A Collaborative Medical Diagnosis System Without Sharing Patient Data

NAN Yucen[1], FANG Minghao[2], ZOU Xiaojing[2], DOU Yutao[3], Albert Y. ZOMAYA[3]

(1. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410003, China;
2. Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China;
3. Center for Distributed and High Performance Computing, University of Sydney, Sydney 2008, Australia)

DOI: 10.12142/ZTECOM.202203002

https://kns.cnki.net/kcms/detail/34.1294.TN.20220901.1320.002.html, published online September 1, 2022

Manuscript received: 2022-06-10

Abstract: As more medical data become digitalized, machine learning is regarded as a promising tool for constructing medical decision support systems. Even with vast medical data volumes, machine learning is still not fully exploiting its potential because the data usually sits in data silos, and privacy and security regulations restrict their access and use. To address these issues, we built a secured and explainable machine learning framework, called explainable federated XGBoost (EXPERTS), which can share valuable information among different medical institutions to improve the learning results without sharing the patients' data. It also reveals how the machine makes a decision through eigenvalues to offer a more insightful answer to medical professionals. To study the performance, we evaluate our approach by real-world datasets, and our approach outperforms the benchmark algorithms under both federated learning and non-federated learning frameworks.

Keywords: explainable machine learning; federated learning; secured data analysis; medical applications

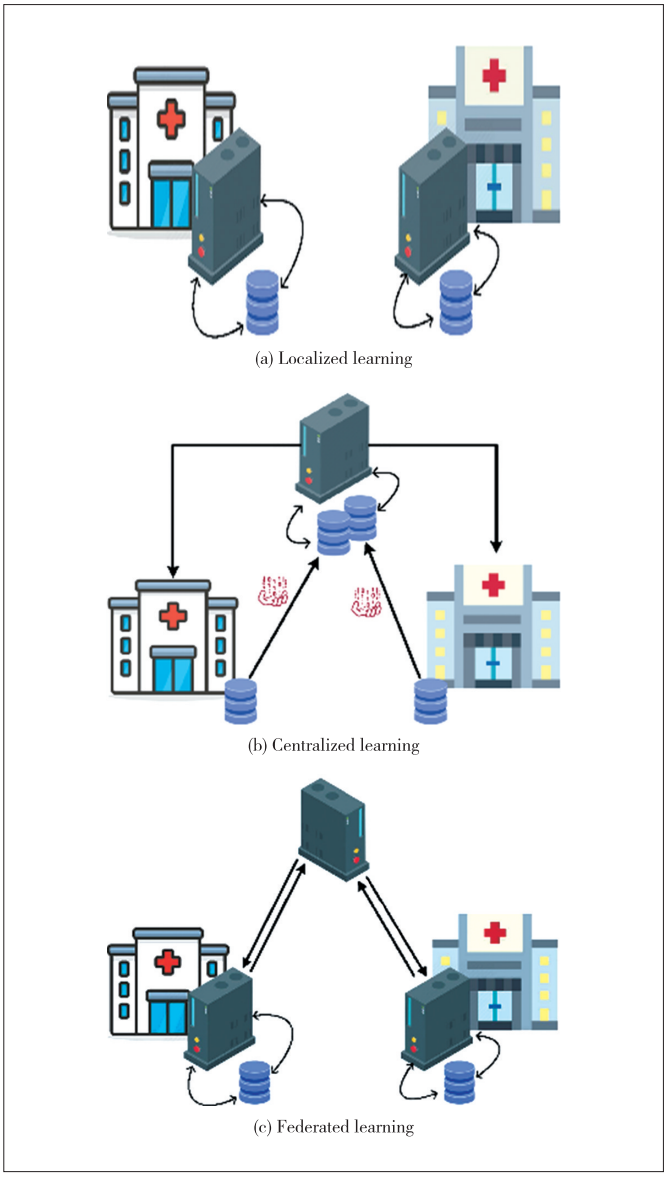
Citation (IEEE Format): Y. C. Nan, M. H. Fang, X. J. Zou, et al., "A collaborative medical diagnosis system without sharing patient data," *ZTE Communications*, vol. 20, no. 3, pp. 3 - 16, Sept. 2022. doi: 10.12142/ZTECOM.202203002.

1 Introduction

Machine learning (ML) has played an important role in the healthcare industry, serving as a decision support system for medical diagnosis, and actively promotes smart medicine development[1-2]. It can be used to complete some laborious and often time-consuming routine tasks for better resource utilization. More importantly, ML can offer meaningful support for clinical decision-making by comprehensively analyzing electronic healthcare records (EHR)[3-5]. More than 70% of medical institutions worldwide have implemented EHR systems, but just 3% of them can exchange data over the network[6]. Without a secured framework for managing the use of EHR[7], patients' information is at a high risk of cyber threats. Meanwhile, the performance of ML could be severely degraded by the limited data available locally.

The data security concerns lead to EHRs which are often used locally for analysis and learning, as depicted in Fig. 1(a). On the other hand, the medical data available in a single place are often not enough to fully exploit the advancement of ML. The lack of data can be solved by using a centralized system as shown in Fig. 1(b) to store the data from multiple sources. However, the security threats to the centralized system come from multiple aspects, such as data transmission over the network, data leaking from the centralized server, and manipulation and misconduct in handling patient data. All these threats pose unique technical and ethical challenges for this solution. Federated learning (FL) is a burgeoning distributed ML paradigm to collectively train a model as a whole without explicitly exchanging data samples[8]. It enables a party (such as medical institutions and organizations) to transparently and securely share knowledge with other parties[9]. FL can be roughly divided into three groups, namely, horizontal FL, vertical FL, and transfer FL[10]. The horizontal FL means that the datasets used for training have the same feature space across all parties. The vertical FL uses different datasets of different feature spaces to jointly train a global model. The transfer FL refers to using transfer learning to utilize a pre-trained model that is trained on a similar dataset for solving a different problem. In this study, the federated learning applied to different medical institutions is horizontal FL[11-12]. Horizontal FL can help solve the problem of lack of data for some hospitals, and only the model parameters are exchanged among parties while developing a global diagnostic model, as shown in Fig. 1(c).

Furthermore, clinical heterogeneity, lack of specific moni-



▲ **Figure 1. Different types of learning**

toring markers, and interpretive uncertainty may lead to misdiagnosis in computer-aided diagnosis. Therefore, in addition to data privacy, a secured medical decision-support system also involves generating reliable and trustful results by providing instrumental clues to medical professionals on why the decision is made, which is also known as explainable/interpretable machine learning[13-15]. Some of the explainable ML techniques are model-dependent, especially for linear models and decision trees, while the others are model agnostic and can be applied to any supervised ML model. The model interpretability is often available for those trained locally, but it is still an open question for FL.

In this study, we develop an explainable XGBoost model (a tree-based extreme gradient boost model) under a horizontal federated learning framework, EXPERTS, to construct a secured medical decision-support system. In the system, we can achieve the desired learning performance without sharing patient data and provide consistent global interpretability among parties. To fully demonstrate and understand the system, we extended the edge analytics framework[16] to collect the same set of features (demographic characteristics, clinical features, vital signals, and laboratory tests) of COVID-19 patients from different places and store the collected data locally. Then EXPERTS is applied to predict the status of the hospitalized COVID-19 patients during their stay. The main contributions of this paper are summarized as follows:

• We propose an explainable XGBoost under the horizontal FL framework, called EXPERTS, to construct a secured medical decision-support system. In this system, only model parameters are shared among parties to build a global model without sharing any patient's data, thereby protecting patients' privacy without losing performance.

• We implement the Shapley value to provide the horizontal FL model interpretability by revealing the detailed feature importance at each party. Within the system, the feature importance is consistent between parties, which means we can provide global model interpretability for all parties.

• We demonstrate the practicality of EXPERTS by a real-world COVID-19 dataset and an open medical dataset named Cerebral Vasoregulation in the elderly with stroke. Our results confirm that EXPERTS can achieve the same performance level as the centralized learning approach.

The remainder of this paper is organized as follows. Section 2 gives the motivation for our work. Section 3 shows the design of our study. Section 4 reveals the experimental results of our design. Section 5 concludes this work and sketches the future work.

2 Related Work

Data-driven ML has emerged as a promising option for developing accurate and efficient diagnostic tools from large volumes of medical data. In Ref. [17], the authors argued that an AI-based tumor detector requires massive and a wide range of data, including possible anatomies, pathologies and many others, to make valuable clinical suggestions, and to be practical and generalizing well to new patients. However, it is impractical to include all of them among medical institutions as the data are highly sensitive and the usage is strictly regulated. Even when the patients are de-identified by removing their personal information, their privacy could still be exposed by reconstructing faces from computed tomography (CT) or magnetic resonance imaging (MRI) data[18].

