Computers in Biology and Medicine 154 (2023) 106619

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Explainable artificial intelligence model for identifying COVID-19 gene biomarkers

Fatma Hilal Yagin [a], İpek Balikci Cicek [a], Abedalrhman Alkhateeb [b,*], Burak Yagin [a], Cemil Colak [a,**], Mohammad Azzeh [c], Sami Akbulut [a,d,e]

[a] *Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, 44280, Malatya, Turkey*
[b] *Software Engineering Department, King Hussein School for Computing Sciences, Amman, Jordan*
[c] *Data Science Department, King Hussein School for Computing Sciences, Amman, Jordan*
[d] *Inonu University, Faculty of Medicine, Department of Surgery, 44280, Malatya, Turkey*
[e] *Inonu University, Faculty of Medicine, Department of Public Health, 44280, Malatya, Turkey*

ARTICLE INFO

Keywords:
COVID-19
Explainable artificial intelligence
LIME
SHAP
XGBoost

ABSTRACT

Aim: COVID-19 has revealed the need for fast and reliable methods to assist clinicians in diagnosing the disease. This article presents a model that applies explainable artificial intelligence (XAI) methods based on machine learning techniques on COVID-19 metagenomic next-generation sequencing (mNGS) samples.
Methods: In the data set used in the study, there are 15,979 gene expressions of 234 patients with COVID-19 negative 141 (60.3%) and COVID-19 positive 93 (39.7%). The least absolute shrinkage and selection operator (LASSO) method was applied to select genes associated with COVID-19. Support Vector Machine - Synthetic Minority Oversampling Technique (SVM-SMOTE) method was used to handle the class imbalance problem. Logistics regression (LR), SVM, random forest (RF), and extreme gradient boosting (XGBoost) methods were constructed to predict COVID-19. An explainable approach based on local interpretable model-agnostic explanations (LIME) and SHAPley Additive exPlanations (SHAP) methods was applied to determine COVID-19-associated biomarker candidate genes and improve the final model's interpretability.
Results: For the diagnosis of COVID-19, the XGBoost (accuracy: 0.930) model outperformed the RF (accuracy: 0.912), SVM (accuracy: 0.877), and LR (accuracy: 0.912) models. As a result of the SHAP, the three most important genes associated with COVID-19 were IFI27, LGR6, and FAM83A. The results of LIME showed that especially the high level of IFI27 gene expression contributed to increasing the probability of positive class.
Conclusions: The proposed model (XGBoost) was able to predict COVID-19 successfully. The results show that machine learning combined with LIME and SHAP can explain the biomarker prediction for COVID-19 and provide clinicians with an intuitive understanding and interpretability of the impact of risk factors in the model.

1. Introduction

The coronavirus (CoV) family of viruses causes symptoms ranging from the common cold to more serious illnesses such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS) (SARS-CoV). The COVID-19 pandemic, caused by the SARS-CoV-2 virus, began on December 31, 2019, in Wuhan, Hubei Province, China, and soon spread worldwide, becoming the world's first coronavirus pandemic. The worry and anxiety generated by the quick transmission of the virus and the steady growth in the number of patients and deaths became an unavoidable concern while new literature was being created about this epidemic, which was encountered for the first time [1–3].

Fever, cough, pneumonia, diarrhea, chest pressure, and shortness of breath are common COVID-19 symptoms, according to the WHO. COVID-19 is difficult to diagnose early since the symptoms are similar to influenza. On the other hand, early diagnosis of positive cases prevents the fast spreading of the disease, endangering the public health system

* Corresponding author.
** Corresponding author.
E-mail addresses: hilal.yagin@inonu.edu.tr (F.H. Yagin), ipek.balikci@inonu.edu.tr (İ.B. Cicek), a.lkhateeb@psut.edu.jo (A. Alkhateeb), burak.yagin@inonu.edu.tr (B. Yagin), cemil.colak@inonu.edu.tr (C. Colak), m.azzeh@psut.edu.jo (M. Azzeh), akbulutsami@gmail.com (S. Akbulut).

https://doi.org/10.1016/j.compbiomed.2023.106619
Received 17 November 2022; Received in revised form 11 January 2023; Accepted 28 January 2023
Available online 1 February 2023

and causing significant repercussions [4]. Due to the rapid development and high mortality rate of COVID-19, investigating the potential risk factors affecting the development of COVID-19 and comorbidities becomes an important research topic. The fact that the disease has reached epidemic level has caused strain on health resources in many countries, making it necessary to evaluate all methods that can guide diagnosis and treatment [5].

The reverse transmission polymerase chain reaction is one of the most common methods for detecting COVID-19 (RT-PCR). The RT-PCR test's sensitivity and accuracy have been questioned in several investigations. They also discovered that the RT-PCR test has a high rate of false negatives and positives [6,7]. In addition to PCR tests, Computed Tomography (CT), chest X-ray, and ultrasound scans can also be used to identify COVID-19 disease. In the literature, many studies based on machine learning can successfully detect COVID-19 using the images of these scans. However, studies that detect COVID-19 from these scans have some limitations. COVID-19 and other types of viral pneumonia share a few common traits. Thus the images from these medical scans may not be able to tell them apart [7,8].

The absence of anomalies in chest X-rays, CT scans, or ultrasound scans, for example, does not guarantee the absence of COVID-19. In addition, there is a scarcity of varied annotated images that can be employed in image-based analytic investigations [7,9]. Large-scale genomics is a powerful technology that has sparked broad interest in biomedical research aimed at discovering biomarkers and unraveling the processes of complicated disorders [10].

In detecting COVID-19, genomic characterization will help accurately describe the virus's origin and evolution. For this reason, although many diagnoses and treatment methods have been reported to detect the virus, revealing the scientific and genetic status of this virus as soon as possible is needed. The genomic structure of the virus, gene regions, protein binding sites, attachment, neutralizing structures, etc., needs to be explored and revealed. For this, it is crucial to isolate the virus first, determine the gene regions at a superficial level, sequence the whole genome at an advanced level, and perform bioinformatics analysis. In conclusion, studies need to define the host response of the virus to identify pathogenicity mechanisms and potential therapeutic targets [11,12].

Emerging pathogen detection using metagenomic next-generation sequencing (mNGS) is agnostic and may be done straight from clinical specimens. mNGS can also identify coinfections affecting illness progression and prognosis and provide valuable information on the microbiome's composition. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and/or other infectious infections can be detected with mNGS [13].

Machine learning (ML) algorithms apply statistical methods to big datasets to uncover correlations between patient features and outcomes, allowing data to be combined to predict results. ML is utilized in various medical disciplines, including diagnosis, class estimate, treatment, disease-related biomarker identification, and medical image/video processing. Also, recently, ML has been used for COVID-19 prediction using data such as clinical, image/video, and genomics [14].

There is still a scarcity of ML research to evaluate COVID-19's prognosis and biomarkers. Furthermore, despite ML's efficacy in prior studies for COVID-19, there is limited evidence for its use in clinical settings and explainable AI models to improve disease prognosis. Clinicians struggle to explain how to make particular patient predictions because of the "black box" nature of ML algorithms. So far, the application of machine learning in medical decision support has been confined by black-box models with little interpretability. Furthermore, one of the most significant barriers to the adoption of ML in the medical profession has been a lack of intuitive understanding of ML models [15].

