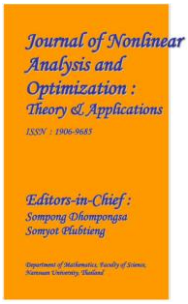
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2, No. 1: 2023
ISSN : **1906-9685**



BIG DATA ANALYTICS FRAMEWORK FOR COVID-19 SEVERITY PREDICTION USING MACHINE LEARNING

N. Hari Priya Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India
Dr.S. Rajeswari Associate Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India

Abstract:
COVID-19 detection has emerged as a high priority for global health systems in order to prevent its spread. Although RT-PCR is still the preferred approach for COVID-19 detection, the usefulness of symptoms data in predictive modelling is being explored steadily. The spread of Coronavirus Disease 19 (COVID-19) over the world is a major health risk that has affected many people. In the field of disease control, rapid diagnosis and precise identification of those who have been affected are critical. The COVID-19 Open Research Dataset Challenge (CORD-19) dataset, downloaded from Kaggle, is used for this research. The proposed approach consists of a data pre-processing phase and a feature selection strategy employing the wrapper method known as SVM-RFE to isolate the most relevant characteristics. The top features selected by SVM-RFE is obtained and they are used by ML models for prediction. Five distinct machine learning algorithms for further classification: Random Forest, SVM, Logistic Regression, XGBOOST and Ensemble Neural Network. Based on experimental data, ensemble neural network with parameter optimization technique surpassed all other classifiers with an accuracy of 97%.

Keywords: Machine Learning, Big data, Feature Selection, COVID-19, Ensemble Neural Networks.

I. INTRODUCTION

The new coronavirus SARS-CoV-2 causes COVID-19 (Coronavirus disease 2019), a highly contagious respiratory illness. In February 2020, the World Health Organization officially named the novel coronavirus-based infectious disease COVID-19 (coronavirus disease 2019), and in March 2020, it was declared as a pandemic [1]. The pandemic has caused an unexpected global health emergency, posing a challenge to governments, healthcare systems, and societies around the globe. Healthcare researchers and practitioners are focusing their efforts on identifying the elements that contribute to the severity of this condition, as millions of people are impacted. Maintaining a responsive healthcare system capable of providing fundamental services has proven challenging for a number of nations [2]. In the face of this urgency, big data analytics has become a viable approach, providing a strong way to extract knowledge, trends, and forecasting models from large and diverse datasets.

In this research, we provide an extensive framework for predicting the severity of COVID-19 by making use of big data analytics. Predicting the severity of an illness inflicted by coronavirus presents significant challenges for conventional methods. However, the use of big data analytics has the potential to reveal hidden connections between demographic, and clinical aspects that contribute to illness development. This approach uses sophisticated analytics methods like machine learning, data mining, and predictive modelling to uncover hidden yet crucial patterns in the collected data [3]**.**

In addition, this concept has wider implications than in patient care. It provides the potential to benefit healthcare practitioners in early identification and risk stratification, hence enabling focused interventions and resource allocation. Furthermore, it may help public health decision-making by shedding light on the causes of regional and population-specific variations in severity. In essence, the fusion of big data analytics with COVID-19 severity prediction stands as a testament to the possibility of technology-driven solutions in fighting global health problems [4]. This framework aims to contribute to a more informed, proactive, and efficient approach to controlling the impact of COVID-19 by making use of information.

The proposed work involves pre-processing the data to manage missing values, reducing redundant values and picking the most informative features. Recursive feature elimination (RFE) is a wrapper method that we employed after pre-processing to select the most important features. Five distinct types of machine learning were utilized for classification, they are Random Forest, SVM, Logistic Regression, XGBOOST, Ensemble Neural Network. Parameter optimization is a technique that can be used to improve the efficiency of machine learning models. In this research, the performance of the machine learning models under consideration has been optimized using grid search CV.