Federated learning[8] is one of the emerging approaches to address security challenges by introducing the idea of sharing the characteristics of the ML model rather than the data itself. More specifically, it keeps the patient data locally for each participant and only transmits the intermediate results of the model at local servers to the centralized server for model iteration and

update, thereby reducing communication intensity and improving data privacy. Since proposed by Google first in 2017[8], FL has attracted more and more attention among researchers and has been widely utilized in various privacy-sensitive domains. It has great potential for medical and healthcare applications[19 - 21].

Nevertheless, there are still several issues that remain in FL. For example, to improve performance, researchers focus heavily on neural networks but ignore other machine learning models, such as decision trees. Not only that, by emphasizing performance using neural networks, researchers also ignore the interpretability of the model, which is crucial for medical professionals to understand what drives the ML to make the decision. The study on the interpretability of FL is very limited. The authors in Ref. [22] studied the model interpretability under the framework of vertical FL. In this work, a party contributes to the vertical FL model by sharing its features with others. The contribution of the party can thus be represented by the combined contributions of its shared features. In other words, the interpretability of the federated model is provided by the group Shapley values instead of the individual ones. To our knowledge, this work is the first interpretability analysis based on original features under the horizontal FL framework, and will not affect the global interpretability for using different data samples stored in medical institutions for model training.

3 Method

The technical detail of EXPERTS is given in this section. To make the entire system clearer, we illustrate the detailed processing flow in Fig. 2. As can be seen, the local data in each hospital will never leave the local physical area. In the local database, data pre-processing is first performed through the patients/variables filter and abnormal removal. Then, we rely on statistical transferring and one-hot sampling to further refine the pre-screened data. After getting the available data, we use the tree-based SHapley Additive exPlanations (SHAP) to rank all the features by correlation, and select the top-20 features for subsequent local learning. Consequently, we perform local fast learning through the initialized model, and upload the local model's parameters to the central processor in the federated node. By applying certain mathematical methods to weight or average the various parameter sets from different hospitals, which generalizes local model parameters to global parameters, we send them back to each local node for model update and learning. In this section, we briefly describe these steps encompassing the learning strategy, like data pre-processing, followed by horizontal federated-XGBoost and model interpretability.

3.1 Data Pre-Processing

Before the data analysis model is performed, the raw data are obtained, organized, and pre-processed locally. The data pre-processing includes variable extraction, unification, artifact removal, feature generation and others. In this regard, we first need to obtain patient data from the local hospital database. The data can be retrieved in different forms and need to be turned into a table-like structure. The steps of data unification involve timestamp unification (unifying time count), unit unification (unifying measurement unit for each variable), categorical variables form unification (converting categorical variables to numeric variables), and representation unification (unifying the name of the feature). In the artifact removal step, multiple procedures are performed to ensure the validity and quality of the data. For example, the timestamp artifact removal procedure is used to remove unrelated medical records. The out-of-range artifact removal procedure is used to remove the values of the feature that greatly exceed its physiological range. The data normalization is still valuable to reduce the adverse effects associated with the use of physiological data. More specifically, Z-score is used to normalize all included features to obtain a normalized version of variables, which is computed by $\text{Normalized}(x) = \frac{x - \bar{x}}{\text{std}(x)}$, where $\bar{x}$ and $\text{std}(x)$ represent the mean and standard deviation of x, respectively. The time-series variables are converted into static features via discretization.

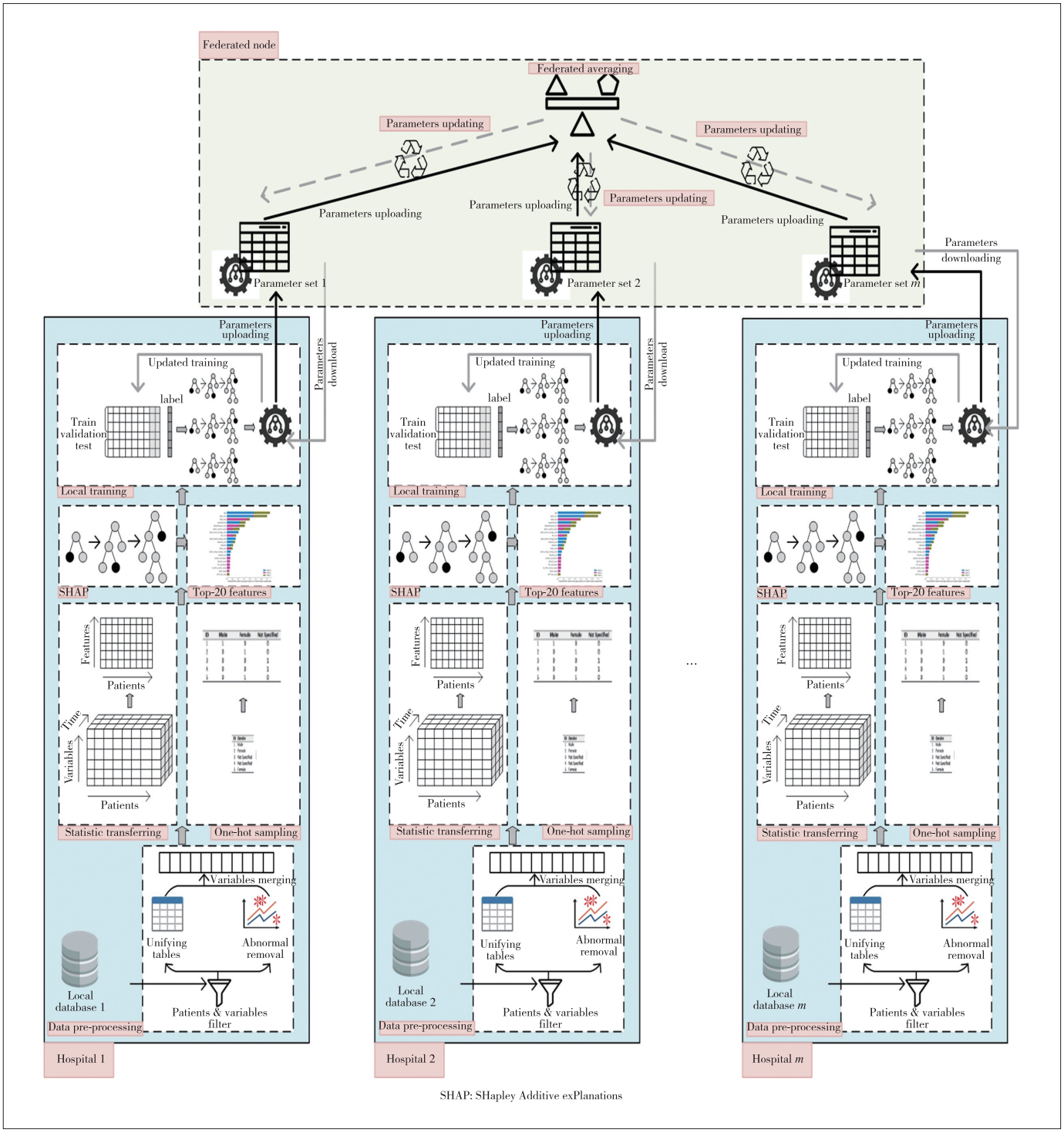
3.2 Horizontal Federated-XGBoost

Although neural networks are currently the most popular ML models, the lack of clear interpretability makes them hard to justify their decisions, which is a prerequisite for the widespread adoption of machine learning approaches by healthcare communities. Instead, decision trees (DT) are regarded as reliable alternatives for balancing accuracy and interpretability. DT is a tree-like ML model that consists of nodes and edges, where the internal nodes present the test instances, the edges present the results, and the leaf nodes present the prediction results. In short, the path from the root to the leaf represents the prediction rule. Although the gradient boosting decision tree (GBDT) has not yet received enough attention under the FL framework, the representative XGBoost is a promising candidate to achieve the desired ML performance.

