection approaches are primarily characterized as either subset selection methods or ranking methods since they seek out the most relevant features for classification(10). Conventional feature selection techniques frequently depend on heuristic or deterministic procedures, which may not consistently produce the optimal subset of characteristics. Hybrid feature selection, which combines the benefits of various approaches, has the ability to overcome the drawbacks of individual approaches.

The Cuckoo Search (CS) technique is an optimization approach that draws inspiration from the brood parasitism behavior exhibited by cuckoo birds(11). Within the domain of machine learning, Cuckoo Search can be utilized as a process for feature selection, with the objective of identifying the most effective subset of features that maximizes the performance of the SVM. The integration of Cuckoo Search with SVM holds potential in developing a resilient and effective framework for predicting COVID-19.

This study presents a novel framework for hybrid feature selection in COVID-19 prediction, utilizing a Cuckoo Search-based Support Vector Machine SVM approach. The objective of this study is to investigate the possibilities of integrating this combination technique in order to improve the predicted accuracy and efficiency of the model.

The primary contributions of this research are as follows:

- To identify the most crucial blood test parameters that potentially leads to the presence of COVID -19 in a patient
- To explore the potential for using different feature selection techniques to identify relevant subsets of clinical blood test parameters
- To utilize the Cuckoo Search algorithm with SVM to explore the feature space efficiently and select features that contribute significantly to the model's performance
- To enhance the accuracy of the machine learning model with the adoption of effective hybrid feature selection method

2 Methodology

The proposed framework is based on hybrid feature selection technique employing both filter and wrapper methods. The two different datasets after undergoing pre-processing are used to calculate the correlation co-efficient of all the features and top 10 features has been extracted. Then a combination of chi-squared and mutual information has been applied on the selected feature set. Finally, cuckoo search based SVM which has been serving as a wrapper method for feature selection has been applied on the dataset and best features has been identified. With the obtained features, various machine learning classifiers has been applied and performances of each of the classifiers has been calculated. Figure 1 shows the entire work flow of the proposed methodology.

2.1 Dataset Description

The dataset consists of blood test data which has been obtained from Zenodo website(12) which has initially 279 COVID-19 cases and 16 features including 2 demographic features as age and gender, 13 blood test indicators and the RT-PCR result from swab test as a target variable and it is shown in Table 1. The dependent variable SWAB has been renamed as "target" and the column named Basophil has been dropped as it has 0 as the majority value in the dataset. The dependent variable "target" is a binary variable which represents 1 for positive COVID-19 case and 0 for negative COVID-19 case. Also, the distribution of COVID-19 cases among male and female patients is shown in Figure 2.

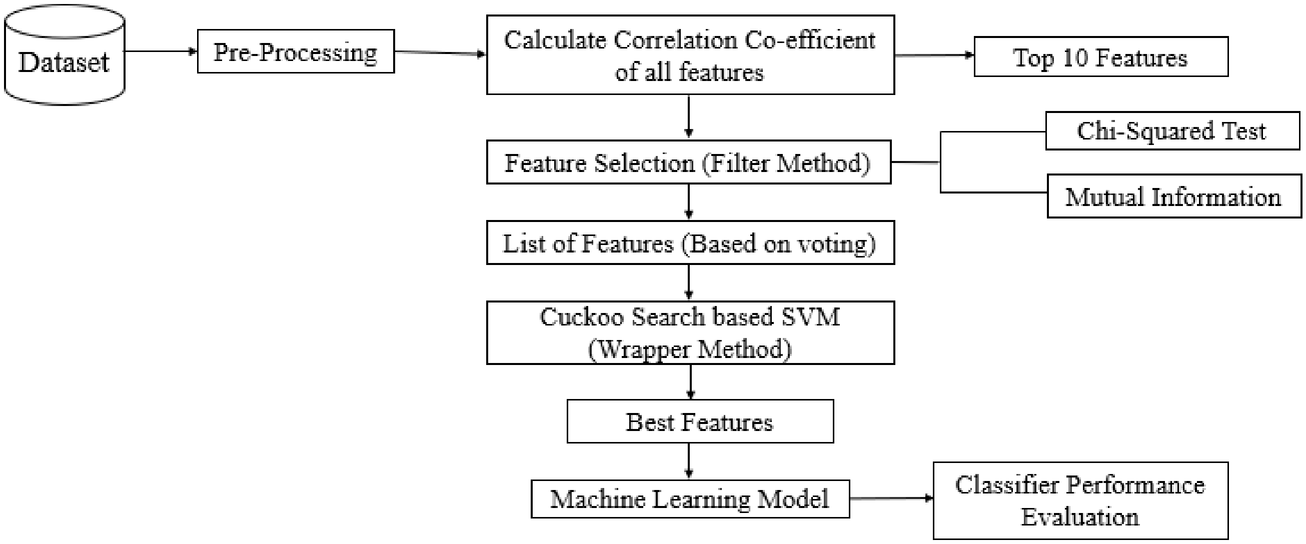

Fig 1. Workflow of the Proposed Methodology

Table 1. L ist of Features (Dataset 1 - Blood Test Data)

