Journal of Information Systems Engineering and Management
2025, 10(24s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing COVID-19 Prediction Using Machine Learning: A Comparative Analysis of Feature Selection and Classification Techniques

L.William Mary[1], Dr. S. Albert Antony Raj[2]
[1]*Research Scholar, Department of Computer Applications, SRM Institute of Science and Technology, Kattankulathur, Chennai.*
[2]*Professor, Deputy Dean, Department of Computer Applications SRM Institute of Science and Technology, Kattankulathur, Chennai.*
wl6649@srmist.edu.in , alberts@srmist.edu.in
Corresponding Author: Dr. S. Albert Antony Raj

ARTICLE INFO

Received: 19 Dec 2024

Revised: 10 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: The early and accurate detection of COVID-19 remains a life-threatening challenge in medical analysis. Machine learning is used for predicting disease outcomes based on clinical parameters. This analysis proposes a comparative analysis of feature selection method and classification techniques to enhance COVID-19 detection accuracy using blood biomarkers. We used a pensourse dataset of 1,724 cases, including 35 features. To improve the model performance data preprocessing process included outlier handling, normalization, and transformation techniques to improve model performance. To identify the relevant features, we employed the three-feature selection methods Chi-Square test, Pearson correlation coefficient, and Random Forest. The model prediction accuracy was enhanced using a stacking ensemble classication techniques. The machine learning based classification models effectively predicted COVID-19 infectious disease using blood biomarkers with optimized feature selection techniques.

Objectives: To enhance the accuracy of COVID-19 prediction using machine learning techniques by applying feature selection and classification techniques on blood biomarkers.

Methods: The comparative analysis utilized a publicly available dataset containing 1,724 cases with 35 attributes. Data preprocessing involved outlier handling, normalization, and transformation techniques. Employed Chi-Square test, Pearson correlation coefficient, and Random Forest feature selection techniques. Stacking ensemble classification algorithm was utilized for the better performance of a model.

Results: The classification models demonstrated efficiency in predicting COVID-19 using blood biomarkers. Optimized feature selection significantly improved predictive accuracy, highlighting the importance of selecting relevant features for model performance enhancement.

Conclusions: This study showcases the potential of ML-driven approaches for COVID-19 detection, emphasizing the role of feature selection in improving classification accuracy. The findings contribute to the advancement of diagnostic tools, offering a data-driven solution for rapid and reliable COVID-19 screening.

Keywords: COVID-19 detection, machine learning, blood biomarkers, feature selection, stacking ensemble, classification models

INTRODUCTION

The COVID-19 pandemic has highlighted the need for rapid and accurate diagnostic methods to facilitate early intervention on healthcare systems. Traditional diagnostic techniques, such as RT-PCR and antigen tests, remain the gold standard methods and they have inherent limitations, including high costs, processing delays, and the potential for false negatives. As a result, there is a growing interest in alternative diagnostic approaches that leverage clinical

and laboratory data for faster and more reliable detection. Machine learning (ML) has emerged as a transformative tool in medical diagnostics, offering the ability to analyze large datasets and extract meaningful patterns for disease prediction. This study investigates the application of ML-based classification models for COVID-19 detection using blood biomarkers. A publicly available dataset of 1,724 cases with 35 attributes was utilized. Data preprocessing techniques, including outlier handling, normalization, and transformation, were applied to improve data quality. Three feature selection methods—Chi-Square test, Pearson correlation coefficient, and Random Forest—were employed to identify the most informative biomarkers. A stacking ensemble classifier was then developed, integrating multiple supervised learning models to enhance predictive accuracy. The findings highlight the potential of AI-driven diagnostic tools for rapid and reliable COVID-19 screening, addressing the limitations of conventional testing methods.

