




OPEN

## Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME

Mariam Laatif<sup>1</sup>, Samira Douzi<sup>2,6</sup>, Hind Ezzine<sup>1,3</sup>, Chadia El Asry<sup>4</sup>, Abdellah Naya<sup>5</sup>, Abdelaziz Bouklouze<sup>6</sup>, Younes Zaid<sup>1,5,7</sup> & Mariam Naciri<sup>1</sup>

The rapid spread of SARS-CoV-2 threatens global public health and impedes the operation of healthcare systems. Several studies have been conducted to confirm SARS-CoV-2 infection and examine its risk factors. To produce more effective treatment options and vaccines, it is still necessary to investigate biomarkers and immune responses in order to gain a deeper understanding of disease pathophysiology. This study aims to determine how cytokines influence the severity of SARS-CoV-2 infection. We measured the plasma levels of 48 cytokines in the blood of 87 participants in the COVID-19 study. Several Classifiers were trained and evaluated using Machine Learning and Deep Learning to complete missing data, generate synthetic data, and fill in any gaps. To examine the relationship between cytokine storm and COVID-19 severity in patients, the Shapley additive explanation (SHAP) and the LIME (Local Interpretable Model-agnostic Explanations) model were applied. Individuals with severe SARS-CoV-2 infection had elevated plasma levels of VEGF-A, MIP-1b, and IL-17. RANTES and TNF were associated with healthy individuals, whereas IL-27, IL-9, IL-12p40, and MCP-3 were associated with non-Severity. These findings suggest that these cytokines may promote the development of novel preventive and therapeutic pathways for disease management. In this study, the use of artificial intelligence is intended to support clinical diagnoses of patients to determine how each cytokine may be responsible for the severity of COVID-19, which could lead to the identification of several cytokines that could aid in treatment decision-making and vaccine development.

Coronavirus disease 2019 (COVID-19) is a global public health emergency with severe consequences for populations, health systems, and the economy; consequently, finding ways to prevent the virus's spread is imperative. This disease is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and its symptoms range from mild fatigue to activation of inflammatory factors due to a large number of inflammatory exudates and erythrocytes entering the alveoli, leading to dyspnea, respiratory failure, and possibly death<sup>1–4</sup>.

Cytokines are water-soluble extracellular polypeptides or glycoproteins ranging in size from 8 to 30 kDa; they are produced by multiple cell types at sites of tissue injury and by immune system cells by activating mitogen-activated protein kinases<sup>5</sup>. In biomedicine, cytokines have risen to prominence as diagnostic, prognostic, and therapeutic agents for human diseases. The “cytokine storm” is a systemic inflammatory response to infections and medications; it causes an excessive activation cascade of immune cells and the production of pro-inflammatory cytokines as a result of an unregulated host immunological response to multiple triggers<sup>6,7</sup> such as infections, malignancies, rheumatic diseases, etc.

The cytokine storm is an inflammatory response to infections and drugs that activates and produces pro-inflammatory cytokines<sup>8,9</sup>. In this context, a number of studies have examined the predictive value of sTREM-1,

<sup>1</sup>Laboratory of Biodiversity, Ecology and Genome, Department of Biology, Faculty of Sciences, Mohammed V University, Rabat, Morocco. <sup>2</sup>IPSS Laboratory, Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco. <sup>3</sup>Public Health International Consultant, Rabat, Morocco. <sup>4</sup>Faculty of Sciences, IPSS Laboratory, Mohammed V University, Rabat, Morocco. <sup>5</sup>Department of Biology, Immunology, and Biodiversity Laboratory, Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco. <sup>6</sup>Laboratory of Pharmacology and Toxicology, Pharmaceutical and Toxicological Analysis Research Team, Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco. <sup>7</sup>Research Center of Abulcasis, University of Health Sciences, Rabat, Morocco. ✉email: s.douzi@um5r.ac.ma

acetylcholine, fatty acids, lipids, IL-1a, IL-1b, TNF, IFN- $\gamma$ , and other mediator biomarkers in patients with COVID-19. These studies suggested that using these biomarkers could improve the timing of clinical and pharmacological interventions in COVID-19 patients<sup>10–12</sup>.

After SARS-CoV-2 attaches to the ACE2 receptor and infects alveolar epithelial cells, specific mechanisms are activated by the secretion of cytokines, resulting in an acute inflammatory response that includes lymphokine (produced by lymphocytes), monokine (produced by monocytes), chemokine (involved in chemotactic activities), and interleukin (produced by leukocyte and acting on other leukocytes), that directly promote the inflammatory process<sup>13</sup>.

Elevated or decreased cytokine levels in affected individuals, particularly critically ill patients, suggest that cytokine storms may play a role in COVID-19 pathophysiology<sup>14,15</sup>. Consequently, the cytokines normally released by the immune system would activate the most diverse arachidonic acid cascades, which produce severity-determining metabolites<sup>16–19</sup>.

Today, there is no cure for severe COVID-19, and few treatments improve clinical outcomes significantly. Even within the scientific community, many individuals are unable to differentiate between infection and inflammation (in this case, the development of the disease) when discussing SARS-CoV-2 infection. However, there are already anti-infection medications<sup>20,21</sup>.

The medical community is evaluating antiviral and immunomodulatory treatments for the disease in the interim. Antiviral and supportive treatments are unquestionably necessary for the treatment of COVID-19 patients, but anti-inflammatory therapy plays a crucial role in the management of COVID-19 patients because of its ability to prevent further harm, organ damage, or organ failure<sup>22</sup>.

Numerous studies have demonstrated the importance of certain treatments, such as dexamethasone, tocilizumab, and Regeneron's monoclonal antibody combination, in reducing the risk of death in hospitalized patients<sup>20–26</sup>. Remdesivir; Sotrovimab; Baricitinib; Evusheld<sup>®</sup> (cilgavimab + tixagevimab); Paxlovid<sup>®</sup> (nirmatadvir + ritonavir); and Molnupiravir<sup>®</sup> were also administered during the COVID-19 crisis<sup>23,24</sup>. The majority of these drugs are based on a variety of cytokines, including IL-6, TNF, IFN, IL-10, IL-1, IL-6, IL-2, IL-8, IL-10, IL-12, and IL-10<sup>27–39</sup>. However, biomarker and immune response analysis are still necessary to better comprehend the pathogenicity of the disease and develop more effective treatments and vaccines.

In addition, as computer technology has progressed, Machine Learning (ML) has become a valuable tool for resolving problems requiring mapping multiple inputs to a desired output. Applying ML techniques to determine if common patterns of cytokines can be identified in COVID-19 patients has shown tremendous promise in recent studies.

Natural Language Processing (NLP) was used by Rahman et al.<sup>33</sup> to extract relevant clinical markers from Electronic Health Records (EHRs). These extracted variables are capable of being modeled to reveal an association between infection severity outcomes and these variables.

Ghazavi et al.<sup>34</sup> use one-way ANOVA and Receiver operating characteristic curve (ROC) to determine the optimal cut-off values of cytokine levels for classifying COVID-19 severity with the highest sensitivity and specificity.

Gao et al.<sup>35</sup> developed a Nano plasmonic digital immunoassay by combining a Machine Learning-assisted nano plasmonic imaging strategy with a microfluidic immunoassay platform that overcomes significant limitations for cytokine profiling in actual patient samples.

Patterson et al.<sup>36</sup> used SMOTE to balance the classes in their Dataset and then developed a random forest classifier to identify relevant cytokines for disease onset.

Cabaro et al.<sup>37</sup>, utilized Machine Learning techniques including Linear Discriminant Analysis (LDA), Tree (CART), and neural networks to develop a cytokine profile of patients with mild and severe COVID-19 symptoms.

Liu et al.<sup>38</sup> used the Mann–Whitney U test and univariate and multivariate logistic regression models to determine the cumulative mortality rate based on the normal range of cytokines in order to examine the effect of COVID-19 on the secretion of cytokines.

Other comparable machine learning-based experiments were discussed in terms of the creation of predictive models that lacked interpretability despite their high performance. In fact, the issue with ML approaches in healthcare applications is their black-box nature<sup>39–42</sup>, in which the process of achieving a particular output is concealed. Incorporating interpretation frameworks could increase the acceptability of an ML technique designed to combat COVID-19 by incorporating interpretation frameworks. The analytical transparency provided by these frameworks exceeds the capabilities of conventional data analysis techniques.

In this paper, we have adopted the SHAP and the LIME explainer. The adoption of these models is a sophisticated method for enhancing the transparency of machine learning (ML) models, as they provide both a global and a local perspective on how each factor influences the final probability associated with the potential development of the pathology.

The practical implications of employing these models can support clinical diagnoses performed on examined patients to determine how each cytokine may be responsible for the possible development of the disease and, therefore, be treated individually, as well as to suggest several cytokines that could aid in treatment decision-making and vaccine development.

Until now, the analytical methods utilized in the study of COVID-19 have always yielded generic results, whereas dissecting the problem and isolating its components could provide clinical operators with useful information.

In this study, the Mice-Forest model was utilized to fill in missing data, followed by the VAE Deep Learning model to generate synthetic data. Then, multiple classifiers were used to forecast COVID-19 outcomes (Healthy, Non-Severe, and Severe). The outputs of the models were then explained globally and locally using SHAP and

LIME. The most predictive attributes for each group (Healthy, Non-Severe, and Severe) were identified and ranked based on the interpretation results.

### Basic concepts

**Shapley additive explanation (SHAP).** The Shapley Additive Explanation (SHAP) algorithm is a technique for explaining the predictions of machine learning models. A game-theoretical approach<sup>43</sup> assigns a value to each input feature to represent its contribution to the prediction while considering all possible feature combinations.

Natural language processing, computer vision, and healthcare have all used the SHAP algorithm. It has been shown that it provides more accurate and comprehensible explanations for complex machine learning model predictions than other techniques<sup>43</sup>.

The SHAP algorithm is as follows:

Given an input  $[x_1, x_2, \dots, x_p]$  with  $p$  as the number of features and a trained model  $f$ , SHAP approximates  $f$  with a simple model  $g$  that can explain the contribution of each feature value<sup>43,44</sup> and thus determine the effect of each feature on the prediction for each possible subset of features. The formula for model  $g$  is as follows:

$$g(z) = \varphi_0 + \sum_{i=1}^M \varphi_i z_i \quad (1)$$

$[z_1, z_2, \dots, z_p]$  is a simplification of the input  $x$ , where the value of  $z$  is 1, corresponding to the features used in the prediction of the data, while its value is 0 the corresponding feature is not used.  $\varphi_i \in \mathbb{R}$  represents the Shapley value of each feature which is a weighted sum of the contributions across all possible subsets, with the weights being proportional to the number of features in each subset.

The  $\varphi_i$  is calculated by the following equation:

$$\varphi_i(f, x) = \sum_{z \subseteq x} \frac{|z|!(p - |z| - 1)!}{p!} [f(z) - f(z \setminus i)] \quad (2)$$

SHAP produces a collection of feature weights<sup>44</sup> that can be utilized to explain the model's predictions. These weights account for the interaction between features and provide a more precise and nuanced explanation of the model's behavior. In other words, SHAP computes the Shapley value of each feature as a player in the learned model, a process that must be performed for all possible permutations of features and requires exponential time. Nevertheless, it is known that SHAP can be efficiently calculated for tree-structured models, and since the learning model used in this work is a gradient-boosting tree, the calculation time for the SHAP algorithm can be decreased (Supplementary Table S1).

**Local interpretable model-agnostic explanations (LIME).** LIME is an interpretable machine learning framework used to explain the independent instance predictions of machine learning models<sup>45</sup>. LIME modifies the feature values of a single data sample and then observes the effect on the output. In accordance with this concept, LIME generates a novel dataset comprised of permuted samples and their respective black box model predictions. Among other techniques, the dataset is created by adding noise to continuous features, removing words (for NLP problems), and concealing portions of an image. On this novel dataset, LIME trains an interpretable model (e.g., a linear regression model, a decision tree, etc.) that is weighted by the proximity of the sampled instances to the required instance and conducts tests to determine what happens to the model's predictions when the data is modified. An explanation is obtained by locally approximating the underlying model with an interpretable model<sup>46</sup>. Local surrogate models with the interpretability requirement can be expressed mathematically as follows:

$$\text{Interpretation}(x) = \arg \min_{v \in V} L(u, v, \pi_x) + w(v) \quad (3)$$

We consider an explainable model  $v$  (e.g., a decision tree) for the sample  $x$  that reduces a loss  $L$  (e.g., binary cross entropy) and measure how close the interpretation is to the value anticipated by the initial model  $u$ . (e.g., a gradient boosting model). This procedure is carried out while minimizing the model complexity  $w(v)$ . Here,  $V$  represents the set of realizable explanations, which, in a hypothetical scenario, could be decision tree models. The closeness measures  $\pi_x$  defines the extent of the locality surrounding sample  $x$  and are considered in the explanation<sup>46</sup>.