This study combines advanced machine learning techniques with a framework based on Local Interpretable Model-Agnostic Explanations (LIME) and SHApley Additive ExPlanations (SHAP) to address the abovementioned drawbacks. This paradigm improves the accuracy of COVID-19 diagnostic prediction and intuitively explains the predictions while considering patient-specific genomic risk factors.

As a result, it aids doctors in better understanding the COVID-19 genomic prediction decision-making process and maximizing patient-specific early detection and treatment options. This will help develop interpretable and individualized COVID-19 predictive models, which will be a significant step forward for machine learning in medicine.

The main findings and contributions of the article are listed below.

- Comparison of results from different machine learning techniques to support the diagnosis of COVID-19 using metagenomic next-generation sequencing data;
- The XGBoost performs better in discriminating patients infected with COVID-19 compared to RF, LR, and SVM models;
- A LIME and SHAP-based methodology to explain the pattern that can assist clinicians in diagnosing COVID-19;
- Patient-specific early detection and treatment opportunities, thanks to individual descriptions of the relative importance of each gene;
- ML and SHAP are useful in diagnosing and treating COVID-19, future therapeutic targets, and personalized medicine applications.
- In future studies, offering genomic predictive qualities in the detection of biomarkers for COVID-19.

2. Literature review for AI-Driven COVID-19

Not only from the medical field but from all areas of science, there is a growing amount of literature on the COVID-19 pandemic. The Kuwait Indicator of Progress (KPI) score was proposed by Al Youha et al. to estimate the severity degree of COVID-19 [16]. This methodology was based on measurable laboratory findings, different other person-reported symptoms, and subjective parameter-based grading systems. If a patient's KPI score is less than −7, they are classified as low-risk, and if it is larger than 16, they are classified as high-risk. The possibility of severity progression in the intermediate group (where patient scores ranged from −6 to 15) was considered doubtful by the authors. This intermediate group is, however, included in several prognostic schemes.

Weng et al. proposed an early estimation score named ANDC to determine the chance of death for COVID-19 patients using a dataset from 301 patients. Using least absolute shrinkage and operator of choice (LASSO) regression, COVID-19 patients' age at admission, neutrophil-to-lymphocyte ratio (NLR), D-dimer, and C-reactive protein (CRP) were identified as predictors of death [17]. They developed a high-performing nomogram as well as an ANDC integrated score with a death probability that matched. They also developed ANDC cut-off values to categorize COVID-19 patients into three risk groups: low, intermediate, and high. In the low, intermediate, and high-risk groups, death rates were 5%, 5–50%, and more than 50%, respectively.

Based on data from 444 patients, Xie et al. developed a predictive model including age, lactate dehydrogenase, SpO2, and lymphocyte count as key biomarkers of COVID-19-concerned death. This model showed good separation for external, internal, and validation, with C statistics of 0.98 and 0.89. Despite the model's promising performance for internal calibration, external validation revealed excessive and low estimation for low-risk and high-risk individuals, respectively [18].

Yan et al. incorporated an ML technique for identifying three COVID-19 biomarkers [lactic dehydrogenase (LDH), lymphocytes, and high-sensitivity C-reactive protein (hs-CRP)] [19]. They used this technique to predict patients' mortality (90% accuracy). They discovered that high LDH levels are crucial in identifying the vast majority of patients who need immediate medical attention.

Because most COVID-19 patients have a lung infection, according to clinical studies, many academics have included X-ray images in their early automated diagnosis algorithms [7,20–22]. Using lung X-ray images, they used several Neural Networks to categorize COVID-19 positive and negative patients. Wang and Wong (2020) applied deep convolutional networks on chest X-ray images to detect patients with COVID-19. The open-source dataset was made publicly available, including 13,975 chest X-ray scans. Majeed et al. constructed a twelve-convolutional neural network using X-ray images. Shi et al. comprehensively review AI techniques for COVID-19 image data [9].

Arentz et al. from Washington State look at 21 COVID-19 patients who are critically ill and some of their characteristics [8]. Wynants et al. assess and critically evaluate 27 academic studies and 31 prediction models. The most relevant predictors for COVID-19 patients were C-reactive proteins, tomography screening features, lactic dehydrogenase, lymphocyte count, age, and gender. They pointed out that all studies have a high risk of bias due to a non-representative selection of control patients and an enormous possibility for model overfitting. According to Yan et al., fever was the most common initial symptom, followed by cough, tiredness, and shortness of breath. They examined over 300 variables and found that lactic dehydrogenase, lymphocytes, and high-sensitivity C-reactive protein are all important clinical markers [23].

Using a decision tree technique, Randhawa et al. evaluated over 5,000 viral genomic sequences, including the COVID-19 virus sequence. Imran et al. devised an artificial intelligence-based method for analyzing the genome sequences of COVID-19 and other viruses (SARS and Ebola, etc.). This method facilitates the extraction of essential data from virus genome sequences. Comparative data analysis is performed by gathering basic information about COVID-19 and other genome sequences, such as nucleotide composition and frequency, amino acid number, alignment between genome sequences, tri-nucleotide compositions, and similar DNA information. They analyzed the genome sequences of these viruses using various visualization approaches and used the support vector machine as a classifier to classify the genome sequences. The algorithm produces good classification results with an accuracy of 97% for COVID-19, 96% for SARS, and 95% for MERS and Ebola genome sequences [24].

This study examines the prediction performance of multiple machine learning models. When understanding model predictions, we go further than most of these papers. It is necessary to employ Shapley values and LIME, which are unavailable in previous papers. Our purpose is entirely practical: we want to know how a machine learning model developed with genetic characteristics predicts a COVID-19 patient's outcome.

3. Material and methods

3.1. Dataset

In this study, the mNGS dataset belonging to open-access COVID-19 positive and negative patients was utilized for the analyses. A cohort study for COVID-19 was undertaken at the University of California, San Francisco (UCSF) and Zuckerberg San Francisco General Hospital, and the dataset was obtained from that study. In the dataset, there are 15,979 genes of 234 patients with COVID-19 negative 141 (60.3%) and COVID-19 positive 93 (39.7%) [25].

The following criteria were used to choose participants: (1) position as a COVID-19 patient under examination, (2) be at least 18 years old, (3) a clinician-ordered SARS-CoV-2 test was performed using RT-PCR from a nasopharyngeal (NP) swab collected with or without an oropharyngeal (OP) swab between 03/10/2020 and 04/07/2020, and (4) For metagenomic sequencing, there was adequate extracted RNA. If more than one sample was collected from a patient later diagnosed with COVID-19, only the first available positive sample was analyzed [25].

3.2. Methods

The binary classification of COVID-19 using the mNGS dataset described in section 1 is the topic of this article. The ML procedure for creating explicable classifiers consists of two main steps: (i) the creation/evaluation of different artificial learning models and (ii) the application of LIME algorithms for local output interpretation with SHAP algorithms for global output interpretation. Fig. 1 provides an overview of the methodology (see Fig. 2 that shows the performance of the ML models).

3.3. Feature selection and data preprocessing

Hundreds of genes were initially included in the COVID-19 mNGS gene expression dataset (called "features" in ML). Feature selection methods can maximally identify the relevant subset of essential features and reduce data redundancy [26,27].