Delivering high-quality services is challenging. Effective illness management and precise diagnosis are critical elements of healthcare. Clinical research shows that variations in blood indicators are linked to the disease severity and the death rates [12]. Machine learning is transforming for the better progression toward prediction. The healthcare industry is adopting machine learning to improve efficiency [13]. This technique is essential in healthcare to detect patterns within large datasets and diagnose the disease. Previous studies have predicted COVID-19 mortality using blood biomarkers and machine-learning approaches. The outcome of the prediction method effectively predicted the non-linear relationships among blood biomarkers [14]. In addition, the prediction includes traditional assessment techniques for monitoring pulmonary diseases, such as X-rays and CT scans [15,16]. Prompt detection and virus diagnosis are essential for infection control and reducing mortality rates. The figure below illustrates a logarithmic scale representing the total monthly deaths from COVID-19 for March 2024.

The study aims to develop a stacking ensemble classifier algorithm designed to accurately predict COVID-19 test outcomes. The approach encompasses data cleaning, preprocessing, feature selection, classification, and comprehensive statistical analysis of the results.

Machine learning (ML) is an important area of computational algorithms replicating human intelligence through the automated learning process. The complex algorithmic techniques developed using a machine learning approach for advanced data analysis. List of machine-learning algorithms:

1. Supervised Learning: It employs labeled data to predict future values. The learning process begins with a training dataset, and targeted strategies are formed to forecast the outcomes for the samplings.
2. Unsupervised Learning: This algorithm utilizes unclassified or unlabelled datasets for the process. The learning process infers a function to uncover hidden insights or patterns within the data that lack labels.
3. Semi-supervised learning: Labeled and unlabelled datasets are utilized in the training process, offering a solution.
4. Reinforcement learning: These approaches deliver feedback to the learning system to detect and correct errors through the error process. The model efficiency and performance are improved by applying this process to the given context.

RELATED WORK

Fernandes et al. developed a model using a database of 1,040 Brazilian patients. This model incorporated routine biomarkers, such as ferritin, CRP, and lymphocytes, with the intensive care unit (ICU) score to predict ICU admission, mechanical ventilation, and mortality [26]. Famiglini et al. developed a model to predict ICU admissions using Complete Blood Count (CBC) data from patients in Italy [27]. Ardabili SF et al. [28] investigated the COVID-19 pandemic by comparing machine learning models with soft computing. According to their investigation, the algorithms Adaptive Neuro-Fuzzy Inference System and the Multilayer Perceptron (MLP) neural network showed strong generalizability for long-term predictions. Deep learning is an advanced approach to COVID-19 detection. Zhang et al. [29] analyzed CT images associated with COVID-19 by applying a SqueezeNet (SN) model with an advanced bypass mechanism. The blood biomarkers Lymphopenia is often observed in COVID-19 diagnosed patients J. Yang et al. [30]. Individuals who test positive significantly reduced calcium levels compared to those who test negative, as analyzed by the authors X. Zhou et al. [31]. C-reactive protein (CRP) is recognized as a critical, independent marker for assessing the disease severity mentioned by the authors Y. Luan et al. [32] and X. Luo et al. [33].

METHODS

Data Collections

Blood biomarkers dataset is collected from a publicly available source Zendo platform. The total number of data comprising 1,724 cases with 35 attributes, was used in this comparative analysis. The dataset includes 814 COVID-19 positive data and COVID-19 910 negative data. The target variable is the numerical values of the confirmed cases.

Handling Outliers and Data Transformation

Z-score Normalization

Outliers affect the performance of the models. The normalization technique is used to rescale the values to avoid the outliers. This technique improves detection to achieve higher accuracy. The normalization procedure validates the scaling attribute values. The normalization technique includes Z-score Normalization, and Min-Max which effectively manage outliers and improve classification performance in datasets. Data points outside these thresholds are replaced with the corresponding boundary values, ensuring the dataset remains within the specified range.

The figure below illustrates the dataset before and after addressing the outliers. Initially, the dataset contained 12 outliers, but after applying the outlier handling process, this number was reduced to zero, as shown in the box plot.