S. No	Attribute	Expansion
1	AGE	Age
2	GENDER	Gender
3	WBC	White Blood Cells
4	PLT	Platelets
5	NEU	Neutrophils
6	LY	Lymphocytes
7	MO	Monocytes
8	EO	Eosinophils
9	CRP	C-Reactive Protein
10	AST	Aspartate aminotransferase
11	ALT	Alanine aminotransferase
12	ALP	Alkaline phosphatase
13	GGT	Gamma glutamyltransferase
14	LDH	Lactate dehydrogenase
15	target	Covid-19 (Positive/Negative)

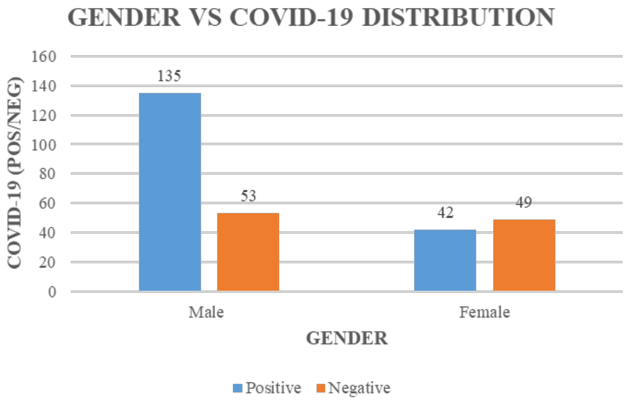

Fig 2. Gender Vs Covid-19 Distribution

2.2 Data Pre-processing

Data pre-processing is an essential phase in the data analysis procedure, as it prepares raw data for subsequent processing and analysis. One common challenge in data pre-processing involves dealing with missing data(13). K-Nearest Neighbors (KNN) imputation is one such technique for handling missing values. This method is potentially more precise than mean, median, or mode imputation and it is compatible with both numeric and qualitative data.

The KNN imputation method estimates missing values using the values of the k-nearest neighbors based on a distance metric (typically Euclidean).

For missing values at a data point:

- It identifies the 'k' data elements closest to the desired point.
- The missing value is then estimated using the mean (for numeric variables) or mode (for categorical variables) of the 'k' nearest neighbors.

2.3 Feature Selection

Feature selection is crucial to the development of accurate and efficient machine learning models, particularly for high-dimensional datasets. In the domain of medical and biological datasets, such as those for COVID-19 prediction, it is crucial to reduce dimensionality without losing information. In this work, a novel hybrid feature selection algorithm that combines the advantages of filter methods with a nature-inspired optimization algorithm, the Cuckoo Search method customized for Support Vector Machines has been proposed.

```
Algorithm: Hybrid Feature Selection using Filter Methods and Cuckoo Search-based SVM
Input:
• Dataset D
• Number of nests: N
• Abandonment probability: $P_a$
• Maximum number of iterations for Cuckoo Search: T
Output:
• Selected features $F_{selected}$
• Performance metrics of the ML Models
Steps:
1. Data Pre-processing and Initial Feature Selection
1.1 Load the dataset D
1.2 Calculate the correlation of each feature with the target variable y:
corr (Xi, y) = Cov ($X_i$, y)/ $\sigma$(Xi) $\sigma$(y)
1.3 Select the top 10 features $F_{corr}$ based on |corr ($X_i$ , y) |
2. Filter Method Feature Selection
2.1 For each feature $f_i$in $F_{corr}$, compute the Chi-Squared score $\chi^2$ ($f_i$, y)
2.2 Rank the features based on $\chi^2$ to get $F_{\chi}$ 2
2.3 For each feature $f_i$ in $F_{corr}$, compute the Mutual Information M I ($f_{i,}$ y)
2.4 Rank the features based on M I to get $F_{MI}$
3. Feature Voting
3.1 Initialize a voting dictionary V with keys as feature names and values as zero
3.2 For each feature $f_i$ in $F_{\chi}$ 2and $F_{MI}$:
V [ $f_i$ ] = V [ $f_i$ ] + 1
3.3 Select features $F_{selected}$ where V [ $f_i$] = 2
4. Cuckoo Search-based Wrapper Method for SVM
4.1 Initialize N nests S1, S2,…, SNwith random feature subsets from $F_{selected}$
4.2 For t=1, 2, …, T:
4.3 Choose the best nest $S^*$ based on the fitness function
5. Machine Learning Model Training and Evaluation
5.1 Train machine learning models using the features from $S^*$
5.2 Evaluate the accuracy Acc on the test set
```