We can conclude that The LIME algorithm is implemented as follows:

- Choose a specific instance of the input data for which the model's prediction is to be explained.
- The selected instance is perturbed by generating a set of slightly modified data points surrounding the original point.
- Predict the output for each perturbed data point using the black-box model and record the corresponding input features.
- Utilize the recorded input features and output values to train a locally interpretable model, such as a linear regression or decision tree, using the recorded input features and output values.
- Explain the prediction of the black-box model on the original data point using the local model.

LIME has been implemented in numerous fields, including natural language processing, computer vision, and healthcare, to explain the predictions of complex machine learning models. It can be used to determine which input features are most influential in determining the model's output for a particular instance of input data.

**Variational AutoEncoder (VAE).** Autoencoders are neural networks that are trained to minimize the reconstruction error between inputs and outputs<sup>47</sup>. The most prevalent constraint consists of reducing the dimensionality of the hidden layers so that the neural network retains only the most essential features required to reconstruct the input.

Variational autoencoder (VAE) inherits the architecture of autoencoder but imposes additional constraints on the bottleneck, thereby transforming conventional deterministic autoencoder into a potent probabilistic model<sup>48</sup>.

As shown in Fig. 1, the VAE is composed of an encoder E and a decoder D; the distributions of the encoder and decoder are denoted by  $q_\phi$  and  $p_\theta$ , respectively. The standard VAE assumes that both X and z adhere to Gaussian distributions, so the encoder does not output z directly, but rather the distribution parameters of z, i.e., the mean and variance of the Gaussian distribution, and z is then reconstructed using reparameterization. The decoder then outputs the mean of the Gaussian distribution with z as the input and sets the variance to a constant value. As the likelihood function of X, this Gaussian distribution is utilized. The model is optimized by calculating the Kullback–Leibler divergence between the prior and posterior distributions of z and the log-likelihood function of X. In conclusion, VAE learns the distribution of z based on X and reconstructs the distribution of X based on z. The following formula expresses the process of encoding and decoding:

$$\begin{cases} \mu, \sigma = E(X, \phi) \\ z = \mu + \sigma \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, 1) \\ \tilde{X} = D(z, \theta) \end{cases} \quad (4)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of z,  $\tilde{X} \in R^L$  is the reconstructed output,  $\phi$  and  $\theta$  are encoder and decoder parameters.

VAE<sup>49</sup> is designed to make the distributions learned by the encoder and decoder as similar as possible. Typically, Kullback–Leibler (KL) divergence is used to describe the proximity of two distributions. The objective function of the VAE, therefore, begins with the KL divergence of the two variational distributions:

$$\begin{aligned} \text{KL}(q_\phi(z|X) || p_\theta(z|X)) &= \text{Eq}_\phi(z|X) [\log q_\phi(z|X) - \log p_\theta(X|z) - \log p_\theta(z)] + \log p_\theta(X) \\ \text{Equivalent to :} & \\ \log p_\theta(X) - \text{KL}(q_\phi(z|X) || p_\theta(z)) &= -\text{Eq}_\phi(z|X) [\log q_\phi(z|X) - \log p_\theta(X|z) - \log p_\theta(z)] \end{aligned} \quad (5)$$

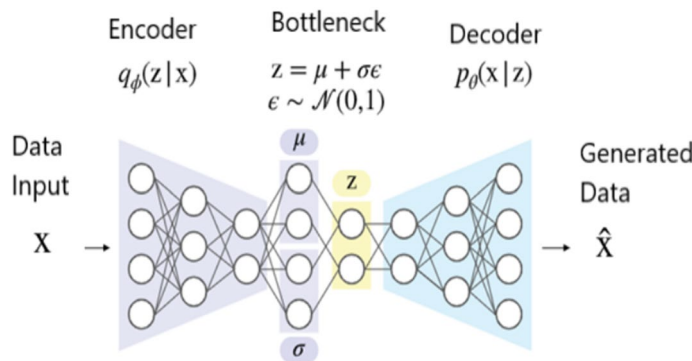
VAE's loss function is expressed as :

$$\begin{aligned} L &= -\text{Eq}_\phi(z|X) [\log q_\phi(z|X) - \log p_\theta(X|z) - \log p_\theta(z)] \\ &= \text{KL}(q_\phi(z|X) || p_\theta(z)) - \text{Eq}_\phi(z|X) [\log p_\theta(X|z)] \end{aligned} \quad (6)$$

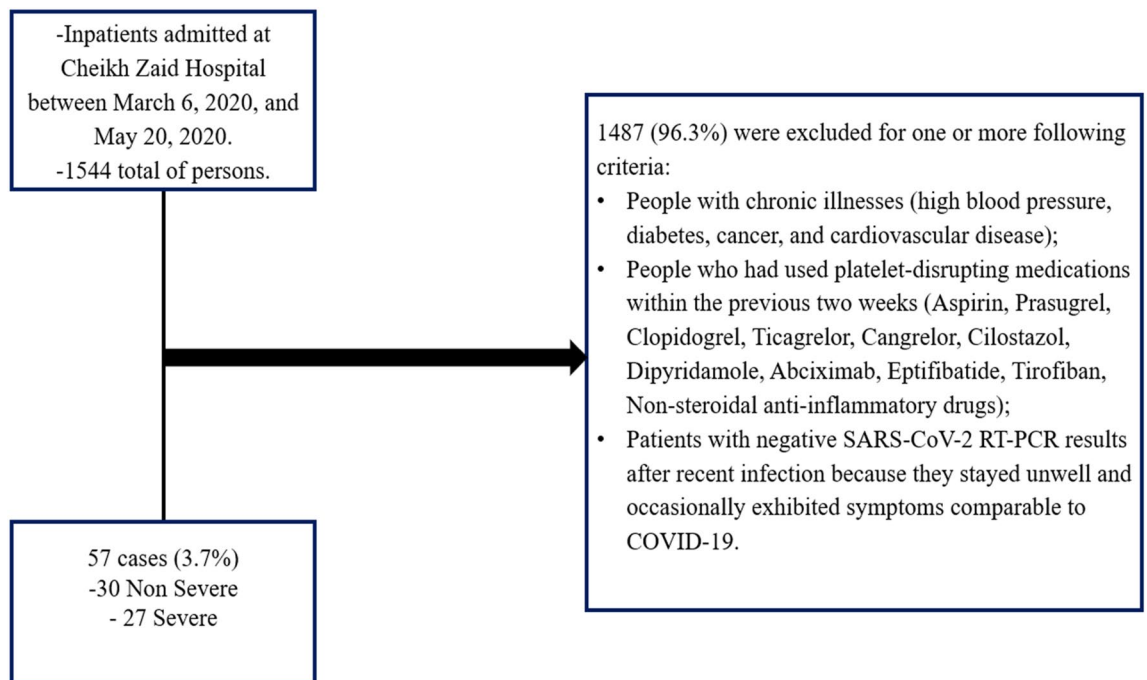
The first term is the regularization error of the posterior ( $|X$ ) and prior  $p_\theta(z)$  distributions. Its goal is to minimize the difference between the posterior and prior distributions (Fig. 2).

The second term is the log-likelihood function of X with respect to ( $|X$ ),  $p_\theta(X|z)$  represents the distribution of X generated by z, and this term computes the difference between X and the reconstructed output  $\tilde{X}$ . Consequently, reconstruction error can be written as:

$$\text{Eq}_\phi(z|X) [\log p_\theta(X|z)] = (X - \tilde{X})^2 \quad (7)$$



**Figure 1.** The standard VAE assumes that both X and z follow Gaussian distributions, so the encoder does not output z directly, but rather the distribution parameters of z, i.e. the mean and variance of the Gaussian distribution, and z is reconstructed using reparameterization. The decoder then outputs the mean of the Gaussian distribution with z as the input and fixes the variance.



**Figure 2.** Flowchart of cases recruitment.

VAE successfully combines a variational inference framework and an autoencoder, enabling the model to extract features and generate data more effectively. Several fields, such as image generation, natural language processing, and chemical design, have already demonstrated the potential of VAE<sup>50</sup>.

## Methods and results

The association between cytokine storm and COVID-19 severity was investigated using advanced machine learning techniques as depicted in Fig. 3. Missing values are filled in after data processing, and synthetic data is generated using the VAE model. Many Classifiers are trained on the synthetic set, and their performance is assessed using real data. Following that, the models' fitting performances are compared, and the best model is chosen. SHAP and LIME analyses are carried out, and their plots are created to reveal the feature impact in all cases. The top five features of the optimal model are chosen based on the overall attribution values. For all analyses, we used Python 3.6.7. The dataset used and the specifics of how the approaches were used are described in the sections that follow.

\*The Recruitment was approved by the Ethics Committee of Cheikh Zaid Hospital (CEFCZ/PR/2020/PR04) and complies with the Declaration of Helsinki. All participants gave their written informed consent and comply with the Declaration of Helsinki.

## Data availability

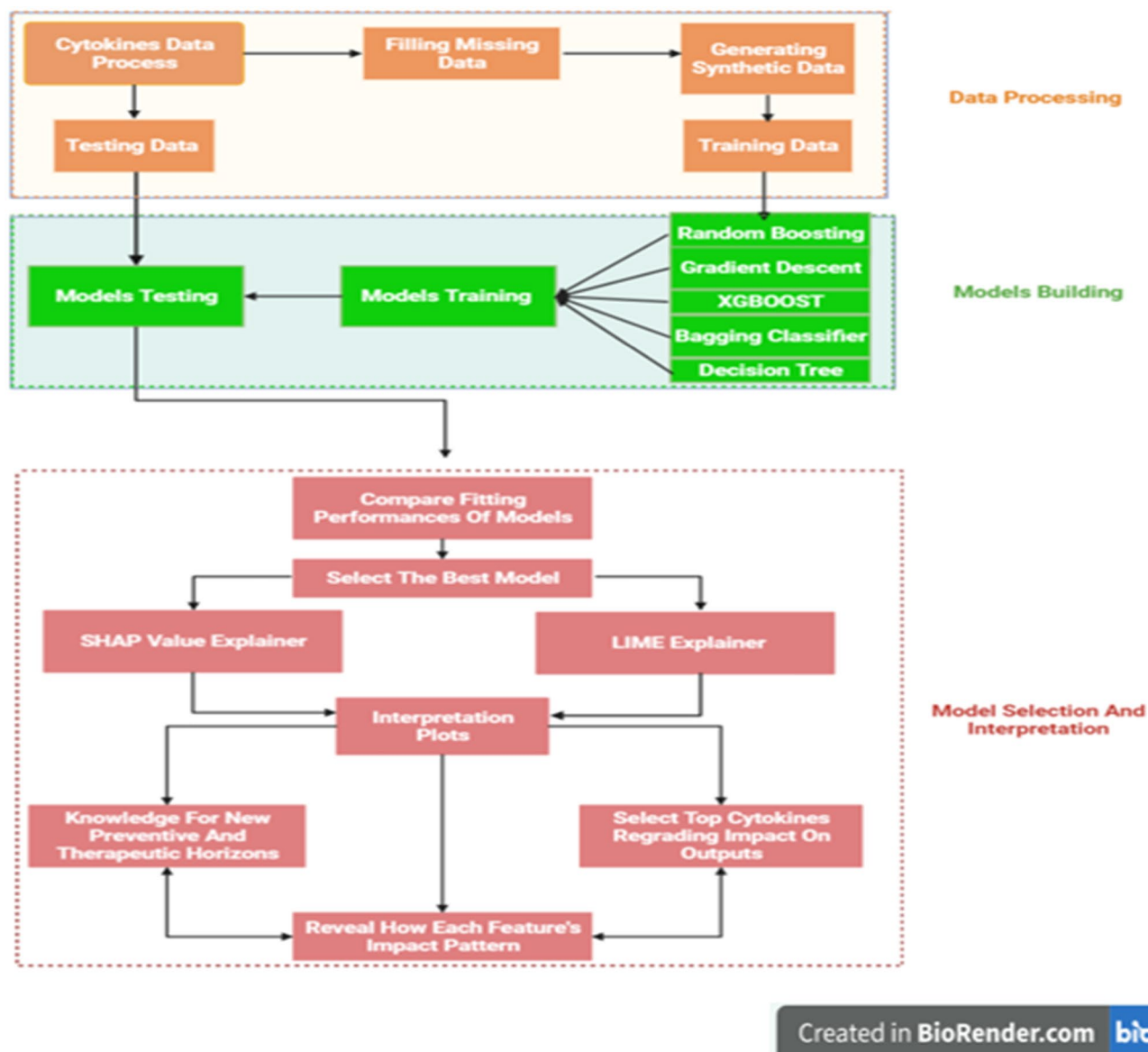
\*The datasets used and/or analyzed during the current study are available from the corresponding author.