In this study, the LASSO method was used for feature selection because it reduces overfitting and is helpful in datasets with few observations. LASSO requires model parameters to have a particular total absolute value (upper bound). To do this, it penalizes the regression variable coefficients by reducing some to zero. Variables with non-zero coefficients are selected during feature selection for the model. In this way, it improves the interpretability of the model and prevents overlearning by removing unnecessary variables. To have the correct type of LASSO regression, the α hyperparameter must be adjusted. The study determined the α value using the iterative 10-fold cross-validation with Grid Search. The optimum value for α was found to be 1.20. The SVM-SMOTE oversampling method was used to balance the COVID-19 positive and COVID-19 negative observations in the dataset. SVM-SMOTE focuses on creating new minority class instances around boundary lines with SVM to assist in building boundaries between classes [28]. The SMOTE algorithm is given below.

Algorithm 1. SMOTE (T, N, k)

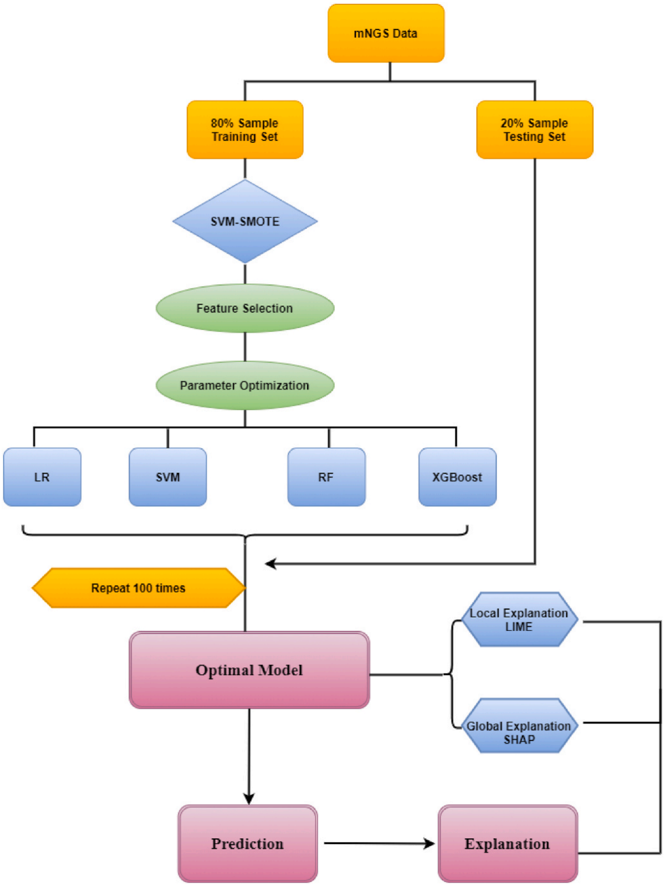


Fig. 1. Diagram of the proposed method combining explainability and classifier.

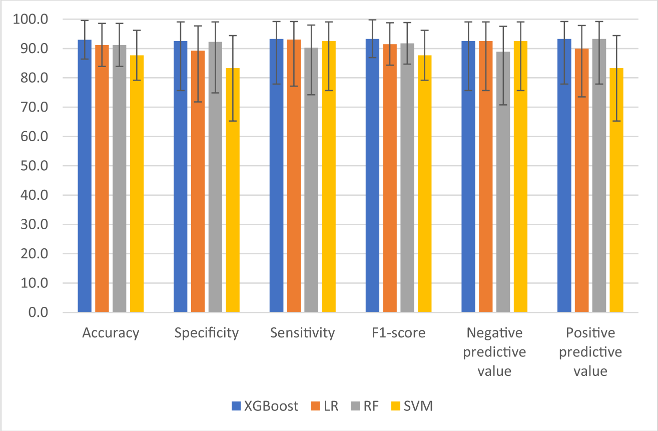


Fig. 2. The plot of results of ML models for COVID-19 (95% confidence interval (CI)).

3.4. *Development and evaluation of predictive models*

Four ML models have been developed to predict COVID-19 based on gene expressions. In the modeling phase, the extreme gradient boosting (XGBoost) method was used in addition to LR, RF, and SVM, which have low interpretability. XGBoost is a high-performance version of the Gradient Boosting technique that has been improved for various configurations. The algorithm's most essential characteristics are its capacity to attain high predictive power, avoid overfitting, and deal with missing variables fast [29,30]. The stratified random sampling method was used to divide 234 patients into a training set and a test set at a ratio of 4:1. The training set was preprocessed using the SVM-SMOTE technique to balance the positive and negative groups. The Grid search method with repeated 10-fold cross-validation was then used to optimize the hyperparameter of the hyperparameters of the ML models (details in Supplementary Table S1). Finally, the performance of each model was evaluated and compared on the test set. To obtain a more robust performance estimate, avoid reporting biased results, and limit overfitting, we repeated the persistence method 100 times with different random seeds and calculated the average performance over these 100 times (Fig. 1). While evaluating the performances of the models, F1-score, accuracy, specificity, sensitivity, negative predictive value, and positive predictive value were used (details in Supplementary Table S3). The best-performing model among the four models employed in classification was chosen for local and global explanations after a thorough evaluation of several performance indicators.

3.5. *Random forest (RF)*

In 2001, Dr. Breiman's RF technique was tested to be a highly effective general-purpose classification and regression tool [31]. The RF algorithm consists of a combination of decision trees, and the trees with the highest accuracy and independence are preferred among the decision trees used. The RF classifier is a meta-estimator that fits a series of decision tree classifications to multiple sub-samples of the dataset, increases prediction accuracy, and controls for overfitting by using the mean [32].

3.6. *Logistic regression (LR)*

Regression approaches analyze explanatory and outcome factors. Typically, the outcome variable has two or more values. LR investigates the cause-effect relationship between explanatory factors and binary, triple, and multiple outcome categories. Risk factors and explanatory variables' effects on dependent variables are calculated as probability LR analysis classifies and assigns. Normality and continuity have no preconditions. Probabilities are assigned to risk factors and explanatory variables' impacts on the dependent variable [33].

3.7. *Support vector machine (SVM)*

SVM is a machine learning model used in regression and classification problems. SVM's primary purpose is to find a hyperplane that accurately identifies the classes corresponding to the target variable. To transform data, SVM employs a technique known as the kernel trick. Kernel trick methods use data transformation models to select the best boundary among possible results. Kernel trick approaches execute sophisticated data transformations first, then determine how to segregate the data depending on defined tags or results [34]. Also, SVM is a powerful technique for detecting subtle patterns in large datasets [35].

3.8. *XGBoost*

Gradient Boosting is an ML technique for regression and classification problems that produces an ensemble form of weak predictive models in a prediction model, typically decision trees. Gradient Boosting is a technique that uses boosting techniques. It seeks to generate a large number of weak learners in order and incorporate them into a complex model since it is based on the boosting method. Compared to other algorithms, XGBoost has a significant speed and performance advantage [36–38].

3.9. *Interpretable machine learning and feature significance*

Because it might be challenging to understand why an algorithm generates correct predictions for a given patient cohort, machine learning models are frequently referred to as "black boxes." Therefore, LIME and SHAP methodologies were used in this study. LIME was used to provide local explanations for four patients with false negative (FN), true negative (TN), false positive (FP), and true positive (TP) in the XGBoost model. This approach generates additional samples around the sample to be explained and uses the old model to estimate the local noise. SHAP was used to provide a broad explanation for our XGBoost prediction model. The relevance of features in the final model was prioritized to find important COVID-19 biomarkers in the patient group. In addition, a Radar plot was drawn for the top five most important biomarkers in a true positive and true negative patient.