II. LITERATURE REVIEW

Using patient-specific clinical data, Nemati et al. [5] built a model to estimate how long COVID-19 inpatients would need to stay in the hospital. The study used dataset of 1182 hospitalized patients available publicly, assembled by researchers from multiple institutions and labs. Different statistical methods and ML approaches were used to create several survival analyses models. The stagewise gradient-boosting survival model yielded the most precise discharge-time estimate (C-index = 71.47). The findings indicated lower discharge probabilities for males and the more senior age groups.

Shoer et al. [6] built a prediction model using responses to nine simple survey questions. The data used in the study was collected from nearly two million Israeli adults who participated in a national survey about their symptoms. Out of a total of 43,752 adults, 498 reported being COVID-19 positive. Participants reported their age, gender, medical history, smoking status, and symptoms including fever, sore throat, cough, shortness of breath, and loss of taste or smell. After being trained with a Logistic regression method, the model achieved an AUC of 0.737.

Five alternative machine learning algorithms, including neural networks, random forest, gradient boost trees, logistic regression, and support vector machines, were used by the authors in [7] to propose a diagnosis of COVID-19. The Albert Einstein Hospital in Brazil contributed a dataset consisting of 235 blood samples and 102 confirmed cases of COVID-19. We selected 15 relevant variables from this dataset for the analysis, and their AUC was 85%, their sensitivity was 68%, and their specificity was 85%.

For primary screening of COVID-19 using routine blood testing, Aljame et al. [8] devised an ensemble learning strategy. The algorithm correctly identified COVID-19 positive cases 99.88% of the time when applied to data from 564 patients at the Albert Einstein Israelita Hospital in Sao Paulo, Brazil.

Five different machine learning models were considered by the authors in order to identify COVID-19 in routine blood samples: k-nearest neighbors, support vector machines (SVM), naïve bayes (NB), logistic regression (LR), and random forests (RF). 1,624 routine blood samples were taken from patients hospitalized to the Italian hospital San Raphael (52% COVID-19 positive). Models achieved between 74% and 88% accuracy, 70% to 89% AUC, 79% to 92% sensitivity, and 74% to 90% specificity [9].

Using information from Chinese hospitals including Wuhan Union Hospital, Bao et al. [10] analysed 294 blood samples using Random Forest and SVM. A total of 208 patients with moderate COVID-19 and 86 patients with mild non-COVID-19 viral pneumonia were used to assess the efficacy of this

technique. SVM achieves a higher rate of accuracy (84%) than the random forest classifier when tested on the same set of fifteen features. They summed up by adding that their findings are consistent with medical and machine learning theories, and that their approach has the potential to add another rapid COVID-19 testing option that can be performed in facilities that are prepared to perform normal blood tests.

III. RESEARCH METHODOLOGY

Today's data-driven environment necessitates the use of machine learning methods for extracting insights and developing predictive models. We have used Python language to implement the work and the whole process flow comprises the complete set of Machine Learning procedure, including data pre-processing, feature selection, hyperparameter tuning, and model evaluation.

As the first step, data pre-processing is carried out using mean imputation method to fill in the missing values and following the partitioning of the dataset into training and testing sets, the dataset is split into feature matrices and target variables. Standard Scaler () uses standardization to perform feature scaling. This standardization assures that features are on the same scale, boosting model performance. For feature selection, Support Vector Machine-Recursive Feature Elimination (SVM-RFE) has been used to find the optimal set of features.

Further, Grid Search CV is used to fine-tune the hyperparameters of an Ensemble Neural Network. The goal of this exhaustive search for optimal hyperparameters is to improve the model's predictive capacity by investigating different sets of input parameters. At last, the models' performance is assessed. The test set is used to evaluate the models' performance on unknown data using both the selected features from SVM-RFE and the tuned Ensemble Neural Network.