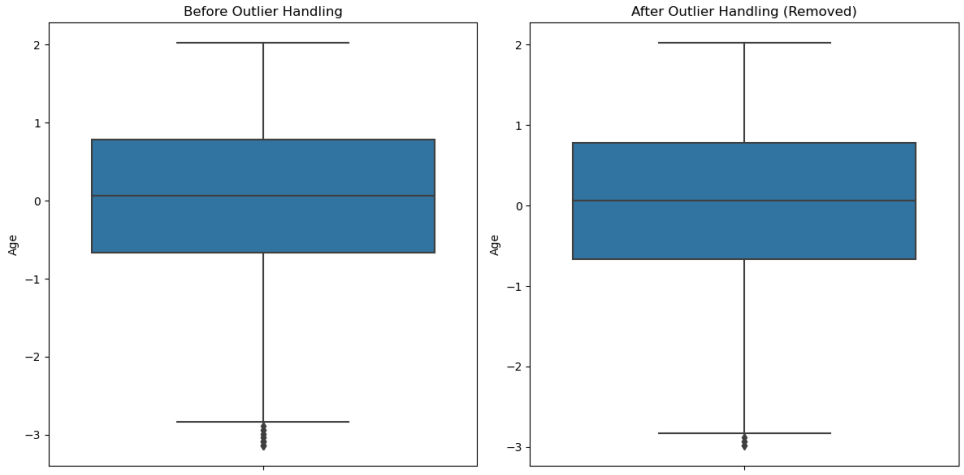

Fig.1 Illustration of before and after outliers using boxplots – Z-score

Data Transformation

Data transformations can significantly impact feature selection. The Yeo-Johnson transformation performs effectively on the applied dataset that handles positive and negative values. Using the same transformation parameter (λ) for both responses provide a flexible way to normalize the data [34]. Data transformation was used to scale and normalize the characteristics in the COVID-19 blood test dataset before analysis. The converted dataset is shown in the table below, which includes the first few rows of processed data:

Table:1 Transformed Dataset Utilizing the Yeo-Johnson Transformation

Gender	Age	CA	CK	CREA	ALP	GGT
0.862219	1.168527	-0.622590	-0.083377	0.503645	0.527782	0.138489
0.862219	-0.624893	-1.422612	1.076256	0.069460	-0.775532	1.100884
0.862219	-0.278070	-0.498112	0.158191	0.150655	0.145013	1.460957
-1.159983	1.168527	0.417498	0.389404	-0.650135	1.089648	1.609782
0.862219	0.968414	-0.749490	-0.530324	1.412117	-0.444731	0.025868

Skewness is a statistical word that describes an asymmetric distribution. This analysis evaluated the skewness of various features in the transformed COVID-19 blood test dataset to understand their distribution characteristics. A skewness value close to zero suggests a symmetrical distribution, while values less than -0.5 or more excellent than

+0.5 indicate moderate to strong skewness. The table below summarizes the skewness values for each feature in the transformed dataset:

Table:2 Skewness of the transformed dataset

Feature	Skewness Value
Gender	-0.298138
Age	-0.056608
CA	-0.476722
CK	-0.068675
CREA	0.068658
ALP	-0.256954
GGT	0.032811
GLU	-0.570121
AST	0.032283
ALT	0.018231
LDH	-0.001766
PCR	-0.096825
KAL	-0.022884
NAT	-0.129594
UREA	-0.003446
WBC	-0.008601
RBC	0.029897
HGB	-0.041138
HCT	-0.006776
MCV	0.222113
MCH	0.363314
MCHC	0.056170
PLT1	0.059178
NE	-0.048072
LY	-0.025680
MO	0.097758
EO	0.276472
BA	0.008532
NET	-0.011013
LYT	-0.012490
MOT	-0.062369
EOT	0.568107
BAT	1.075967
UREA	-0.003446
WBC	-0.008601
RBC	0.029897

Optimizing Features

To detect COVID-19 positive patients the current study employed three distinct feature selection methods, the Chi-Square test, Pearson correlation coefficient, and Random Forest. We computed feature importance scores for each feature using these techniques and established an average score as a threshold for feature selection. Constructed the classification model based on the selected features that exceeded this threshold across all methods. The table below presents the list of features along with their Chi-Square test scores and corresponding P-values.