3.10. *SHApley Additive ExPlanations (SHAP)*

SHAP is a new way to explain diverse black box ML models developed by Lundberg and Lee as a unified framework for interpreting ML predictions. Compared to other methods, SHAP can achieve both local and global interpretability simultaneously and has a robust theoretical base [39,40].

SHAP calculates the impact of each feature on the learned model's predictions. SHAP approximates f with a simple model g that can simply explain the contribution of each feature value given an input $x = x_1, x_2, \ldots, x_p$ and a trained model f. The following is a formula for the g model.

$$g(z) = \varnothing_0 + \sum_{i=1}^{p} \varnothing_i z_i$$

Here, p is the number of features and $z = [z_1, z_2, \ldots, z_p]^T$ is a simplification of the input x, with z relating to the features utilized in the data prediction being 1 and z corresponding to the features not used is 0.

$$\varnothing_i(f, x) = \sum_{z \subseteq x} \frac{|z|!(p - |z| - 1)!}{p!} [f(z) - f(z \backslash i)]$$

Tree SHAP generates a N × M matrix with SHAP values using tree-based models and an input dataset X of size N × M. (N represents the

Algorithm 1 SMOTE (T, N, k)

- **Input:** Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k
- **Output:** $(N/100) * T$ synthetic minority class samples. If N is less than 100%, randomize the minority class samples as only a randompercent of them will be SMOTEd.

```
if $N \leq 100$ then
    Randomize the T minority class samples
    $T = (N/100) * T$
    $N = 100$
end if
$N = (int)(N/100)$ (The amount of SMOTE is assumed to be in integralmultiples of 100.)
$k$ = Number of nearest neighbors
$numattrs$ = Number of attributes
Sample [ ][ ]: array for original minority class samples new index: keepsa count of number of synthetic samples generated, initialized to 0
Synthetic [ ][ ]: array for synthetic samples (Compute k nearest neighborsfor each minority class sample only.)
for $i \leftarrow 1 : T$ do
    Compute $k$ nearest neighbors for $i$, and save the indices in the nnarray
    Populate ($N$, $i$, nnarray)
end for
Populate ($N$, $i$, nnarray) (Function to generate the synthetic samples.)
for $N \nvDash 0$ do
    Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of $i$.
    for $attr \leftarrow 1 : numattrs$ do
        Compute: dif = Sample [nnarray[nn]][attr]  Sample[i][attr]
        Compute: gap = random number between 0 and 1
        Synthetic [newindex][attr] = Sample[i][attr] + gap dif
    end for
    $newindex++$
    $N = N - 1$
end for
return (End of Populate)
End of Pseudo-Code.
```

number of samples here). The SHAP interaction values ensure that individual prediction explanations of interaction effects are consistent. Global and local interpretability are two unique advantages of SHAP values. Unlike other essential features in machine learning models, SHAP can evaluate whether each input characteristic has a positive or negative influence [41].

3.11. Local interpretable model-agnostic annotations (LIME)

LIME is a popular method for making black-box machine learning algorithms more understandable. To create an explanation for a single prediction by any ML model, LIME learns a more simply interpretable model around the prediction, generates simulated data around the sample with random perturbation, and extracts feature importance by applying some type of feature selection [42]. LIME can determine how much each variable in the data contributes to each (patient-specific) prediction in the model. Using the LIME method, it can be determined which variables affect each prediction in the model to what degree and in what direction or which variable has a greater effect on each prediction's outcome than other variables [43]. Pseudo codes of the LIME algorithm are given below.

```
LIME Explainer Object (sample Instance, black box model) {
For given sample instance, do:
1. Create a new dataset around an observation by sampling from distribution learned on training data.
2. Calculate the distance between permutations and original observations.
3. Use a black box model to predict probability on new points.
4. Pick m features best describing the complex model outcome from the permuted data.
5. Fit a linear model on data in m dimensions weighted by similarity.
6. Weights of the linear model are used as an explanation of decision.

return Explanation (sample instance)
}
```

3.12. *Statistical analysis*

Quantitative data are summarized by median (minimum-maximum). Normal distribution was evaluated with the Kolmogorov-Smirnov test. In terms of input variables, the existence of a statistically significant difference and the relationship between the categories of the output variable, "positive " and "negative" groups, were examined using the Mann-Whitney *U* test. Expression levels for the top five most important genes identified by xAI method were presented in box plots, along with the median and interquartile range. $p < 0.05$ values were considered statistically significant. American Psychological Association (APA) 6.0 style was used to report statistical differences [44]. All statistical analyses were performed using IBM SPSS Statistics for Windows version 28.0 (New York; USA) software and graphs using GraphPad Prism 9.4.1 software (Details in Supplementary Tables S4 and S5).

4. Results

The COVID-19 gene expression dataset based on mNGS initially had 15,979 genes. LASSO feature selection method was applied to eliminate the high dimensionality problem in the dataset. Thirty-one genes associated with COVID-19 were selected after LASSO (Details are in Supplementary Table S1). In addition, Table 1 shows the results of COVID-19 classification using accuracy, positive predictive value, negative predictive value, specificity, sensitivity, and F1-score.

The XGBoost model achieved an Accuracy of 0.93 (95% CI: 0.864–0.996) and an F1 score of 0.933 (95% CI: 0.869–0.998) for COVID-19. These values were higher compared to the corresponding values in the other three models. These values were higher compared to the corresponding values in the other three models. Among the performance measures with XGBoost, sensitivity was relatively high at 0.933 (95% CI: 0.779–0.992) and specificity at 0.926 (95% CI: 0.757–0.991).

Fig. 3 A shows the importance of biomarker candidate genes for model decisions using global SHAP values reflecting their positive or negative contribution to the prediction of the optimal model. A positive SHAP number indicates that the contribution to the target variable is positive, whereas a negative SHAP value indicates that the contribution is negative. These importances are shown in descending order, showing that IFI27, LGR6, and FAM83A are the three most important genes contributing to the target variable. In addition to this, the dots on the graph are colored according to the normalized values of the patient's gene expression levels, such as IFI27. The gene expression level value decreases as it gets closer to blue and increases as it gets closer to pink. Therefore, a higher level of IFI27, LGR6, FAM83A, TBCE, and BACH2, as well as a lower level of GLTPD2, DCUN1D3, SCG83A1, VSIG1, METRNL, RASL11A, STK, ALOX15B, DUSP6, ITGB1BP2, ERVMER341, PCDHB9, RTN2, and TPT1 can be said to increase the risk.