**Clinical data.** We used a case–control study for our research. All patients were recruited from Rabat's Cheikh Zaid University Center Hospital. COVID-19 hospitalizations occurred between March 6, 2020, and May 20, 2020, and were screened using clinical features (fever, cough, dyspnea, fatigue, headache, chest pain, and pharyngeal discomfort) and epidemiological histology. Any patient admitted to Cheikh Zaid Hospital with a positive PCR-RT for SARS-CoV-2 was considered a COVID-19 case. According to the severity, the cases were divided into two categories: Cases with COVID symptoms and a positive RT-PCR test requiring oxygen therapy are considered severe. Case not requiring oxygen therapy: any case with or without COVID symptoms, normal lung CT with positive RT-PCR. The Controls were selected from Cheikh Zaid Hospital employees (two to three per week) who exhibited no clinical signs of COVID-19 and whose PCR-RT test was negative for the virus. People with chronic illnesses (high blood pressure, diabetes, cancer, and cardiovascular disease) and those who had used platelet-disrupting medications within the previous two weeks (Aspirin, Prasugrel, Clopidogrel, Ticagrelor, Cangrelor, Cilostazol, Dipyridamole, Abciximab, Eptifibatide, Tirofiban, Non-steroidal anti-inflammatory drugs) are excluded from our study (Fig. 2).

Consequently, a total of 87 participants were selected for this study and divided as follows: 57 Patients infected with SARS-CoV-2: Thirty without severe COVID-19 symptoms, twenty-seven with severe symptoms requiring hospitalization, and thirty healthy controls. Table 1 displays patients' basic demographic and clinical information.

The cytokines investigated in our study are displayed in Table 2, it consists of two panels, the first one contains 48 cytokines, while the second panel contains only 21 cytokines.





**Figure 3.** This flowchart depicts the various machine learning models used in this study, beginning with the filling in of missing values, followed by the generation of data, and concluding with the application of predictive models that have proven effective. The best of these models was chosen for interpretation, and LIME and SHAPE were used to interpret the model's decisions and determine the cytokines that influence the severity of covid 19 disease.

**Missing data handling process.** A data imputation procedure was considered for filling in missing values in entries. In fact, 29 individuals in our dataset had a missingness rate of more than 50 percent for their characteristics (cytokines), therefore our analysis will be significantly impacted by missing values. The most prevalent method for dealing with incomplete information is data imputation prior to classification, which entails estimating and filling in the missing values using known data.

There are a variety of imputation approaches, such as mean, k-nearest neighbors, regression, Bayesian estimation, etc. In this article, we apply the iterative imputation strategy Multiple imputation using chained equations Forest (Mice-Forest) to handle the issue of missing data. The reason for this decision is to employ an imputation approach that can handle any sort of input data and makes as few assumptions as possible about the data's structure<sup>55</sup>. the chained equation process is broken down into four core steps which are repeated until optimal results are achieved<sup>56</sup>. The first step involves replacing every missing data with the mean of the observed values for the variable. In the second phase, mean imputations are reset to "missing." In the third step, the observed values of a variable (such as 'x') are regressed on the other variables, with 'x' functioning as the dependent variable and the others as the independent variables. As the variables in this investigation are continuous, predictive mean matching (PPM) was applied.

Index	Healthy donors	COVID-19 non-severe	COVID-19 severe	p-value		
				Healthy vs non-severe	Healthy vs severe	Non-severe vs severe
N° of patients	30	30	27			
Female/Male	15/15	14/16	14/13	–	–	–
Age, years	54.32 ± 9.26	58.12 ± 8.60	61.15 ± 17.82	0.74	0.49	0.96
Weight, kg	87.79 ± 14.53	79.41 ± 16.38	75.68 ± 8.91	0.28	0.37	>0.99
Platelet number × 10 <sup>9</sup> /L	234 ± 63.07	242 ± 56.20	229 ± 37.64	0.56	>0.99	0.54
ALT, U/L	12.22 ± 4.53	12.47 ± 6.62	19.15 ± 5.08	>0.99	≤0.05	≤0.05
AST, U/L	10.95 ± 5.35	17.31 ± 4.26	25.69 ± 6.40	≤0.05	≤0.05	≤0.05
LDH, U/L	325.77 ± 83.46	449.66 ± 83.92	458.50 ± 102.11	≤0.05	≤0.05	>0.99
C-reactive protein, mg/L	5.45 ± 3.57	12.16 ± 6.77	19.94 ± 4.88	≤0.05	≤0.05	≤0.05
D-dimers, mg/L	0.36 ± 0.43	0.81 ± 0.48	0.88 ± 0.78	≤0.05	≤0.05	0.94

**Table 1.** Patients' data is presented as mean standard deviation. ALT stands for alanine aminotransferase; AST stands for aspartate aminotransferase; and LDH stands for lactate dehydrogenase. Unpaired statistical analysis P values were calculated using the student t-test. Statistical significance is defined as p less than 0.05.

Panel	Number of cytokines	Number of patients	Included cytokines
Panel-1	48	58	RANTES, sCD40L, EGF, Eotaxin, FGF-2, FLT-3L, MIG/CXCL9, IL-1b, Fractalkine, G-CSF, GM-CSF, GROa, IFN-a2, IFNy, IL-1a, PDGF-AB/BB, IL1-ra, IL-2, IL-3, VEGF-A, IL-4, IL-10, IL-5, IL-6, TGFa, IL-7, IL-8, IL-17A, IL-9, IL-12p40, MCP-1, IL-12p70, IL-13, MIP-1a, IL-15, IL-17E/IL-25, IL-17F, IL-18, IL-22, IL-27, IP-10, MCP-3, M-CSF, MDC, TNF, MIP-1b, PDGF-AA, <sup>51-54</sup>
Panel-2	21	29	GM-CSF, IL-8, GROa, IL-3, IFN-a2, IFNy, IL-1a, IL-9, IL-1b, IL-2, IL-4, IL-6, G-CSF, IL-7, IL-10, IL-12p40, IL-5, IL-12p70, IL-13, IL-15, IL1-ra

**Table 2.** Cytokines contained in each panel.

The fourth stage involves replacing the missing data with the regression model's predictions. This imputed value would subsequently be included alongside observed values for other variables in the independent variables. An iteration is the recurrence of steps '2' through '4' for each variable with missing values. After one iteration, all missing values are replaced by regression predictions based on observed data. In the present study, we examined the results of 10 iterations.

The convergence of the regression coefficients is ideally the product of numerous iterations. After each iteration, the imputed values are replaced, and the number of iterations may vary. In the present study, we investigated the outcomes of 10 iterations. This is a single "imputation." Multiple imputations are performed by holding the observed values of all variables constant and just modifying the missing values to their appropriate imputation predictions. Depending on the number of imputations, this leads to the development of multiply imputed datasets (30, in this study). The number of imputations depends on the values that are missing. The selection of 30 imputations was based on the White et al.<sup>57</sup> publication. The fraction of missing data was around 30%. We utilized the version 5.4.0 of the miceforest Python library to impute missing data. The values of the experiment's hyper-parameters for the Mice-Forest technique are listed in Table 3, and Fig. 4 illustrates the distribution of each imputation comparing to original data (in red).

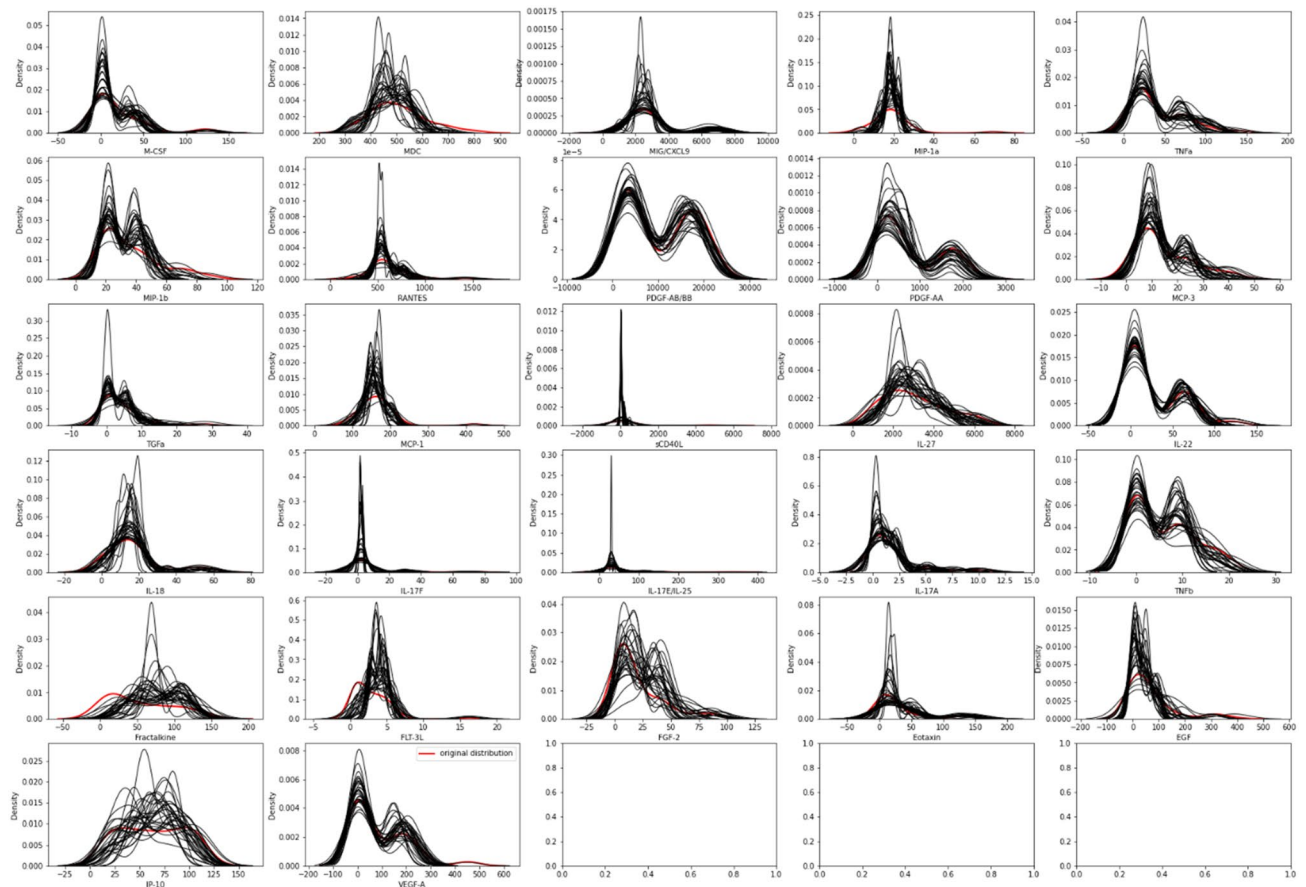
**Synthetic data generation.** Machine learning frameworks have demonstrated their ability to deal with complex data structures, producing impressive results in a variety of fields, including health care. However, a large amount of data is required to train these models<sup>58</sup>. This is particularly challenging in this study because available datasets are limited (87 records and 48 attributes) due to acquisition accessibility and costs, such limited data cannot be used to analyze and develop models.

To solve this problem, Synthetic Data Generation (SDG) is one of the most promising approaches (SDG) and it opens up many opportunities for collaborative research, such as building prediction models and identifying patterns.

