IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON COMMUNICATIONS IN HARSH ENVIRONMENTS

Received December 23, 2020, accepted December 30, 2020, date of publication January 12, 2021, date of current version February 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051215

Ensemble Learning Based Postpartum Hemorrhage Diagnosis for 5G Remote Healthcare

YAWEI ZHANG, XIN WANG, NINGYU HAN, AND RONG ZHAO

Obstetrics Department, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100026, China

Corresponding author: Xin Wang (wx1501@ccmu.edu.cn)

This work was supported by the Beijing Obstetrics and Gynecology Hospital, Capital Medical University, under Grant FCYY201914.

ABSTRACT The fifth-generation (5G) communications enables various promising applications that was once impossible, e.g. remote healthcare with the help of fast and reliably delivery of medical data. Post-partum hemorrhage (PPH) refers to the massive blood loss after the birthing stage (within 24 hours), i.e. >500ml for the vaginal delivery, and >1000ml for the cesarean section. PPH is by far the most common cause of the mortality rate of pregnant women, as well as a primary cause of current pregnant mortality in China. Despite the great potential of prediction of PPH, there is currently no effective tool based on the limited raw data from the clinical trials. In the study, we retrospectively study the 3842 vaginal delivery cases in 2017 collected from Beijing Obstetrics and Gynecology Hospital, Capital Medical University. In particular, we obtain the prediction based diagnostic model relying on machine learning, and we adopt the ensemble learning to accomplish this task, by combining the results of various candidate methods. According to the experimental results, the accuracy of correct PPH diagnosis would approach 96.7%; the total disseminated intravascular coagulation (DIC) prediction accuracy approaches 90.3%. In this regard, we may conclude the proposed model based on machine learning would allow us to predict successfully the risk of PPH, and assess the critical level of PPH patient. We anticipate our study results would contribute to the reduction the mortality of pregnant women.

INDEX TERMS Postpartum hemorrhage, prediction model, automated diagnosis, ensemble learning, random forest, machine learning.

I. INTRODUCTION

The fifth-generation (5G) communications allow for high-speed and ultra-reliable data transmissions [1]–[3], which would boost various new demands and emerging applications [4], [5]. Recently, the Internet of Medical Things (IMT) has attached general interests in both academy and industry [6]. Combined with artificial intelligence (AI) and machine learning (ML), the remote healthcare thus provides the great promise to the remote medical decision in many critical situations, e.g. remote patient monitoring and remote medical learning [7]. Among them, the postpartum hemorrhage (PPH) diagnosis is one of such important scenarios.

The American College of Obstetricians and Gynecologists (ACOG) defines PPH as a blood loss of > 500 mL in the case of vaginal delivery, or > 1000 mL in the cesarean section within 24 hours [8]. The World Health Organization (WHO) statistical analysis suggests that PPH remains the leading direct cause of maternal death worldwide, contributing to 27.1% of total maternal deaths [9]. Moreover, PPH causes several serious complications, such as shock, disseminated intravascular coagulation (DIC), and so on. Most of the PPH induced maternal deaths are relevant to the delay of clinical diagnosis. To this end, the early stage prediction of PPH and its complications is of great importance to reduce the maternal death ratio for obstetricians, which allows the obstetricians to provide the timeliest medical treatments for pregnant patients with a high risk of PPH. ACOG also suggested to use the risk assessment tool to predict the occurrence of PPH [10].

The risk associated factors attributed to PPH have been extensively researched based on conventional statistical methods [11]–[18]. Logistic regression model or lasso regression model was used for the PPH risk prediction [11], [12]. In particular, lasso/logistic regression models show low-but-good discriminative ability [12]. Based on maternal clinical characteristics and medical history, a risk score was used for prediction of PPH [13], [14]. Although the conventional prediction methods are proved to be effective, the performance

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

of prediction is unsatisfied: only 60% of women with high PPH risk are identified, whilst the other 40% of women who had PPH are not identified in advance [19].

With the recent advance of artificial intelligence (AI), various machine learning-based models have been developed in the medical fields. For instance, the performance of ensemble learning was assessed for automated diagnosis of breast cancer using an open access dataset in [20]. Predictive models have been studied for cancer diagnosis using Support Vector Machines (SVMs) are developed in [21]. Prediction of heart disease risk using classification and regression tree (CART) was developed in [22]. A homogeneous ensemble is then created from the different CART models using an accuracy based weighted aging classifier ensemble, which is a modification of the weighted aging classifier ensemble (WAE). In our concerned PPH prediction field, only a few machines learning based methods were studied [23]–[25]. A fuzzy expert system to predict the risk of developing PPH was developed in [23]. Besides, another Mamdani inference was used to simulate the experts reasoning and thereby enables the predictionary analysis. According to the previous study on this topic, we unfortunately find that the attained PPH prediction recall (sensitivity) would even approach 87.48%, which is far from reliable for automated analysis and practical deployment in the medical diagnosis. It is reported that XGBoost would perform better than logistic regression and Artificial Neural Network (ANN) for the repeat cesarean delivery in [24], [25]. However, their datasets were only for cesarean delivery and did not cover the vaginal delivery, and the ensemble learning method was not explored.