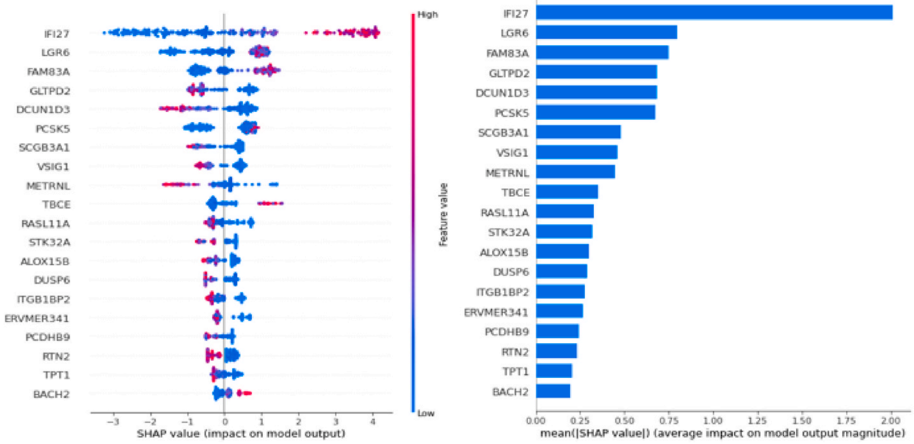
When the normalized SHAP values in Table 2 are examined, the five most important risk factors for COVID-19 are IFI27, LGR6, FAM83A, GLTPD2, and DCUN1D3. The percentages of these risk factors increasing the risk of COVID-19 are 19.84%, 10.41%, 8.31%, 6.88%, and 6.41%, respectively. Fig. 4 shows the variation of the expression levels of these genes in the 6.41%, respectively. Fig. 4 shows the variation in expression levels of these genes in the groups, and Fig. 5 shows the Radar plot of these genes.

Fig. 6 A, B, C, and D show local explanation examples for four patients estimated as false positive, true positive, false negative, and true negative using Local Interpretable Model-agnostic Explanations (LIME). In Fig. 6 A, B, C, and D, right-pointing bars show features that are positively correlated with output, while left-hand bars show negatively correlated with output.

Therefore, for the patient in Fig. 6 A, low IFI27 values (IFI27 value between 185 and 534) and high DCUN1D3 values (DCUN1D3 value greater than 372) have a negative correlation with positive COVID-19 according to the LIME descriptions. For the patient in Fig. 6C, high IFI27 values (IFI27 value between 534 and 1395) and high LGR6 values (LGR6 value greater than 65) have a positive correlation with positive COVID-19 according to the LIME descriptions. According to the LIME statements, especially it can be said that the high level of IFI27 gene expression contributes to increasing the probability of the positive class.

Table 1
Results of ML models for COVID-19 (95% confidence interval (CI)).

Score/Model	XGBoost	LR	RF	SVM
Accuracy	0.93 (0.864–0.996)	0.912 (0.839–0.986)	0.912 (0.839–0.986)	0.877 (0.792–0.962)
Specificity	0.926 (0.757–0.991)	0.893 (0.718–0.977)	0.923 (0.749–0.991)	0.833 (0.653–0.944)
Sensitivity	0.933 (0.779–0.992)	0.931 (0.772–0.992)	0.903 (0.742–0.98)	0.926 (0.757–0.991)
F1-score	0.933 (0.869–0.998)	0.915 (0.843–0.988)	0.918 (0.847–0.989)	0.877 (0.792–0.962)
Negative predictive value	0.926 (0.757–0.991)	0.926 (0.757–0.991)	0.889 (0.708–0.976)	0.926 (0.757–0.991)
Positive predictive value	0.933 (0.779–0.992)	0.9 (0.735–0.979)	0.933 (0.779–0.992)	0.833 (0.653–0.944)



(A) Global SHAP (B) Global SHAP

Fig. 3. Interpretation of the XGBoost model. (A): Ranking the importance of the top 20 risk factors with stability and interpretation using the optimal model. (B): The order of importance of the first 20 variables according to the mean (|SHAP value|); the higher the SHAP value of a trait is given, the higher the probability that the patient will be COVID-19 positive.

Table 2
Importance of risk factors (genes) for COVID-19.

Genes	Feature importance (normalize-shap values)
IFI27	0.19849
LGR6	0.10410
FAM83A	0.0831121
GLTPD2	0.068832
DCUN1D3	0.064159
PCSK5	0.060126
SCG83A1	0.05948
VSIG1	0.033271
METRNL	0.029042

5. Discussion

In this study, we aimed to study the molecular pathogenesis of SARS-CoV-2 and developed an ML-based classification model for interpretable prediction of COVID-19 based on host gene expression in patients with acute respiratory disease. LASSO feature selection, GridSearchCV for hyperparameter optimization, and SVM-SMOTE for resampling methods were combined with advanced ML algorithms. Of the four ML classifiers, XGBoost performed the best with fast computation and strong generalization ability; therefore, the XGBoost model was used for COVID-19 prediction. With an accuracy of 0.93 [0.864–0.996], F1-score of 0.933 [0.869–0.998], a sensitivity of 0.933 [0.779–0.992], and specificity of 0.926 [0.757–0.991], the ML model significantly outperformed the other available prediction models. A higher sensitivity value means a lower false negative (FN) value. False-positive and false-negative errors are common in comparative biological investigations. Therefore determining the probability of a real impact being significant is crucial [45]. A lower FN value is an encouraging result for COVID-19 cases. This result is very important because minimizing missed COVID-19 cases (false negatives) is one of the main goals of this research.

Furthermore, using SHAPley values and SHAP plots, we proved that our approach could demonstrate the key features and interpretations of ML results. The results of the SHAP method showed that the 20 genes associated with COVID-19 and most important for the model decision were IFI27, LGR6, FAM83A, GLTPD2, DCUN1D3, PCSK5, SCGB3A1, VSIG1, METRNL, TBCE, RASL11A, STK32A, ALOX15B, DUSP6, ITGB1BP2, ERVMER341, PCDHB9, RTN2, TBT1, and BACH2. The SHAPley value assesses the importance of the output, including all feature combinations, and provides consistent and locally accurate attribute values for each feature in the prediction model. This annotation method is applied to XGBoost's black box tree integration model to help users better understand the model's decision-making process. The detailed information disclosed in the results and descriptions of biomarker candidate genes provides further insights to help doctors trust the results of the algorithm or model and make more informed decisions. Finally, visualized descriptions of domain-specific cumulative trait importance and trait importance can contribute to physicians' intuitive understanding of the key features of the XGBoost model and its prediction results. After examining the holistic descriptions of the model with SHAP plots, we used the LIME approach to examine genomic biomarkers on a patient-by-patient basis and to interpret the model, providing explanations for examples predicted as a false positive, false negative, true positive, and true negative. As a result of LIME, it was discovered that the model uses several factors to "diagnose" each patient. This revealed that COVID-19 affects several genes and that the relevance of these genes varies depending on the individual. In summary, considering key genomic biomarkers, our approach can intuitively explain to clinicians which specific characteristics of COVID-19 patients predispose them to a higher (or lower) disease risk. When it comes down to it, such a predictive approach has potential in clinical practice by personalizing disease prevention and strengthening potential therapeutic strategies.

Another study determined that the FAM83A (ENSG000001147689) and LGR6 (ENSG00000133067) genes are potential biomarkers to identify the affected upper respiratory tract tissues of COVID-19 patients. The authors found that FAM83A was essential in distinguishing upper respiratory tract specimens infected with SARS-CoV-2 from healthy controls or other infections. In our study, we found that FAM83A may be a potential biomarker for COVID-19 and that a higher FAM83A level may be associated with the risk of COVID-19. The literature has reported that LGR6 is associated with several viruses, including SARS-CoV-2 [46]. The current research results showed that the LGR6 gene is an important biomarker and that increased expression levels of this gene are associated with COVID-19. A different study showed that IFI27 is highly expressed in the lymphocytes of COVID-19 patients but only partially expressed in the lymphocytes of controls. IFI27 was found to be a biomarker candidate for SARS-CoV-2 infection [47]. In current research, we found that an increase in the level of IFI27 for COVID-19 increases the risk of disease. In the literature, they reported that SARS-CoV-2 infection suppressed the expression of the DCUN1D3 gene and the disease was associated with this gene [48]. Our research found that low levels of the DCUN1D3 gene are associated with COVID-19.