Synthetic Data is artificial data generated by a model trained or built to imitate the distributions (i.e., shape and variance) and structure (i.e., correlations among the variables) of actual data<sup>59,60</sup>. It has been studied for

Techniques	Hyper-parameters
Mice-Forest	iterations = 10, imputation = 10 estimators = 50

**Table 3.** Parameters of mice-forest.



**Figure 4.** The distribution of each imputation compared to the original data (in red).

several modalities within healthcare, including biological signals<sup>61</sup>, medical pictures<sup>62</sup>, and electronic health records (EHR)<sup>63</sup>.

In this paper, a VAE network-based approach is suggested to generate 500 samples of synthetic cytokine data from real data. VAE's process consists of providing labeled sample data ( $X$ ) to the Encoder, which captures the distribution of the deep feature ( $z$ ), and the Decoder, which generates data from the deep feature ( $z$ ) (Fig. 1).

The VAE architecture preserved each sample's probability and matched the column means to the actual data. Figure 5 depicts this by plotting the mean of the real data column on the  $X$ -axis and the mean of the synthetic data column on the  $Y$ -axis.

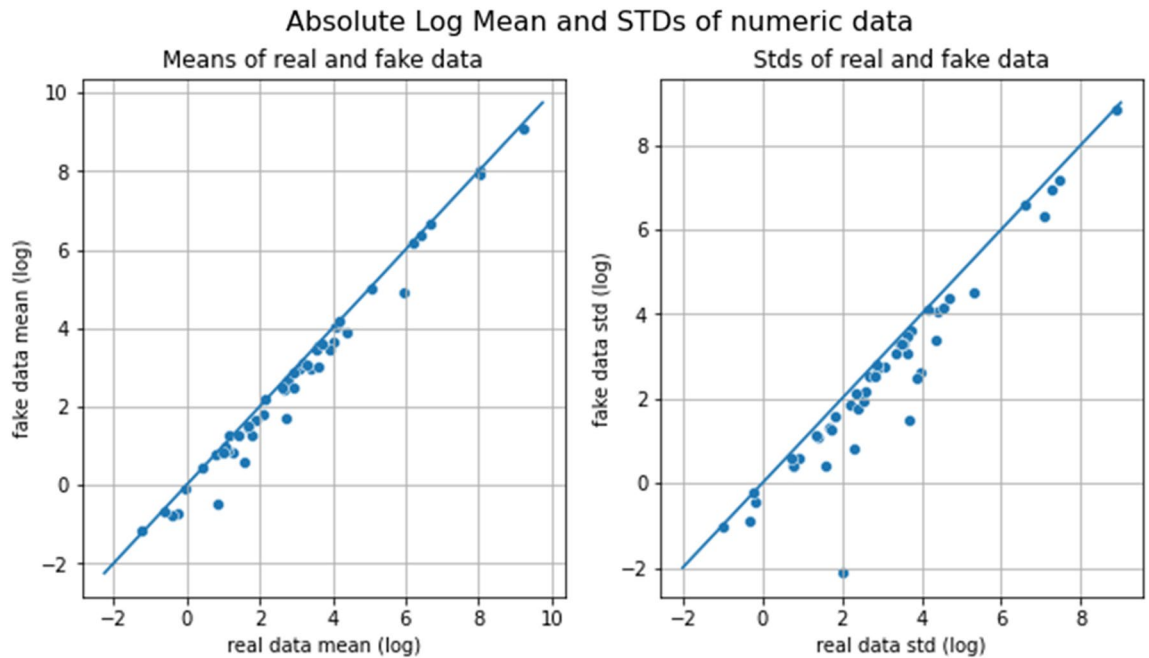
The cumulative feature sum is an extra technique for comparing synthetic and real data. The feature sum can be considered as the sum of patient diagnosis' values. As shown in Fig. 6, a comparison of the global distribution of feature sums reveals a significant similarity between the data distributions of synthetic and real data.

**Classification.** Five distinct models are trained on synthetic data (Random Forest, XGBoost, Bagging Classifier, Decision Tree, and Gradient boosting Classifier). Real data is used for testing, and three metrics were applied to quantify the performance of fitting: precision, recall, F1 score, and confusion matrix.

As shown in Figs. 7, 8, 9, 10 and 11 the performance of the Gradient Boosting Classifier proved to be superior to that of other models, with higher Precision, Recall, and F1 score for each class, and a single misclassification. Consequently, we expect that SHAP and LIME's interpretation of the Gradient Boosting model for the testing set will reflect accurate and exhaustive information for the cytokines data set.

**Explanations with LIME and SHAP models.** Explaining a prediction refers to the presentation of written or visual artifacts that enable qualitative knowledge of the relationship between the instance's components and the model's prediction. We suggest that if the explanations are accurate and understandable, explaining predictions is an essential component of convincing humans to trust and use machine learning effectively<sup>43</sup>. Figure 12 depicts the process of explaining individual predictions using LIME and SHAP as approaches that resemble the classifier's black box to explain individual predictions. When explanations are provided, a doctor is clearly in a much better position to decide using a model. Gradient Boosting predicts whether a patient has an acute case of COVID-19 in our study, whereas LIME and SHAP highlight the cytokines that contributed to this prediction.





**Figure 5.** Each point represents a column mean in the real and synthetic data. A perfect match would be indicated by all the points lying on the line  $y = x$ .

*Explanation of SHAP model.* The SHAP explanation utilized in this study is the Kernel Explainer, a model-agnostic approach that produces a weighted linear regression depending on the data, predictions, and model<sup>64</sup>. It examines the contribution of a feature by evaluating the model output if the feature is removed from the input for various (theoretically all) combinations of features. The Kernel Explainer makes use of a backdrop dataset to demonstrate how missing inputs are defined, i.e., how a missing feature is approximated during the toggling process.

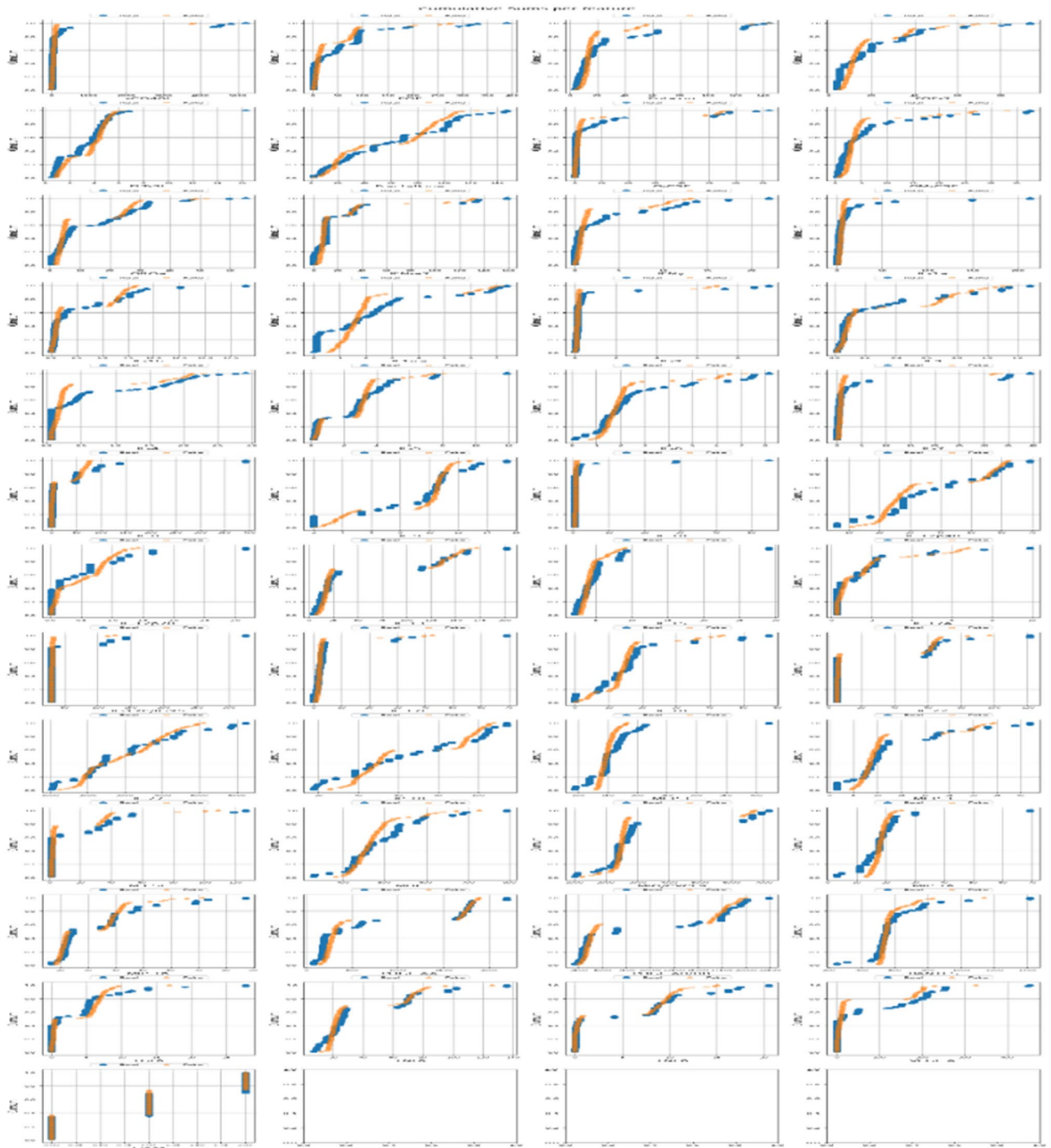
SHAP computes the impact of each characteristic on the learned system's predictions. Using gradient descent, SHAP values are created for a single prediction (local explanations) and multiple samples (resulting in global explanations).

Figure 13 illustrates the top 20 SHAP value features for each class in the cytokine data prediction model (Healthy, Severe, and Non-Severe classes). The distribution of SHAP values for each feature is illustrated using a violin diagram. Here, the displayed characteristics are ordered by their highest SHAP value. The horizontal axis represents the SHAP value. The bigger the positive SHAP value, the greater the positive effect of the feature, and vice versa. The color represents the magnitude of a characteristic value. The color shifts from red to blue as the feature's value increases and decreases. For example, Mip-1b in Figure 8, the positive SHAP value increases as the value of the feature increases. This may be interpreted as the probability of a patient developing COVID-19, severity increasing as MIP-1b levels rise.

In the situation of a healthy patient, TNE, IL-22, and IL-27 are the most influential cytokines, as shown in Fig. 14's first SHAP diagram (from left). The second diagram is for a patient with severity, and we can observe that the VEGF-A cytokine's value is given greater weight. This can be viewed as an indication that the patient got a serious COVID-19 infection due to the increase in this cytokine.

The last SHAP diagram depicts an instance of a non-Severe patient, and we can see that the higher the feature value, the more positive the direction of IL-27. On the other hand, MDC, PDGF-AB/BB, and VEGF-A cytokines have a deleterious effect. The levels of MDC and PDGF-AB/BB cytokines suggest that the patient may be recovering, however, the presence of VEGF-A suggests that the patient may develop a severe case of COVID-19, despite being underweight.