Meanwhile, it should be noted that such machine learning models also incurs the stringent requirements on the amount of experimental data. Unlike the other fields whereby the observation of data is relatively less expensive, for the considered PPH problem the collection of clinical data is extremely time consuming. When the dataset is not large enough, then we can expect the deduced prediction model would be less reliable for practical applications.

In this paper, we study the ensemble learning in the context of PPH and thereby construct the complication prediction model, enabled by the recent advances on 5G communication and machine learning. The 5G communications would greatly facilitate the high-speed and highly reliable data collection via remote monitoring, which the new ML paradigm inspire us to develop new efficient methods on highly accurate automatic diagnosis. The main contribution of this work contain two folds. First, we collect a total of 3842 vaginal delivery cases in 2017 from Beijing Obstetrics and Gynecology Hospital, Capital Medical University. This large dataset potentially allows us to derive a reliable machine learning prediction model. Second, targeting at improving the PPH and its complication prediction performance of base learners, we carefully adapt an ensemble learning scheme to handle realistic challenges in the model accuracy, especially for the dataset with imbalanced samples as in our study (e.g., the positive samples are dramatically smaller than the negative ones). In particular, the selection of base learners in ensemble learning (e.g. ANN, SVM, regression, etc.), which is of great importance for the performance, has been rarely exploited in the literatures on PPH data analysis. To address this practical difficulty, we construct different EL schemes for both the PPH and DIC tasks, based on their features and limitations.

We collected 23 PPH-relevant features of each patient as the input for our PPH prediction model. The importance of the input features is also studied and ranking of the features is obtained. For the base learners, after carefully evaluating the most popular ML methods, we use the Random Forest (RF), Extreme Gradient Boosting (XGB), Gradient Boosting Decision Tree (GBDT) and support vector machine (SVM) as the base learners,[1] based on the 3842 records dataset. On this basis, we explored the use of averaging and voting ensembles to improve predictive performance. In addition, the prediction performance of PPH complications, namely DIC, is evaluated. As demonstrated by our experimental results, the accuracy of correct PPH predictive diagnosis would surpass 90%; the total DIC prediction accuracy approaches 90.3%. Other performance metrics used for our imbalanced samples, e.g. the recall ratio, the F-measure as well as the Matthews correlation coefficient (MCC), are also investigated. In this regard, we would conclude our proposed model based on well-designed ensemble learning allows us to predict successfully the risk of PPH, and assess the critical level of PPH patients. We anticipate our study results would contribute to the reduction the mortality of pregnant women.

In conventional prediction methods, many of the risk factors have a low value and may not be effective among the hybrid risk factors [26]. However, we find that ensemble learning can avoid this problem, a feature with binary value (1 or 0) among other features with large values can also be identified as the most important feature to predict PPH. It should be emphasized that, although the multiple learners-based ensemble learning is not a new idea for the ML community, the application of it to such new medical diagnosis problems, especially with the collected real PPH dataset, has not been reported in this field. As shown by our performance comparison with regards to ANN and logistic regression, we finally valid the advantage of our suggested models in two specific medical diagnosis tasks (PPH and DIC predictive classification), and we anticipate this could provide insights to widespread medical diagnosis applications.

The rest of this paper is organized as follows. In Section II, the collect dataset and the selected feature (for both PPH and DIC) are described. In Section III, the base learners used in our ensemble learning to predict PPH and DIC are shortly introduced and evaluated. Then, in Section IV we describe a modeling framework for the ensemble learning. In Section V we provide the numerical evaluations results of both classical

[1] We have also studied the popular artificial neural network (ANN), i.e. the multiple layer perceptron (MLP). Unfortunately, for this specific problem, we found it was not applicable (due to the poor performance) and hence was excluded from the final ensemble learning.

base learners and the designed ensemble learning, and compare our proposed scheme with the previously used methods (such as ANN and logistic regression). Finally, we conclude our study in this work.

II. DATASET FEATURE SELECTION

In the work, our study aims to construct a prediction model of PPH and its complications based on the method of ensemble learning. Our prediction results are expected to give the obstetricians necessary time to deal with the potential PPH, and therefore the proactive treatment can be carried out in advance, such as appropriate hemostasis, timely fluid resuscitation, massive transfusion protocol, and tranexamic acid.

There are dozens of (or even more) associative features relevant to each patient. In practical clinic trails, collecting all the potential features for the prediction model would become rarely feasible for our large dataset involving 3842 records. Thus, the associative features shall be carefully selected, to balance the potential information loss and the complexity.

The widely accepted rule in feature engineering is that more features do not necessarily leads to the improved prediction accuracy.

A. PPH FEATURES

For the primary goal of PPH prediction, our clinical team maintain a large dataset consisting of 3842 patient records, with 361 PPH records which were collected during 2017 at the Beijing Obstetrics and Gynecology Hospital. The dataset, as common cases, is also characterized by the significant imbalance positive/negative samples. I.e., the true PPH occurrence ratio accounts for 9.4%, which is a moderate value for the patient class of vaginal delivery.