Table:3 Analyzing feature importance with Chi-Square Scores and P-values

Feature	Chi-Square Test	P-Value
LDH	318.119746	3.719333e-71
BA	286.423439	2.991390e-64
CA	270.088184	1.085650e-60

EO	268.998331	1.875945e-60
AST	255.582609	1.575553e-57
LYT	197.948025	5.856391e-45
PCR	171.652208	3.223426e-39
MOT	171.576357	3.348755e-39
ALT	147.149301	7.279364e-34
WBC	124.842868	5.508865e-29
BAT	124.148900	7.815326e-29
EOT	107.737924	3.067760e-25
CK	90.563699	1.791166e-21
GGT	67.940283	1.685229e-16
NE	48.075967	4.100215e-12
NAT	38.644784	5.083895e-10
NET	37.064800	1.142680e-09
HGB	33.554569	6.929243e-09
LY	32.223820	1.373960e-08
RBC	29.117469	6.811997e-08
HCT	27.033208	1.999900e-07
ALP	19.444370	1.035727e-05
PLT1	19.185339	1.186209e-05
GLU	17.712166	2.569778e-05
MO	17.465624	2.925499e-05
MCHC	10.292700	1.335575e-03
CREA	6.827556	8.976200e-03
MCV	4.370262	3.657137e-02
KAL	4.257041	3.908796e-02
UREA	1.131593	2.874358e-01
AGE	0.393773	5.303227e-01
MCH	0.223118	6.366744e-01

The table above presents insights into feature importance analysis based on Chi-Square Scores and associated P-Values for the COVID-19 blood test dataset.

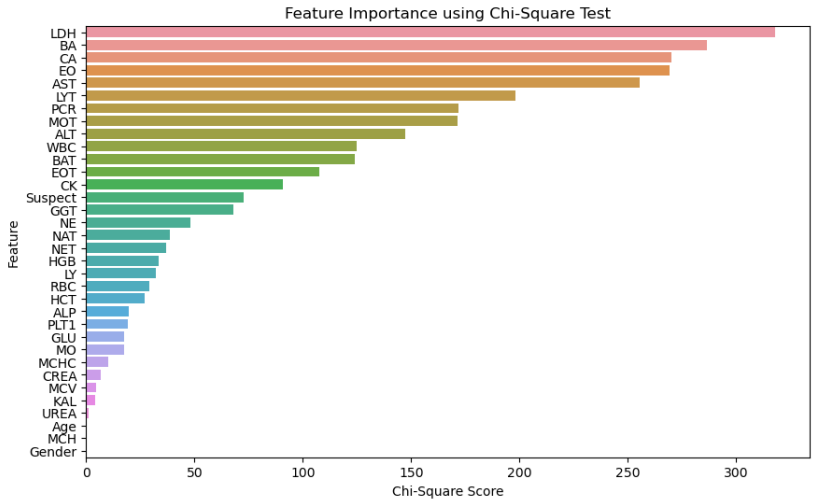

Fig.2 Measuring Feature Importance through Chi-Square Analysis

The Chi-Square test analyse the target variable and expected values relationships. This approach is based on the premise that features independent of the target variable contribute minimal information for classification.