DATASET DESCRIPTION

The research dataset was obtained from Kaggle [11] and is part of the COVID-19 Open Research Dataset Challenge (CORD-19). It comprises of 127 instances of patient's data. Age, gender, body temperature, dry cough, sore throat, weakness, breathing problem, drowsiness, fever, history of travel to infected countries, diabetes, stroke, chest pain, high blood pressure, loss of smell, change of appetite are the features of the dataset. At the pre-processing stage, we removed unnecessary data pieces and missing values are handled by mean imputation approach. Age and Body Temperature distribution of all patients in the dataset is depicted in Figure 1. As a part of descriptive analytics, for all the features, bivariate analysis is performed and it is shown in Figure 2.

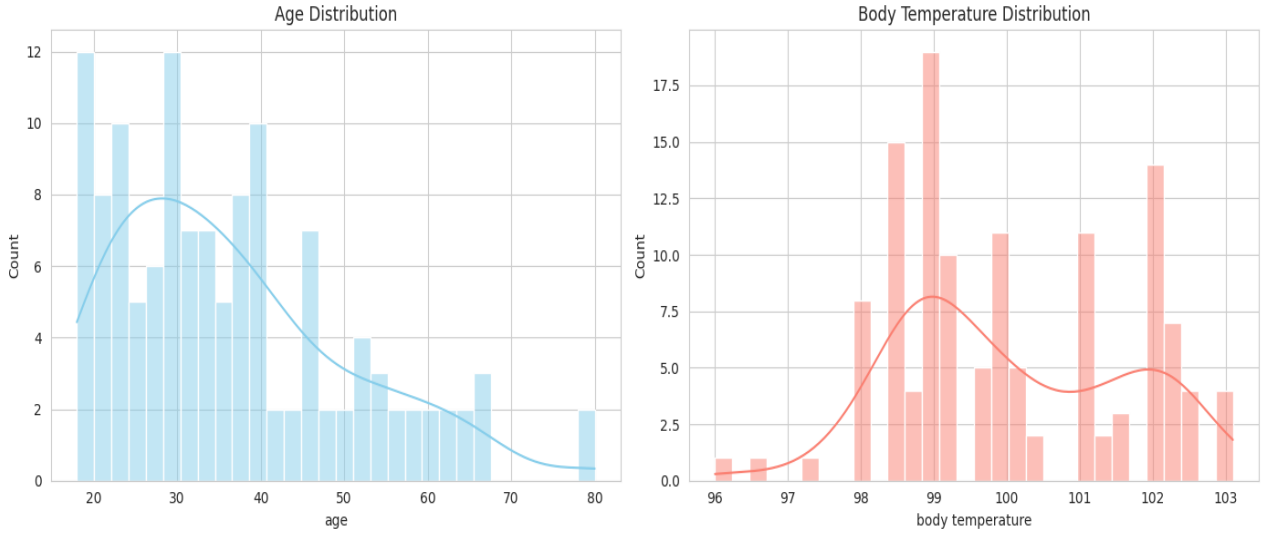


Figure 1. Age and Body Temperature Distribution

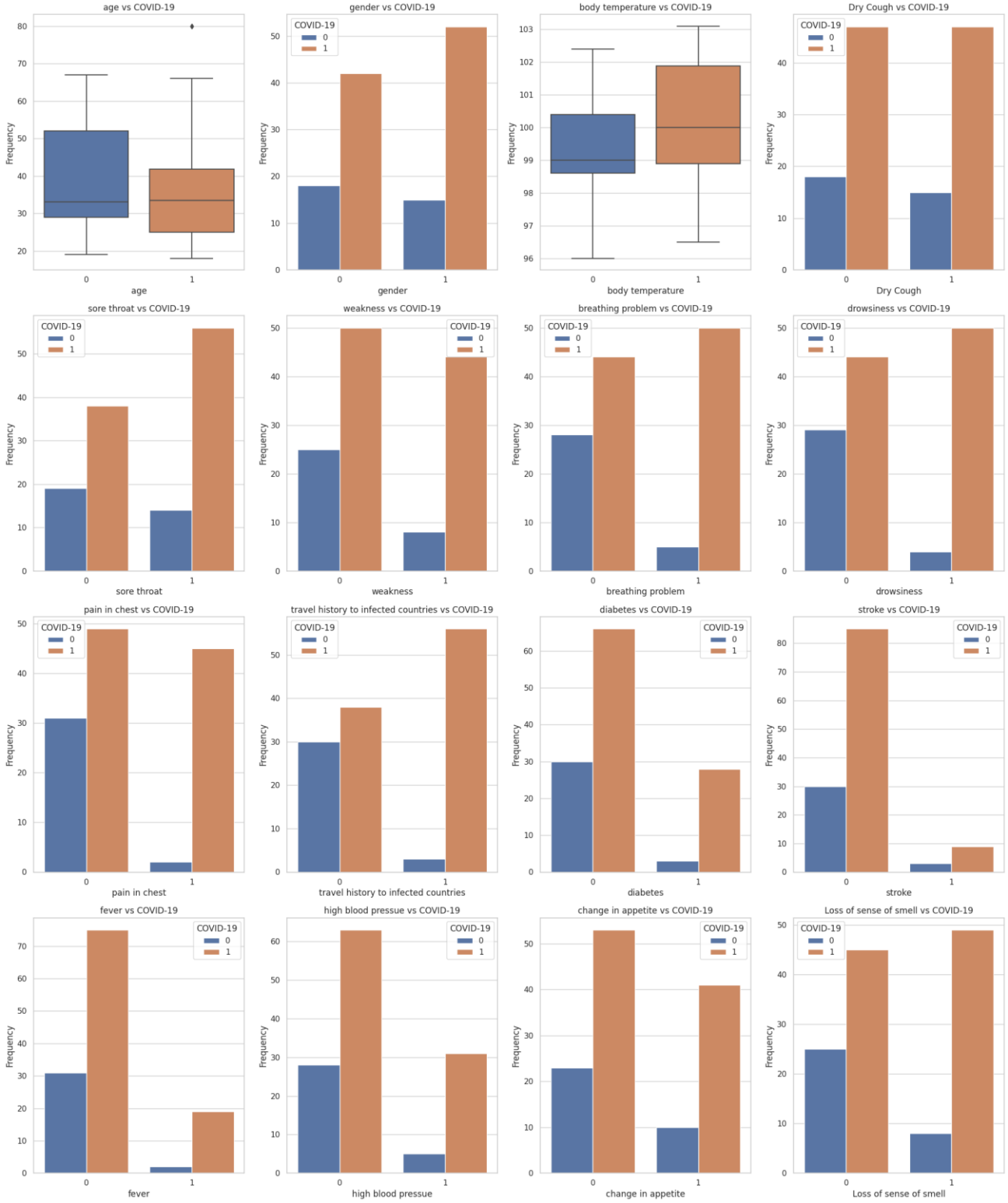


Figure 2. Bivariate Analysis of all Features

FEATURE SELECTION

The process of feature selection is a critical component within the field of machine learning, as it entails the selection of the most pertinent features in order to construct robust models. Selecting the most important features helps machine learning models perform better and provides new insights into the data's underlying structure and relationships. Recursive Feature Elimination (RFE), a wrapper method based on support vector machines, was utilized to calculate the linear correlation between the variables and find the most important features in the dataset. Features are chosen according to how

well they perform in a particular machine learning method. The SVM method is used to prioritize characteristics for inclusion in the model, and the model's performance is measured based on the features that were prioritized.

The features that are most significantly associated with the presence of COVID-19 are ranked first, as illustrated in Figure 3. Prior to implementing the feature selection method, the dataset was divided into training and test sets.

