






Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task

Han Wu , Wenjie Ruan , *Member, IEEE*, Jiangtao Wang , Dingchang Zheng , *Member, IEEE*, Bei Liu, Yayuan Geng, Xiangfei Chai, Jian Chen, Kunwei Li, Shaolin Li, and Sumi Helal , *Fellow, IEEE*

Abstract—The black-box nature of machine learning models hinders the deployment of some high-accuracy medical diagnosis algorithms. It is risky to put one's life in the hands of models that medical researchers do not fully understand or trust. However, through model interpretation, black-box models can promptly reveal significant biomarkers that medical practitioners may have overlooked due to the surge of infected patients in the COVID-19 pandemic. This research leverages a database of 92 patients with confirmed SARS-CoV-2 laboratory tests between 18th January 2020 and 5th March 2020, in Zhuhai, China, to identify biomarkers indicative of infection severity prediction. Through the interpretation of four machine learning models, decision tree, random forests, gradient boosted trees, and neural networks using permutation feature importance, partial dependence plot, individual conditional expectation, accumulated local effects, local interpretable model-agnostic explanations, and Shapley additive explanation, we identify an increase in N-terminal pro-brain natriuretic peptide, C-reaction protein, and lactic dehydrogenase, a decrease in lymphocyte is associated with severe infection and an increased risk of death, which is consistent with recent medical research on COVID-19 and other research using dedicated models. We further validate our methods on a large open dataset with 5644 confirmed patients from the Hospital Israelita Albert Einstein, at São Paulo, Brazil

from Kaggle, and unveil leukocytes, eosinophils, and platelets as three indicative biomarkers for COVID-19.

Impact Statement—The pandemic is a race against time. We seek to answer the question, how can medical practitioners employ machine learning to win the race in the pandemic? Instead of targeting at a high-accuracy black-box model that is difficult to trust and deploy, we use model interpretation that incorporates medical practitioners' prior knowledge to promptly reveal the most important indicators in early diagnosis, and thus, win the race in the pandemic.

Index Terms—Artificial intelligence in health, artificial intelligence in medicine, interpretable machine learning.

I. INTRODUCTION

THE sudden outbreak of COVID-19 has caused an unprecedented disruption and impact worldwide. With more than 100 million confirmed cases as of February 2021, the pandemic is still accelerating globally. The disease is transmitted by inhalation or contact with infected droplets with an incubation period ranging from 2 to 14 days [1], making it highly infectious and difficult to contain and mitigate.

With the rapid transmission of COVID-19, the demand for medical supplies goes beyond hospitals' capacity in many countries. Various diagnostic and predictive models are employed to release the pressure on healthcare workers. For instance, a deep learning model that detects abnormalities and extract key features of the altered lung parenchyma using chest CT images is proposed [2]. On the other hand, Rich Caruana *et al.* [3] exploit intelligible models that use generalized additive models with pairwise interactions to predict the probability of readmission. To maintain both interpretability and complexity, DeepCOVIDNet is present to achieve predictive surveillance that identifies the most influential features for the prediction of the growth of the pandemic [4] through the combination of two modules. The embedding module takes various heterogeneous feature groups as input and outputs an equidimensional embedding corresponding to each feature group. The DeepFM [5] module computes second and higher order interactions between them.

Models that achieves high accuracy provide fewer interpretations due to the tradeoff between accuracy and interpretability [6]. To be adopted in healthcare systems that require both interpretability and robustness [7], the multitree XGBoost algorithm is employed to identify the most significant indicators in COVID-19 diagnosis [8]. This method exploits the recursive

Manuscript received 12 November 2020; revised 7 February 2021 and 30 April 2021; accepted 8 June 2021. Date of publication 25 June 2021; date of current version 21 July 2023. This work was supported in part by HY Medical Technology, Scientific Research Department, Beijing, CN. The work of Han Wu and Wenjie Ruan was supported by Offshore Robotics for Certification of Assets (ORCA) Partnership Resource Fund (PRF) on Towards the Accountable and Explainable Learning-Enabled Autonomous Robotic Systems (AELARS) under Grant EP/R026173/1. The authors would like to thank the anonymous reviewers and Associate Editor of this manuscript for their comments to help us strengthen this work. This paper was recommended for publication by Associate Editor D. Zhao upon evaluation of the reviewers comments. (*Corresponding author: Wenjie Ruan; Jiangtao Wang; and Shaolin Li.*)

Han Wu and Wenjie Ruan are with the University of Exeter, EX4 4PY Exeter, U.K. (e-mail: hw630@exeter.ac.uk; W.Ruan@exeter.ac.uk).

Jiangtao Wang and Dingchang Zheng are with the Coventry University, CV1 5FB Coventry, U.K. (e-mail: jiangtao.wang@coventry.ac.uk; ad4291@coventry.ac.uk).

Bei Liu is with the Department of Gastroenterology, 910 Hospital of PLA, Beijing, China (e-mail: liubei0927@outlook.com).

Yayuan Geng and Xiangfei Chai are with the Scientific Research Department Beijing, HY Medical Technology, Beijing 100192, China (e-mail: gengyayuan@huiyihuiying.com; chaixiangfei@huiyihuiying.com).

Jian Chen and Kunwei Li are with the Department of Radiology, Hospital of Sun Yat-sen University, Zhuhai 519000, China (e-mail: drchenj@126.com; likunwei@mail.sysu.edu.cn).

Shaolin Li is with the Department of Radiology, and Guangdong Provincial Key Laboratory of Biomedical Imaging, Hospital of Sun Yat-sen University, Zhuhai 519000, China (e-mail: lishlin5@mail.sysu.edu.cn).

Sumi Helal is with the University of Florida, Gainesville, FL 32611 USA (e-mail: helal@acm.org).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TAI.2021.3092698>.

Digital Object Identifier 10.1109/TAI.2021.3092698

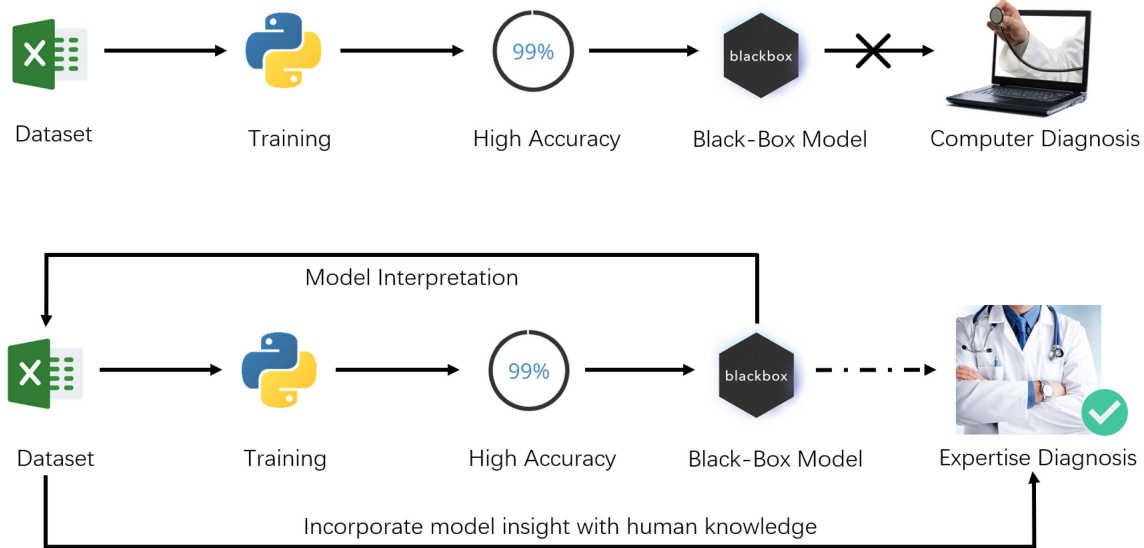


Fig. 1. Difference between the usual workflow of machine learning, and our approach.

tree-based decision system of the model to achieve high interpretability. On the other hand, a more complex convolutional neural network (CNN) model can discriminate COVID-19 from non-COVID-19 using chest CT image [9]. It achieves interpretability through gradient-weighted class activation mapping to produce a heat map that visually verifies where the CNN model is focusing.

Besides, several model-agnostic methods have been proposed to peek into black-box models, such as partial dependence plot (PDP) [10], individual conditional expectation (ICE) [11], accumulated local effects (ALE) [12], permutation feature importance [13], local interpretable model-agnostic explanations (LIME) [14], Shapley additive explanation (SHAP) [15], and anchors [16]. Most of these model-agnostic methods are reasoned qualitatively through illustrative figures and human experiences. To quantitatively measure their interpretability, metrics such as faithfulness [17] and monotonicity [18] are proposed.

In this article, instead of targeting a high-accuracy model, we interpret several models to help medical practitioners promptly discover the most significant biomarkers in the pandemic, as illustrated in Fig. 1.

Overall, this article makes the following contributions.