A recent paper has made a dual classification of COVID-19 and other

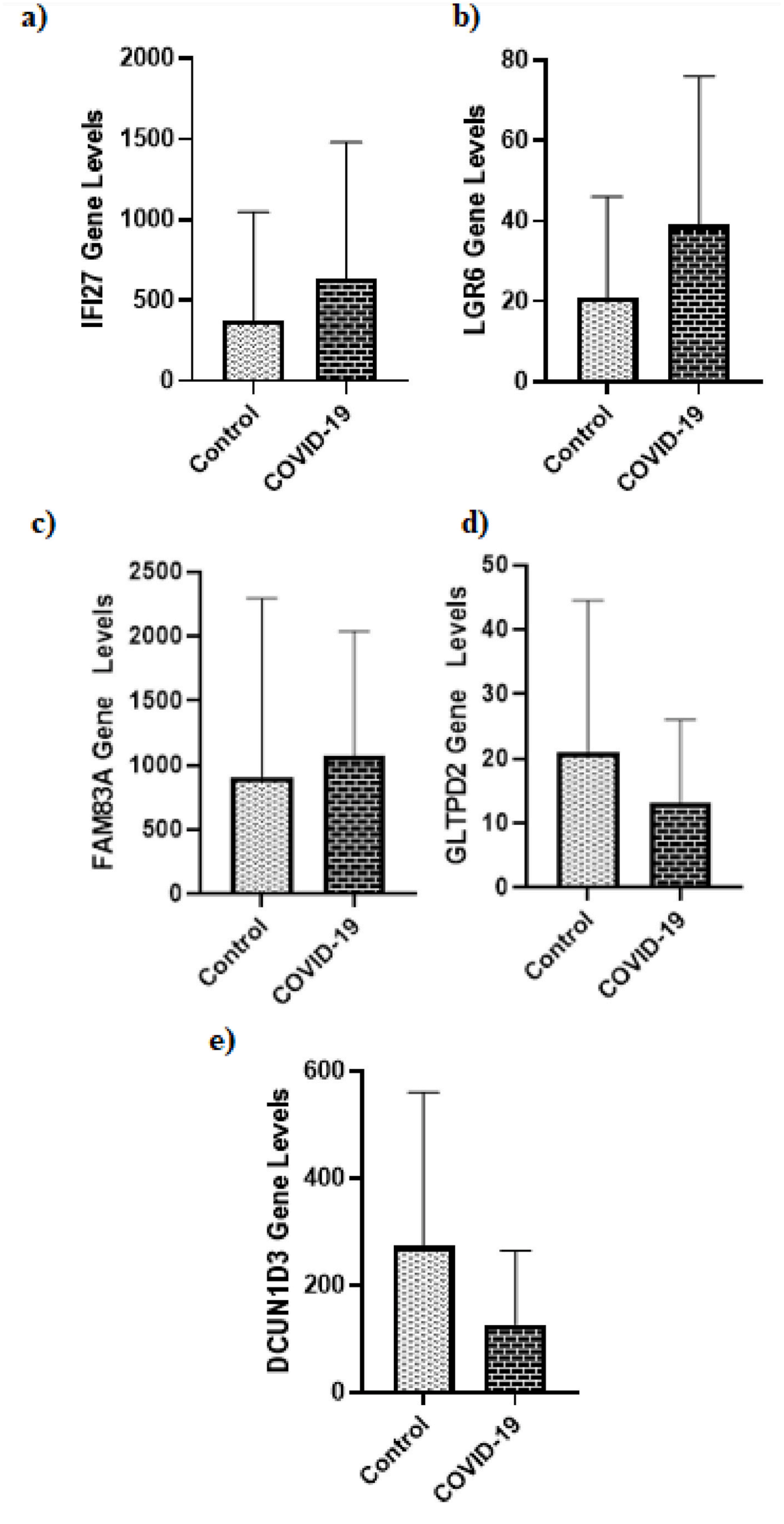


Fig. 4. Variation of the five most important gene expression levels between groups.

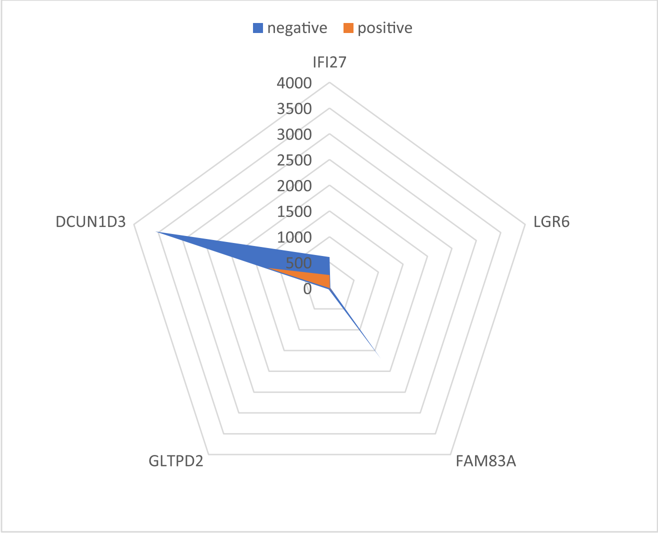


Fig. 5. Radar plot of the five most important genes. - It is a COVID-19 patient appearing in orange on the Radar plot and a patient in the control group appearing in blue.

coronavirus types by extracting features from genome sequences. For this purpose, four different models were created with support vector machines, K-nearest neighbor, pure Bayesian, and random forest methods, and 93% accuracy was obtained with the decision tree method [49]. The mutation rate was studied in genomic sequences collected from GenBank data on COVID-19 patients. Genomes were the first place where the rate of missense nucleotide mutation and the rate of codon mutation was discovered [50]. Genomes were the first place where the rate of missense nucleotide mutation and the rate of codon mutation was discovered. A recurrent neural network-based long short-term memory (LSTM) model was then used to estimate the virus's future mutation rate. The study's authors focused solely on base substitution mutation rates, disregarding insertion and deletion rates. In addition, techniques for tracking SARS-CoV-2 genetic variations were developed [51]. A WCGFVL network, a wavelet-coupled random vector functional link (RVFL) network, was also presented for modeling and forecasting COVID-19 spread in the top 5 worst-hit countries (India, Brazil, Russia, Peru, and the United States) [52].