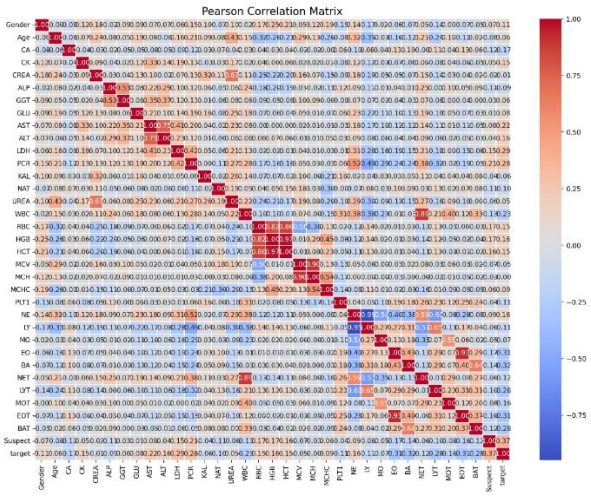

Fig.3 Evaluation of the Pearson Correlation Matrix for blood test data related to COVID-19

The study using Pearson correlation method to evaluate the correlations between the variables and determines the linear relationships ranges from -1 to 1. It is widely used in statistics to identify correlations, trends amid the largest data. This method is represented by the r-value, which ranges from -1 to +1. The r-value of +1 indicates a positive relationship, 0 shows no relationship between the variables, and -1 specifies a negative one. This study examined disease-positive and disease-negative elements in the dataset as variables. The correlation coefficient was computed for all aspects, as shown in the picture.

R-values greater than 0.05 indicate a positive relationship, where an increase in one factor may lead to a rise in another. Conversely, r-values below 0 imply an inverse relationship, where a decrease in one factor may cause a reduction in another. The r-value of 0.85 is the strong positive correlation between the two variables, indicating a significant direct link [35].

Savitzky–Golay filter – Smoothing Techniques

Smoothing filters are critical tools for decreasing noise and improving data quality in analysis. Using approaches like the Savitzky-Golay filter, we can get a more detailed depiction of the dataset's underlying trends. The Savitzky-Golay filter is helpful since it smoothes data while keeping crucial features like peaks and troughs. This technique is beneficial in contexts where maintaining the shape of the data distribution is vital [38].

The table below presents the smoothed data obtained using the Savitzky–Golay filter, showcasing the first few rows of the processed dataset:

Table:4 Data Smoothing with the Savitzky-Golay Filter

S. No	Age	Gender	CA	RBC	HGB
0	0.644987	0.951049	2.026224	4.602517	14.498252
1	0.535730	0.925874	2.076126	4.497128	13.854452
2	0.431769	0.882051	2.117403	4.407337	13.299308
3	0.333103	0.819580	2.150056	4.333145	12.832821
4	0.239732	0.738462	2.174084	4.274550	12.454988

Reducing Dimensionality

To effectively reduce the dimensionality of the blood test dataset, we utilized the method called PCA. It helps machine learning algorithms work faster and more efficiently [39]. PCA is a statistical method that changes correlated variables into uncorrelated ones through orthogonal transformation. Although reducing the number of variables may impact accuracy, PCA is particularly useful for high-dimensional datasets where there are more columns than rows [40].

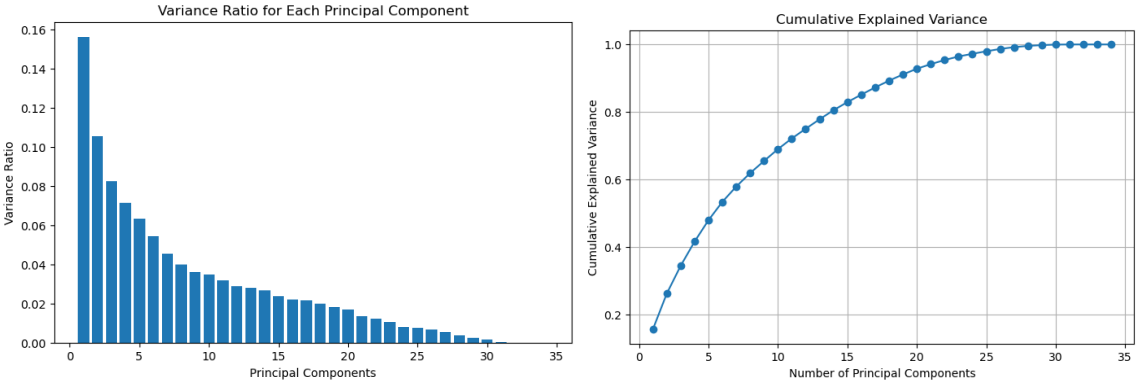

Fig.4 Dimensionality reduction using PCA

In the analysis, we found that 22 components are needed to explain 95% of the variance in the dataset.