- 1) *Evaluation*: A systematic evaluation of the interpretability of machine learning models that predict the severity level of COVID-19 patients. We experiment with six interpretation methods and two evaluation metrics on our dataset and receive the same result as research that uses a dedicated model. We further validate our approach on a dataset from Kaggle.
- 2) *Implication*: Through the interpretation of models trained on our dataset, we reveal N-terminal probrain natriuretic peptide (NTproBNP), C-reaction protein (CRP), lactic dehydrogenase (LDH), and lymphocyte (LYM) as the most indicative biomarkers in identifying patients' severity level. Applying the same approach on the Kaggle dataset,

we further unveil three significant features, leukocytes, eosinophils, and platelets.

- 3) *Implementation*: We design a system that healthcare professionals can interact with its AI Models to incorporate model insights with medical knowledge. We release our implementation, models for future research and validation.¹

II. PRELIMINARY OF AI INTERPRETABILITY

In this section, six interpretation methods, partial dependence plot, individual conditional expectation, accumulated local effects, local interpretable model-agnostic explanations, and Shapley additive explanation are summarized. We also summarize two evaluation metrics, faithfulness, and monotonicity.

A. Model-Agnostic Methods

In healthcare, restrictions to using only interpretable models bring many limitations in adoption while separating explanations from the model can afford several beneficial flexibilities [19]. As a result, model-agnostic methods have been devised to provide interpretations without knowing model details.

- 1) *Partial Dependence Plot*: PDPs reveal the dependence between the target function and several target features. The partial function $f_{x_s}(x_s)$ is estimated by calculating averages in the training data, also known as the Monte Carlo method. After setting up a grid for the features we are interested in (target features), we set all target features in our training set to be the value of grid points, then make predictions and average them all at each grid. The drawback of PDP is that one target feature produces 2D plots and two produce 3D plots while it can be pretty hard

¹Our source code and models are available at <https://github.com/wuhanstudio/interpretable-ml-covid-19>.

for a human to understand plots in higher dimensions

$$\hat{f}_{x_s}(x_s) = \frac{1}{n} \sum_1^n \hat{f}(x_s, x_c^i). \quad (1)$$

- 2) *Individual Conditional Expectation*: ICE is similar to PDP. The difference is that PDP calculates the average over the marginal distribution while ICE keeps them all. Each line in the ICE plot represents predictions for each individual. Without averaging on all instances, ICE unveils heterogeneous relationships but is limited to only one target feature since two features result in overlay surfaces that cannot be identified by human eyes [20].
- 3) *Accumulated Local Effects*: ALE averages the changes in the predictions and accumulate them over the local grid. The difference with PDP is that the value at each point of the ALE curve is the difference to the mean prediction calculated in a small window rather than all of the grid. Thus ALE eliminates the effect of correlated features [20] which makes it more suitable in healthcare because it is usually irrational to assume young people having similar physical conditions with the elderly.
- 4) *Permutation Feature Importance*: The idea behind permutation feature importance is intuitive. A feature is significant for the model if there is a noticeable increase in the model's prediction error after permutation. On the other hand, the feature is less important if the prediction error remains nearly unchanged after shuffling.
- 5) *Local Interpretable Model-Agnostic Explanations*: LIME uses interpretable models to approximate the predictions of the original black-box model in specific regions. LIME works for tabular data, text, and images, but the explanations may not be stable enough for medical applications.
- 6) *Shapley Additive Explanation*: SHAP borrows the idea of Shapley value from game theory [21], which represents contributions of each player in a game. Calculating Shapley values is computationally expensive when there are hundreds of features, thus Lundberg and Lee [15] proposed a fast implementation for tree-based models to boost the calculation process. SHAP has a solid theoretical foundation but is still computationally slow for a lot of instances.

To summarize, PDP, ICE, and ALE only use graphs to visualize the impact of different features while permutation feature importance, LIME, and SHAP provide numerical feature importance that quantitatively ranks the importance of each feature.

B. Metrics for Interpretability Evaluation

Different interpretation methods try to find out the most important features to provide explanations for the output. But as Doshi-Velez and Kim [6] questioned, "Are all models in all defined-to-be-interpretable model classes equally interpretable?" And how can we measure the quality of different interpretation methods?

Faithfulness: Faithfulness incrementally removes each of the attributes deemed important by the interpretability metric, and evaluate the effect on the performance. Then it calculates the

correlation between the weights (importance) of the attributes and corresponding model performance and returns correlation between attribute importance weights and the corresponding effect on classifier [17].

Monotonicity: Monotonicity incrementally adds each attribute in order of increasing importance. As each feature is added, the performance of the model should correspondingly increase, thereby resulting in monotonically increasing model performance, and it returns true or false [18].

In our experiment, both faithfulness and monotonicity are employed to evaluate the interpretation of different machine learning models.

III. EMPIRICAL STUDY ON COVID

In this section, features in our raw dataset and procedures of data preprocessing are introduced. After preprocessing, four different models: Decision tree, random forest, gradient boosted trees, and neural networks are trained on the dataset. Model interpretation is then employed to understand how different models make predictions, and patients that models make false diagnoses are investigated respectively.

A. Dataset and Preprocessing

The raw dataset consists of patients with confirmed SARS-CoV-2 laboratory tests between 18th January 2020 and 5th March 2020, in Zhuhai, China. Our Research Ethics Committee waived written informed consent for this retrospective study that evaluated deidentified data and involved no potential risk to patients. All the data of patients have been anonymized before analysis.

Tables in the Appendix list all 74 features in the raw dataset consisting of body mass index (BMI), complete blood count (CBC), blood biochemical examination, inflammatory markers, symptoms, anamneses, among others. Whether or not health care professionals will order a test for patients is based on various factors such as medical history, physical examination, and etc. Thus, there is no standard set of tests that are compulsory for every individual which introduces data sparsity. For instance, left ventricular ejection fraction (LVEF) are mostly empty because most patients are not required to take the color doppler ultrasound test.

After pruning out irrelevant features, such as patients' medical numbers that provide no medical information, and features that have no patients' records (no patient took this test), 86 patients' records with 55 features are selected for further investigation. Among those, 77 records are used for training, cross-validation, and 9 reserved for testing. The feature for classification is Severity01 which indicates normal with 0, and severe with 1. More detailed descriptions about features in our dataset are listed in the Appendix.

Feature engineering is applied before training and interpreting our models, as some features may not provide valuable information or provide redundant information.

First, constant and quasi-constant features were removed. For instance, the two features, PCT2 and Stomachache, have the

TABLE I
FEATURE CORRELATION

Feature 1	Feature 2	Correlation
cTnICKMBOrdinal1	cTnICKMBOrdinal2	0.853741
LDH	HBDH	0.911419
NEU2	WBC2	0.911419
LYM2	LYM1	0.842688
NTproBNP	N2L2	0.808767
BMI	Weight	0.842409
NEU1	WBC1	0.90352

same value for all patients providing no valuable information in distinguishing normal and severe patients.

Second, correlated features were removed because they provide redundant information. Table I lists all correlated features using Pearson's correlation coefficient.

- 1) There is strong correlation between cTnICKMBOrdinal1 and cTnICKMBOrdinal2 because they are the same test among a short range of time which is the same for LYM1 and LYM2.
- 2) LDH and HBDH levels are significantly correlated with heart diseases, and the HBDH/LDH ratio can be calculated to differentiate between liver and heart diseases.
- 3) Neutrophils (NEU1/NEU2) are all correlated to the immune system. In fact, most of the white blood cells that lead the immune system's response are neutrophils. Thus, there is a strong correlation between NEU1 and WBC1, NEU2 and WBC2.
- 4) In the original dataset, there is no much information about N2L2 which is correlated with NTproBNP, thus NTproBNP remains.
- 5) The correlation between BMI and weight is straight forward because BMI is a person's weight in kilograms divided by the square of height in meters.

Third, statistical methods that calculate mutual information are employed to remove features with redundant information.

Mutual information is calculated using (2) that determines how similar the joint distribution $p(X, Y)$ is to the products of individual distributions $p(X)p(Y)$. Univariate test measures the dependence of two variables, and a high p-value indicates a less similar distribution between X and Y .

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \quad (2)$$

After feature engineering, there are 37 features left for training and testing.

B. Training Models

Machine learning models outperform humans in many different areas in terms of accuracy. Interpretable models such as the decision tree are easy to understand, but not suitable for large scale applications. Complex models achieve high accuracy while giving less explanation.

For healthcare applications, both accuracy and interpretability are significant. Four different models are selected to extract

TABLE II
FEATURES WITH MUTUAL INFORMATION

Statistical Methods	Removed Features
Mutual Information Univariate	Height, CK, HiCKMB, Cr, WBC1, Hemoptysis Weight, AST, CKMB, PCT1, WBC2

TABLE III
STRUCTURE OF NEURAL NETWORKS

Layer Type	Output Shape	Param
Dense	(None, 10)	370
Dropout	(None, 10)	0
Dense	(None, 15)	165
Dropout	(None, 15)	0
Dense	(None, 5)	80
Dropout	(None, 5)	0
Dense	(None, 1)	6

information from our dataset: Decision tree, random forests, gradient boosted trees, and neural networks.