We first give a recap of the the XGBoost algorithm. For a given set of n independently identically distributed and labeled examples $\{(x_i, y_i),\ i = 0,\cdots,n\}$, where $\boldsymbol{X} \in \mathcal{R}^{n \times d}$ and d represents the feature dimension. The goal of XGBoost is to train a learning model with a set of parameters to minimize the objective loss function for the K iterations, which can be represented as follows:

$$\text{Objective} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{1}$$

where $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ is the total training loss after K iterations to measure how well the model fits. More specifically, y_i is the real label, and $\hat{y}_i$ presents the predicted output for the i-th



▲ **Figure 2. Complete workflow of EXPERTS**

data sample after K iterations through using K CARTs, which can be calculated as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ . \tag{2}$$

$\sum_{k=1}^{K} \Omega(f_k)$ in Eq. (1) is the regularization term to measure the complexity of the model, and $\Omega(f_k)$ can be depicted as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||\omega||^2, \tag{3}$$

with the component ω_j of ω being the score/weight of the j-th leaf node of the tree. T is the number of leaf nodes, and $\frac{1}{2}\lambda||\omega||^2$ is the L2 regularization term of the leaf node score. The score of each leaf node is increased by L2 smoothing to prevent overfitting. In short, by minimizing the objective function of Eq. (1), both the accuracy and stability of the model can be considered, and it is the balance between the deviation and the variance.

Moreover, XGBoost is an additive model and the newly generated tree needs to fit the last predicted residual, which means the objective is no longer to directly optimize the entire objective function, but to optimize the objective function step by step from the first tree to the K-th tree. Then, $\hat{y}_i$ can be rewritten as $\hat{y}_i^k = \hat{y}_i^{(k-1)} + f_k(x)$ for the k-th iteration.

After that, we need to find the best split of samples of the tree from root to leaf. By using the greedy algorithm to search for the best split which aims to maximize the learning gain at each iteration, the gain can be calculated as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right], \tag{4}$$

where I_L and I_R represent the left and right sets of data sample indices. When searching for the best split point, instances g_i and h_i in the left and right space will be calculated for getting the value of Gain. When a CART structure is fixed, the weight ω_j of a leaf node j is calculated by:

$$\omega_j^{\dagger} = -\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}. \tag{5}$$

Considering the generality, we apply a particular logistic loss function $l\left(y_i, \hat{y}_i^{(t-1)}\right) = y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i)\ln(1 + e^{\hat{y}_i})$ as our picked loss function. Then, the first and second order gradient of the loss function can be derived as:

$$g_i = \frac{1}{1 + e^{-\hat{y}^{(k-1)}}} - y_i, \tag{6}$$

and

$$h_i = \frac{1}{1 + e^{-\hat{y}^{(k-1)}}} \times \left(1 - \frac{1}{1 + e^{-\hat{y}^{(k-1)}}}\right), \tag{7}$$

separately.

In this work, we study the horizontally partitioned data for different nodes, which means the nodes have the same feature dimension and each node holds the entire features of an instance. For better understanding, we modeled this method as the following. Assuming there are L distributed parties $P_0,\cdots,P_L$ that hold sample sets $X_0,\cdots,X_L$, where each $X_l = \begin{bmatrix} X_{l0} \\ \cdots \\ X_{lm} \end{bmatrix}$ involves m samples in the l-th party, entities accompanied with the label involved in the l-th party can be shown as $\left[(x_{l0}^0,\cdots,x_{l0}^d, y_{l0}),\cdots,(x_{lm}^0,\cdots,x_{lm}^d, y_{lm})\right]$. To implement the XGBoost under the federated-learning framework, the key idea is to calculate the parameters g_i and h_i at each local party discussed in Eqs. (6) and (7), and then pass them to the central aggregator to determine an optimal split through iterative model averaging to further update the model. In short, XGBoost under the FL framework is summarized as follows:

- Each party downloads the latest XGBoost model from the central aggregate server.
- Each party uses local data to train the downloaded XGBoost model and uploads the gradient to the central aggregate server, and the server aggregates the gradient of each user to update the model parameters.
- The central aggregate server distributes the updated model to each party.
- Each party updates the local model accordingly.

3.3 Model Interpretability

When the final model updates, we will conduct a feature importance analysis on local nodes and compare the explanation from each local node to check the robustness of our model.

In the past, people used Gain[23] or Split[24] to explain the model as they could summarize a complicated ensemble model and provide insight into what features drive the model's prediction. However, it cannot be ignored that in some cases, the rankings of Gain and Split are often inconsistent even for the same features. To solve the problem of inconsistency in the feature attribution method, we choose Shapley value as an explanatory tool for our model. As defined in Ref. [25], the Shapley value for the j-th feature is a solution concept in the cooperative game theory, which can be obtained by:

$$\phi_j(\text{val}) = \sum_{\mathcal{S}\subseteq\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_d\}\setminus\{\boldsymbol{X}_j\}} \frac{|\mathcal{S}|!\left(d - |\mathcal{S}| - 1\right)!}{d!}\left[\text{val}\left(\mathcal{S}\cup\left\{\boldsymbol{X}_j\right\}\right) - \text{val}(\mathcal{S})\right], \tag{8}$$

where $\mathcal{S}$ is the sub-set of features used in the model, $\boldsymbol{X}$ is the vector of features of the instance to be explained, and d is the number of features defined above. $\text{val}_X(\mathcal{S})$ is the prediction of the eigenvalues in the set S, and the features excluded in the set S are marginalized as:

$$\mathrm{val}_X(\mathcal{S}) = \int \hat{f}(x_1,\cdots,x_d)\mathrm{d}P_{X\notin\mathcal{S}} - E_X(\hat{f}(X)), \tag{9}$$

which performs multiple integrations for each excluded feature. The Shapley value obtained in this way can satisfy efficiency, symmetry, dummy, and additivity[26] at the same time, which can be regarded as the definition of fair expenditures.

However, the exact Shapley value must be estimated using the j-th feature and all possible subsets that exclude the j-th feature. As more features are involved, the computational complexity of the accurate solution to this problem increases exponentially. To reduce the complexity, we adopt SHapley Additive exPlanations as an alternative. SHAP is the Shapley value estimate based on the game theory, and it has two variants, namely KernelSHAP and TreeSHAP. The computation cost of KernelSHAP is very high as it aims for serving all ML models, so it can only approximate the actual Shapley value. TreeSHAP is fast; it can calculate the accurate Shapley value, and even correctly estimate the Shapley value when the features are correlated.

The TreeSHAP value is defined below:

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^{M}\phi_j z_j', \tag{10}$$

where $f(x)$ represents the predicted value of the sample in the decision tree, $z_j' \in \{0,1\}^M$ represents how many features of all d features are included in the decision path where the sample is located. For example, if the feature k is not in its decision path, the SHAP value of the corresponding feature is 0, that is, $\phi_k = 0$, which means that the feature k will not contribute to the final predicted value. Moreover, ϕ_i is represented as below:

$$\phi_j = \sum_{\mathcal{S}\subseteq N\{j\}} \frac{|\mathcal{S}|!(M-|S|-1)!}{M!}\left[f_x(\mathcal{S}\cup\{j\}) - f_x(\mathcal{S})\right], \tag{11}$$

where N is the collection of all the features in the training set, and its dimension is M; $\mathcal{S}$ is a subset extracted from N and its dimension is $|\mathcal{S}|$.

The pseudo code of our proposed algorithm EXPERTS is provided in Algorithm 1, which is formed by the above process.

Algorithm 1: EXPERTS

Input: each party P_l inputs m samples, and each sample has all d features and the corresponding label y_{lm}
Output: K decision trees with global feature interpretability
1: Perform pre-processing steps discussed in Section 3.1 in each local party for every sample
2: **Aggregate server**
3: **for** each round $t = 1, 2, \cdots$ **do**
4: set L local parties with hyper-parameters
5: for the maximal score, aggregate server sends gain to other local Ps
6: **end for**
7: **Local client:**
8: **for** $l = 1 \rightarrow L$ **do**
9: split samples in P_l into Ω batches
10: receive default hyper-parameter from aggregate server
11: **for** each local epoch from 1 to E **do**
12: **for** batch $\omega \in \Omega$ **do**
13: P_l initializes $\{\hat{y}\}_{ml}$ with hyper-parameters
14: **end for**
15: **end for**
16: **end for**
17: **for** $k = 1 \rightarrow K$ **do**
18: **for** $l = 1 \rightarrow L$ **do**
19: P_l computes g_i and h_i described in Eqs. (6) and (7)
20: **end for**
21: **for** each node in the current tree **do**
22: **for** $j = 1 \rightarrow d$ **do**
23: P_l run Eq. (4) for split
24: **end for**
25: for the maximal score, P_l sends gain to other Ps
26: **end for**
27: update $y_0, \cdots, y_d$ based on the weights in Eq. (5)
28: calculate the approximate Shapley value through Eq. (10)
29: **end for**

4 Performance Evaluation

This section first presents our experiments' setup, which involves both a testbed study and a numerical study. The testbed is a real-world prototype for COVID-19 diagnosis. To study the flexibility of the proposed framework EXPERTS, we also used a publicly available dataset for stroke in our experiments. For each medical application, we treated the data collected from two different hospitals but the approach can be extended to multiple parties. The framework's performance was comprehensively evaluated using multiple metrics, including accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve, and Precision-Recall (PR) curve. We also implemented numerous benchmark algorithms involving the federated learning framework and its counterparts, namely, federated-multilayer perceptron (MLP), XGBoost, MLP, and Random forest. All algorithms are performed after the missing values imputed with the mean, except for EXPERTS and XGBoost.