Table:5 Reduced dataset - generated through PCA

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC22
-0.061752	0.962991	-0.893140	2.759184	-1.164073	-0.940033	-0.554908	-0.556493
0.145286	-1.557350	2.083196	-0.164105	0.546332	-0.708482	-0.343300	0.332257
0.649736	-2.138058	0.095132	-0.503751	-0.074827	-0.287020	0.046354	0.397289
-2.699535	1.447374	1.428619	1.368538	1.932497	0.173227	0.137474	0.107161
-1.063620	2.276196	-4.392559	1.628325	1.747231	0.624216	-2.683068	-0.729635

RESULTS

Comparative Analysis of Machine Learning Algorithms

The accuracy of predictive model's depends on the volume and high-quality data [47]. Both regularized versions of Logistic Regression performed similarly, achieving an accuracy of 0.82. These models are known for their interpretability, which makes them useful in therapeutic contexts where understanding feature contributions is critical. XGBoost and Gradient Boosting Machine algorithms predicts with accuracies of 0.81 and 0.80, respectively. These techniques are very useful when dealing with large datasets. While DNN and ANN recorded lower accuracies of 0.79 and 0.78, they still demonstrated reasonable recall rates. In contrast, Naive Bayes and Decision Tree algorithms were the least effective in this analysis, with accuracies of 0.77 and 0.75, respectively. Their lower precision and F1 scores may not be as suitable for this classification task.

Table:6 Comparison of Classic Algorithms Accuracy

Techniques	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Random Forest	0.82	0.83	0.79	0.81	0.89
LightGBM	0.82	0.84	0.77	0.81	0.89
Support Vector Machine	0.82	0.82	0.80	0.81	0.86
Logistic Regression with Regularization Lasso	0.82	0.83	0.78	0.81	0.88
Logistic Regression with Regularization Ridge	0.82	0.83	0.78	0.80	0.88
Logistic Regression	0.81	0.79	0.80	0.80	0.86
XGBoost	0.81	0.82	0.77	0.80	0.89
AdaBoost	0.81	0.81	0.77	0.79	0.86

Multilayer Perceptron	0.81	0.84	0.74	0.79	0.87
Gradient Boosting Machine	0.80	0.82	0.76	0.78	0.89
DNN	0.79	0.75	0.81	0.78	0.84
K-Nearest Neighbors	0.78	0.76	0.77	0.77	0.84
ANN	0.78	0.77	0.77	0.77	0.85
Naive Bayes	0.77	0.75	0.77	0.76	0.84
Decision Tree	0.75	0.72	0.76	0.74	0.75

The findings propose that Random Forest and LightGBM are the best.

DISCUSSION

The results of this study demonstrate that ML-based classification models can effectively predict COVID-19 infection using blood biomarkers. The application of feature selection techniques significantly improved model accuracy by reducing irrelevant or redundant attributes, thereby enhancing interpretability and efficiency. Among the three feature selection methods tested, the Random Forest approach provided the most robust results, indicating its ability to identify key biomarkers with strong predictive power. The stacking ensemble classifier further enhanced performance by combining the strengths of multiple supervised learning models. This approach leverages the diverse learning capabilities of individual classifiers, reducing bias and improving overall generalizability. Compared to the traditional ML models, the ensemble classifier exhibited higher accuracy, sensitivity, and specificity, reinforcing its potential for real-world clinical applications. The dataset used in this study was publicly available. Future research should also investigate the integration of deep learning models and hybrid ML techniques to enhance diagnostic accuracy. This study contributes the effectiveness of feature selection and ensemble learning for COVID-19 detection.