To achieve better prediction performance, the assessment indicators of PPH and its complication DIC have been systematically reviewed, by comprehensively exploring their relationship with the blood loss [27], [28]. On this basis, we have further studied and selected 23 features with high relation to the occurrence of PPH. These features are categorized into ''Present Gestation'' and ''Factors related to delivery'', as shown in Table 1. The data formats and categories of such selected features are described by Table 1. We use f0 to f22 to denote the 23 features in Table 1 sequentially. It should be noted that, according to the ranking analysis of such features, we believe such features constitute a relatively complete description or representation of PPH, which are expected to produce the good prediction diagnosis model.

B. DIC FEATURES

DIC is a typical complication developed following PPH. Based on the 361 patients who have PPH, a total number of 212 PPH patients were included in the DIC dataset, among which only 7 PPH patients presented the DIC complication.

For the DIC dataset, we have also carefully selected 19 features, as shown in Table 2. We use f0 to f18 to denote the 19 features shown in Table 2 sequentially.

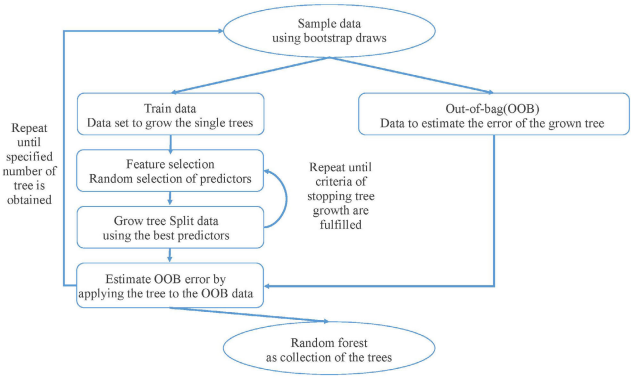

FIGURE 1. Random forest algorithm.

III. BASE LEARNERS FOR ENSEMBLE LEARNING

As discussed, the machine learning algorithms are utilized to perform automated diagnosis for PPH and DIC. To be specific, we adopt the ensemble learning in this paper, with four base learners of random forest (RF), gradient boosting decision tree (GBDT), XGBoost (XGB) and SVM. As such, our ensemble learning collects different merits of four methods, and enable the more reliable prediction model. The well-known ANN and logistical regression (LR) are not included in this section, but we have tested their performances in Section V.

A. RANDOM FOREST

Random forest model is widely used for classification. In principle, such a random forest model is one bagging-type ensemble (collection) of decision trees, which trains several trees in parallel and thereby uses the majority decision of the trees as the final decision of whole forest model. Usually, individual decision tree model is easy to interpret, but the whole model is nonunique and exhibits high variance.

Random Forest combines the two concepts of Bagging and Random Selection of Features by generating a set of T regression trees. The algorithm flow of RF algorithm is illustrated in Fig. 1, whereby the sample data is obtained using bootstrap and the splitting predictor is selected from a randomly selected subset of predictors [29]. Each tree thus constitutes a standard classification or regression tree (CART) which uses the so-called decrease of Gini impurity (GI) as a splitting criterion. In practice, the GI is computed by:

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = 1 - \sum_{i=1}^{J} p_i^2, \quad (1)$$

where J is the total number of classes, p_i is the fraction of items labeled with class i in the set $\mathcal{T}$, $i \in \{1, 2, \ldots, J\}$. Meanwhile, the out-of-bag (OOB) error serves as an important feature of RF, which is simply the average error frequency obtained when the observations from the data set are predicted using the trees for which they are OOB – they are not used to construct the trees.

TABLE 1. Selected risk factors causing PPH.

Category	Present Gestation						
Variable	Age	Number of pregnancies	Number of deliveries	Gestational age	Anemia		Hemoglobin(g/L)
Variable Value	-	-	-	-	No (≥110.0 g/L)	Yes (<110.0 g/L)	-
ARV (Actual Range of Variable)	-	-	-	-	0	1	-

Category	Factors related to delivery				
Variable	The character of amniotic fluid				
Variable Value	Clear amniotic fluid	amniotic fluid meconium I	amniotic fluid meconium II	amniotic fluid meconium III	Bloody amniotic fluid
ARV	0	1	2	3	4

Category	Factors related to delivery					
Variable	Methods of induced labor				Frequency of vaginal medication	Days of oxytocin application
Variable Value	No	Prostaglandins	Oxytocin	Prostaglandins+ Oxytocin	-	-
ARV	0	1	2	3	-	-

Category	Factors related to delivery								
Variable	Application time of Oxytocin (min)	Premature rupture of membrane		Artificial rupture of membranes		Labor analgesia		Pethidine hydrochloride	
Variable Value	-	No	Yes	No	Yes	No	Yes	No	Yes
ARV	-	0	1	0	1	0	1	0	1

Category	Factors related to delivery					
Variable	Application of Oxytocin during labor				Time of Oxytocin use during labor (min)	First stage of labor (min)
Variable Value	No	Latent phase	Active phase	Discontinuous application	-	-
ARV	0	1	2	3	-	-

Category	Factors related to delivery							
Variable	Second stage of labor (min)	Total stage of labor (min)	Method of delivery			Uterine inertia		Neonatal weight (g)
Variable Value	-	-	Damage of soft birth canal	Episiotomy	Forceps delivery	No	Yes	-
ARV	-	-	0	1	2	0	1	-

TABLE 2. Selected risk factors causing DIC.