4.1 Experiment Setting

To evaluate the performance of our proposed algorithm and to conduct a fair comparison, all data analytics were carried out on the same setting servers, which was a laptop with a 2.3 GHz Intel Core i5 CPU and 8 GB memory. Additionally, all computational steps involved in this study, such as preprocessing and learning, and the proposed algorithm, along with the selected benchmark algorithms, were all implemented in Python 3.8 with PyTorch and TensorFlow.

4.1.1 Dataset

• Real-world dataset: Our study was performed at two designated hospitals for treating COVID-19 patients during the outbreak. We retrospectively analyzed 1 012 and 1 642 hospitalized patients separately, involving patients with the mild symptom, severe symptom, and critical symptom diagnosed according to WHO interim guidance[27]. Laboratory confirmation of SARS-CoV-2 infection was performed by the local health authority[1]. In total, 24 items within CBC (shown in Table 1), two demographic variables (gender and age), five types of comorbidities (including hypertension, coronary heart disease, diabetes, stroke, and cancer), and five time-series vital signals (breath, blood pressure, SpO2, pulse, and temperature) were used to represent the physical condition of patients in this study.

▼ **Table 1. Feature abbreviation checklist within complete blood count (CBC) test**

Abbreviation	Full Name
EON (#)	Eosinophils (#)
EON (%)	Eosinophils (%)
EOP (#)	Basophils (#)
EOP (%)	Basophils (%)
HCT	Hematocrit
HGB	Hemoglobin
LYM (#)	Lymphocyte (#)
LYM (%)	Lymphocyte (%)
MCH	Mean corpuscular hemoglobin
MCHC	Mean corpuscular hemoglobin concentration
MCV	Mean corpuscular volume
MONON (#)	Monocyte (#)
MONON (%)	Monocyte (%)
MPV	Mean platelet volume
NEU (#)	Neutrophils (#)
NEU (%)	Neutrophils (%)
PCT	Procalcitonin
PDW	Platelet distribution width
P-LCR	Platelet-large cell ratio
PLT	Platelet
RBC	Red blood cell
RDW-CV	Red blood cell distribution width CV
RDW-SD	Red blood cell distribution width SD
WBC	White blood cell

• Open dataset: We also used a public non-image based real-world dataset, known as Cerebral Vasoregulation in the Elderly with Stroke[2][28] in our experiment. Cerebral Vasoregulation in the Elderly with Stroke with numerous feature values can be identified as a binary category (stroke or non-stroke). This dataset involves 164 patient instances and contains a large number of missing values, since it was produced from the data collected in a real medical care environment after a long period of time. To simulate the framework of federated learning, we first split this data into 70% training set, 20% validation set, and 10% test set. Then we randomly split the training set to simulate data from two different institutions. In this study, we divided the training set into 65% and 35% for performance evaluation.

4.1.2 Benchmark Algorithm

To quantitatively evaluate the performance of our EXPERTS algorithm, we implemented multiple algorithms as our performance benchmarks, including:

• Federated-MLP (with mean value imputation): a multilayer perceptron model under the federated learning framework. Additionally, MLP cannot handle the missing values, so we used the mean value for the imputation.

• XGBoost (without data imputation): an extreme tree-based model under a non-federated learning framework.

• MLP (with mean value imputation): MLP model under non-federated learning framework for data processing and the mean value is used for the imputation.

• Random forest (with mean value imputation): a tree-based model under the non-federated learning framework. Like MLP, the random forest cannot handle the missing values as well, so we used the mean value for imputation.

4.2 Results Analysis

4.2.1 Performance Evaluation of Federated-XGBoost

In these tests, we evaluated the performance of EXPERTS on the COVID-19 dataset. We used 70% of the samples as the training set, 20% of the samples as the validation set, and the rest as the testing set. Our approach can achieve 93% accuracy in predicting the patients' clinical courses. However, accuracy is not always enough to evaluate the clinical performance of the algorithm. We also employed the averaged Precision, Recall, and F1-score as performance metrics in our experiments. Precision and recall are both used to evaluate the quality of classification to show the accuracy of the model. More specifically, precision indicates the percentage of the relevant results retrieved, and recall refers to the percentage of the total relevant results correctly classified. The F1-Score is the harmonic mean of Precision and Recall. The results of predicting COVID-19 patients' clinical course are shown in Table 2.

Moreover, the results of the tests are plotted in Fig. 3, and Fig. 3(a) shows the areas under the receiver operator curves (AUROCs) for different COVID-19 patients. The accuracy for the mild, severe and critical patients reaches 0.994, 0.981 and

1.The studies involving human participants were reviewed and approved by the ethical committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, China. Informed patient consent was waived by the Ethics Commission due to the retrospective and observational nature of this study.
2.The dataset is available on the website: https://physionet.org/content/cves/1.0.0/.

▼Table 2. Classification results for EXPERTS

	Precision	Recall	F1-Score
Mild	0.96	0.91	0.93
Severe	0.94	0.96	0.95
Critical	0.89	0.87	0.88
Macro average	0.93	0.91	0.92
Weighted average	0.93	0.93	0.93

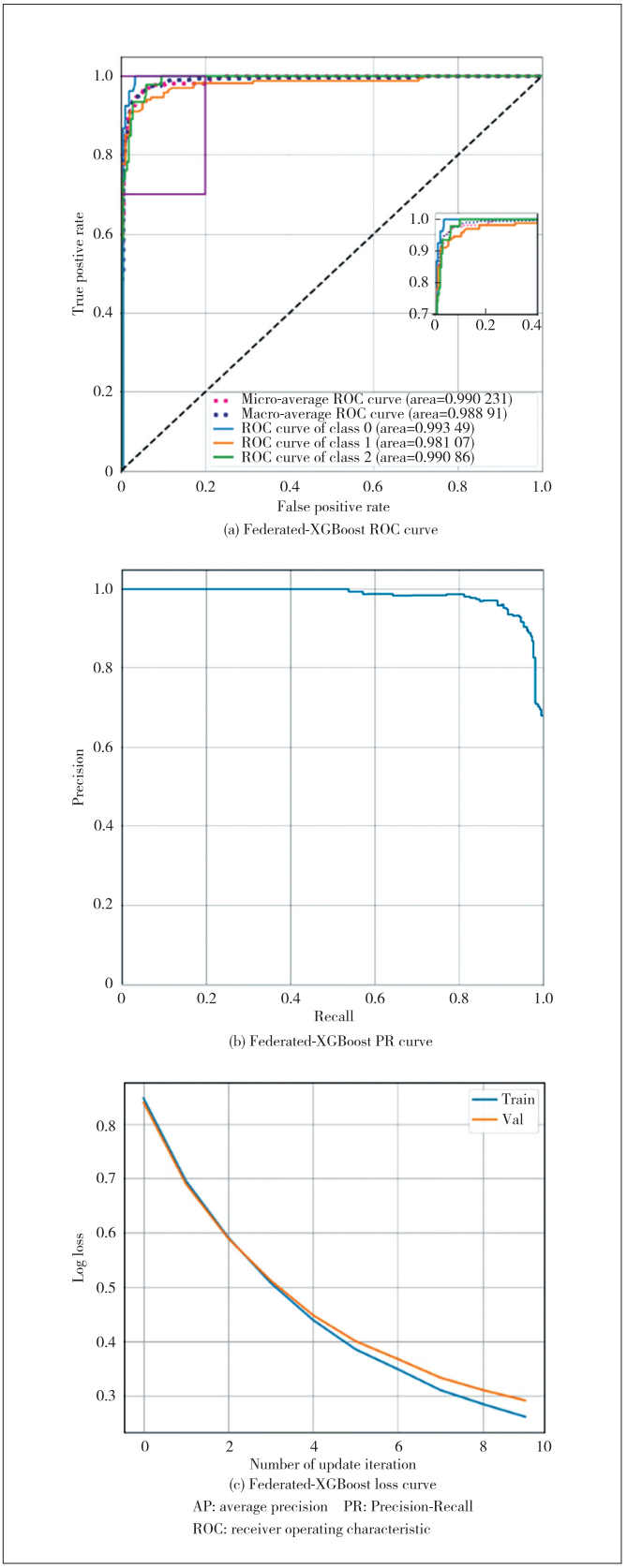
0.991, respectively. The embedded figure enlarges the details of the top left corner of Fig. 3(a). The PR curve is a non-decreasing function of the true positive rate (TPR) with respect to the false positive rate (FPR). The PR curve is shown in Fig. 3(b) with Precision as the *Y*-axis and Recall as the *X*-axis. It can be seen that when Recall is less than 0.5, Precision is always 1; while Recall is greater than 0.5, Precision gradually decreases from 1 to around 0.7.