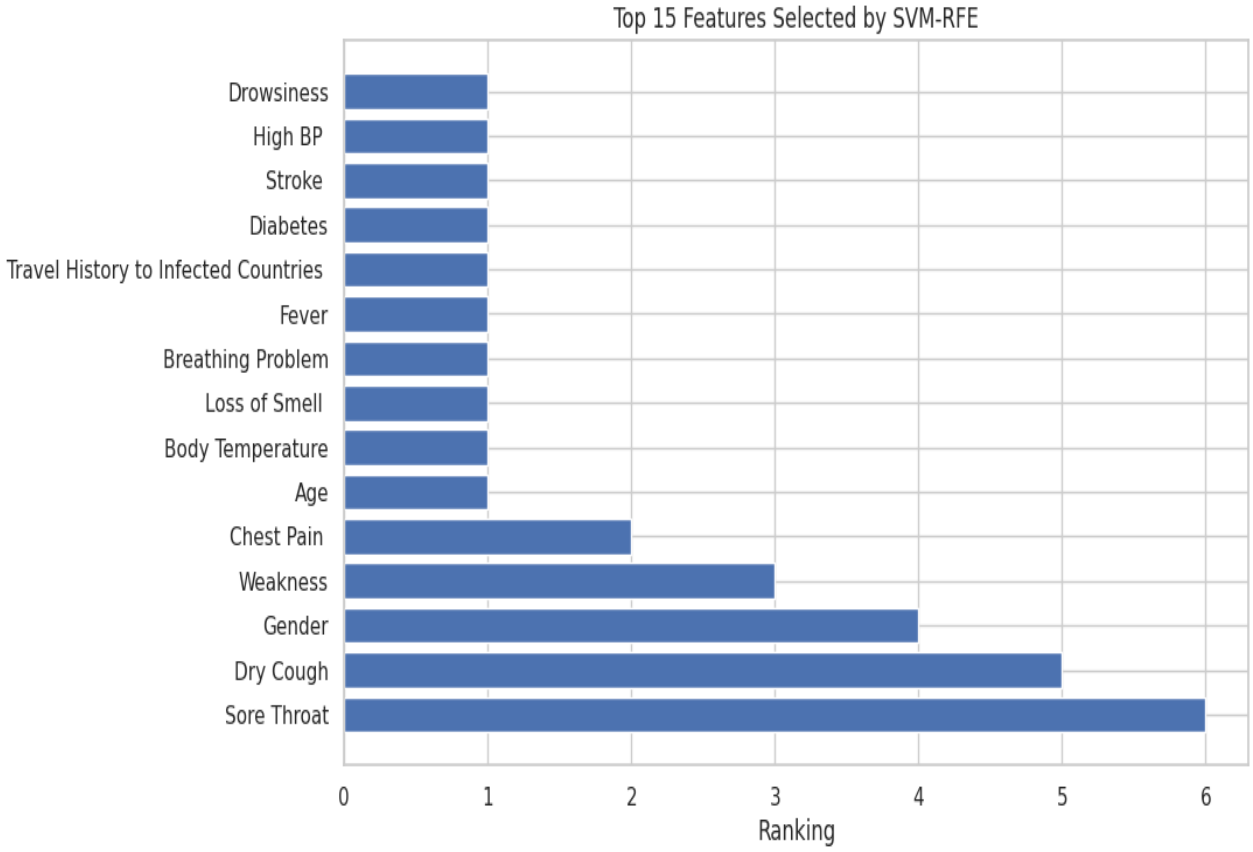


Figure 3. Visualization of Feature Ranking based on SVM-RFE

MACHINE LEARNING ALGORITHMS

In this study, five classification algorithms were used: Random Forest, SVM, Logistic Regression, XGBOOST, Ensemble Neural Network.

Evaluation Metrics

To compute the performance of the different ML models, we evaluated true positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN).

TP: A true positive is that a person who has Covid-19 is correctly identified as positive by a diagnostic test.

FP: A false positive would occur if a person who does not have Covid-19 is incorrectly identified as positive by a diagnostic test.

TN: A true negative would occur if a person who does not have Covid-19 is correctly identified as negative by a diagnostic test.

FN: A false negative would occur if a person who actually has Covid-19 is incorrectly identified as negative by a diagnostic test.