2.4 Machine Learning Algorithms

Machine Learning (ML) is one of the most important tools for the classification task and is used to forecast the number of potential confirmed cases and deaths. Machine learning models can help us understand, predict, and handle COVID-19 datasets in numerous ways. Various research is currently being conducted to forecast COVID-19 cases using various ML models (14). We examined several well-known forecasting models from the pool of machine learning (ML) models, including Support Vector Machine (SVM) (15), Random Forest (RF) (12), Naive Bayes (NB) (16), Logistic Regression (LR) (17). These models were evaluated based on the performance metrics such as accuracy, precision, recall and F1 score.

3 Results and Discussions

This section describes the findings regarding the top features selected by filter method and the proposed hybrid approach also the performance of each model using the best features, as well as the experimental values of our assessment measures. This hybrid approach signifies a comprehensive feature selection process by combining the robustness of filter methods with the adaptability of the Cuckoo Search optimization algorithm. Initial selection via filter methods rapidly reduces dimensionality, while subsequent Cuckoo Search fine-tunes the selection in order to optimize SVM performance. This two-tiered methodology is anticipated to produce models with increased accuracy and decreased overfitting, particularly in intricate data sets such as those used for COVID-19 forecasting.

3.1 Experimental results of Blood test data

The top 10 features selected by individually applying Chi-squared and Mutual information on a dataset are LDH, LY, MO, EO, GENDER, CRP, AST, ALT, WBC and GGT. The features selected by these methods were same but there exists a difference in the rank obtained by each feature which is clearly shown in Figure 3. In order to arrive at a conclusion, we have combined these two methods to find the final set of features selected by the filter method.

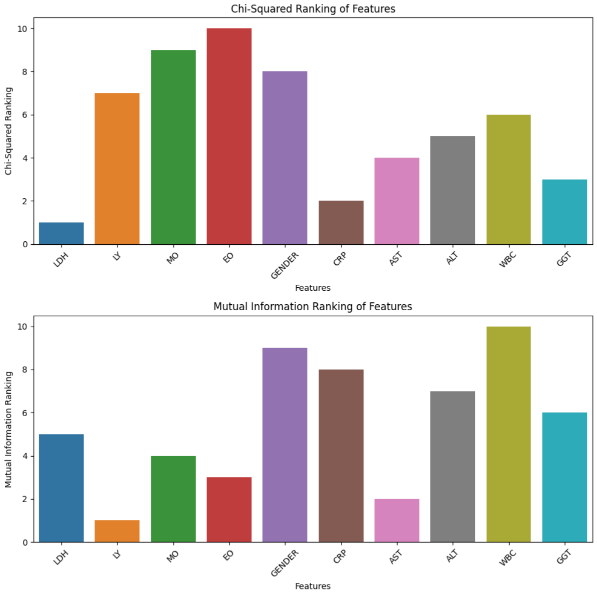

Fig 3. Feature Ranking by Chi-squared and Mutual Information

Figure 4 shows the list of features obtained when both the two types of filter methods (chi-squared and Mutual information) were combined. This is achieved by employing a voting mechanism in which a voting dictionary has been created to store the votes of each feature and based on the threshold value top features were identified. With this approach, the resultant features were the same set of features as all the features has scored a vote of 2 and considered pertinent by the filter methods.