The loss value of our iteration-like method is shown in Fig. 3(c). The *x*-axis of Fig. 3 (c) is the number of update iterations of the training, and the *y*-axis represents the loss function of our model. It can be seen from the figure that the loss value drops greatly in the initial iteration of the training stage, indicating that the learning rate is appropriate and the gradient descent process is carried out. After the six-th iteration, it can be seen obviously that the loss curve tends to be stable, and the change in the loss was not as obvious as in the beginning.

4.2.2 Feature Importance

In this experiment, we studied the model interpretability of EXPERTS by identifying the important features. We performed the averaging on the aggregated server to update the parameters gathered from local parties and returned the updated parameters to each party for the next iteration. In our tests, we found that the feature importance of EXPERTS derived from each party was the same no matter how the data varied. This suggests that EXPERTS can address the unevenly distributed datasets and avoid using the local optima for a global explanation.

Fig. 4 shows the top 20 important features and their individual contribution to the final diagnosis results. Figs. 4(a), 4 (b), and 4(c) represent that the summary plot of COVID-19 patients is in mild, severe, and critical status. In these figures, the *y*-axis lists the features in the reverse order of their importance from top to bottom, and the *x*-axis represents the SHAP value. Besides, the features that drive the prediction toward positive are in red, and those pushing the prediction negative are in blue. By reviewing the influence of the selected features in the model, it is obvious that age plays a crucial role for mild-symptom patients shown in Fig. 4(a) and severe-symptom patients shown in Fig. 4(b). The elder the age, the less likelihood for those patients to be less affected by COVID-19, and they will develop into a severe or worse situation. For the comorbidities, cancer could be a useful bio-marker to identify the risk of COVID-19 patients being mild shown in Fig. 4(a) or severe shown in Fig. 4(b). Cancer will increase the chance of poor



▲Figure 3. Performance metrics for federated XGBoost on COVID-19 cases

prognosis, but comorbidities are not the main factors for critically ill patients. The vital signal SpO2 also plays a key role in our experiments. The greater the minimum value of SpO2 in an observation period, the more likely for the patients to stay in a mild condition shown in Fig. 4(a). Otherwise, the possibility of turning severe condition is higher shown in Fig. 4(b). For those critically ill patients in Fig. 4(c), the SpO2 current reading becomes more important. Our findings are consistent with the earlier medical studies[29 - 30]. In summary, Fig. 4 shows the top 20 important features among all features in descending order of their mean absolute SHAP values, and plots their distribution across all predictions accordingly.

4.2.3 Model Generalization

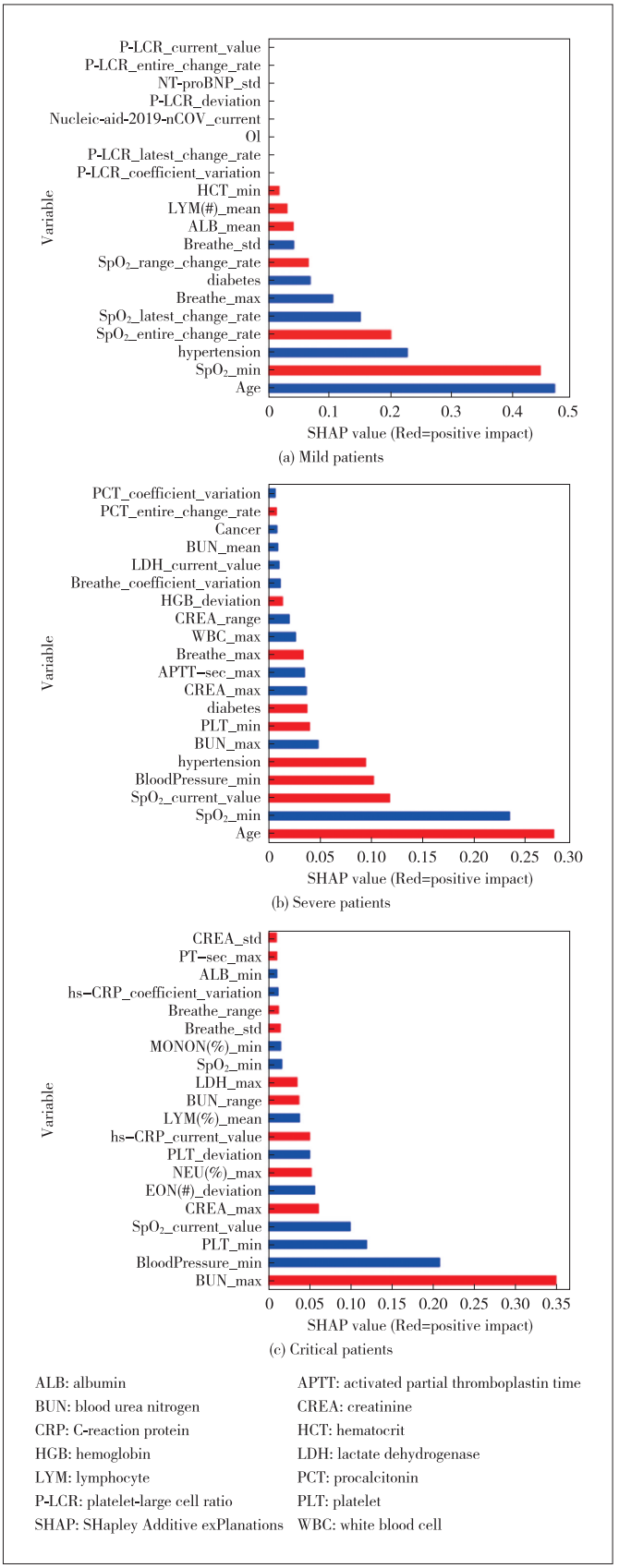
To prove the generalization of EXPERTS, we also tested it on a public dataset, called Cerebral Vasoregulation in the Elderly with Stroke. In Fig. 5(a), we can see that the area under the curve (AUC) achieves 70.8% after ten times the model update, and the PR curve can be seen in Fig. 5(b). Furthermore, in Fig. 5(c), it is obvious that within the process of the ten times iteration, the loss curve shows a non-increasing trend. EXPERTS cannot achieve the same performance level of the COVID-19 case as ML is restricted by the relatively small sample size. Meanwhile, we also showed the top 20 important features in Fig. 5 (d).

4.2.4 Performance Comparison with Benchmark Algorithms

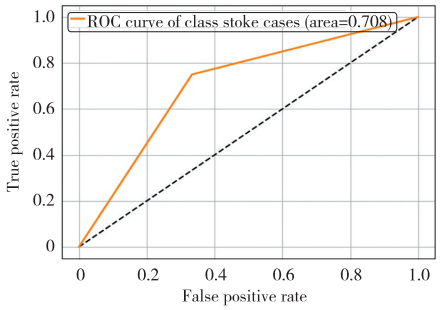
In this part, we compared EXPERTS with the selected benchmark algorithms, and all the experiments were performed on the COVID-19 dataset.