Decision Tree: DT is a widely adopted method for both classification and regression. It's a nonparametric supervised learning method that infers decision rules from data features. The decision tree try to find decision rules that make the best split measured by Gini impurity or entropy. More importantly, the generated decision tree can be visualized, thus easy to understand and interpret [22].

Random Forest: RF is a kind of ensemble learning method [23] that employs bagging strategy. Multiple decision trees are trained using the same learning algorithm, and then predictions are aggregated from the individual decision tree. Random forests produce great results most of the time even without much hyperparameter tuning. As a result, it has been widely accepted for its simplicity and good performance. However, it is rather difficult for humans to interpret hundreds of decision trees, so the model itself is less interpretable than a single decision tree.

Gradient Boosted Trees: Gradient boosted trees is another ensemble learning method that employs boosting strategy [24]. Through sequentially adding one decision tree at one time, gradient boosted trees combine results along the way. With fine-tuned parameters, gradient boosting can result in better performance than random forests. Still, it is tough for humans to interpret a sequence of decision trees, and thus, considered as black-box models.

Neural Networks: Neural networks could be the most promising model in achieving a high accuracy and even outperforms humans in medical imaging [25]. Though the whole network is difficult to understand, deep neural networks are stacks of simple layers, thus can be partially understood through visualizing outputs of intermediate layers [26].

As for the implementation, there is no hyperparameter for the decision tree. For random forests, 100 trees are used during the initialization. The hyperparameters for gradient boosted trees are selected according to prior experience. The structure for neural networks is listed in Table III. All these methods are implemented using scikit-learn [27], Keras, and python3.6.

TABLE IV
CLASSIFICATION RESULTS ON OUR DATASET

Classifier	CV		Test Set		95% confidence interval
	F1	Precision	Recall	F1	
Decision Tree	0.55	0.67	0.50	0.57	0.31
Random Forest	0.62	0.67	0.50	0.57	0.31
Gradient Boosted Trees	0.67	0.78	1.00	0.80	0.27
Neural Networks	0.58	0.78	1.00	0.80	0.27

TABLE V
FIVE MOST IMPORTANT FEATURES

Model	Most Important Features
Decision Tree	NTproBNP, CRP2, ALB2, Temp, Symptom
Random Forest	CRP2, NTproBNP, cTnI, LYM1, ALB2
Gradient Boosted Trees	CRP2, cTnITimes, LYM1, NTproBNP, Phlegm
Neural Networks	NTproBNP, CRP2, CRP1, LDH, Age

After training, gradient boosted trees and neural networks achieve the highest precision on the test set. Among 9 patients in our test set, four of them are severe. Both the decision tree and random forests fail to identify two severe patients, while gradient boosted trees and neural networks find all of the severe patients.

C. Interpretation (Permutation Feature Importance)

First, we use permutation feature importance to find the most important features in different models. In Table V, CRP2 and NTproBNP are recognized as most important features by most models.

According to medical knowledge, CRP, which increases when there's inflammation or viral infection in the body. CRP levels are positively correlated with lung lesions and could reflect disease severity [28]. NTproBNP refers to N-terminal prohormone of brain natriuretic peptide, which will be released in response to changes in pressure inside the heart. The CRP level in severe patients rises due to viral infection, and patients with higher NT-proBNP (above 88.64 pg/mL) level had more risks of in-hospital death [29].

D. Interpretation (PDP, ICE, ALE)

After recognizing the most important features, PDP, ICE, and ALE are employed to further visualize the relationship between CRP and NTproBNP.

In the PDPs, all of the four models indicate a higher risk of turning severe with the increase of NTproBNP and CRP which is consistent with the retrospective study on COVID-19, as depicted in Fig. 2. The difference is that different models have different tolerances and dependence on NTproBNP and CRP. Averagely, the decision tree has less tolerance on a high level of NTproBNP (>2000 ng/ml), and gradient boosted trees give a much higher probability of death as CRP increases. Since PDPs only calculate an average of all instances, we use ICEs to identify heterogeneity as illustrated in Fig. 3.

ICE reveals individual differences. Though all of the models give a prediction of a higher risk of severe as NTproBNP and CRP increase, some patients have a much higher initial probability which indicates other features have an impact on overall predictions. For example, elderly people have higher NTproBNP than young people and have a higher risk of turning severe as illustrated in Fig. 4.

In the ALEs, as NTproBNP and CRP get higher, all of the four models give a more positive prediction of turning severe, which coincides with medical knowledge.

E. Misclassified Patients

Even though the most important features revealed by our models exhibit medical meaning, some severe patients fail to be recognized. Both gradient boosted trees and neural networks recognize all severe patients and yield a recall of 1.00, while the decision tree and random forests fail to reveal two of them.

Patient No. 2 (normal) is predicted with a probability of 0.53 of turning severe which is around the boundary (0.5). While for patient No. 5 (severe), the model gives a relatively low probability of turning severe (0.24).

F. Interpretation (False Negative)

Suppose different models represent different doctors, then the decision tree and random forests make the wrong diagnosis for patient no. 5. The reason human doctors classified the patient as severe is that he actually needed a respirator to survive. To further investigate why the decision tree and random forests make wrong predictions, LIME, and SHAP are employed.

LIME: Features in green have a positive contribution to the prediction (increasing the probability of turning severe), and features in red have a negative effect on the prediction (decreasing the probability of turning severe).

SHAP: Features pushing the prediction to be higher (severe) are shown in red, and those pushing the prediction to be lower (normal) are in blue.

1) *Wrong Diagnoses*: Take the decision tree as an example, in the Fig. 5(a), the explanation by LIME illustrates that NTproBNP and CRP are two features (in green) that have a positive impact on the probability of turning severe. Even though patient No. 5 is indeed severe, the decision tree gives an overall prediction of normal (false negative). Thus, we would like to investigate features that have a negative impact on the probability of turning severe.

In the Fig. 6(c), the explanation by SHAP reveals that the patient is diagnosed as normal by the decision tree because the patient has no symptom. Even though the patient has a high NTproBNP and CRP, having no symptom makes it less likely to classify him as severe. The record was taken when the patient came to the hospital for the first time. It is likely that the patient developed symptoms later and turned severe.

However, both gradient boosted trees and neural networks are not deceived by the fact the patient has no symptom. Their predictions indicate that the patient is likely to turn severe in the future.

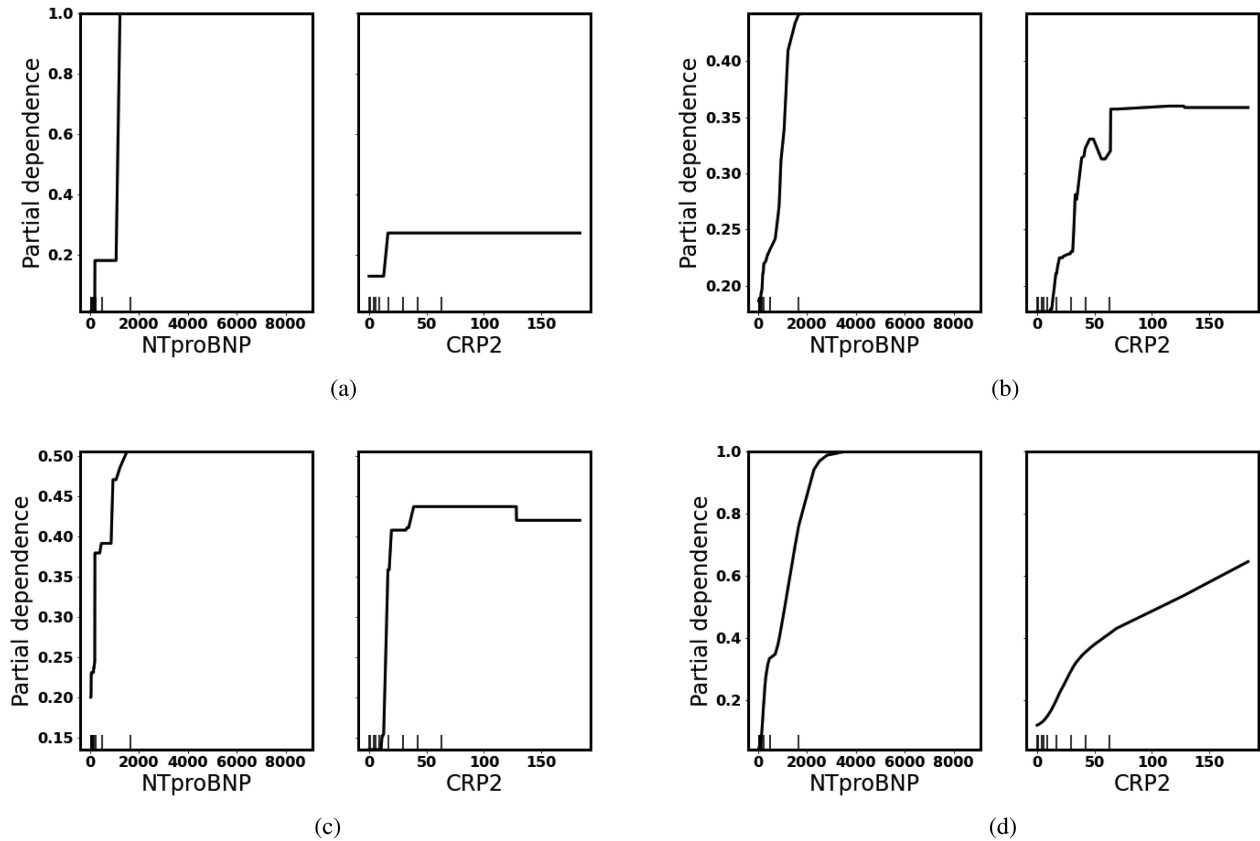


Fig. 2. Partial dependence plot: There is a positive correlation between the level of NTproBNP/CRP and the probability of turning severe because as NTproBNP/CRP increases, the average possibility (y-axis) of turning severe increases. (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