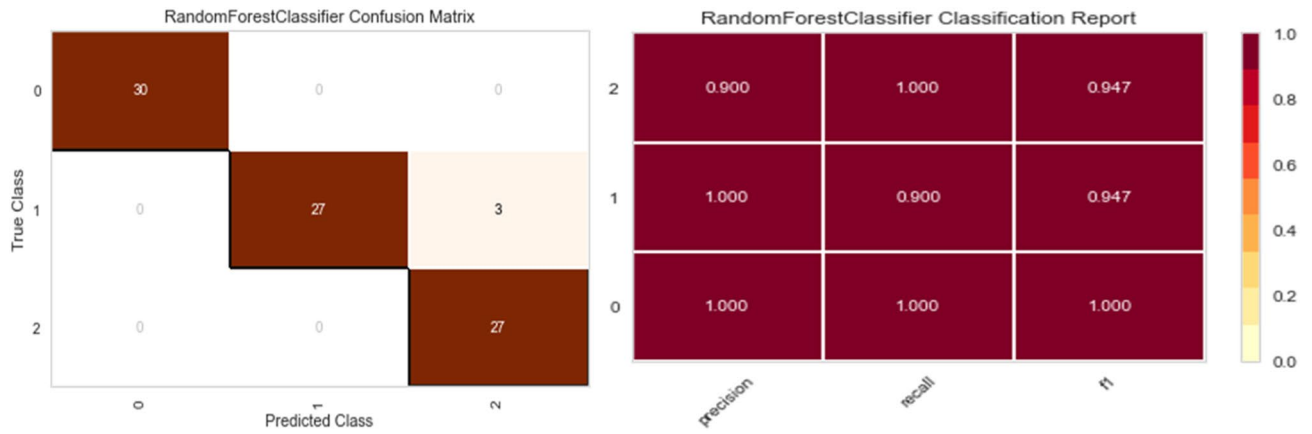
*Explanation of the LIME model.* LIME is a graphical approach that helps explain specific predictions. It can be applied to any supervised regression or classification model, as its name suggests. Behind the operation of LIME is the premise that every complex model is linear on a local scale and that it is possible to fit a simple model to a single observation that mimics the behavior of the global model at that locality. LIME operates in our context by sampling the data surrounding a prediction and training a simple interpretable model to approximate the black box of the Gradient Boosting model. The interpretable model is used to explain the predictions of the black-box model in a local region surrounding the prediction by generating explanations regarding the contributions of the features to these predictions. As shown in Fig. 15, a bar chart depicts the distribution of LIME values for each feature, indicating the relative importance of each cytokine for predicting Severity in each instance. The order of shown features corresponds to their LIME value.



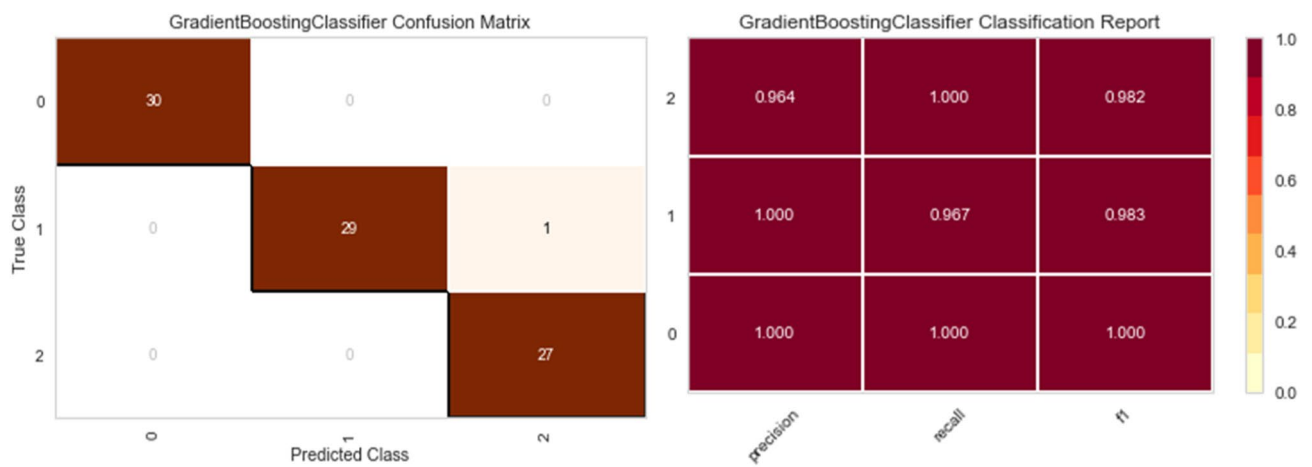
**Figure 6.** Plots of each feature in our actual dataset demonstrate the similarity between the synthesized and actual datasets.

In the illustrations explaining various LIME predictions presented in Fig. 16. We note that the model has a high degree of confidence that the condition of these patients is Severe, Non-Severe, or Healthy. In the graph where the predicted value is 2, indicating that the expected scenario for this patient is Severe (which is right), we can see for this patient that Mip-1b level greater than 41 and VEGF-A level greater than 62 have the greatest influence on severity, increasing it. However, MCP-3 and IL-15 cytokines have a negligible effect in the other direction.

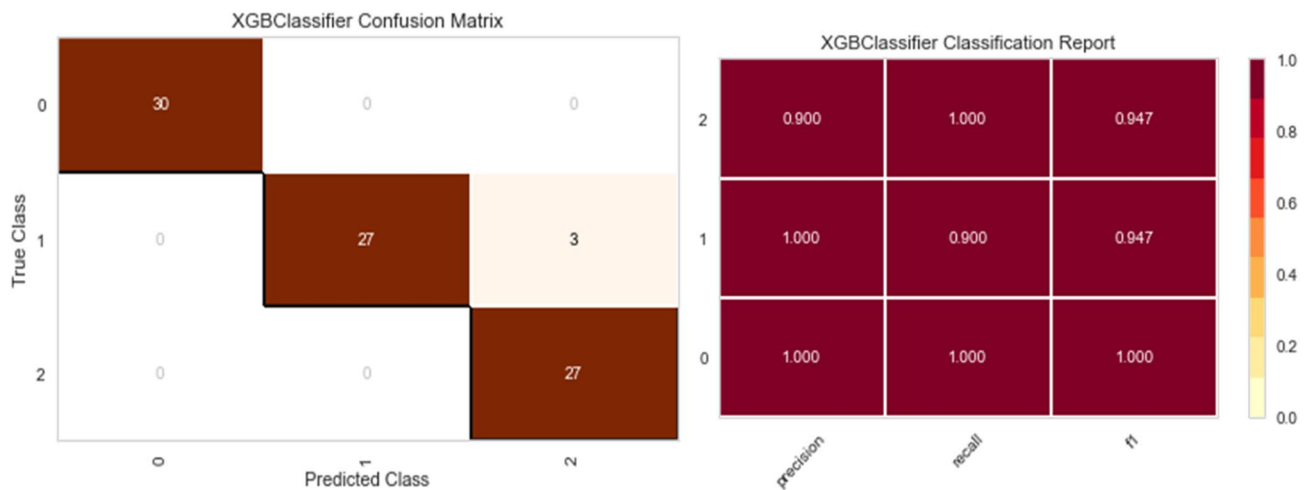
Alternatively, there are numerous cytokines with significant levels that influence non-Severity. For example, IL-27 and IL-9, as shown in the middle graph in Fig. 14. and that IL-12p40 below a certain value may have the opposite effect on model decision-making. RANTES levels less than 519, on the other hand, indicate that the patient is healthy, as shown in Fig. 16.



**Figure 7.** Matrix confusion and Report Classification of Random Forest.

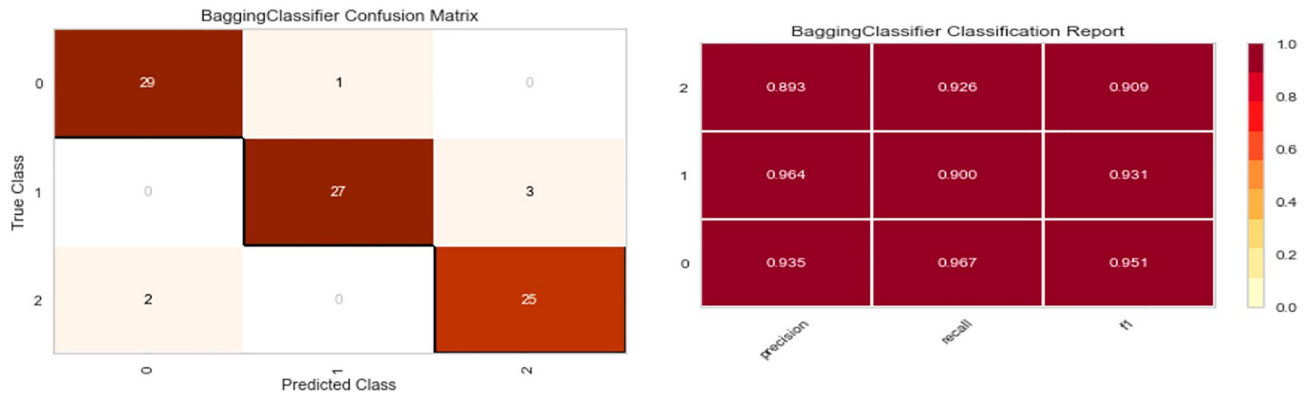


**Figure 8.** Matrix confusion and Report Classification of Gradient Boosting.

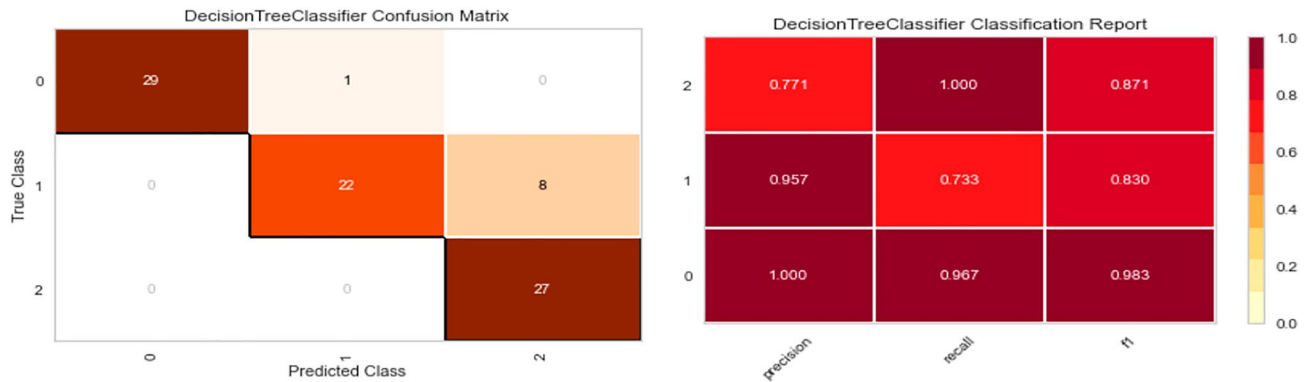


**Figure 9.** Matrix confusion and Report Classification of XGB Classifier.

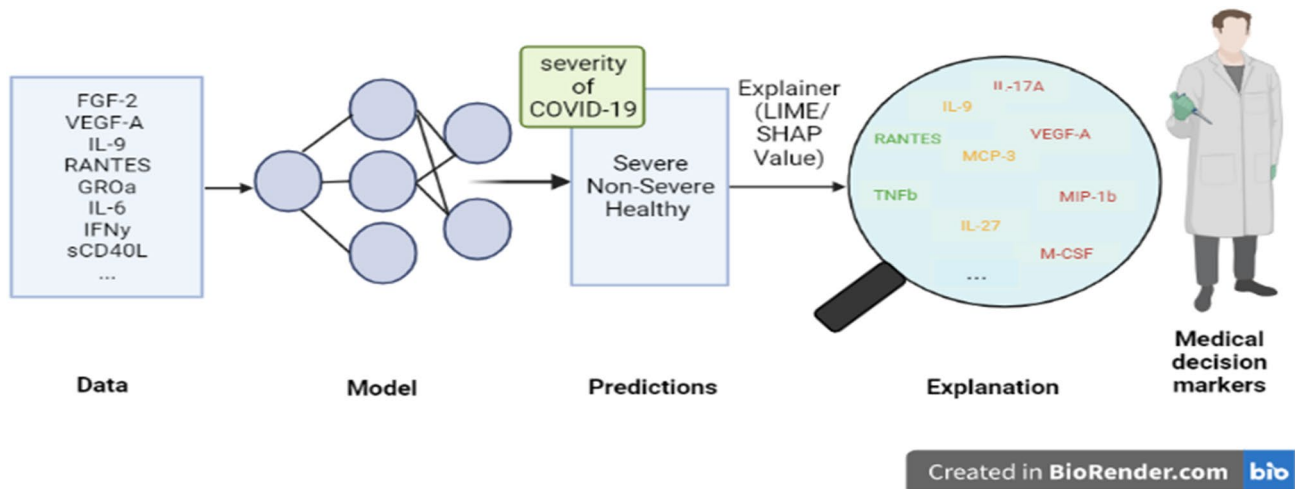
By comparing the individual’s explanation of SHAP values to the individual’s explanation of LIME values for the same patients, we may be able to determine how these two models differ in explaining the Severity results of the Gradient descent model. As a result, we can validate and gain insight into the impact of the most significant factors. To do so, we begin by calculating the frequency of the top ten features among all patients for each Explainer. We only consider features that appear in the top three positions, as we believe this signifies the feature’s