In our tests, the best accuracy of Federated-MLP can only reach 78% on the COVID-19 dataset, which lags far behind EXPERTS. A possible reason for this limited performance is that MLP cannot properly handle a large number of missing data. As we used the mean imputation method, it might change the distribution of the original data and affect the final performance. Another possible cause is that MLP is a relatively simple neural network, so its capacity is not as strong as the complex deep neural networks. Fig. 6 shows the learning results based on a fully connected neural network MLP which is typically simple under the framework of federated learning. As can be seen from Fig. 6(a), for different COVID-19 patients, their AUC can only reach 89%, 87%, and 92% respectively. Fig. 6(b) shows its PR curve, which is a non-increasing curve. As can be seen, within the interval of Recall from 0 to 1, the value of Precision drops from 1 to 0.3. Fig. 6(c) shows the loss curve of Federated-MLP within the process of iteration and model update, and the loss gradually decreases from 0.78 to 0.56 in these continuous update iterations.

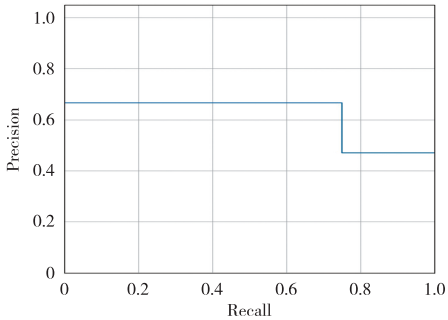
We also implemented three non-federated benchmark algorithms in our experiments, including two tree-based methods and one neural network method. We evaluated their perfor-



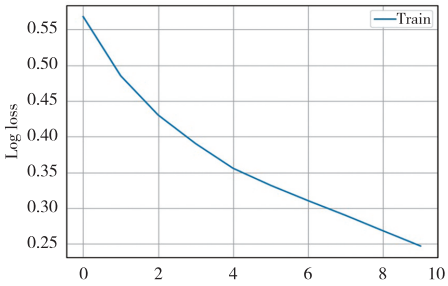
▲ **Figure 4. Feature importance among different patients types**



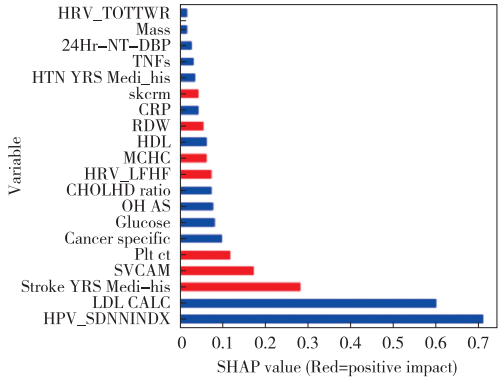
(a) Federated-XGBoost ROC curve



(b) Federated-XGBoost PR curve



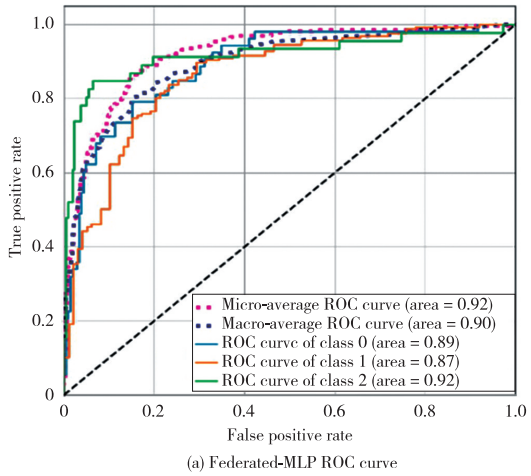
(c) Federated-XGBoost loss curve



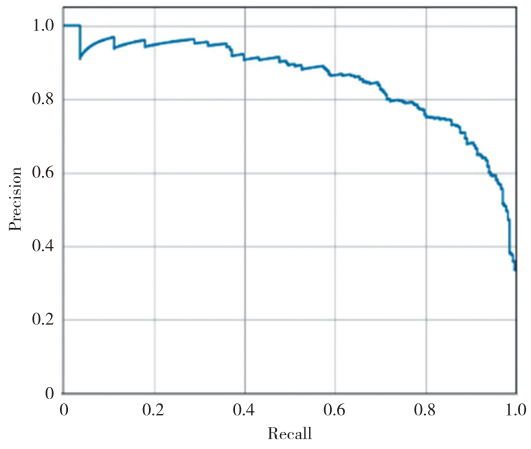
(d) Federated-XGBoost feature importance

CHOL: total cholesterol
CRP: C-reaction protein
HDL: high density lipoprotein
ROC: HRV: heart rate variability
LDL: low density lipoprotein
MCHC: mean corpuscular hemoglobin concentration
PLT: platelet
PR: Precision-Recall
ROC: receiver operating characteristic

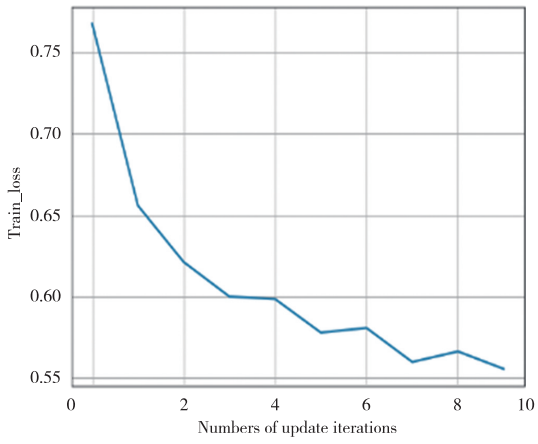
▲ **Figure 5. Performance evaluation of EXPERS on the stroke cases**



(a) Federated-MLP ROC curve



(b) Federated-MLP PR curve

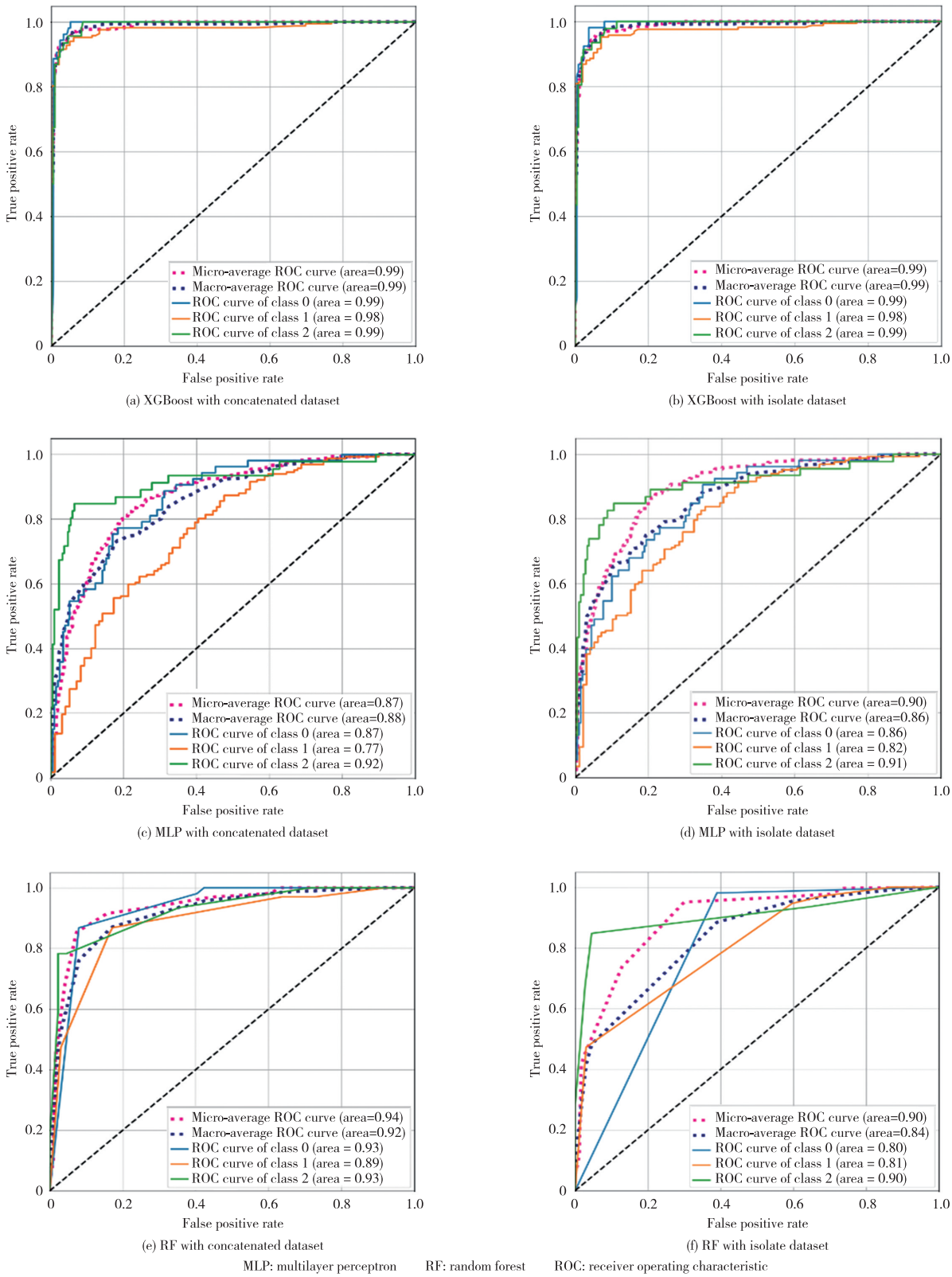


(c) Federated-MLP loss function

AP: average precision
MLP: multilayer perceptron
PR: Precision-Recall
ROC: receiver operating characteristic