2) *Correct Diagnoses*: In the Fig. 6(c) and (d), gradient boosted trees and neural networks do not prioritize the feature symptom. They put more weight on test results (NTproBNP and CRP). Thus they make correct predictions based on the fact that the patient's test results are serious.

Besides, neural networks notice that the patient is elderly (Age = 63). If we calculate the average age in different severity levels, it is noticeable that elderly people are more likely to deteriorate.

Gradient boosted trees and neural networks make correct predictions because they trust more in test results, while the decision tree relies more on whether or not a patient has symptoms. As a result, gradient boosted trees and neural networks are capable of recognizing patients that are likely to turn severe in the future while the decision tree makes predictions relying more on patients' current situation.

Medical research is a case-by-case study. Every patient is unique. It is strenuous to find a single criterion that suits every patient, thus it's important to focus on each patient and make a diagnosis accordingly. This is one of the benefits of using interpretable machine learning. It unveils the most significant features for most patients and provides the interpretation for each patient as well.

TABLE VI
MISCLASSIFIED PATIENTS

No	Class	Probability of Severe	Prediction	Type
2	Normal	0.53	Severe	False Positive
5	Severe	0.24	Normal	False Negative

TABLE VII
AVERAGE AGE IN DIFFERENT SEVERITY LEVELS

Severity Level	Average Age
0	36.83
1	47.45
2	54.31
3	69.40

G. Interpretation (False Positive)

With limited medical resources at the initial outbreak of the pandemic, it is equally important to investigate false positive, so that valuable resources can be distributed to patients in need.

In Table VI, patient 2 is normal, but all of our models diagnose the patient as severe. To further explain the false positive prediction, Table VIII lists anonymized medical records for patient 2 (normal) and patient 5 (severe) for comparison.

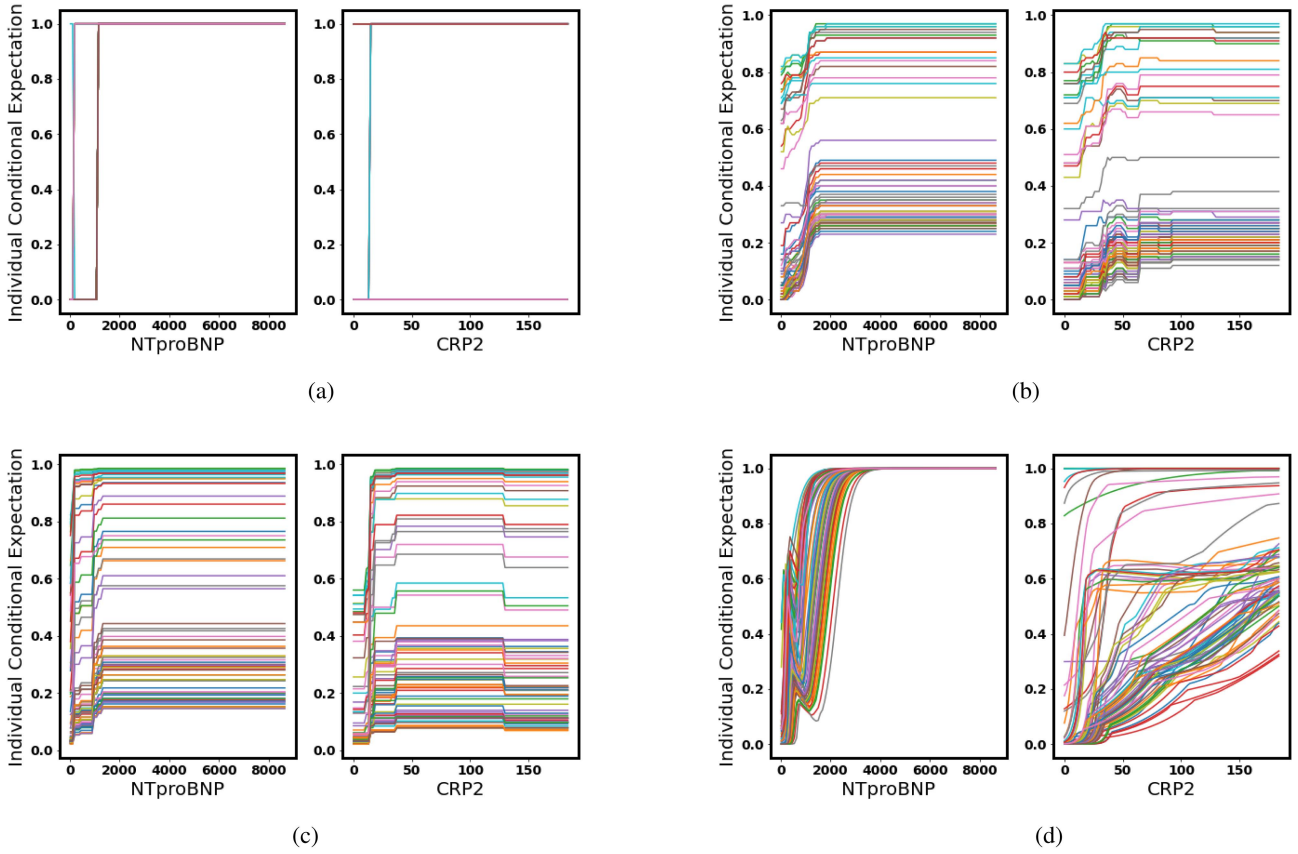


Fig. 3. Individual conditional expectation: Each line in different colors represents a patient. As we increase NTproBNP/CRP while keeping other features the same, the probability of turning severe increases for each individual, but each patient has a different starting level because their other physical conditions differ. (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

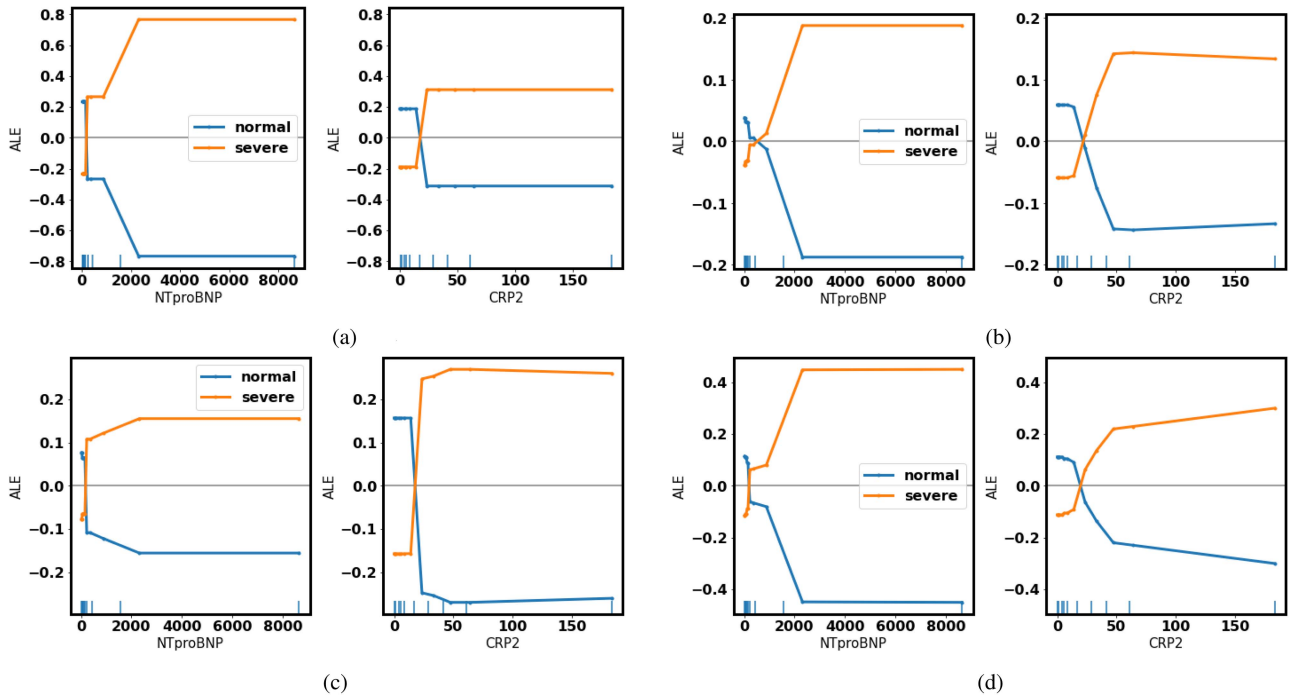


Fig. 4. Accumulated local effects: As the level of NTproBNP/CRP increases, the possibility of turning severe (yellow) goes above the average. (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