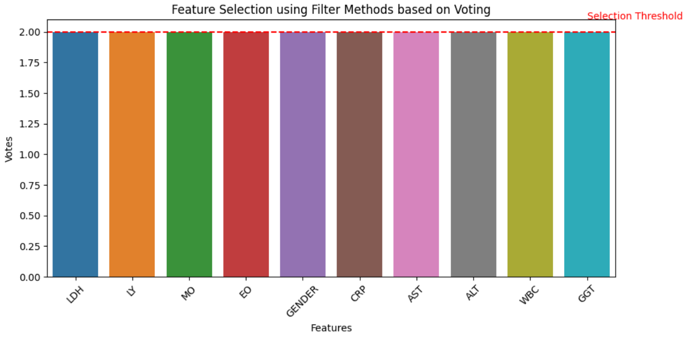

Fig 4. Feature Selection using filter method based on voting mechanism

The proposed feature selection technique combining both the filter and wrapper methods outperformed well with the datasets utilized in this study. In the dataset, the most crucial features pertaining to the COVID-19 severity has been identified. The selected features from the filter methods were further refined using the proposed Cuckoo search based SVM model. This proposed method finds the optimal set of features for predicting the severity in a COVID-19 patient. The proposed method makes use of nests and iteratively draw the best features from the feature subset. From the blood test data, features such as LDH, LY, EO, Gender, CRP were found by the proposed hybrid cuckoo-search method as the most severe indicators in predicting COVID-19 severity and it is depicted in Figure 5.

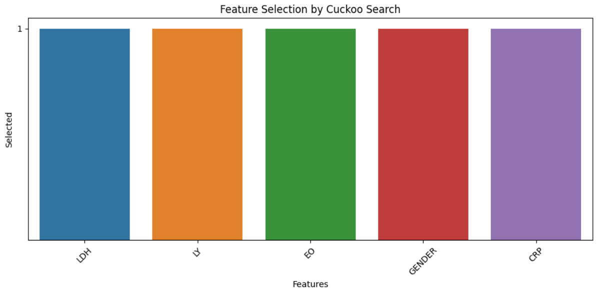

Fig 5. Feature Selection by Cuckoo-Search based SVM (Proposed method)

The selected features from the dataset have been applied separately on different machine learning models and the various performance metrics were calculated. SVM attained highest accuracy when compared with other ML models. Also, comparing with the existing methodologies, the proposed hybrid framework achieved the higher accuracy rate of 92% using the blood test dataset which is shown in Table 2.

Table 2. Performance metrics using blood test data

Classifiers	Accuracy	Precision	Recall	F1 Score
SVM	92%	90%	88%	89%
RF	77%	77%	72%	76%
NB	80%	78%	73%	78%
Logistic Regression	81%	80%	78%	80%

Table 3 shows the research findings of existing methodologies which showed the accuracy of SVM as 80% and 88%. Comparing the accuracy of SVM with the proposed approach, the proposed hybrid feature selection with cuckoo search based SVM method acquired 92 % accuracy. Among the existing methodologies, feature selection was adopted by (17) and some works performed diagnosis without feature selection techniques. Feature selection technique not only helps in increasing the accuracy of a machine learning model but also helps to identify the top features which has high impact with the target variable.

Table 3. Performance of Existing Techniques

References	Research Findings of Existing Techniques
Brinati et al. (2020) (17)	Accuracy of SVM in COVID-19 prediction using blood test ranges from 69% to 80%
Cabitza et al. (2020) (18)	The study used different blood test data with SVM-RFE as a feature selection method and the accuracy of SVM model was 88%
Hany et al. (2021) (19)	SVM achieved 88% accuracy

As Feature selection is applied in case of dealing with high dimensional data, there are two aspects which we have identified and found that applying feature selection in this dataset is necessary also provides better results. In the existing study, they haven't adopted any separate method for feature selection as they have calculated feature importance scores using the machine learning model. To make fair comparisons with the existing study, we have used the same dataset which they have used and found the optimal set of features in predicting the target variable. Also, we have experimented the same work without using feature selection technique and have used all the features to train the ML model. From the results, we could see that without feature selection, the accuracy of the ML model has been reduced when all the features have been used.

In the case of randomly generated data, an accuracy score of 60% and above would indicate that the algorithm is overfitting the data. Overfitting occurs when a model learns the training data too well, including the random noise present in the data, and performs poorly on unseen data. In the case of randomly generated data, achieving high accuracy may not be meaningful, as the data has no underlying patterns or relationships for the model to learn. It's important to evaluate the performance of ML algorithms on real-world or meaningful datasets to draw accurate conclusions about their capabilities. In our case, we have considered the real dataset collected from a hospital. In this regard, the performance measure of 80 % to 90% after training a machine learning model on this dataset may simply indicate that the model has effectively learned patterns and relationships within the medical data, leading to accurate predictions on the given dataset. Also, the quality of the dataset after applying pre-processing and feature selection steps also influence the model performance.

4 Conclusion

The proposed hybrid feature selection technique combines the benefits of deterministic statistical measurements and nature-inspired optimization. The approach adapted for SVM is a potential way for dealing with its application in forecasting diseases like COVID-19 might be enormous in healthcare decision-making. According to experimental results, the SVM model outperforms other classifiers achieving an accuracy of 92%. The top most features selected by the proposed method were LDH, LY, EO, Gender and CRP. Also, the execution time of the proposed algorithm is 0.6 seconds which proves the efficiency of the proposed algorithm. For the future work, it would be better to consider parameters like patient's medical history and their symptoms. Also, for the diagnosis and prediction of COVID-19, deep learning techniques could be utilized.