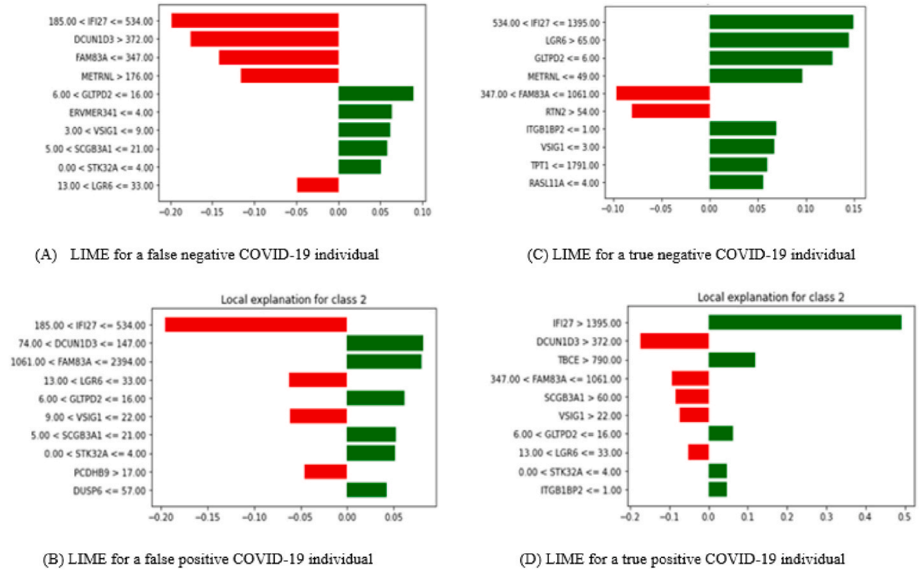
In summary, the proposed model (XGBoost) successfully predicted and classified COVID-19. The results show that ML combined with LIME and SHAP can explain the biomarker prediction for COVID-19 and provide clinicians with an intuitive understanding and interpretability of the impact of risk factors in the model. The results' precise information and risk factor explanations provide clinicians with more information, allowing them to make better decisions rather than relying just on the algorithm's conclusions. Individual explanations can also assist clinicians in understanding why the model generates certain high-risk judgment suggestions. Given the major risk variables, the model can intuitively explain to clinicians which patient characteristics predispose them to a higher (or lower) risk of COVID-19. This topic-specific estimation can help guide and reinforce treatment techniques in clinical practice.

5.1. Limitation and future works

First, our study lacked external validation by an independent cohort, which could provide further evidence to confirm the superiority of the proposed prediction model. It is essential to expand the current study further to include multicenter trials in future studies or to use the related data from different centers for external validation. In addition, in this study, we analyzed patients' genomic data and predicted COVID-19 based on these data. More detailed research is needed to integrate relevant clinical risk factors, environmental factors, lifestyles, and other factors to improve future predictions and examine the impact of confounding factors. Better predictive results can be obtained if patients' clinical research information and multi-omics information are combined. Finally, the extracted biomarkers can be validated by pharmacoproteomic techniques including sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) [53].

5.2. Conclusion

This study's proposed ML approach-based XGBoost algorithm could accurately classify and assess COVID-19 patients via the selected genomic biomarkers. A combination of ML and xAI might provide a clear interpretation of the individualized and overall risk estimation for COVID-19, allowing physicians to intuitively understand the impact of key genomic features in the suggested model (Fu et al., 2018).



(A) LIME for a false negative COVID-19 individual

(C) LIME for a true negative COVID-19 individual

(B) LIME for a false positive COVID-19 individual

(D) LIME for a true positive COVID-19 individual

Fig. 6. Local interpretable model-agnostic explanations.

Author contributions

Conceptualization, Fatma Hilal Yagin, İpek Balikci Cicek, Burak Yagin, Cemil Colak and Sami Akbulut; Data curation, Fatma Hilal Yagin and Mohammad Azzeh; Formal analysis, Fatma Hilal Yagin, İpek Balikci Cicek, Abedalrhman Alkhateeb and Sami Akbulut; Funding acquisition, Abedalrhman Alkhateeb; Investigation, Abedalrhman Alkhateeb, Burak Yagin, Cemil Colak and Mohammad Azzeh; Methodology, Fatma Hilal Yagin, İpek Balikci Cicek, Burak Yagin, Cemil Colak and Sami Akbulut; Project administration, Cemil Colak; Resources, İpek Balikci Cicek.

Funding

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Institute's Clinical Research Ethics Committee (protocol code = 2022/4040). This work has been partially supported by King Abdullah I School of Graduate Studies and Scientific Research at the Princess Sumaya University for Technology with grant number 2021/2022 –25 (16), recieved by Abedalrhman Alkhateeb.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.106619.

References

[1] M. Smith, F. Alvarez, Identifying mortality factors from Machine Learning using Shapley values–a case of COVID19, Expert Syst. Appl. 176 (2021), 114832.

[2] J. Wu, J. Shen, M. Xu, M. Shao, A novel combined dynamic ensemble selection model for imbalanced data to detect COVID-19 from complete blood count, Comput. Methods Progr. Biomed. (2021), 106444.

[3] A. Humayun, S. Shahabuddin, S. Afzal, A.A. Malik, S. Atique, U. Iqbal, Healthcare strategies and initiatives about COVID19 in Pakistan: telemedicine a way to look forward, Comput. Methods Progr. Biomed.Update 1 (2021), 100008.

[4] R. Padmanabhan, H.S. Abed, N. Meskin, T. Khattab, M. Shraim, M.A. Al-Hitmi, A review of mathematical model-based scenario analysis and interventions for COVID-19, Comput. Methods Progr. Biomed. (2021), 106301.

[5] A. Ravizza, F. Sternini, F. Molinari, E. Santoro, F. Cabitza, A proposal for COVID-19 applications enabling extensive epidemiological studies, Procedia Comput. Sci. 181 (2021) 589–596.

[6] A.J. Rufaidah Dabbagh, M.-H. Temsah, J.H.B. Masud, M. Titi, Y. Amer, M. Alkubeyyer, T. Alhazmi, F. Baothman, L. Hneiny, Machine learning models for predicting diagnosis or prognosis of COVID-19: a systematic review, Comput. Methods Progr. Biomed. 205 (2021), 105993.

[7] F.H. Yağin, E. Güldoğan, H. Ucuzal, C. Çolak, A computer-assisted diagnosis tool for classifying COVID-19 based on chest X-ray images, Konuralp Med. J., 13 438-445..

[8] M. Arentz, E. Yim, L. Klaff, S. Lokhandwala, F.X. Riedo, M. Chong, M. Lee, Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State, JAMA 323 (2020) 1612–1614.

[9] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Rev. Biomed. Eng. 14 (2020) 4–15.

[10] Q. Yang, B. Li, P. Wang, J. Xie, Y. Feng, Z. Liu, F. Zhu, LargeMetabo: an out-of-the-box tool for processing and analyzing large-scale metabolomic data, Briefings Bioinf. 23 (2022).

[11] M.N. Hoque, M.S. Rahman, R. Ahmed, M.S. Hossain, M.S. Islam, T. Islam, M. A. Hossain, A.Z. Siddiki, Diversity and genomic determinants of the microbiomes associated with COVID-19 and non-COVID respiratory diseases, Gene Rep. 23 (2021), 101200.

[12] Y. Zhang, Y. Pan, X. Zhao, W. Shi, Z. Chen, S. Zhang, P. Liu, J. Xiao, W. Tan, D. Wang, Genomic characterization of SARS-CoV-2 identified in a reemerging COVID-19 outbreak in Beijing's Xinfadi market in 2020, Biosaf.Health 2 (2020) 202–205.

[13] H.H. Mostafa, J.A. Fissel, B. Fanelli, Y. Bergman, V. Gniazdowski, M. Dadlani, K. C. Carroll, R.R. Colwell, P.J. Simner, Metagenomic next-generation sequencing of nasopharyngeal specimens collected from confirmed and suspect COVID-19 patients, mBio 11 (2020) e01969, 01920.