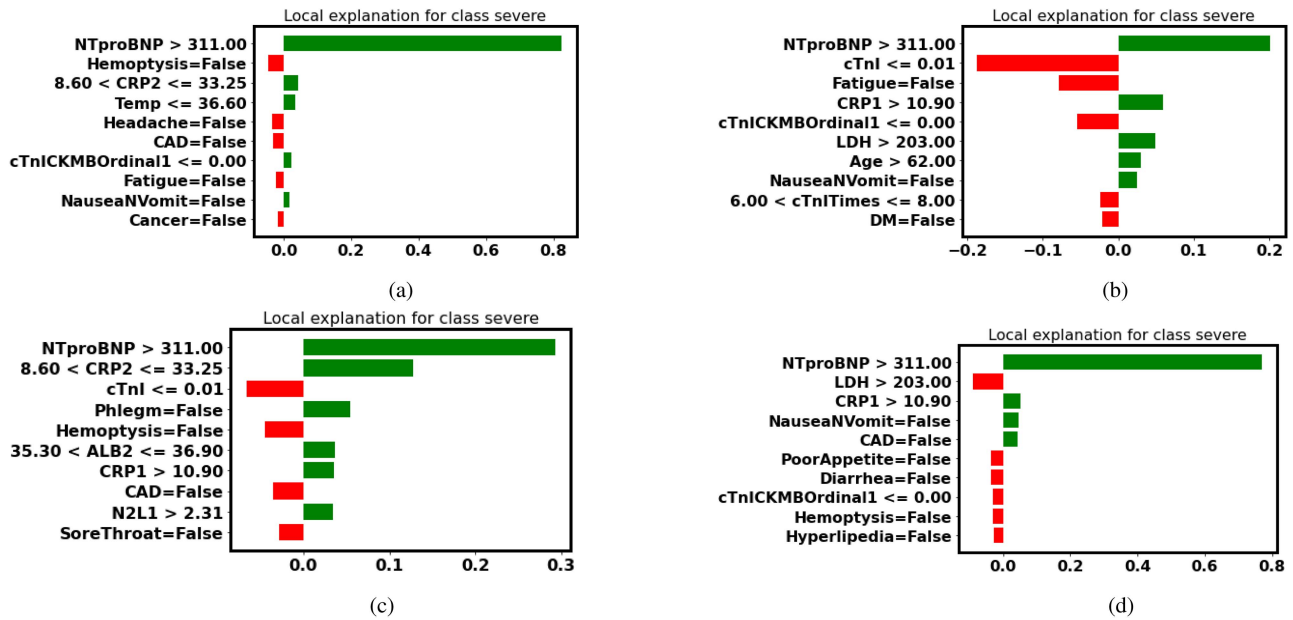


Fig. 5. LIME explanation (false-negative patient No. 5): Features in green have a positive contribution to the prediction (increasing the probability of turning severe), and features in red have a negative effect on the prediction (decreasing the probability of turning severe). (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

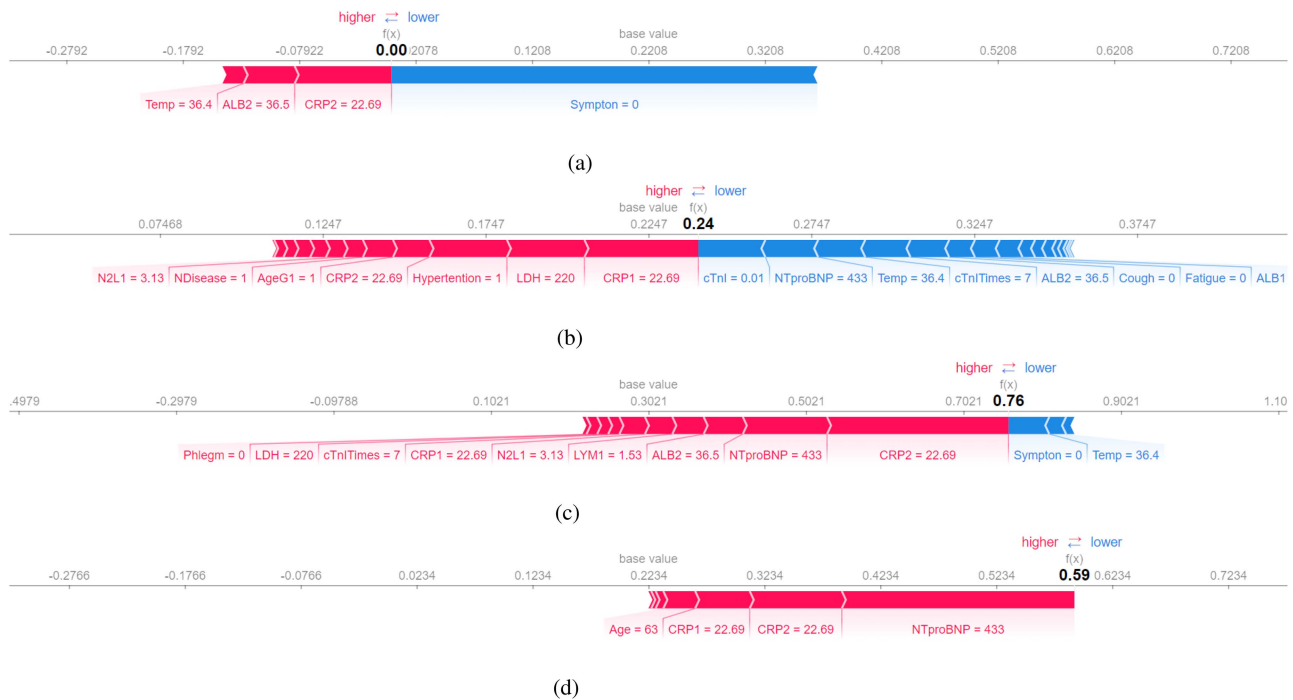


Fig. 6. SHAP explanation (false-negative patient No.5): Features pushing the prediction to be higher (severe) are shown in red, and those pushing the prediction to be lower (normal) are in blue. (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

1) *Doctors' Diagnoses:* We present the test results of both patients to doctors without indicating which patient is severe. All doctors mark patient No. 2 as more severe which is the same as our models. Doctors' decisions are based on the COVID-19 diagnosis and treatment guide in China. The increased level in

CRP, LDH, decreased level in LYM are associated with severe COVID-19 infection in the guideline, and patient 2 has a higher level of CRP and LDH, a lower level of LYM than patient 5. As a result, doctors' diagnoses are consistent with models' predictions

TABLE VIII
RECORD OF THE FALSE POSITIVE PATIENT 2

Feature	Patient 5 (Severe)	Patient 2 (Normal)
Sex	1.00	1.00
Age	63.00	42.00
AgeG1	1.00	0.00
Temp	36.40	37.50
cTnITimes	7.00	8.00
cTnI	0.01	0.01
cTnICKMBOrdinalI	0.00	0.00
LDH	220.00	263.00
NTproBNP	433.00	475.00
LYM1	1.53	1.08
N2L1	3.13	2.16
CRP1	22.69	36.49
ALB1	39.20	37.60
CRP2	22.69	78.76
ALB2	36.50	37.60
Symptoms	None	Fever
NDisease	Hypertention	Hypertention, DM, Hyperlipedia

TABLE IX
MOST IMPORTANT FEATURES FROM LIME AND SHAP

Model	LIME
Decision Tree	NTproBNP, CRP2, NauseaNVomit
Random Forest	NTproBNP, CRP2, CRP1
Gradient Boosted Trees	NTproBNP, CRP2, LYM1
Neural Networks	NTproBNP, CRP2, PoorAppetite
Model	SHAP
Decision Tree	CRP2, NTproBNP, ALB2
Random Forests	CRP2, CRP1, LDH
Gradient Boosted Trees	CRP2, NTproBNP, LDH
Neural Networks	CRP2, NTproBNP, CRP1

TABLE X
FAITHFULNESS EVALUATION

Models	LIME	SHAP
Random Forests	0.37	0.59
Gradient Boosted Trees	0.46	0.49
Neural Networks	0.45	0.33

TABLE XI
MONOTONICITY EVALUATION

Models	LIME	SHAP
Random Forests	False	False
Gradient Boosted Trees	22% True	22% True
Neural Networks	False	False

2) *Models' Diagnoses*: Even though all of the four models make the same predictions as human doctors, it is important to confirm models' predictions are in accordance with medical knowledge. Table IX lists the three most important features in the interpretation of LIME and SHAP. More detailed interpretations are illustrated in the Figs. 7 and 8.