▲ **Figure 6. Performance metrics for federated MLP**

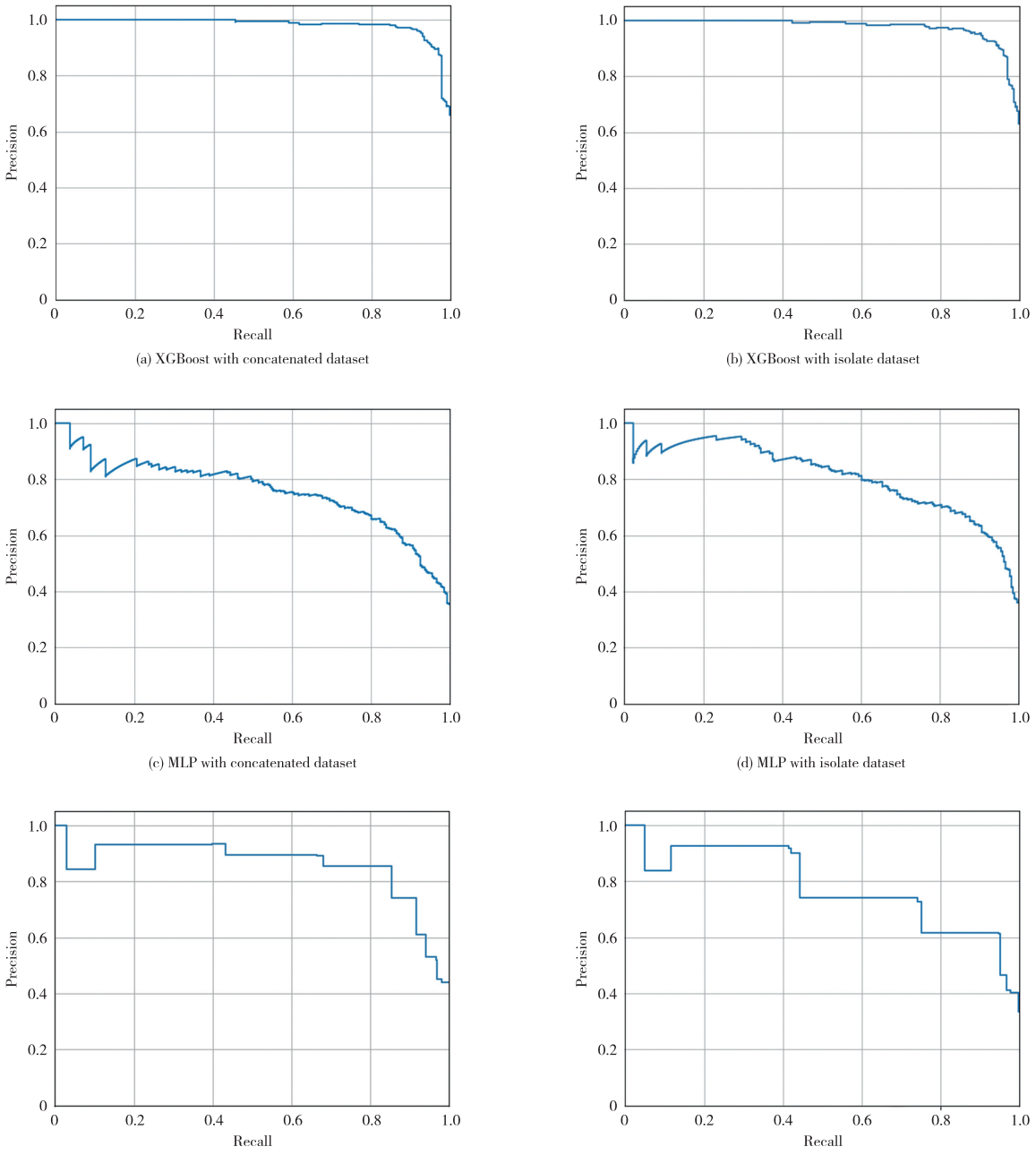
mance on the centralized dataset (bigger size) and the distributed local datasets (smaller size). Fig. 7 presents all ROC



▲**Figure 7. Comparison among ROCs**

curves for the benchmarks. Fig. 7(a) shows the result of the XGBoost algorithm on the centralized dataset. Correspondingly, Fig. 7(b) shows the result of the XGBoost algorithm on the distributed local dataset. It is easy to observe that XGBoost works well in both cases. Figs. 7(c) and 7(d) respectively show the performance of MLP on the centralized data set and the distributed local datasets. We can see that the learning effect of the distributed data sets is slightly worse than the centralized data set. Finally, in Figs. 7(e) and 7(f), we studied the performance difference of the random forest on different dataset cases. It shows that the performance of the random forest significantly drops when the size of the data reduces.

Fig. 8 shows the PR curves of all the benchmark algorithms without running under the federated learning framework, and



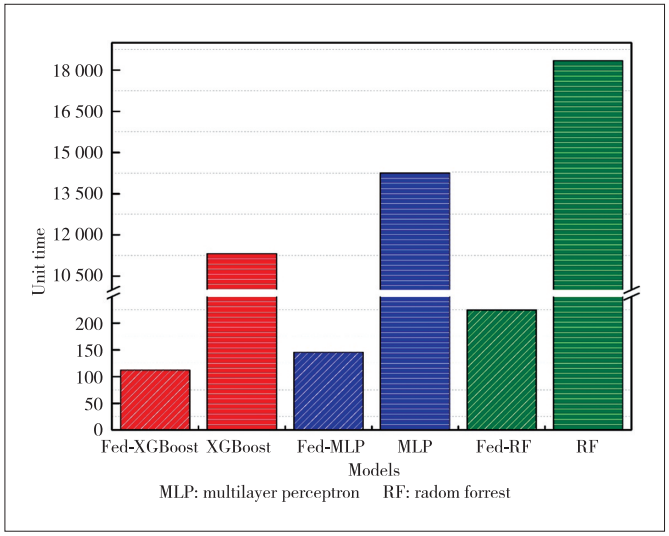
(a) XGBoost with concatenated dataset
(b) XGBoost with isolate dataset
(c) MLP with concatenated dataset
(d) MLP with isolate dataset
(e) RF with concatenated dataset
(f) RF with isolate dataset

AP: average precision MLP: multilayer perceptron RF: random forest

▲ **Figure 8. Comparison among Precision-Recall (PR) curves**

it can be seen that they are all non-increasing curves. Within the range of Recall from 0 to 1, Precision all reduces from 1 to less than 0.5 except for the two XGBoost cases. In addition, comparing them with the ROC curves shown in Figs. 3(a) and Fig 6(a) under the federated learning framework, we find that EXPERTS inherits the legacy of the XGBoost and can achieve the same performance of the XGBoost under the centralized data setting. This would suggest that EXPERTS can achieve the desired performance without sharing patients' data.

In addition to the excellent performance of learning, another advantage of federated learning is that it consumes fewer transmission resources. As depicted, the difference between federated learning and traditional centralized learning is that the original data transmission is replaced by the transmission of model parameters only, which greatly reduces the overload of data transmission and effectively improves the overall performance. From the perspective of time consumption, the results are shown in Fig. 9. It can be seen that no matter what model is used, the federated model is far superior to the one without the federated architecture.



▲**Figure 9. Overhead comparison among different models on COVID-19 cases**

5 Conclusions

The effective application of federated learning in the medical field is essential to address the security threats of personal medical data and the resource imbalance at all levels of hospitals. We show that EXPERTS could build a global model for COVID-19 patients' diagnoses without sharing their data. It also gives the leading factors of the COVID-19 patients in different statuses. To test the flexibility of EXPERTS that handles different medical applications, our system has also been verified with an open dataset for stroke. EXPERTS can adapt to the new application with traceable decision support, making it more suitable than static scoring that requires manual processing.