**Figure 10.** Matrix confusion and Report Classification of Bagging Classifier.



**Figure 11.** Matrix confusion and Report Classification of Decision Tree.

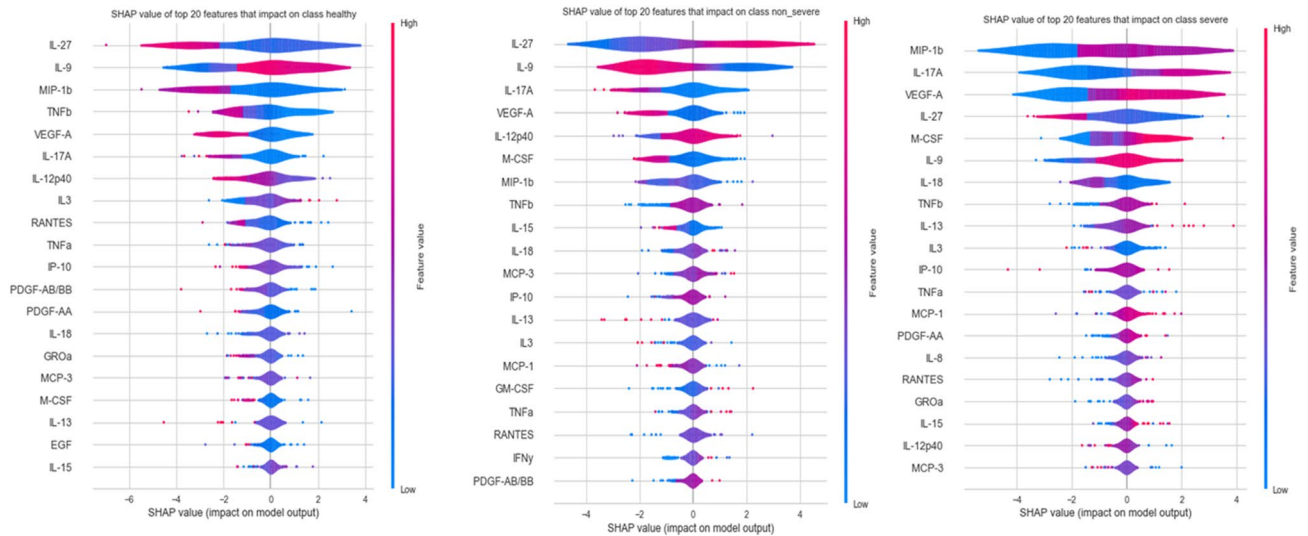


**Figure 12.** The Flow chart demonstrates how Machine learning can be used to make medical decisions. We entered cytokine data from severe, non-severe, and healthy patients, trained predictive models on cytokine data, and then used LIME and SHAP to explain the most important cytokine for each class of patients (Fig. 12).

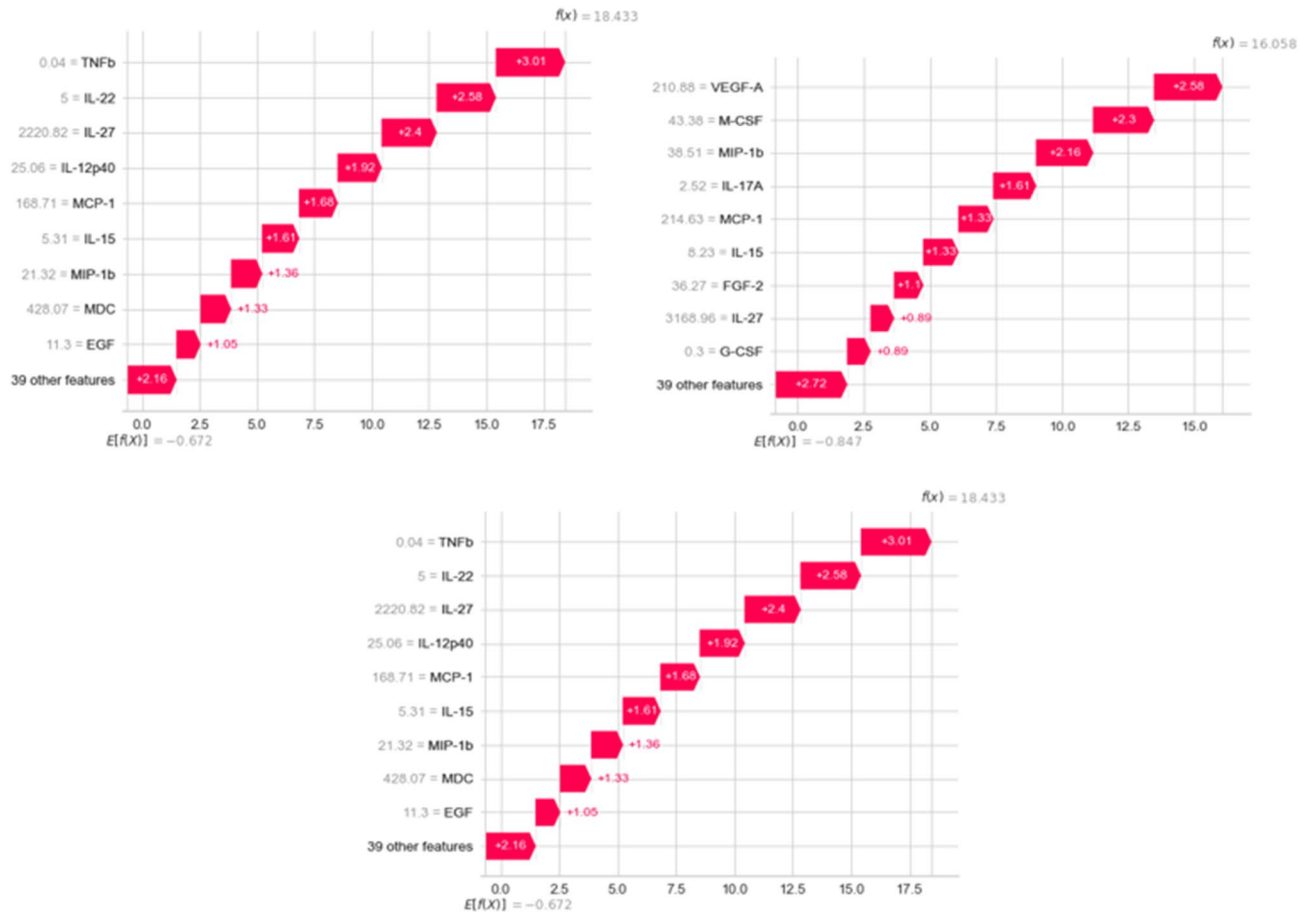
high value, and we only consider the highest-scoring features that appear at least ten times across all SHAP or LIME explanations (Tables 4, 5, and 6).

Table 4 demonstrates that MIP-1b, VEGF-A, and IL-17A have Unanimous Importance according to the SHAP Value and LIME. In addition, we can remark that M-CSF is necessary for LIME but is ranks poor.

In the instance of non-Severity, Table 5 reveals that IL-27 and IL-9 are essential in both explanatory models for understanding non-Severity in patients. We can see that IL-12p40 and MCP-3 are also essential for LIME and are highly ranked; hence, we add these two characteristics to the list of vital features for the non-Severity



**Figure 13.** Examples of SHAP values computed for individuals' predictions (local explanations) for Healthy, Non-Sever, and Sever patients.

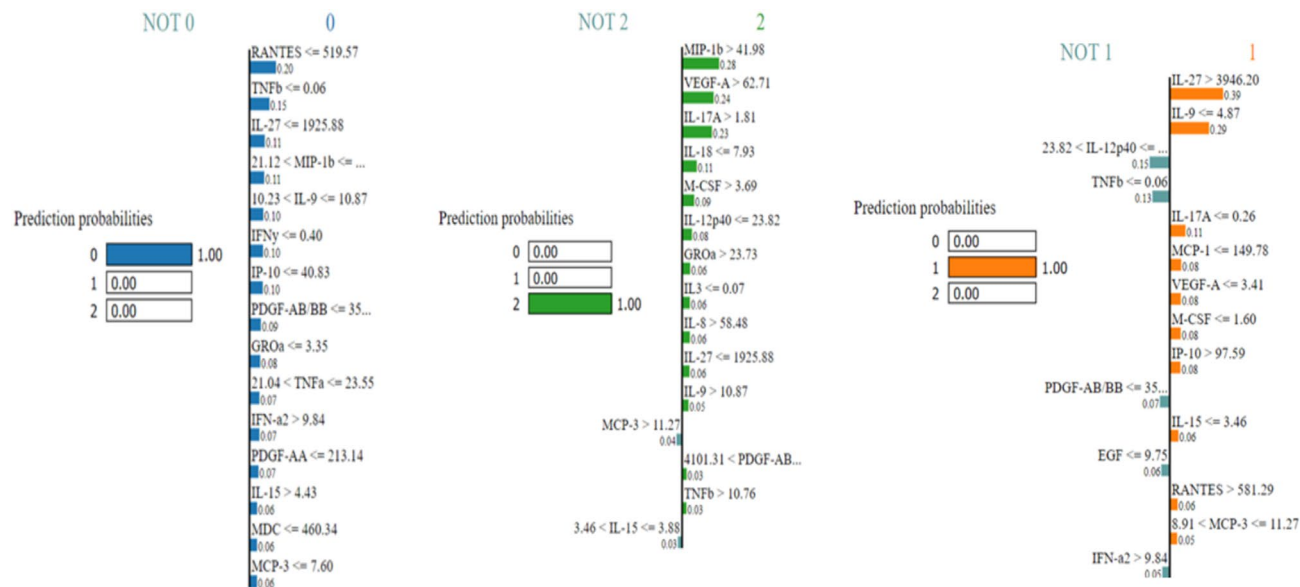


**Figure 14.** SHAP diagrams of characteristics with varying conditions: Healthy, Severe, and Non-Severe, respectively.

instance. RANTES, TNF, IL-9, IL-27, and MIP-1b are the most significant elements in the Healthy scenario, according to Table 6.

The elements that explain the severity of the COVID-19 sickness are summarized in Table 7.





**Figure 16.** Explaining individual predictions of Gradient descent classifier by LIME.

Cytokine	Appearance	
	SHAP value	LIME
MIP-1b	23	14
VEGF-A	24	27
IL-17A	23	23
M-CSF	9	15
IL-9	9	0
IL-12p40	4	1
IL-18	8	4
IL-8	2	0

**Table 4.** The top selected feature in the case of Severity for both explainers.

Cytokine	Appearance	
	SHAP value	LIME
IL-27	23	24
IL-9	27	22
IL-17A	9	12
VEGF-A	8	3
IL-12p40	9	23
MIP-1b	7	2
IP-10	5	6
MCP-3	6	16

**Table 5.** The top selected feature in the case of non-Severity for both explainers.

### Discussion

The severity-defining cytokines are VEGF-A, IL-17A, and MIP-1b, as shown in Table 6 and Figs. 13, 14. The non-Severity was linked to IL-27, IL-9, IL-12p40, and MCP-3, and RANTES, TNE, IL-9, IL-27, and MIP-1b were correlated with healthy cases. In addition, the levels of these cytokines are identified based on thresholds detected in the plasma of patients; in the case of Severity, the VEGF-A concentration was found to be greater than<sup>65,66</sup>. VEGF-A, which is essential for vascular endothelial homeostasis, is present in numerous cells and tissues. According to Zhang et al.<sup>67</sup>, VEGF-A plays an essential role in the activation of endothelial cells by binding

Cytokine	Appearance	
	SHAP value	LIME
RANTES	6	21
TNF	7	23
IL-9	21	18
IL-27	26	22
MIP-1b	22	16
IP-10	8	7

**Table 6.** The top selected feature in the case of Healthy for both explainers.

Selected cytokines		
Severity	Non-Severity	Healthy
MIP-1b	IL-27	RANTES
VEGF-A	IL-9	TNF
IL-17A	IL-12p40	IL-9
-	MCP-3	IL-27
-	-	MIP-1b

**Table 7.** The most significant features selected by SHAP Value and LIME.

to cell surface receptors, and the integrity of the endothelial barrier in lung tissue is essential for the regulation of alveolar immune function. In COVID-19, severe lung inflammation and associated immune responses induce apoptosis of epithelial and endothelial cells, which augments VEGF-A production and worsens edema and immune cell extravasation<sup>67</sup>. VEGF-A effects on vascular permeability and neo-angiogenesis<sup>68,69</sup> is responsible for this factor's pathogenic properties. Anti-VEGF-A therapy has the potential to be a miracle cure for reducing the severity of the disease in patients.