[14] Y. Bouchareb, P.M. Khaniabadi, F. Al Kindi, H. Al Dhuhli, I. Shiri, H. Zaidi, A. Rahmim, Artificial intelligence-driven assessment of radiological images for COVID-19, Comput. Biol. Med. (2021), 104665.

[15] F.K. Došilović, M. Brčić, N. Hlupić, Explainable Artificial Intelligence: A Survey, 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2018, pp. 210–215.

[16] S. Al Youha, S.A. Doi, M.H. Jamal, S. Almazeedi, M. Al Haddad, M. AlSeaidan, A. Y. Al-Muhaini, F. Al-Ghimlas, S.K. Al-Sabah, Validation of the Kuwait progression indicator score for predicting progression of severity in COVID19, medRxiv (2020).

[17] Z. Weng, Q. Chen, S. Li, H. Li, Q. Zhang, S. Lu, L. Wu, L. Xiong, B. Mi, D. Liu, ANDC: an early warning score to predict mortality risk for patients with coronavirus disease 2019, J. Transl. Med. 18 (2020) 1–10.

[18] J. Xie, D. Hungerford, H. Chen, S.T. Abrams, S. Li, G. Wang, Y. Wang, H. Kang, L. Bonnett, R. Zheng, Development and External Validation of a Prognostic Multivariable Model on Admission for Hospitalized Patients with COVID-19, 2020.

[19] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, An interpretable mortality prediction model for COVID-19 patients, Nat. Mach. Intell. 2 (2020) 283–288.

[20] I.D. Apostolopoulos, T.A. Mpesiana, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci.Med. 43 (2020) 635–640.

[21] A. Narin, C. Kaya, Z. Pamuk, Automatic Detection of Coronavirus Disease (Covid-19) Using X-Ray Images and Deep Convolutional Neural Networks, Pattern Analysis and Applications, 2021, pp. 1–14.

[22] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, Viral Pneumonia Screening on Chest X-Ray Images Using Confidence-Aware Anomaly Detection, 2020 arXiv preprint arXiv:2003.12338.

[23] L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, S. Li, M. Zhang, Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan, medRxiv (2020).

[24] I. Ahmed, G. Jeon, Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses, Interdiscipl. Sci. Comput. Life Sci. (2021) 1–16.

[25] E. Mick, J. Kamm, A.O. Pisco, K. Ratnasiri, J.M. Babik, C.S. Calfee, G. Castañeda, J. L. DeRisi, A.M. Detweiler, S. Hao, Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2, medRxiv (2020).

[26] J. Tang, M. Mou, Y. Wang, Y. Luo, F. Zhu, MetaFS: performance assessment of biomarker discovery in metaproteomics, Briefings Bioinf. 22 (2021) bbaa105.

[27] J. Fu, Y. Zhang, J. Liu, X. Lian, J. Tang, F. Zhu, Pharmacometabonomics: data processing and statistical analysis, Briefings Bioinf. 22 (2021) bbab138.

[28] D. Liang, B. Yi, W. Cao, Q. Zheng, Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and SMOTE, Expert Syst. Appl. 188 (2022), 116051.

[29] K. Dalakleidi, K. Zarkogianni, A. Thanopoulou, K. Nikita, Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications, Expet Syst. 34 (2017), e12214.

[30] A. Ogunleye, Q.-G. Wang, XGBoost model for chronic kidney disease diagnosis, IEEE ACM Trans. Comput. Biol. Bioinf 17 (2019) 2131–2140.

[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[32] R. Yilmaz, F.H. Yağin, Early detection of coronary heart disease based on machine learning methods, Med. Record 4 (2022) 1–6.

[33] E. Ürük, İstatistiksel Uygulamalarda Lojistik Regresyon Analizi, Marmara Universitesi (Turkey), 2007.

[34] W. Xu, J. Zhang, Q. Zhang, X. Wei, Risk prediction of type II diabetes based on random forest model, in: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), IEEE, 2017, pp. 382–386.

[35] Q. Yang, Y. Li, B. Li, Y. Gong, A novel multi-class classification model for schizophrenia, bipolar disorder and healthy controls using comprehensive transcriptomic data, Comput. Biol. Med. 148 (2022), 105956.

[36] J. Dikker, Master Thesis Boosted Tree Learning for Balanced Item Recommendation in Online Retail, 2017.

[37] Z. Salam Patrous, Evaluating XGBoost for User Classification by Using Behavioral Features Extracted from Smartphone Sensors, 2018.

[38] S. Akbulut, F.H. Yagin, C. Colak, Prediction of breast cancer distant metastasis by artificial intelligence methods from an epidemiological perspective, Istanb. Med. J. 23 (2022).

[39] K. Wang, J. Tian, C. Zheng, H. Yang, J. Ren, Y. Liu, Q. Han, Y. Zhang, Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP, Comput. Biol. Med. 137 (2021), 104813.

[40] L. Antwarg, R.M. Miller, B. Shapira, L. Rokach, Explaining anomalies detected by autoencoders using Shapley Additive Explanations, Expert Syst. Appl. 186 (2021), 115736.

[41] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.

[42] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ECGs, Comput. Biol. Med. 133 (2021), 104393.

[43] M.R. Zafar, N.M. Khan, DLIME: a Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems, 2019 arXiv preprint arXiv:1906.10263.

[44] B. Yağin, F.H. Yağin, H. Gözükara, C. Colak, A web-based software for reporting guidelines of scientific researches, J.Cognit. Syst. 6 (2021) 39–43.

[45] F. Li, Y. Zhou, X. Zhang, J. Tang, Q. Yang, Y. Zhang, Y. Luo, J. Hu, W. Xue, Y. Qiu, SSizer: determining the sample sufficiency for comparative biological study, J. Mol. Biol. 432 (2020) 3411–3421.

[46] Y.-H. Zhang, H. Li, T. Zeng, L. Chen, Z. Li, T. Huang, Y.-D. Cai, Identifying transcriptomic signatures and rules for SARS-CoV-2 infection, Front. Cell Dev. Biol. 8 (2021), 627302.

[47] L. Huang, Y. Shi, B. Gong, L. Jiang, Z. Zhang, X. Liu, J. Yang, Y. He, Z. Jiang, L. Zhong, Dynamic blood single-cell immune responses in patients with COVID-19, Signal Transduct. Targeted Ther. 6 (2021) 1–12.

[48] M.N. Hoque, M. Sarkar, M. Hasan, M. Khan, M. Hossain, M. Hasan, M. Rahman, M. Habib, S. Akter, T.A. Banu, Differential gene expression profiling reveals potential biomarkers and pharmacological compounds against SARS-CoV-2: insights from machine learning and bioinformatics approaches, Front. Immunol. (2022) 3875.

[49] H. Arslan, Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data, Multidisciplinary Digital Publishing Institute Proceedings, 2021, p. 20.

[50] R.K. Pathan, M. Biswas, M.U. Khandaker, Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model, Chaos, Solit. Fractals 138 (2020), 110018.

[51] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent Individualized Feature Attribution for Tree Ensembles, 2018 arXiv preprint arXiv:1802.03888.

[52] B.B. Hazarika, D. Gupta, Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks, Appl. Soft Comput. 96 (2020), 106626.

[53] J. Fu, J. Tang, Y. Wang, X. Cui, Q. Yang, J. Hong, F. Zhu, Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification, Front. Pharmacol. 8 (2018) 681.