In Table IX, NTproBNP, CRP, LYM, LDH are the most common features that are deemed crucial by all different models. The three features, CRP, LYM, LDH, are listed as the most indicative biomarkers in the COVID-19 guideline. While the correlation

TABLE XII
PATIENT NO. 0 IN THE KAGGLE DATASET

Feature	Value
SARS-Cov-2 test result	1
Patient Age Quantile	14.00
Hematocrit	0.92
Platelets	-1.26
Mean platelet volume	0.79
Mean corpuscular hemoglobin concentration (MCHC)	-0.65
Leukocytes	-1.47
Basophils	-1.14
Eosinophils	-0.83
Monocytes	0.96
Proteina C reativa mg/dL	0.236

TABLE XIII
CLASSIFICATION RESULTS (KAGGLE)

Classifier	CV	Test Set			95% confidence interval
	F1	Precision	Recall	F1	
Decision Tree	0.37	0.88	0.75	0.71	0.098
Random Forests	0.37	0.90	0.50	0.67	0.089
Gradient Boosted Trees	0.56	0.90	0.75	0.59	0.089
Neural Networks	0.38	0.90	0.50	0.67	0.089

between NTproBNP and COVID-19 are investigated in a article from world health organization (WHO) global literature on coronavirus disease, that reveals elevated NTproBNP is associated with increased mortality in patients with COVID-19 [30].

As a result, the prediction of false-positive is consistent with doctors' diagnoses. Patient 2 who is normal is diagnosed as severe by both doctors and models. One possibility is that even though the patients' test results are not optimistic, he did not require a respirator to survive when he came to the hospital for the first time, so he was classified as normal. In this way, models' predictions can act as a warning. If a patient is diagnosed as severe by models, and the prediction is in accordance with medical knowledge, but the patient feels normal, we can suggest to the patient to put more attention on his health condition.

In conclusion, as illustrated previously in the explanation for patient 5 (false negative), every patient is unique. Some patients are more resistant to viral infection, while some are more vulnerable. Pursuing a perfect model is tough in healthcare, but we can try to understand how different models make predictions using interpretable machine learning to be more responsible with our diagnoses.

H. Evaluating Interpretation

Though we do find some indicative symptoms of COVID-19 through model interpretation, they are confirmed credible because these interpretations are corroborated by medical research. If we use the interpretation to understand a new virus at the early stage of an outbreak, there will be less evidence to support our interpretation. Thus we use monotonicity and faithfulness

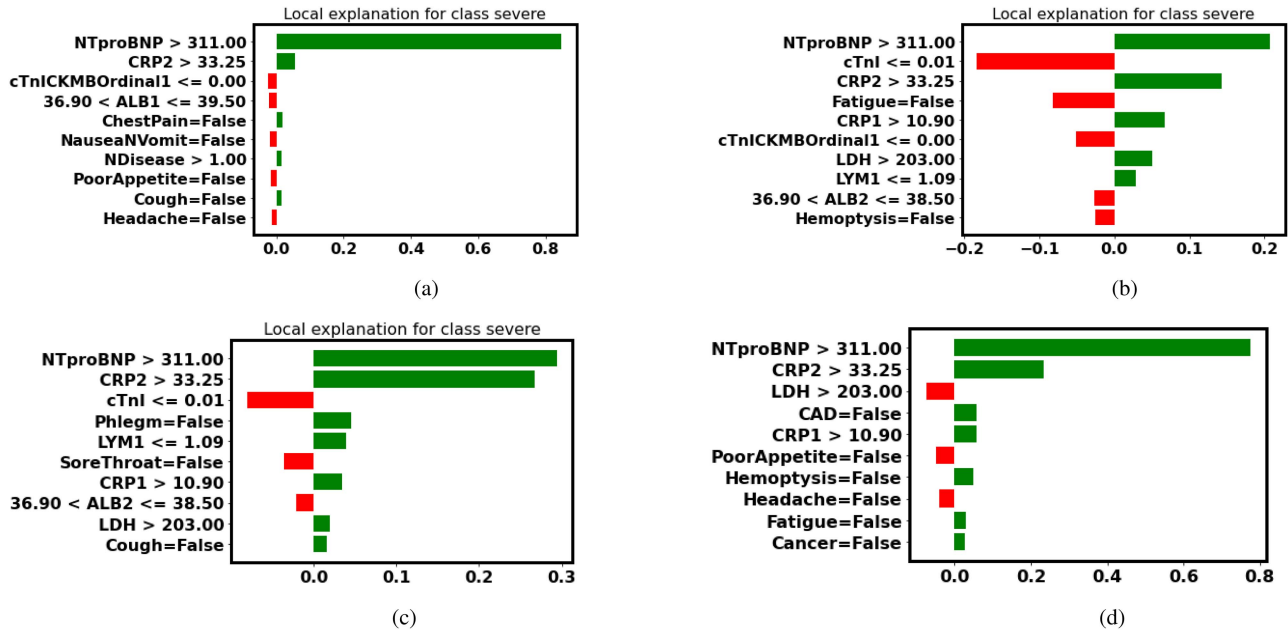


Fig. 7. LIME explanation (false-positive patient No.2): Features in green have a positive contribution to the prediction (increasing the probability of turning severe), and features in red have a negative effect on the prediction (decreasing the probability of turning severe). (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

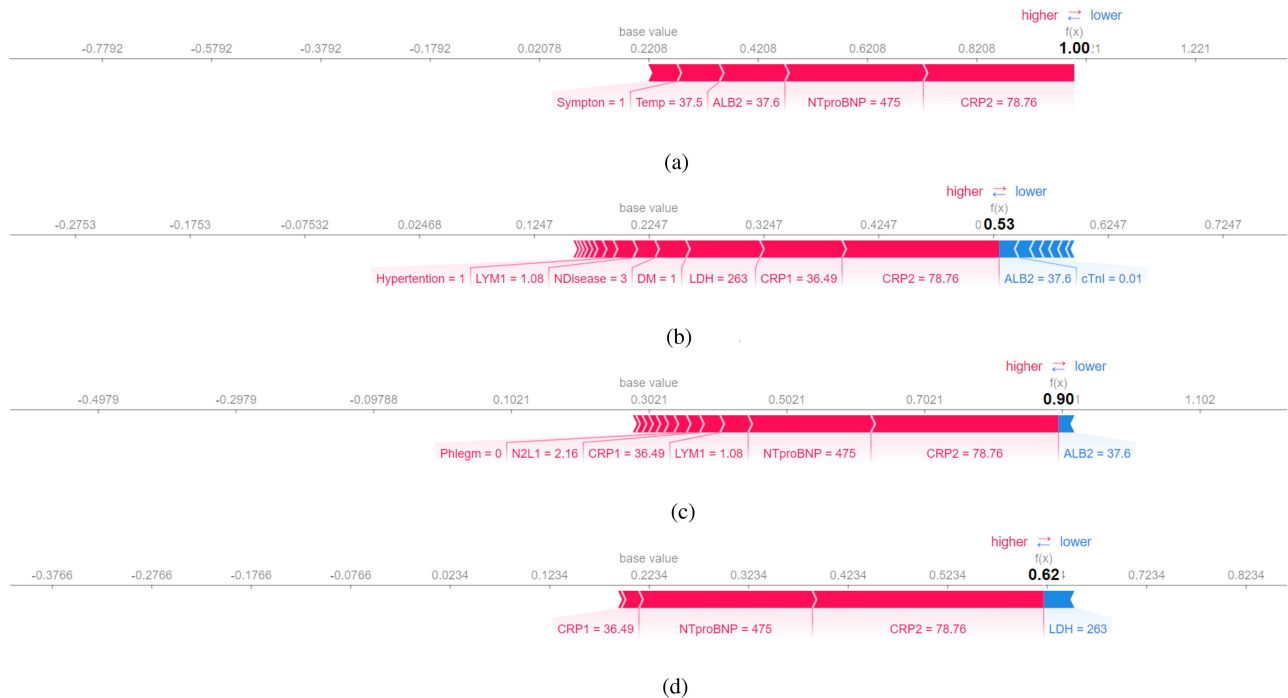


Fig. 8. SHAP explanation (false-positive patient No.2): Features pushing the prediction to be higher (severe) are shown in red, and those pushing the prediction to be lower (normal) are in blue. (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

to evaluate different interpretations using IBM AIX 360 toolbox [31]. The decision tree only provides a binary prediction (0 or 1) rather than a probability between 0 and 1, so it cannot be evaluated using monotonicity and faithfulness.