There are several limitations to this study. Our study is now designed as retrospective ones, and we will extend our framework to prospective studies. EXPERTS is tested as a horizontal federated learning model, and vertical federated learning should also be considered to further prove the reliability of the model since the features collected in different hospitals are usually not the same.

Acknowledgement

We are grateful to AMD Product (China) Co., Ltd. and the Sugon Information Industry Co., Ltd. for its X785-g30 series GPU server.

References

[1] DA SILVA D B, SCHMIDT D, DA COSTA C A, et al. DeepSigns: a predictive model based on deep learning for the early detection of patient health deterioration [J]. Expert systems with applications, 2021, 165: 113905. DOI: 10.1016/j.eswa.2020.113905
[2] YE B, YUAN X X, CAI Z C, et al. Severity assessment of COVID-19 based on feature extraction and V-descriptors [J]. IEEE transactions on industrial informatics, 2021, 17(11): 7456 – 7467. DOI: 10.1109/TII.2021.3056386
[3] NAPI N M, ZAIDAN A A, ZAIDAN B B, et al. Medical emergency triage and patient prioritisation in a telemedicine environment: a systematic review [J]. Health and technology, 2019, 9(5): 679 – 700. DOI: 10.1007/s12553-019-00357-w
[4] CORDERO A, GARCÍA-ACUÑA J M, RODRÍGUEZ-MAÑERO M, et al. Prevalence, long-term prognosis and medical alternatives for patients admitted for acute coronary syndromes and prasugrel contraindication [J]. International journal of cardiology, 2018, 270: 36 – 41. DOI: 10.1016/j.ijcard.2018.06.057
[5] FERRONI P, ZANZOTTO F M, RIONDINO S, et al. Breast cancer prognosis using a machine learning approach [J]. Cancers, 2019, 11(3): 328. DOI: 10.3390/cancers11030328
[6] CHOUDHURY O, GKOULALAS-DIVANIS A, SALONIDIS T, et al. Differential Privacy-enabled Federated Learning for Sensitive Health Data [EB/OL]. (2019-10-07)[2022-02-28]. https://arxiv.org/abs/1910.02578v2
[7] LIANG W, LI K C, LONG J, et al. An industrial network intrusion detection algorithm based on multifeature data clustering optimization model [J]. IEEE transactions on industrial informatics, 2020, 16(3): 2063 – 2071. DOI: 10.1109/TII.2019.2946791
[8] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. AISTATS, 2017
[9] LI L, FAN Y X, TSE M, et al. A review of applications in federated learning [J]. Computers & industrial Engineering, 2020, 149: 106854
[10] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions [J]. IEEE signal processing magazine, 2020, 37 (3): 50 – 60. DOI: 10.1109/MSP.2020.2975749
[11] BRISIMI T S, CHEN R D, MELA T, et al. Federated learning of predictive models from federated Electronic Health Records [J]. International journal of medical informatics, 2018, 112: 59 – 67. DOI: 10.1016/j.ijmedinf.2018.01.007
[12] KAISSIS G A, MAKOWSKI M R, RÜCKERT D, et al. Secure, privacy-preserving and federated machine learning in medical imaging [J]. Nature machine intelligence, 2020, 2(6): 305 – 311. DOI: 10.1038/s42256-020-0186-1
[13] DOSILOVIC F K, BRCIC M, HLUPIC N. Explainable artificial intelligence: a survey [C]//Proceedings of 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

IEEE, 2018. DOI: 10.23919/mipro.2018.8400040
[14] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions [EB/OL]. [2022-02-28]. https://arxiv.org/abs/1705.07874
[15] STOJIĆ A, STANIĆ N, VUKOVIĆ G, et al. Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition [J]. Science of the total environment, 2019, 653: 140 – 147. DOI: 10.1016/j.scitotenv.2018.10.368
[16] NAN Y C, LI W, LU F, et al. Developing practical multi-view learning for clinical analytics in P4 medicine [J]. IEEE transactions on emerging topics in computing, 2021: 1. DOI: 10.1109/tetc.2021.3054761
[17] RIEKE N, HANCOX J, LI W Q, et al. The future of digital health with federated learning [J]. NPJ digital medicine, 2020, 3(1): 119. DOI: 10.1038/s41746-020-00323-1
[18] SCHWARZ C G, KREMERS W K, THERNEAU T M, et al. Identification of anonymous MRI research participants with face-recognition software [J]. The New England journal of medicine, 2019, 381(17): 1684 – 1686. DOI: 10.1056/nejmc1908881
[19] SHELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data [J]. Scientific reports, 2020, 10(1): 12598. DOI: 10.1038/s41598-020-69250-1
[20] SILVA S, GUTMAN B A, ROMERO E, et al. Federated learning in distributed medical databases: meta-analysis of large-scale subcortical brain data [C]// IEEE 16th International Symposium on Biomedical Imaging. IEEE, 2019: 270 – 274. DOI: 10.1109/ISBI.2019.8759317
[21] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics [J]. Journal of healthcare informatics research, 2021, 5(1): 1 – 19. DOI: 10.1007/s41666-020-00082-4
[22] WANG G. Interpret federated learning with shapley values [EB/OL]. [2022-02-28]. https://arxiv.org/abs/1905.04519
[23] RYZIN J V, BREIMAN L, FRIEDMAN J H, et al. Classification and regression trees [J]. Journal of the American statistical association, 1986, 81(393): 253. DOI: 10.2307/2288003
[24] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785 – 794. DOI: 10.1145/2939672.2939785
[25] SHAPLEY L S. A value for n-person games [EB/OL]. [2022-02-28]. https://www.rand.org/pubs/papers/P295.html
[26] LUNDBERG S M, ERION G G, LEE S I. Consistent individualized feature attribution for tree ensembles [EB/OL]. [2022-02-28]. https://arxiv. org/abs/1802.03888
[27] WHO. Coronavirus disease 2019 (COVID-19): situation report [R], Geneva, Switzerland: WHO, 2020
[28] NOVAK V, HU K, DESROCHERS L, et al. Cerebral flow velocities during daily activities depend on blood pressure in patients with chronic ischemic infarctions [J]. Stroke, 2010, 41(1): 61 – 66. DOI: 10.1161/STROKEAHA.109.565556
[29] FERRARI D, MOTTA A, STROLLO M, et al. Routine blood tests as a potential diagnostic tool for COVID-19 [J]. Clinical chemistry and laboratory medicine (CCLM), 2020, 58(7): 1095 – 1099. DOI: 10.1515/cclm-2020-0398
[30] AN X S, LI X Y, SHANG F T, et al. Clinical characteristics and blood test results in COVID-19 patients [J]. Annals of clinical and laboratory science, 2020, 50(3): 299 – 307

Biographies

NAN Yucen (yucen.nan@sydney.edu.au) received her PhD and MPhil degrees from University of Sydney, Australia in 2022 and 2017. She is currently a lecturer in the College of Intelligence Science and Technology, National University of Defense Technology, China. Her current research interests are in the area of edge computing and the Internet of Things.

FANG Minghao received his MD degree from Huazhong University of Science and Technology, China in 2006. He is currently an associate professor with Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. He has worked in emergency and critical care medicine for 20 years. His research interests include the diagnosis and treatment of critical respiratory and cardiovascular disease.

ZOU Xiaojing received her MD degree from Tongji Medical College, Huazhong University of Science and Technology, China in 2011. She is an associate chief physician in Emergency Department and Intensive Care Unit of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. Her research interests are in sepsis and the application of artificial intelligence in critical diseases.

DOU Yutao received his BE degree in software engineering from University of Canberra, Australia in 2020. He is currently working toward the master of philosophy degree with the University of Sydney, Australia. His research interests mainly include distributed computing, bioinformatics, and artificial intelligence.

Albert Y. ZOMAYA is the chair professor of high performance computing & networking in the School of Computer Science and Director of the Center for Distributed and High Performance Computing at the University of Sydney, Australia. He has published more than 600 scientific papers and is the (co-)author/editor of more than 30 books. As a sought-after speaker, he has delivered more than 190 keynote addresses, invited seminars, and media briefings. His research interests span several areas in parallel and distributed computing and complex systems. He is currently the Editor in Chief of the ACM Computing Surveys and served in the past as Editor in Chief of the *IEEE Transactions on Computers* (2010—2014) and the *IEEE Transactions on Sustainable Computing* (2016—2020).