References

1) Meraihi Y, Gabis AB, Mirjalili S, Ramdane-Cherif A, Alsaadi FE. Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. *SN Computer Science*. 2022;3(4):1–35. Available from: https://doi.org/10.1007/s42979-022-01184-z.
2) Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*. 2021;24:1–17. Available from: https://doi.org/10.1016/j.imu.2021.100564.
3) Zhou K, Sun Y, Li L, Zang Z, Wang J, Li J, et al. Eleven routine clinical features predict COVID-19 severity uncovered by machine learning of longitudinal measurements. *Computational and Structural Biotechnology Journal*. 2021;19:3640–3649. Available from: https://doi.org/10.1016/j.csbj.2021.06.022.
4) Singh V, Poonia RC, Kumar S, Dass P, Agarwal P, Bhatnagar V, et al. Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. *Journal of Discrete Mathematical Sciences and Cryptography*. 2020;23(8):1583–1597. Available from: https://doi.org/10.1080/09720529.2020.1784535.
5) Bao FS, He Y, Liu J, Chen Y, Li Q, Zhang CR, et al. Triaging moderate COVID-19 and other viral pneumonias from routine blood tests. 2020. Available from: https://doi.org/10.48550/arXiv.2005.06546.
6) de Moraes Batista AF, Miraglia JL, Donato THR, Filho ADPC. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *MedRxiv*. 2020;p. 1–8. Available from: https://www.medrxiv.org/content/10.1101/2020.04.04.20052092v2.full.pdf.
7) Shoer S, Karady T, Keshet A, Shilo S, Rossman H, Gavrieli A, et al. A Prediction Model to Prioritize Individuals for a SARS-CoV-2 Test Built from National Symptom Surveys. *Med*. 2021;2(2):196–208. Available from: https://doi.org/10.1016/j.medj.2020.10.002.
8) Nemati M, Ansary J, Nemati N. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns*. 2020;1(5):1–8. Available from: https://doi.org/10.1016/j.patter.2020.100074.
9) Zhang J, Xiong Y, Min S. A new hybrid filter/wrapper algorithm for feature selection in classification. *Analytica Chimica Acta*. 2019;1080:43–54. Available from: https://doi.org/10.1016/j.aca.2019.06.054.
10) Singh N, Singh P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemometrics and Intelligent Laboratory Systems*. 2021;217:104396. Available from: https://doi.org/10.1016/j.chemolab.2021.104396.
11) Khan A, Khan A, Bangash JI, Subhan F, Khan A, Khan A, et al. Cuckoo Search-based SVM (CS-SVM) Model for Real-Time Indoor Position Estimation in IoT Networks. *Security and Communication Networks*. 2021;2021:1–7. Available from: https://doi.org/10.1155/2021/6654926.
12) Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems*. 2020;44(8):1–12. Available from: https://doi.org/10.1007/s10916-020-01597-4.
13) Karrar AE. The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*. 2022;10(2):375–384. Available from: http://section.iaesonline.com/index.php/IJEEI/article/viewFile/3730/697.
14) Ahmad HF, Khaloofi H, Azhar Z, Algosaibi A, Hussain J. An Improved COVID-19 Forecasting by Infectious Disease Modelling Using Machine Learning. *Applied Sciences*. 2021;11(23):1–38. Available from: https://doi.org/10.3390/app112311426.
15) Chauhan VK, Dahiya K, Sharma A. Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*. 2019;52(2):803–855. Available from: https://doi.org/10.1007/s10462-018-9614-6.
16) Chen S, Webb GI, Liu L, Ma X. A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*. 2020;192:105361. Available from: https://doi.org/10.1016/j.knosys.2019.105361.
17) Nusinovici S, Tham YC, Yan MYC, Ting DSW, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*. 2020;122:56–69. Available from: https://doi.org/10.1016/j.jclinepi.2020.03.002.
18) Cabitza F, Campagner A, Ferrari D, Resta CD, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2020;59(2):421–431. Available from: https://doi.org/10.1515/cclm-2020-1294.
19) Hany N, Atef N, Mostafa N, Mohamed S, Elsahhar M, Abdelraouf A. Detection COVID-19 using Machine Learning from Blood Tests. In: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). IEEE. 2021;p. 229–234. Available from: https://doi.org/10.1109/MIUCC52538.2021.9447639.