Faithfulness (ranging from -1 to 1) reveals the correlation between the importance assigned by the interpretability algorithm and the effect of each attribute on the performance of the model. All of our interpretations receive good faithfulness

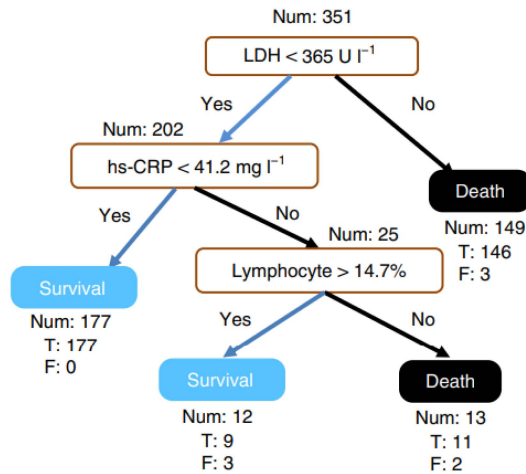


Fig. 9. Decision rule using three key features and their thresholds in absolute value. Num, the number of patients in a class; T, the number of correctly classified; F, the number of misclassified patients. [8].

scores, and SHAP receives a higher faithfulness score than LIME on average. The interpretation by SHAP receives better results because the Shapley value is calculated by removing the effect of specific features which is similar to how faithfulness is computed, so SHAP is more akin to faithfulness.

As for monotonicity, most interpretation methods receive a false though we do find valuable conclusions from interpretations. The difference between faithfulness and monotonicity is that faithfulness incrementally removes each of the attributes, while monotonicity incrementally adds each of the attributes. By incrementally adding each attribute, initially, the model may not be able to make correct predictions with only one or two features, but this does not mean these features are not important. Evaluation metrics for different interpretation methods is still an active research direction, and our results may hopefully stimulate further research on the development of better evaluation metrics for interpreters.

I. Summary

In this section, the interpretation of four different machine learning models reveals that NTproBNP, CRP, and LDH, and LYM are the four most important biomarkers that indicate the severity level of COVID-19 patients. In the next section, we further validate our methods on two datasets to corroborate our proposal.

IV. VALIDATION ON OTHER DATASETS

At the initial outbreak of the pandemic, our research leverages a database consisting of patients with confirmed SARS-CoV-2 laboratory tests between 18th January 2020, and 5th March 2020, in Zhuhai, China, and reveals that an increase in NT-proBNP, CRP, and LDH, and a decrease in lymphocyte count indicates a higher risk of death. However, the dataset has a limited record of 92 patients which may not be enough to support our proposal. Luckily, and thanks to global cooperation, we do

have access to larger datasets. In this section, we further validate our methods on two datasets, one with 485 infected patients in Wuhan, China [8], and the other with 5644 confirmed cases from the Hospital Israelita Albert Einstein, at São Paulo, Brazil from Kaggle.

A. Validation on 485 Infected Patients in China

The medical record of all patients in this dataset was collected between 10th January and 18th February 2020, within a similar date range as our dataset. Yan *et al.* construct a dedicated simplified and clinically operable decision model to rank 75 features in this dataset, and the model demonstrates that three key features, LDH, LYM, and high-sensitivity CRP (hs-CRP) can help to quickly prioritize patients during the pandemic, which is consistent with our interpretation in Table V.

Findings from the dedicated model are consistent with current medical knowledge. The increase of hs-CRP reflects a persistent state of inflammation [32]. The increase of LDH reflects tissue/cell destruction and is regarded as a common sign of tissue/cell damage, and the decrease of lymphocyte is supported by the results of clinical studies [33].

Our methods reveal the same results without taking efforts to design a dedicated interpretable model but can be more prompt to react to the pandemic. During pandemic outbreak, a prompt reaction that provides insights on the new virus could save lives and time.

B. Validation on 5644 Infected Patients in Brazil

Our approach obtains the same result on the dataset with 92 patients from Zhuhai, China, and a medium-size dataset with 485 patients from Wuhan, China. Besides, we further validate our approach on a larger dataset with 5644 patients in Brazil, from Kaggle.

This dataset consists of 111 features including anonymized personal information, laboratory virus tests, urine tests, venous blood gas analysis, arterial blood gases, blood routine test, among other features. All data were anonymized following the best international practices and recommendations. The difference between this dataset and ours is that all data are standardized to have a mean of zero and a unit standard deviation, thus the original data range that contains clinical meaning is lost. Still, the most important medical indicators can be extracted using interpretation methods.

Following the same approach, a preprocessing is applied on the dataset that removes irrelevant features such as patients' intention to the ward level, and features that have less than 100 patient's records, for instance, urine tests and arterial blood gas tests. On the other hand, patients that have less than 10 records are dropped, because these records do not provide enough information. After preprocessing, we have a full record of 420 patients with 10 features.

After training and interpreting four different models, decision tree, random forests, gradient boosted trees, and neural networks, the most important features are identified and listed

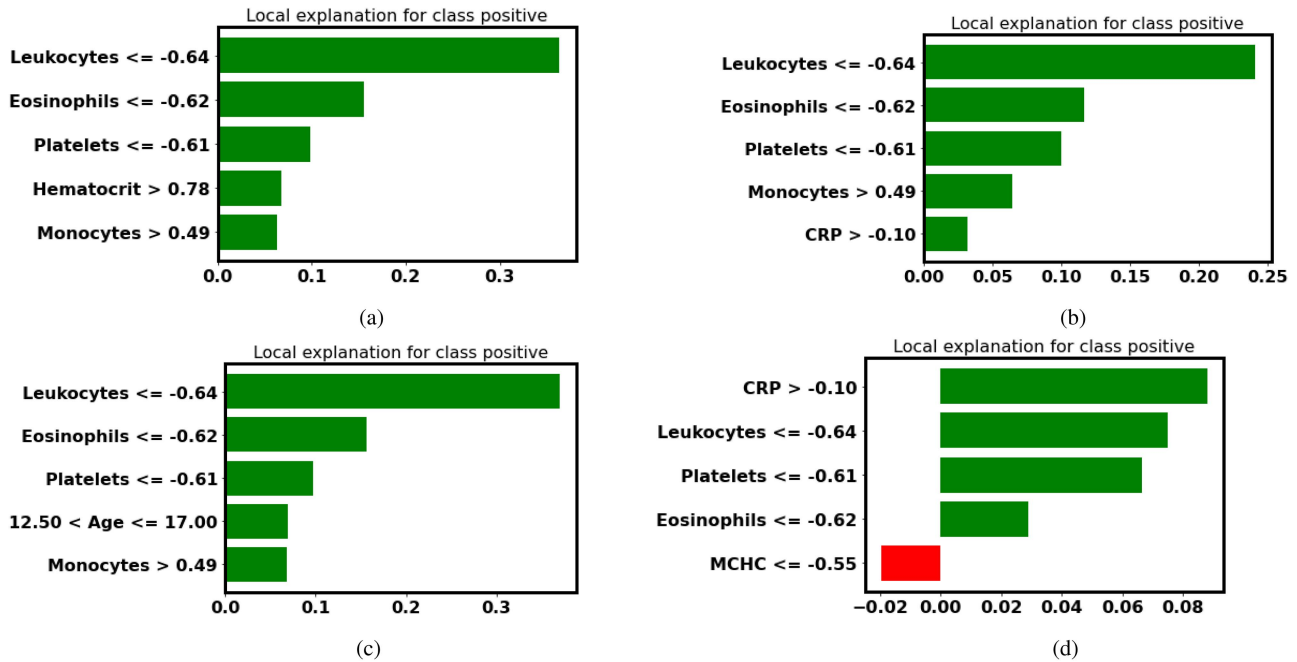


Fig. 10. LIME explanation (Kaggle patient 0): Features in green have a positive contribution to the prediction (increasing the probability of turning severe), and features in red have a negative effect on the prediction (decreasing the probability of turning severe). (a) Decision Tree. (b) Random Forest. (c) Gradient Boosted Trees. (d) Neural Networks.

TABLE XIV
FIVE MOST IMPORTANT FEATURES (KAGGLE)

Model	Most Important Features
Decision Tree	Leukocytes, Eosinophils, Patient age quantile
Random Forest	Leukocytes, Eosinophils, Platelets
Gradient Boosted Trees	Patient age quantile, Hematocrit, Platelets
Neural Networks	Leukocytes, Platelets, Monocytes

in Table XIV. The three most common indicative features are leukocytes, eosinophils, and platelets.

According to medical research, patients with increased leukocyte count are more likely to develop critical illness, more likely to admit to an ICU, and have a higher rate of death [34]. Du *et al.* noted that at the time of admission, 81% of the patients had absolute eosinophil counts below the normal range in the medical records of 85 fatal cases of COVID-19 [35]. Wool and Miller [36] discovered that COVID-19 is associated with increased numbers of immature platelets which could be another mechanism for increased clotting events in COVID-19.

In addition, the two datasets collectively reveal that elderly people are more susceptible to the virus. The significant feature NTproBNP in the Chinese dataset is often used to diagnose or rule out heart failure which is more likely to occur in elderly people. And patients that have abnormally low levels of platelets are more likely to be older, male as well [36].