Variable	Amount of bleeding (ml)	Hematocrit (HCT) before delivery (%)	Hemoglobin (HGB) before delivery (g/L)	Platelet before delivery ($\times 10^9$/L)	Fibrinogen before delivery (g/L)	D-Dimer before delivery (ng/mL)
Variable Value	-	-	-	-	-	-
ARV	-	-	-	-	-	-

Variable	Packed Red Blood Cell (mL)	Plasma(mL)	Platelet	Fibrinogen (g/L)	Prothrombin complex (IU)	Massage of uterus		Carboprost Tromethamine injection (ug)
Variable Value	-	-	-	-	-	No	Yes	-
ARV	-	-	-	-	-	0	1	-

Variable	Tranexamic acid (mL)	Hem coagulase (U)	Carbetoxin (mL)	Crystal liquid (mL)	Colloidal liquid (mL)	Liquid consumption (mL)
Variable Value	-	-	-	-	-	-
ARV	-	-	-	-	-	-

B. GRADIENT BOOSTING

Gradient enhancement is a kind of machine learning techniques used for regression and classification problems, its weak prediction model (usually the decision tree) generated forecast model in the form of collection. It likes other strengthening methods, building the model in the form of stage, and by allowing the optimization of the loss function of arbitrary separable variables to a generalized model. Thus, the generic gradient boosting model is specifically described in ***Algorithm 1*** [30]. The negative gradient $g_t(x)$ along the

observed data is denoted as:

$$g_t(x) = E_y\left[\frac{\partial \psi(y, f(x))}{\partial f(x)}\right]_{f(x)=\widehat{f}_{t-1}(x)} \quad (2)$$

where $f_t(x)$ is the estimated function at the t-th iteration. Because it is rarely feasible to identify one general solution for a boost increment in the functional space, one may alternatively choose another new base-learner function $h(x,\theta)$ increment which is expected to be mostly correlated with $-g_t(x)$. According to eq. (2), we find the best gradient descent step-size ρ_t, and then the function estimate is updated by eq. (3).

Algorithm 1 Gradient Boosting

Input:
input data $(x, y)_{i=1}^{N}$
number of iterations M
choices of the loss-function $\psi(y, f)$
choice of the base-learner model $h(x, \theta)$
Algorithm:
1. initialize $\hat{f}_0$ with a constant
2. **for** t=*1* to M **do**
3. compute the negative gradient $g_t(x)$
4. fit a new base-learner function h(x, θ_t)
5. find the best gradient descent step-size ρ_t:

$$\rho_t = arg \min_{\rho} \sum_{i=1}^{N} \psi[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_i)] \quad (3)$$

6. Update the function estimate:

$$\widehat{f}_t \leftarrow \widehat{f}_{t-1} + \rho h(x_i, \theta_t) \quad (4)$$

7. **end for**

C. EXTREME GRADIENT BOOSTING

Extreme Gradient Boosting (XGBoost or XGB) is one popular supervised learning algorithm that implements a process called boosting to yield accurate models. Boosting refers to the ensemble learning technique of building many models sequentially, with each new model attempts to correct for the deficiencies in the previous attained model. In the tree boosting, each new model that is added to the ensemble is a decision tree. XGBoost thus provides the parallel tree boosting (also known as GBDT, GBM) that is efficient to many data science problems in a fast and accurate way. For many problems, XGBoost is one of the best gradient boosting machine (GBM) frameworks today.

As shown in Fig. 2, at each iteration of gradient boosting, the residual will be used to correct the previous predictor that the specified loss function can be optimized [31]. Since the base model is decision tree, the output of model $\widehat{y}_i$ is obtained by a collection F of k trees:

$$\widehat{y}_i = \sum_{i=1}^{k} f_k(x_i), f_k \in F. \quad (5)$$

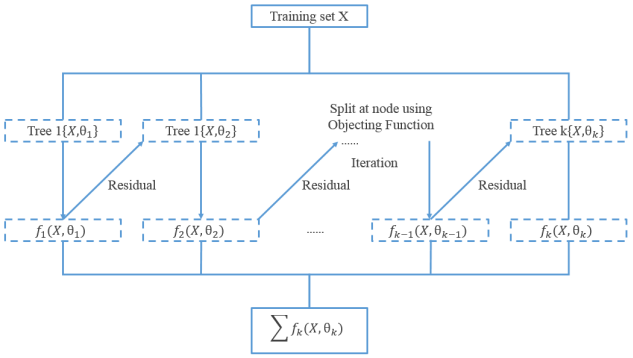

FIGURE 2. Flow chart of extreme gradient boosting.