The different performance metrics considered in this study are listed below;

Accuracy: This is a measure of how often a classifier is correct. It is calculated as the ratio of the number of correct predictions to the total number of predictions made.
The formula for accuracy is:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision: This is a measure of how many of the positive predictions made by a classifier are correct. It is calculated as the ratio of the number of true positive predictions to the total number of positive predictions made.
The formula for precision is:

$$Precision = TP / (TP + FP)$$

Recall: This is a measure of how many of the actual positive cases a classifier is able to correctly identify. It is calculated as the ratio of the number of true positive predictions to the total number of actual positive cases.
The formula for recall is:

$$Recall = TP / (TP + FN)$$

F1 Score: This is a measure of the overall accuracy of a classifier that takes both precision and recall into account. It is calculated as the harmonic mean of precision and recall.
The formula for F1 score is:

$$F1\ Score = 2 * (Precision * Recall)/ (Precision + Recall)$$

IV. RESULTS AND DISCUSSION

In this study, we found that the features that has a rank 1 are the best predictors of COVID-19 illness severity. The features with rank 1 are drowsiness, high BP, stroke, diabetes, travel history to infected countries, fever, breathing problem, loss of smell, body temperature and age. To further identify the indications of COVID -19 severity and to increase the prediction accuracy, we have used ensemble neural network with SVM. The symptoms data are used to make predictions about COVID-19 using different machine learning models. Five different classification-based ML models such as Random Forest, SVM, Logistic Regression, XGBoost and Ensemble Neural Network were used in this study, along with the wrapper feature selection method known as SVM-RFE. Among the different ML algorithms, ensemble neural network achieved the highest accuracy of 97%. The performance comparison of algorithms is shown in Table 1.

Table 1. Performance Metrics of ML Algorithms

Algorithm/Measures	Accuracy	Precision	Recall	F1 Score
Random Forest	91%	92%	88%	89%
SVM	89%	91%	87%	86%
Logistic Regression	90%	91%	89%	90%
XGBoost	86%	87%	85%	85%
Ensemble Neural Network	97%	96%	95%	96%

V. CONCLUSION

In pandemic scenarios, it is critical to identify persons who are vulnerable to infection and disease spread in order to devise treatment and prevention strategies. The most essential symptoms were discovered in the current study. The top features selected by SVM-RFE is obtained and they are used by ML models for prediction. According to the experimental results, ensemble neural network

classifier outperforms the other classifiers used in this work for COVID-19 diagnosis. When the system is overburdened as a result of congestion, this strategy can help hospitals and medical institutions decide which patients require immediate attention before other patients, hence avoiding delays in providing the necessary care.

REFERENCES

[1] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health, 20, 100178.
[2] Saadatmand, S., Salimifard, K., Mohammadi, R., Kuiper, A., Marzban, M., & Farhadi, A. (2023). Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. Annals of Operations Research, 328(1), 1043-1071.
[3] Zhou, K., Sun, Y., Li, L., Zang, Z., Wang, J., Li, J., ... & Guo, T. (2021). Eleven routine clinical features predict COVID-19 severity uncovered by machine learning of longitudinal measurements. Computational and structural biotechnology journal, 19, 3640-3649.
[4] Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. Journal of big Data, 9(1), 3.
[5] Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. Patterns, 1(5), 100074.
[6] Shoer, S., Karady, T., Keshet, A., Shilo, S., Rossman, H., Gavrieli, A., ... & Segal, E. (2021). A prediction model to prioritize individuals for a SARS-CoV-2 test built from national symptom surveys. Med, 2(2), 196-208.
[7] de Moraes Batista, Andre Filipe, et al. "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach." MedRxiv (2020): 2020-04.
[8] AlJame, Maryam, et al. "Deep forest model for diagnosing COVID-19 from routine blood tests." Scientific reports 11.1 (2021): 16682.
[9] Cabitza, Federico, 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) 59.2 (2021): 421-431.
[10] Bao, Forrest Sheng, et al. "Triaging moderate COVID-19 and other viral pneumonias from routine blood tests." arXiv preprint arXiv:2005.06546 (2020).
[11] https://www.kaggle.com/datasets/bitsofishan/covid19-patient-symptoms