To further validate our interpretation, faithfulness and monotonicity are calculated and listed in Tables XV and XVI. Similarly, our interpretations are consistent with medical knowledge and receive a good faithfulness score, but receive a worse score on monotonicity because the calculation procedure of monotonicity is contrary to faithfulness.

TABLE XV
FAITHFULNESS EVALUATION (KAGGLE)

Models	LIME	SHAP
Random Forests	0.71	0.82
Gradient Boosted Trees	0.61	0.72
Neural Networks	0.25	0.42

TABLE XVI
MONOTONICITY EVALUATION (KAGGLE)

Models	LIME	SHAP
Random Forests	False	False
Gradient Boosted Trees	True	False
Neural Networks	False	False

V. CONCLUSION

In this article, through the interpretation of four different machine learning models, we reveal that NTproBNP, CRP, LDH, and LYM are the four most important biomarkers that indicate the severity level of COVID-19 patients. Our findings are consistent with medical knowledge and recent research that exploits dedicated models. We further validate our methods on a large open dataset from Kaggle and unveil leukocytes, eosinophils, and platelets as three indicative biomarkers for COVID-19.

The pandemic is a race against time. Using interpretable machine learning, medical practitioners can incorporate insights from models with their prior medical knowledge to promptly reveal the most significant indicators in early diagnosis and hopefully win the race in the fight against the pandemic.

APPENDIX

TABLE XVII
DIAGNOSES

Feature	Comments
Severity03	Severe (3) - Normal (0)
Severity01	Severe (1), Normal (0)

TABLE XVIII
PERSONAL INFO

Feature	Comments
MedNum	Medical Number
No	Patient No.
Sex	Man (1), Woman(0)
Age	-
AgeG1	Age > 50(1), Age ≤ 50(0)
Height	-
Weight	-
BMI	Body Mass Index

TABLE XIX
COMPLETE BLOOD COUNT

Feature	Comments
WBC1	White Blood Cell (first time)
NEU1	Neutrophil Count (first time)
LYM1	Lymphocyte Count (first time)
N2L1	-
WBC2	White Blood Cell (second time)
NEU2	Neutrophil Count (second time)
LYM2	Lymphocyte Count (second time)
N2L2	-

TABLE XX
INFLAMMATORY MARKERS

Feature	Comments
PCT1	Procalcitonin (first time)
CRP1	C-Reactive Protein (first time)
PCT2	Procalcitonin (second time)
CRP2	C-Reactive Protein (second time)

TABLE XXI
BIOCHEMICAL EXAMINATION

Feature	Comments
AST	Aspartate aminotransferase
LDH	Lactate Dehydrogenase
CK	Creatine Kinase
CKMB	The amount of an isoenzyme of creatine kinase (CK)
HBDH	Alpha-Hydroxybutyrate Dehydrogenase
HiCKMB	Highest CKMB
Cr	Serum Creatinine
ALB1	Albumin Count (first time)
ALB2	Albumin Count (second time)

REFERENCES

[1] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *Indian J. Pediatrics*, vol. 87, no. 4, pp. 281–286, 2020.

TABLE XXII
SYMPTOMS AND ANAMNESES

Feature	Comments
Symptom	-
Fever	-
Cough	-
Phlegm	-
Hemoptysis	-
SoreThroat	-
Catarrh	-
Headache	-
ChestPain	-
Fatigue	-
SoreMuscle	-
Stomachache	-
Diarrhea	-
PoorAppetie	-
NauseaNVomit	-
Hypertention	-
Hyperlipedia	-
DM	Diabetic Mellitus
Lung	Lunge Disease
CAD	Coronary Heart Disease
Arrhythmia	-
Cancer	-

TABLE XXIII
OTHER TEST RESULTS

Feature	Comments
Temp	Temperature
LVEF	Left Ventricular Ejection Fraction
Onset2Admi	Time from onset to admission
Onset2CT1	Time from onset to CT test
Onset2CTPositive1	Time from onset to CT test positive
Onset2CTPeak	Time from onset to CT peak
cTnITimes	When was cTnI tested
cTnI	Cardiac Troponin I
cTnICKMBOrdinal1	The value when hospitalized
cTnICKMBOrdinal2	The maximum value when hospitalized
CTScore	Peak CT Score
AIVolumeP	Peak Volume
SO2	Empty
PO2	Empty
YHZS	Empty
RUL	Empty
RML	Empty
RLL	Empty
LUL	Empty
LLL	Empty

- [2] S. Basu, S. Mitra, and N. Saha, "Deep learning for screening COVID-19 using chest X-Ray images," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2020, pp. 2521–2527.
- [3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-Day readmission," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1721–1730.
- [4] A. Ramchandani, C. Fan, and A. Mostafavi, "Deepcovidnet: An interpretable deep learning model for predictive surveillance of COVID-19 using heterogeneous features and their interactions," *IEEE Access*, vol. 8, pp. 159915–159930, 2020.
- [5] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: A factorization-machine based neural network for CTR prediction," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 8, 2017, pp. 1725–1731.
- [6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
- [7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

- [8] L. Yan *et al.*, “An interpretable mortality prediction model for COVID-19 patients,” *Nature Mach. Intell.*, vol. 2, no. 5, pp. 283–288, May 2020.
- [9] E. Matsuyama *et al.*, “A deep learning interpretable model for novel coronavirus disease (COVID-19) screening with chest CT images,” *J. Biomed. Sci. Eng.*, vol. 13, no. 07, p. 140, 2020.
- [10] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [11] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” *J. Comput. Graphical Statist.*, vol. 24, no. 1, pp. 44–65, 2013.
- [12] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models,” *J. Roy. Stat. Soc. Ser. B*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020.
- [13] A. Fisher, C. Rudin, and F. Dominici, “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously,” *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.
- [14] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should i trust you?”: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2016, pp. 1135–1144.
- [15] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. H. Bengio R. Wallach Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774.
- [16] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proc. Assoc. Adv. Artif. Intell.*, 2018, pp. 1527–1535.
- [17] D. Alvarez-Melis and T. S. Jaakkola, “Towards robust interpretability with self-explaining neural networks,” 2018, *arXiv:1806.07538*.
- [18] R. Luss *et al.*, “Generating contrastive explanations with monotonic attribute functions,” 2019, *arXiv:1905.12698*.
- [19] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-agnostic interpretability of machine learning,” 2016, *arXiv:1606.05386*.
- [20] C. Molnar, *Interpretable Machine Learning*. Morrisville, NC, USA: Lulu.com, 2020.
- [21] L. S. Shapley, “A value for n-person games,” in *Contributions to the Theory of Games (AM-28)*, vol. II, Princeton, NJ, USA: Princeton Univ. Press, pp. 307–318, 1953.
- [22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth Int. group, 1984.
- [23] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [24] S. Schaal and C. C. Atkeson, “From isolation to cooperation: An alternative view of a system of experts,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 8, 1996, pp. 605–611.
- [25] A. Maier, C. Syben, T. Lasser, and C. Riess, “A gentle introduction to deep learning in medical image processing,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.
- [26] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.
- [27] F. Pedregosa *et al.*, “scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [28] L. Wang, “C-reactive protein levels in the early stage of COVID-19,” *Med. Mal. Infect.*, vol. 50, no. 4, pp. 332–334, 2020.
- [29] L. Gao *et al.*, “Prognostic value of NT-proBNP in patients with severe COVID-19,” *Respir. Res.*, vol. 21, no. 1, 2020, Art. no. 83.
- [30] R. Pranata, I. Huang, A. A. Lukito, and S. B. Raharjo, “Elevated n-terminal pro-brain natriuretic peptide is associated with increased mortality in patients with COVID-19: Systematic review and meta-analysis,” *Postgraduate Med. J.*, vol. 96, no. 1137, pp. 387–391, 2020.
- [31] V. Arya *et al.* “One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques,” 2019, *arXiv:1909.03012*.
- [32] E. K. Bajwa, U. A. Khan, J. L. Januzzi, M. N. Gong, B. T. Thompson, and D. C. Christiani, “Plasma C-reactive protein levels are associated with improved outcome in ARDS,” *Chest*, vol. 136, no. 2, pp. 471–480, Aug. 2009.
- [33] N. Chen *et al.*, “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study,” *Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.
- [34] K. Zhao *et al.*, “Clinical features in 52 patients with COVID-19 who have increased leukocyte count: A retrospective analysis,” *Eur. J. Clin. Microbiology Infect. Diseases*, vol. 39, no. 12, pp. 2279–2287, 2020.
- [35] Y. Du *et al.*, “Clinical features of 85 fatal cases of COVID-19 from Wuhan. A retrospective observational study,” *Amer. J. Respiratory Crit. Care Med.*, vol. 201, no. 11, pp. 1372–1379, 2020.
- [36] G. D. Wool and J. L. Miller, “The impact of COVID-19 disease on platelets and coagulation,” *Pathobiology*, vol. 88, no. 1, pp. 14–26, 2021.