REFERENCES

[1] Ali N. (2020). Elevated level of C-reactive protein may be an early marker to predict risk for severity of COVID-19. J Med Virol. 92(11):2409-24.doi: 10.1002/jmv.26097.

[2] Janairo GIB, Yu DEC, Janairo JIB. (2021). A machine learning regression model for the screening and design of potential SARS-CoV-2 protease inhibitors. Netw Model Anal Health Inform Bioinform. 10(1):51. doi: 10.1007/s13721-021-00326-2.

[3] Zuin G, Araujo D, Ribeiro V, (2022). Prediction of SARS-CoV-2-positivity from million-scale complete blood counts using machine learning. Commun Med (Lond). 2:72. doi: 10.1038/s43856-022-00129-0.

[4] Qomariyah, Nunung Nurul, Purwita, et al.(2021). A tree-based mortality prediction model of COVID-19 from routine blood samples, in: 2021 International Conference on ICT for Smart Society (ICISS), IEEE, doi:10.1109/ICISS53185.2021.9533219.

[5] Dabbagh R, Jamal A, Bhuiyan Masud JH, et al.(2023). Harnessing Machine Learning in Early COVID-19 Detection and Prognosis: A Comprehensive Systematic Review. Cureus.15(5):e38373. doi: 10.7759/cureus.38373.

[6] Fernandes, F.T., de Oliveira, T.A., Teixeira, C.E. *et al.*(2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Sci Rep* **11**, 3343. doi.org/10.1038/s41598-021-82885-y

[7] L. Famiglini, A. Campagner, A. Carobene, et al.(2022). A robust and parsimonious machine learning method to predict ICU admission of COVID-19 patients. Med Biol Eng Comput. 1–13. doi: 10.1007/s11517-022-02543-x.

[8] Ardabili SF, Mosavi A, Ghamisi P, et al.(2020). Covid-19 outbreak prediction with machine learning. Algorithms.doi.org/10.3390/a13100249

[9] Zhang Y, Khan MA, Zhu Z, et al. (2023): SqueezeNet-guided ELM for COVID-19 recognition. Comput Syst Sci Eng.46(1):13.

[10] J. Yang, M. Zhong, E. Zhang, et al. (2021). Broad phenotypic alterations and potential dysfunction of lymphocytes in individuals clinically recovered from COVID-19, J. Mol. Cell Biol., pp. 197-209.
[11] X. Zhou, D. Chen, L. Wang, et al. (2020). Low serum calcium: a new, important indicator of COVID-19 patients from mild/moderate to severe/critical Biosci. Rep, p. 40
[12] Y. Luan, C. Yin, Y. Yao (2021). Update advances on C-reactive protein in COVID-19 and other viral infections Front. Immunol.,12:720363. doi: 10.3389/fimmu.2021.720363.
[13] X. Luo, W. Zhou, X. Yan, et al. (2020). Prognostic value of C-reactive protein in patients with coronavirus 2019 Clin. Infect. Dis., 71, pp. 2174-2179.
[14] Riani, M., Atkinson, A.C. & Corbellini, A. (2023). Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression. *Stat Methods Appl* 32, 75–102. doi:10.1007/s10260-022-00640-7
[15] Islam, M.N., Islam, M.S., Shourav, N.H. *et al.* (2024). Exploring post-COVID-19 health effects and features with advanced machine learning techniques. *Sci Rep* 14, 9884. doi:10.1038/s41598-024-60504-w
[16] Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627-1639.
[17] Brownlee, J. *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End* (Machine Learning Mastery, 2016).
[18] Douglass, M. J. J. Book review: Hands-on machine learning with scikit-learn, keras, and tensorflow, 2nd edition by Aurélien géron. Phys. Eng. Sci. Med. 43(3), 1135–1136 (2020).
[19] Yury V. Kistenev, Denis A. Vrazhnov, et al (2022) Zuhayri,Predictive models for COVID-19 detection using routine blood tests and machine learning,Heliyon,Volume 8, Issue 10, e11185,doi:10.1016/j.heliyon.2022.e11185