Researchers have linked COVID-19-related lung inflammation to increased plasma levels of a variety of proinflammatory cytokines, including IL-17A<sup>70,71</sup>. These results are similar with our own, where we discovered higher IL-17A levels in the peripheral blood of infected patients (Table 6 and Figs. 14, 16). As IL-17A promotes the production of other pro-inflammatory cytokines, such as IL-1, IL-6, and TNF, this finding clearly suggests that IL-17A plays an amplifying role in the inflammatory response. Moreover, the increase in IL-17 cytokines seen in these individuals lends credence to the theory that an immunological response leads to severe inflammation<sup>2</sup>. Furthermore, Previous research<sup>72</sup> indicates that COVID-19-infected patients with severe acute respiratory syndrome had higher levels of circulating IL-17A. According to this idea, the notion of a direct relationship between elevated IL-17A levels and the progression of illness severity becomes more consistent, and our findings indicate that patients develop disease severity when IL-17A concentrations exceed 1.81.

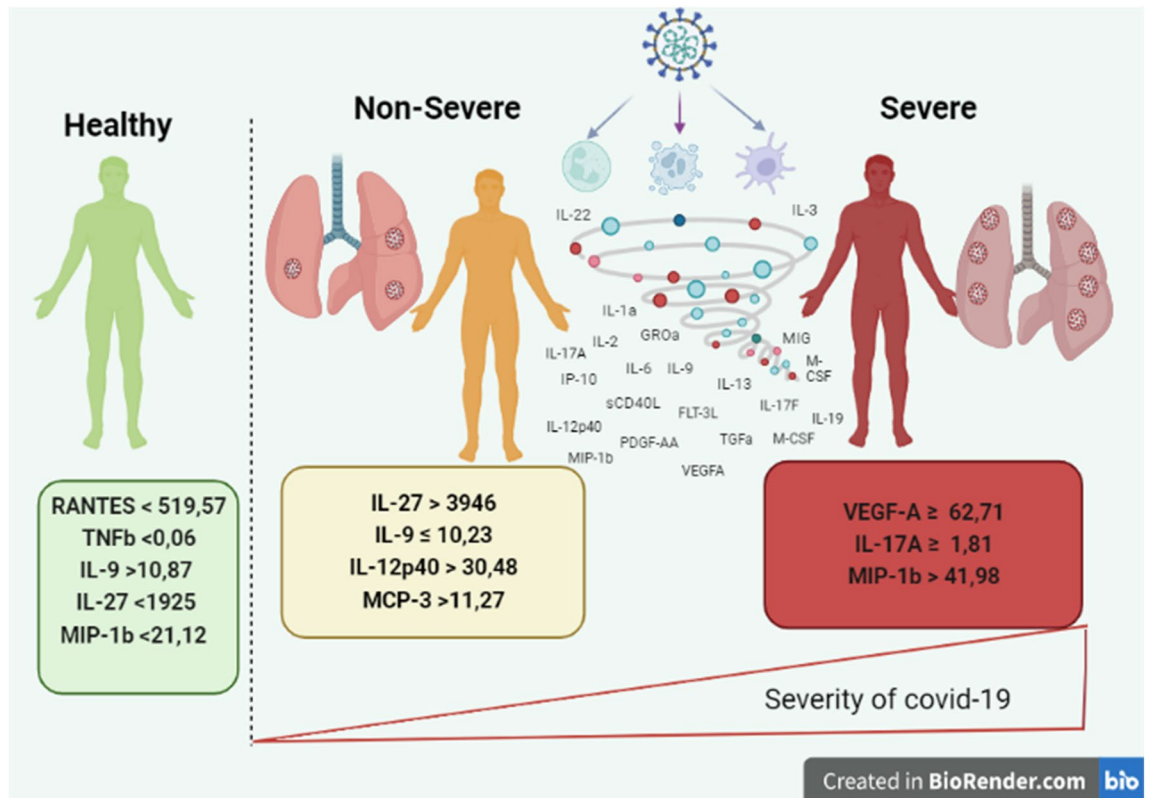
MIP-1b is likewise one of the cytokines with the highest risk when its concentration surpasses 41.98. Multiple disorders, particularly COVID-19, have been shown to exhibit an increase in MIP-1b protein<sup>65,73–75</sup>. The chemokine MIP-1b is a cytokine with the potential to attract monocytes and T cells and may be essential for the recruitment of inflammatory cells to damaged regions<sup>76,77</sup>. Directing inflammatory cells to the airways may result in severe illness or death<sup>77,78</sup>. In addition, MIP-1b was found to be substantial in healthy instances, albeit at a lesser concentration than in severe cases, with MIP-1b levels below 21.12. This result has never previously been reported.

The identification of VEGF-A, MIP-1b, and IL-17A as biomarkers that identify patients with a severe form of SARS-CoV-2 and the determination of their boundaries would enable researchers to more effectively triage and treat patients through the development of therapies and vaccines.

IL-27, IL-9, IL-12p40, and MCP-3 seemed significant in non-Severe samples, and other studies have reported aberrant levels of these cytokines in COVID-19 patients<sup>79–81</sup>.

According to Table 6 and Figs. 14, 16, IL-27 is more relevant in situations that are not severe and exceed 3946. IL-27 is involved in the development of Th1 cells and is a reliable predictive biomarker for COVID-19. On the other hand, we observe that IL-27 levels are low in Healthy cases, which is consistent with research indicating that patients who tested positive for SARS-COV-2 had an immunological imbalance with elevated levels of IL-17A and low levels of IL-27 compared to Healthy sample<sup>82</sup>.

IL-9 is also an indicator cytokine for non-Severity, with a threshold of 10.23 or below. Moreover, the results for the Healthy group indicated that IL-9 appeared significant, but with a higher threshold (IL-9 > 10.87). Consistent with prior research by Ghazavi et al., the concentration of serum IL-9 in COVID-19 patients did not differ significantly from that of the healthy group<sup>34</sup>. In other investigations, cytokine IL-9 levels were shown to be higher in COVID-19 patient groups compared to healthy controls<sup>66,83</sup>, contradicting our findings. The cytokine IL-12p40 with a threshold greater than 30.48 (IL-12p40 > 30.48) was a crucial bioactive in the non-Severe form. IL-12p40 is a macrophage chemoattractant that increases the migration of dendritic cells triggered by bacteria. Our



**Figure 15.** Different Cytokine Profiles Associated with the Progression of COVID-19 Severity.

results demonstrated that COVID-19-infected individuals express less IL-12p40 than healthy persons, and that IL-12p40 levels decreased as disease severity increased. These results are comparable to those of prior research that showed a decrease in IL-12p40 in intensive care patients<sup>66</sup>.

MCP-3 cytokine is also crucial in mild forms with a threshold greater than 11.27 ( $MCP-3 > 11.27$ ). MCP-3 was discovered by Yang et al.<sup>66</sup>, as an excellent predictor of COVID-19 progression and may serve as a starting point for therapy studies. Non-Severe patients had the highest amount of MCP-3 expression compared to the severe group<sup>84</sup>. Figure 15 summarizes our findings demonstrating the presence of specific biomarkers in healthy samples: RANTES with a threshold of less than 519.57 and TNF with a threshold of less than 0.06. These two biomarkers could be protective factors for these individuals, and research into these biomarkers could be a means for scientists to treat sick people more successfully.

## Conclusion

Cytokines are polypeptide signaling molecules that regulate multiple biological processes via cell surface receptors, such as those involved in adaptive and innate immunity for pro-inflammatory, interleukin, and anti-inflammatory cytokines. Using Machine Learning, we compared the cytokine profiles in this study to determine their significance in the disease's development. The SHAP and LIME models were used to evaluate the experimental findings regarding the relationship between cytokine storm and the severity of COVID-19 in patients, as well as the influence of various cytokines on severity.

We demonstrated that certain cytokines are likely produced in COVID-19-infected patients and that there are significant increases and decreases in the levels of these cytokines. Significant cytokines were identified as VEGF-A, MIP-1b, IL-17A, M-CSF, IL-27, IL-9, IL12p40, RANTES, and TNF in severe, non-severe, and healthy cases, respectively. These findings suggest that these cytokines may be disease promoters and open new avenues for disease prevention and treatment, this would contribute to a reduction in disease burden in terms of morbidity and mortality. Furthermore, the development of such treatments should consider the cost-benefit ratio, particularly for low-income countries.

Received: 27 November 2022; Accepted: 14 March 2023

Published online: 04 April 2023

## References

- Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**(10223), 507–513 (2020).
- Hoffmann, M. *et al.* SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**(2), 271–280 (2020).