Compared with the general gradient boosting method, XGBoost has more complex objective function and, in particular, it involves the additional regularization to further improve performance, which is given by:

$$\begin{aligned} J^{(t)} &= \sum_{i=1}^{n} L(y_i, \widehat{y}_i) + \sum_{k=1}^{t} \Omega(f_k), \\ &= \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \end{aligned} \quad (6)$$

where $J^{(t)}$ denotes the objective function at the t time iteration, and the n is number of predictions. As described in [31], the objective function will be further written as:

$$\begin{aligned} J^{(t)} &\approx \sum_{i=1}^{n}\left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega_{q(x_i)}^2\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T}\left[(\sum_{i\in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)\omega_j^2\right] + \gamma T, \end{aligned} \quad (7)$$

where g_i and h_i denote first and second-order Taylor expansion coefficients, respectively. The number of leaf nodes is T, and the decision tree is composed of a vector of values $w \in \mathrm{R}^T$ corresponding to all leaf node, $\mathcal{I}_j = \{i|q(x_i) = j\}$, is defined as the set of all training samples divided into leaf nodes j. Hence, the optimization of objective function can be transformed into a problem of finding the minimum of a quadratic function.

Owing to the regularization when optimizing the objective function, a trained predictive classifier is very robust to the overfitting.

D. SUPPORT VECTOR MACHINE

Another popular method is the SVM, which has been demonstrated to be effective in classification/prediction problem. Unlike the aforementioned base learning methods, SVM aims to minimize the empirical risk in deriving the classification bound in the high- dimensional feature space. In particular, SVM constitutes the following quadratic optimization

problem:

$$W(\alpha) = -\sum_{i=1}^{k}\alpha_i + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} y_i y_j \alpha_i \alpha_j h\left(\mathbf{x}_i, \mathbf{x}_j\right)$$
$$s.t. \sum_{i=1}^{k} y_i \alpha_i = 0, \quad \forall i : 0 \le \alpha_i \le C \tag{8}$$

Here, α is the vector of k variables, which is determined by minimizing the cost $W(\alpha)$; each element α_i corresponds to a training example $(\mathbf{x}_i, y_i)$; $h(\mathbf{x}_i, \mathbf{x}_j)$ is the nonlinear kernel function, such as the Gaussian function (in the simplest linear case, $h(\mathbf{x}_i,\mathbf{x}_j)$ gives the dot product of $\mathbf{x}_i$, and $\mathbf{x}_j$.

In this way, SVM searches for a hypersurface to maximize the minimal distance to different samples. As a result, SVM is capable of minimizing the empirical risk on training data, and thereby obtains the optimal accuracy on test data ever saw.

IV. ENSEMBLE LEARNING

In order to attain the highest accurate mode which outputs the highest probability of the positive class prediction, we combine the results based on a well-known *voting* principle [32]. That means, based on the common voting rule, the final output corresponds to the dominant output results of the RF, GBDT, XGB and SVM base learners. In this way, we can avoid the less convincing result and attain the highest probability of the positive class prediction. That means, if the proposed automated diagnosis model gives a ''1'' prediction for a patient, then the probability that PPH occurs on this patient with the probability approximating 1. For each base learner (RF, GBDT, XGB or SVM), we used a *grid-search* method to obtain the optimal value for the main hyperparameters.

For PPH prediction, since the prediction target is the blood loss volume of each patient, each base learner (RF, GBDT or XGB) outputs a blood loss volume prediction result. If these blood loss volume prediction results are combined softly first and then compared with the PPH threshold (500 mL), we call it Softly Combined Ensemble Leaning (EL-SC); otherwise, if the blood loss volume prediction results of the base learners are decided to be true or false hardly using the PPH threshold (500 mL) and then the resultant binary prediction results are combined, we call it Hardly Combined Ensemble Leaning (EL-HC). We compare the performances of two kinds of EL combination scheme in Section V.

For SVM, only binary PPH results are used as the training data and test data, which means the blood loss volume data are compared with the threshold (500 mL) first.

For the DIC prediction, since the prediction target is that whether DIC occurs or not, the binary prediction results of all base learner are combined directly.

V. EXPERIMENTAL RESULTS

The effectiveness of our designed ensemble learning method for both the PPH and DIC prediction is verified by our collected clinical datasets in Beijing Obstetrics and Gynecology Hospital. The numerical evaluation is based on the Python 3.8 platform.

A. PERFORMANCE EVALUATION FOR PPH PREDICTION

Our collected PPH dataset consists of 3842 records. In the analysis, 3500 records that are 65% of all records are used as training dataset; while the remaining 1342 records are used as the test dataset.

Among the 3842 records, there are 361 positive PPH patients. The training dataset contains 2500 records with 242 positive PPH instances, whilst the testing dataset contains 1342 records with total 119 positive PPH instances. The ratios of the positive instances for the training dataset (8.87%) and the testing dataset (9.68%) are approximately equal.

In order to evaluate the performance of the compared base learning methods and ensemble learning, two well-known performance measures in classification were used, especially for the imbalance data set as in our cases. These are the commonly used classification accuracy, F-measure and MCC.

Classification accuracy (A) is the ratio of true positives and true negatives obtained by the designed classifier over the total number of records in the test dataset, as given by (9)

$$A = \frac{TN + TP}{TP + FP + FN + TN} \tag{9}$$

Here, TN, TP, FP and FN denote the number of true negatives, true positives, false positives and false negatives, respectively, as also shown in Table 3.

TABLE 3. Description of TN, TP, FP and FN.