3. Channappanavar, R. & Perlman, S. Pathogenic human coronavirus infections causes and consequences of cytokine storm and immunopathology. *Semin. Immunopathol.* **39**, 529–539 (2017).
4. Djomkam, A. L. Z., Ochieng'Olwal, C., Sala, T. B. & Paemka, L. Commentary: SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Front. Oncol.* **14**, 1–3 (2020).
5. Mehta, P. *et al.* COVID-19: Consider cytokine storm syndromes and immunosuppression. *Lancet* **395**(10229), 1033–1034 (2020).
6. Behrens, E. M. & Koretzky, G. A. Cytokine storm syndrome: Looking toward the precision medicine era. *Arthritis Rheumatol.* **69**(6), 1135–1143 (2017).
7. Tisoncik, J. R. *et al.* Into the eye of the cytokine storm. *Microbiol. Mol. Biol. Rev.* **76**(1), 16–32 (2012).
8. Cron, R. Q. (2019). IL-1 family blockade in cytokine storm syndromes. *Cytokine Storm Syndr.* 549–559.
9. Moore, J. B. & June, C. H. Cytokine release syndrome in severe COVID 19. *Science* **368**(6490), 473–474 (2020).
10. Pérez, M. M. *et al.* Acetylcholine, fatty acids, and lipid mediators are linked to COVID-19 severity. *J. Immunol.* **209**(2), 250–261 (2022).
11. Archambault, A. S. *et al.* High levels of eicosanoids and docosanoids in the lungs of intubated COVID-19 patients. *FASEB J.* **35**(6), 1–11 (2021).
12. Zaid, Y. *et al.* Chemokines and eicosanoids fuel the hyperinflammation within the lungs of patients with severe COVID-19. *J. Allergy Clin. Immunol.* **148**(2), 368–380 (2021).
13. Ferreira, A. C. *et al.* SARS-CoV-2 engages inflammasome and pyroptosis in human primary monocytes. *Cell Death Discovery* **7**(1), 1–12 (2021).
14. Laatifi, M. *et al.* Machine learning approaches in Covid-19 severity risk prediction in Morocco. *J. Big Data* **9**(1), 1–21 (2022).
15. He, L. *et al.* Expression of elevated levels of pro-inflammatory cytokines in SARS-CoV-infected ACE2+ cells in SARS patients: Relation to the acute lung injury and pathogenesis of SARS. *J. Pathol. J. Pathol. Soc. Great Br. Irel.* **210**(3), 288–297 (2006).
16. Onuk, S., Sipahioğlu, H., Karahan, S., Yeşiltepe, A., Kuzugüden, S., Karabulut, A., Akın, A. *et al.* Cytokine levels and severity of illness scoring systems to predict mortality in COVID-19 infection. In *Healthcare*, Vol. 11, No. 3, 387 (Multidisciplinary Digital Publishing Institute, 2023).
17. Kalinina, O. *et al.* Cytokine storm signature in patients with moderate and severe COVID-19. *Int. J. Mol. Sci.* **23**(16), 8879 (2022).
18. Ishay, Y. *et al.* A digital health platform for assisting the diagnosis and monitoring of COVID-19 progression: An adjuvant approach for augmenting the antiviral response and mitigating the immune-mediated target organ damage. *Biomed. Pharmacother.* **143**, 112228 (2021).
19. Ramatillah, D. L. *et al.* Impact of cytokine storm on severity of COVID-19 disease in a private hospital in West Jakarta prior to vaccination. *PLoS ONE* **17**(1), e0262438 (2022).
20. RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. *N. Engl. J. Med.* **384**(8), 693–704 (2021).
21. Tomazini, B. M. *et al.* COVID-19-associated ARDS treated with DEXamethasone (CoDEX): Study design and rationale for a randomized trial. *Rev. Bras. Ter. Intensiva* **32**, 354–362 (2020).
22. Tanaka, T. *et al.* T-705 (Favipiravir) suppresses tumor necrosis factor  $\alpha$  production in response to influenza virus infection: A beneficial feature of T-705 as an anti-influenza drug. *Acta Virol.* **61**(1), 48–55 (2017).
23. Lester, M., Sahin, A. & Pasyar, A. The use of dexamethasone in the treatment of COVID-19. *Ann. Med. Surg.* **56**, 218 (2020).
24. [https://www.recoverytrial.net/files/recovery\\_dexamethasone\\_statement\\_160620\\_v2final.pdf](https://www.recoverytrial.net/files/recovery_dexamethasone_statement_160620_v2final.pdf)
25. <https://www.recoverytrial.net/files/recovery-monoclonal-antibodies-press-release-final.pdf>
26. Tocilizumab reduces deaths in patients hospitalised with COVID-19
27. Dimopoulos, G. *et al.* Favorable anakinra responses in severe Covid-19 patients with secondary hemophagocytic lymphohistiocytosis. *Cell Host Microbe* **28**(1), 117–123 (2020).
28. Aouba, A. *et al.* Targeting the inflammatory cascade with anakinra in moderate to severe COVID-19 pneumonia: Case series. *Ann. Rheum. Dis.* **79**(10), 1381–1382 (2020).
29. Ozcicek, F., Kara, A. V., Akbas, E. M., Kurt, N., Yazici, G. N., Cankaya, M., & Suleyman, H. *et al.* Effects of anakinra on the small intestine mucositis induced by methotrexate in rats. *Exp. Anim.* 19-0057 (2019).
30. Sugiyama, K. *et al.* Differing effects of clarithromycin and azithromycin on cytokine production by murine dendritic cells. *Clin. Exp. Immunol.* **147**(3), 540–546 (2007).
31. Aghai, Z. H. *et al.* Azithromycin suppresses activation of nuclear factor-kappa B and synthesis of pro-inflammatory cytokines in tracheal aspirate cells from premature infants. *Pediatr. Res.* **62**(4), 483–488 (2007).
32. Tkalčević, V. I. *et al.* Anti-inflammatory activity of azithromycin attenuates the effects of lipopolysaccharide administration in mice. *Eur. J. Pharmacol.* **539**(1–2), 131–138 (2006).
33. Rahman, A., Kriak, J., Meyer, R., Goldblatt, S., & Rahman, F. A machine learning based modeling of the cytokine storm as it relates to COVID-19 using a virtual clinical semantic network (vCSN), in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3803–3810 (IEEE, 2020).
34. Ghazavi, A., Ganji, A., Keshavarzian, N., Rabiemajd, S. & Mosayebi, G. Cytokine profile and disease severity in patients with COVID-19. *Cytokine* **137**, 155323 (2021).
35. Gao, Z. *et al.* Machine-learning-assisted microfluidic nanoplasmonic digital immunoassay for cytokine storm profiling in COVID-19 patients. *ACS Nano* **15**(11), 18023–18036 (2021).
36. Patterson, B. K. *et al.* Immune-based prediction of COVID-19 severity and chronicity decoded using machine learning. *Front. Immunol.* **12**, 2520 (2021).
37. Cabaro, S. *et al.* Cytokine signature and COVID-19 prediction models in the two waves of pandemics. *Sci. Rep.* **11**(1), 1–11 (2021).
38. Liu, Q. Q. *et al.* Cytokines and their relationship with the severity and prognosis of coronavirus disease 2019 (COVID-19): A retrospective cohort study. *BMJ Open* **10**(11), e041471 (2020).
39. Khadem, H., Nemat, H., Eissa, M. R., Elliott, J. & Benaissa, M. COVID-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework. *Comput. Biol. Med.* **144**, 105361. <https://doi.org/10.1016/j.compbiomed.2022.105361> (2022).
40. Doshi-Velez, F., & Kim, B. Towards a rigorous science of interpretable machine learning. arXiv preprint [arXiv:1702.08608](https://arxiv.org/abs/1702.08608) (2017).
41. Ekanayake, I. U., Meddage, D. P. P. & Rathnayake, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud. Constr. Mater.* **16**, e01059 (2022).
42. Futagami, K., Fukazawa, Y., Kapoor, N. & Kito, T. Pairwise acquisition prediction with SHAP value interpretation. *J. Finance Data Sci.* **7**, 22–44 (2021).
43. Molnar, C. *Interpretable Machine Learning*. Lulu.com (2020).
44. Magesh, P. R., Myloth, R. D. & Tom, R. J. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput. Biol. Med.* **126**, 104041 (2020).
45. Deotare, U., Al-Dawsari, G., Couban, S. & Lipton, J. H. G-CSF-primed bone marrow as a source of stem cells for allografting: Revisiting the concept. *Bone Marrow Transpl.* **50**(9), 1150–1156 (2015).
46. Root, R. K. & Dale, D. C. Granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor: Comparisons and potential for use in the treatment of infections in nonneutropenic patients. *J. Infect. Dis.* **179**(Supplement\_2), S342–S352 (1999).
47. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006).

48. Higgins, I. *et al.* Beta-vae: Learning basic visual concepts with a constrained variational framework (2016).
49. Kingma, D. P., Max, W. Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) (2013).
50. Pol, A. A. *et al.* Anomaly detection with conditional variational autoencoders, in *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)* (IEEE, 2019).
51. Monserrat, J. *et al.* Role of innate and adaptive cytokines in the survival of COVID-19 patients. *Int. J. Mol. Sci.* **23**(18), 10344 (2022).
52. Mendy, J. F. (2019). *Analysis of Ex Vivo Host Biomarkers in Sputum Samples for Diagnosis of Pulmonary Tuberculosis* (Doctoral dissertation, Stellenbosch: Stellenbosch University).
53. Que, Y. *et al.* Cytokine release syndrome in COVID-19: A major mechanism of morbidity and mortality. *Int. Rev. Immunol.* **41**(2), 217–230 (2022).
54. Sanz, J. M., Gómez Lahoz, A. M. & Martín, R. O. Role of the immune system in SARS-CoV-2 infection: Immunopathology of COVID-19. *Medicine (Madr)* **13**(33), 1917–1931 (2021).
55. Stekhoven, D. J. & Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**(1), 112–118 (2012).
56. Mungkasi, S., & Dong, Z. Y. in *The 6th International Conference on Computer Science and Computational Mathematics (ICCSM)*, 2017).
57. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **493**, 28–45 (2022).
58. Plesovskaya, E. & Ivanov, S. An empirical analysis of KDE-based generative models on small datasets. *Procedia Comput. Sci.* **193**, 442–452 (2021).
59. Hernandez-Matamoros, A., Fujita, H. & Perez-Meana, H. A novel approach to create synthetic biomedical signals using BiRNN. *Inf. Sci.* **541**, 218–241 (2020).
60. Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Nakayama, H. *et al.* GAN-based synthetic brain MR image generation, in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 734–738 (IEEE, 2018).
61. Guan, J., Li, R., Yu, S., & Zhang, X. Generation of synthetic electronic medical record text, in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 374–380 (IEEE, 2018).
62. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. Modeling tabular data using conditional gan. *Adv. Neural Inform. Process. Syst.* **32** (2019).
63. Kellner, L., Stender, M., Polach, F. V. B. & Ehlers, S. Predicting compressive strength and behavior of ice and analyzing feature importance with explainable machine learning models. *Ocean Eng.* **255**, 111396 (2022).
64. Romero Starke, K. *et al.* The age-related risk of severe outcomes due to COVID-19 infection: A rapid review, meta-analysis, and meta-regression. *Int. J. Environ. Res. Public Health* **17**(16), 5974 (2020).
65. Baggiolini, M., Dewald, B. & Moser, B. Interleukin-8 and related chemotactic cytokines—CXC and CC chemokines. *Adv. Immunol.* **55**, 97–179 (1993).
66. Yang, Y. *et al.* Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. *J. Allergy Clin. Immunol.* **146**(1), 119–127 (2020).
67. Zhang, R. *et al.* COVID-19: Melatonin as a potential adjuvant treatment. *Life Sci.* **250**, 117583 (2020).
68. Alkharsah, K. R. VEGF upregulation in viral infections and its possible therapeutic implications. *Int. J. Mol. Sci.* **19**(6), 1642 (2018).
69. Jamilloux, Y. *et al.* Should we stimulate or suppress immune responses in COVID-19? Cytokine and anti-cytokine interventions. *Autoimmun. Rev.* **19**(7), 102567 (2020).
70. Honore, P. M. *et al.* Inhibiting IL-6 in COVID-19: We are not sure. *Crit. Care* **24**(1), 1–3 (2020).
71. Orlov, M., Wander, P. L., Morrell, E. D., Mikacenic, C. & Wurfel, M. M. A case for targeting Th17 cells and IL-17A in SARS-CoV-2 infections. *J. Immunol.* **205**(4), 892–898 (2020).
72. McManus, C. M., Brosnan, C. F. & Berman, J. W. Cytokine induction of MIP-1 $\alpha$  and MIP-1 $\beta$  in human fetal microglia. *J. Immunol.* **160**(3), 1449–1455 (1998).
73. Zaid, Y. *et al.* Platelets can associate with SARS-Cov-2 RNA and are hyperactivated in COVID-19. *Circ. Res.* **127**(11), 1404–1418 (2020).
74. Heimfarth, L., Serafini, M. R., Martins-Filho, P. R., Quintans, J. D. S. S. & Quintans-Junior, L. J. Drug repurposing and cytokine management in response to COVID-19: A review. *Int. Immunopharmacol.* **88**, 106947 (2020).
75. Krzysiek, R. *et al.* Antigen receptor engagement selectively induces macrophage inflammatory protein-1 $\alpha$  (MIP-1 $\alpha$ ) and MIP-1 $\beta$  chemokine production in human B cells. *J. Immunol.* **162**(8), 4455–4463 (1999).
76. Sheahan, T. *et al.* MyD88 is required for protection from lethal infection with a mouse-adapted SARS-CoV. *PLoS Pathog.* **4**(12), e1000240 (2008).
77. Tamayo-Velasco, Á. *et al.* HGF, IL-1 $\alpha$ , and IL-27 are robust biomarkers in early severity stratification of COVID-19 patients. *J. Clin. Med.* **10**(9), 2017 (2021).
78. Burgos-Blasco, B. *et al.* Hypercytokinemia in COVID-19: tear cytokine profile in hospitalized COVID-19 patients. *Exp. Eye Res.* **200**, 108253 (2020).
79. Liao, H. H. *et al.* Down-regulation of granulocyte-macrophage colony-stimulating factor by 3C-like proteinase in transfected A549 human lung carcinoma cells. *BMC Immunol.* **12**(1), 1–9 (2011).
80. Yendo, T. M. *et al.* Impact of inflammatory immune dysfunction in psoriasis patients at risk for COVID-19. *Vaccines* **9**(5), 478 (2021).
81. Darden, D. B., Hawkins, R. B., Larson, S. D., Iovine, N. M., Prough, D. S., & Efron, P. A. The clinical presentation and immunology of viral pneumonia and implications for management of coronavirus disease 2019. *Crit. Care Explor.* **2**(4) (2020).
82. Blanco-Melo, D. *et al.* Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**(5), 1036–1045 (2020).
83. Trombetta, A. C. *et al.* Severe COVID-19 recovery is associated with timely acquisition of a myeloid cell immune-regulatory phenotype. *Front. Immunol.* **12**, 2346 (2021).
84. Lee, J. *et al.* IL-17E, a novel proinflammatory ligand for the IL-17 receptor homolog IL-17Rh1. *J. Biol. Chem.* **276**(2), 1660–1664 (2001).

### Author contributions

M.L., H.E., Z.Y., N.A., M.N.: Collection of data, Data analysis and interpretation, Writing the article, reviewed the manuscript, Final approval of the article. S.D., C.E.A., A.B.: Research concept and design, development of Models, Writing the article, reviewed the manuscript, Final approval of the article.

### Competing interests

The authors declare no competing interests.



### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-31542-7>.

**Correspondence** and requests for materials should be addressed to S.D.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023