True Positive (TP)	False Negative (FN)	Actual Positive (TP+FN)
False Positive (FP)	True Negative (TN)	Actual Negative (FP+TN)
Predicted Positive (TP+FP)	Predicted Negative (FN+TN)	

Recall ratio (R) is defined as the proportion of the true positives against the true positives and false negatives, which is given by (10),

$$R = \frac{TP}{TP + FN} \tag{10}$$

Precision ratio (P) accounts for the proportion of the true positives against the true positives and false positives, which is given by (11),

$$P = \frac{TP}{TP + FP} \tag{11}$$

For the most dataset with imbalance class samples, the F-measure is more useful as it acts as the harmonic mean of precision and recall which is given by (12). In practice, the F-measure takes values in [0, 1] interval and the values of

F-measure closer to 1 should indicate the better classification performance.

$$F = \frac{2PR}{P+R} \tag{12}$$

Matthews correlation coefficient (MCC) is in essence a correlation coefficient between the observed and predicted binary classifications, which is given by (13); it returns a value between -1 and $+1$. Using MCC as a metric would be more meaningful for our imbalanced dataset.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{13}$$

Since a high precision value (P) results in a small value of FP, which means no false positive (FP) instances in the prediction results. This would be very useful to allow the obstetricians to accept the prediction results with a high probability.

In Fig. 3, we provide the blood loss prediction results as well as the actual binary test dataset. In the subplot from (a) to (c), we respectively show the results of the used basis methods, i.e. RF, GBDT and XGB. It can be seen that such three multi-classifiers can give the approximate blood loss estimate as the true blood loss. In this way, the predicted blood loss is supposed to provide the obstetricians with more information on the critical levels of blood loss.

Fig. 4 shows the binary blood loss prediction results comparison with the actual binary test dataset. In the actual dataset, if blood loss $>$ 500 ml, it is labelled as 1; and otherwise it is labelled as 0. For the purpose of illustration, in the predicted results if the blood loss $>$ 500 mL, then

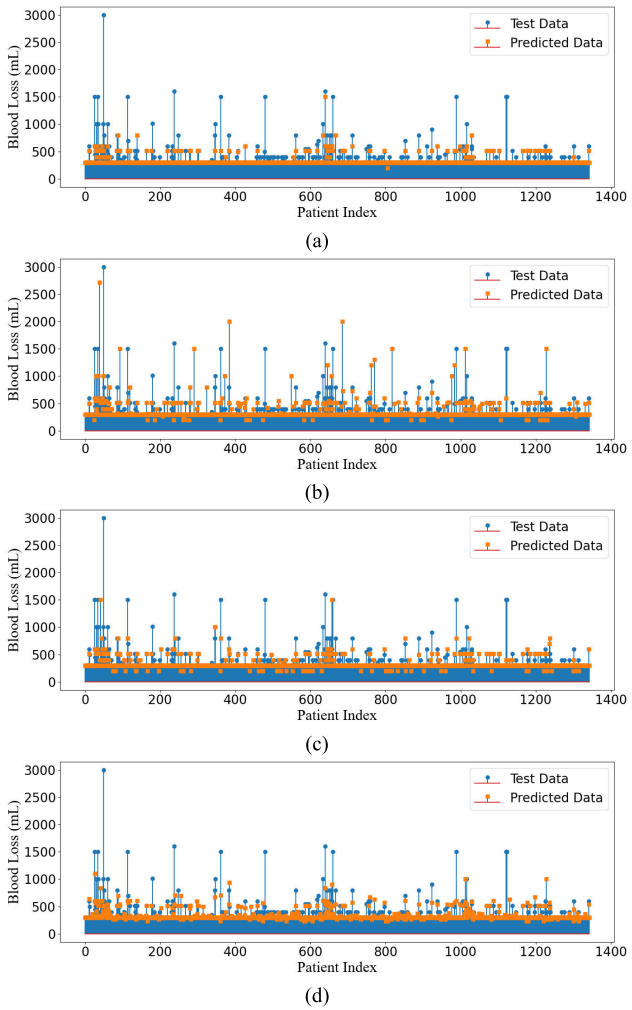

FIGURE 3. Prediction results comparison with the actual test dataset. (a) Random forest, (b) GBDT (c) XGB and (d) ensemble learning (EL-SC).

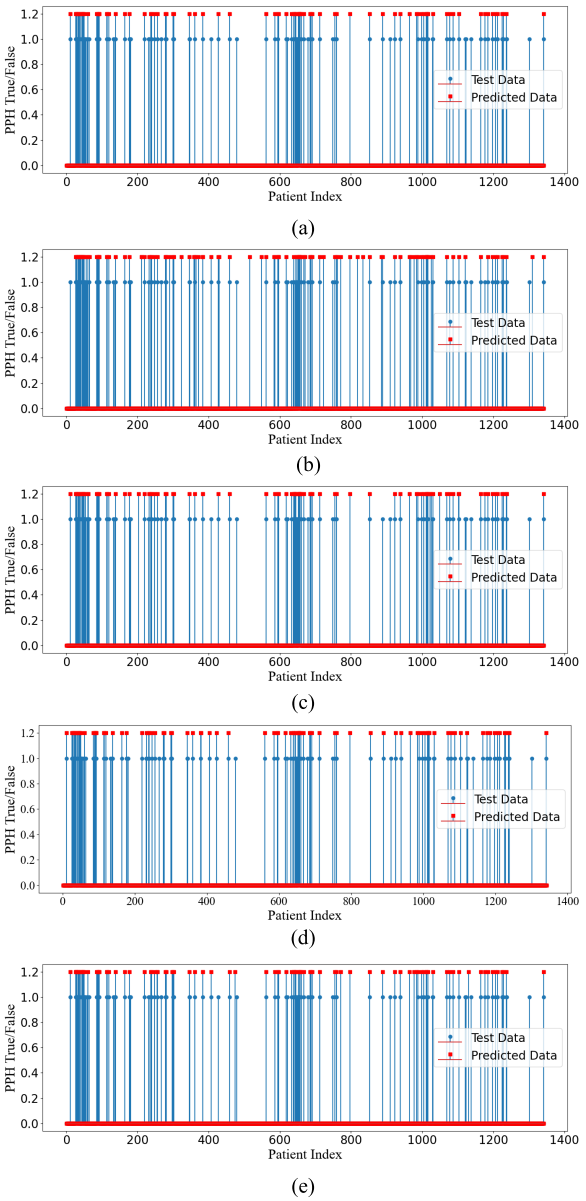

FIGURE 4. Prediction results (binary) comparison with the actual test dataset (binary) (a) random forest, (b) GBDT, (c) XGB, (d) SVM and (e) ensemble learning (EL-SC).

TABLE 4. Accuracy, Precision, Recall, F-measure and MCC for PPH prediction.

Methods	*A*	*R*	*P*	*F-measure*	*MCC*
RF	97.2%	74.7%	93.7%	0.83	0.81
GBDT	94.9%	65.5%	73.6%	0.69	0.67
XGB	96.6%	73.9%	87.8%	0.80	0.76
SVM	97.3%	74.8%	93.7%	0.83	0.82
EL-HC	96.7%	65.6%	95.3%	0.78	0.78
EL-SC	96.2%	69.7%	84.7%	0.76	0.69

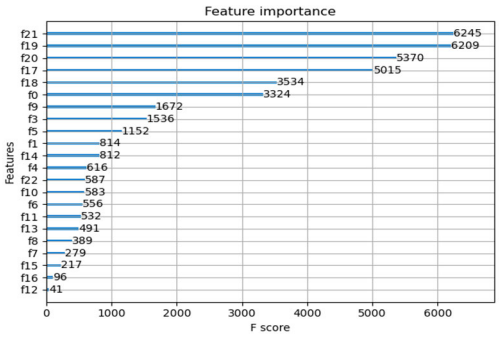

FIGURE 5. Prediction results comparison with the actual test dataset for PPH prediction.

it is denoted by 1.2 (for comparison purpose); otherwise it is denoted by 0. The subplots (a) to (d) give the analysis results for the basis methods, i.e. RF, GBDT, XGB and SVM respectively.

Table 4 summarizes the classification accuracy, precision, recall, F-measure and MCC for the PPH prediction model. In Table 4, the EL-HC results are obtained by a voting policy that all "1" ballot from RF, GBDT, XGB and SVM can give an EL result "1". It is shown that the SVM method achieves the highest F-measure and MCC, while EL-HC obtains the similar performance. RF achieves almost the same performance as SVM. Note that, we have also evaluated ANN base learner for PPH prediction; but unfortunately, we found such an ANN method cannot achieve satisfied prediction performance for our dataset (as seen lately in the Table 6). For EL-SC, the prediction result is attained by averaging the blood loss volume results of the 3 base learners (RF, GBDT and XGB) first, then compare with the PPH threshold (500 mL). We can see that the performance of EL-HC is higher than EL-SC.

Fig.5 shows the feature importance of PPH features given by XGB. The feature f21 (uterine inertia) shows the highest importance (F score). The F score is computed using the *plot_importance* function of the *xgboost* module. It should be noted that f21 (uterine inertia) is a binary feature with small values 0 or 1. However, XGB can avoid the unbalanced value problem, that is, a binary feature among features with large values (thousands) can also be identified as the most important feature. This figure shows all the features that are important to instruct the obstetricians for prompt intervention of PPH.

B. PERFORMANCE EVALUATION FOR DIC DATASET

For the DIC prediction model, there are 212 records in the dataset. 150 records are used as the training dataset, and the other 62 records are used as the testing dataset. Among the 62 testing records, a total number of 7 positive instances (DIC complication) occur.

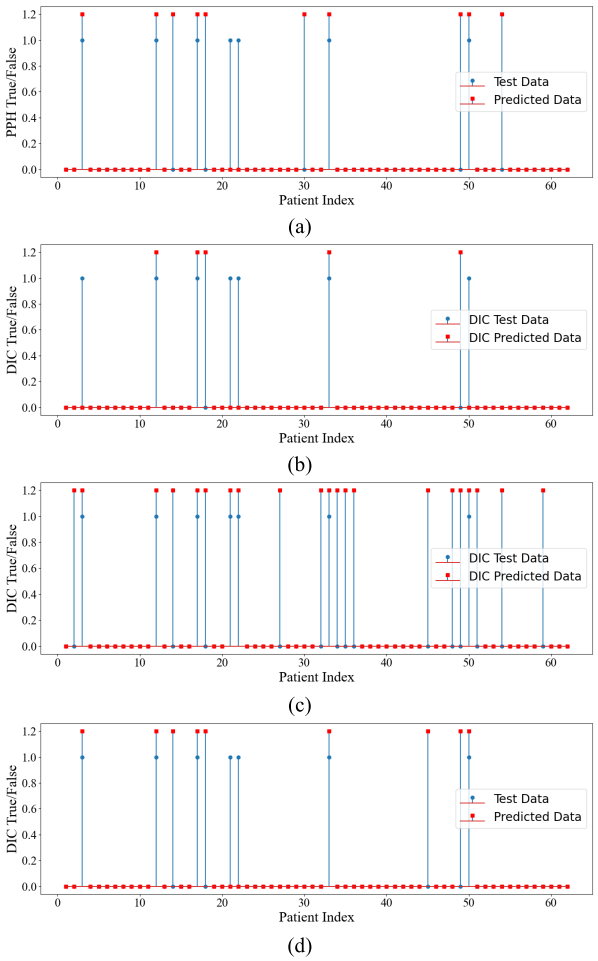

FIGURE 6. DIC prediction results comparison with the actual test dataset. (a) random forest, (b) GBDT (c) XGB and (d) ensemble learning.

Fig. 6 shows DIC prediction results comparison with the actual test dataset. It is shown EL and RF achieves better prediction performance.

Table 5 shows Accuracy, Precision, Recall, F-measure and MCC for DIC prediction using only 3 base learners (RF, GBDT and XGB). In Table 5, the EL results are obtained by a voting policy that a majority "1" ballot (2/3) can give an EL "1". It can be seen that the highest *F-measure* is obtained for the DIC prediction using the RF base learner. EL achieves identical MCC, F-measure, *A*, *R*, and *P* as the RF base learner. That means performance of EL would not descend compared

TABLE 5. Accuracy, Precision, Recall, F-measure and MCC for DIC prediction.

Methods	*A*	*R*	*P*	*F-measure*	*MCC*
RF	90.3%	71.4%	55.6%	0.63	0.58
GBDT	90.3%	42.9%	60.0%	0.50	0.45
XGB	77.4%	100.0%	33.3%	0.50	0.50
EL	90.3%	71.4%	55.6%	0.63	0.58

to a base learner. It is shown that both F-measure and MCC of the DIC case are much lower than that of PPH prediction, due to the much imbalanced dataset of DIC.

Fig.7 shows the feature importance of DIC features given by XGB. The feature f16 (Crystal liquid) shows the highest importance (F score). The F score is computed using the *plot_importance* function of the *xgboost* module. This figure shows all the features that are important to instruct the obstetricians for early intervention of DIC.

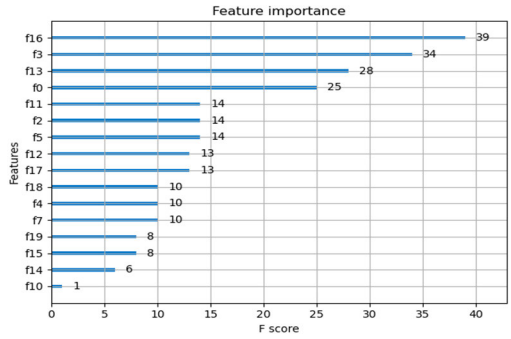

FIGURE 7. Prediction results comparison with the actual test dataset for DIC prediction.

Note that, in this dataset, we find both the ANN method and the other SVM method (with both linear kernel and nonlinear Gaussian kernels) may acquire the less accurate performance (as seen in Table 7), which hence are excluded from the base learner in our EL scheme for DIC prediction.

C. PERFORMANCE COMPARISON

As discussed, logistical regression (LR) and artificial neural networks (ANN) are two popular methods used for medical diagnosis. In the last numerical experiments, we further compared the proposed EL method with LR and ANN, on both the PPH and DIC prediction. The prediction results of PPH are summarized in Table 6, and the results of DIC are summarized in Table 7.

For the PPH prediction task (i.e. the binary classification), we find that, when using the ANN method, the maximum F-measure is only 0.55 (based on a grid search of hidden layers and number of neutrons), which is much smaller than the F-measure of EL-HC (0.78).

For the other DIC prediction task, we show that both the SVM method (with the linear/nonlinear kernels) attains the less accurate performance, i.e. the F-measure is only about 0.50. Another logistic repression method abstains also the uncompetitive result, e.g. its obtained F-measure is only about 0.42, which is far less than our proposed EL method (with an improved F-measure of 0.63).

TABLE 6. Performance comparison of different methods for the PPH prediction.

Methods	*A*	*R*	*P*	*F-measure*
ANN	83.4%	49.6%	61.5%	0.55
EL-HC	96.7%	65.6%	95.3%	0.78

TABLE 7. Performance comparison of different methods for the DIC prediction.

Methods	*A*	*R*	*P*	*F-measure*
LR	84.3%	36.36%	50%	0.42
SVM	83.87%	71.4%	38.64%	0.50
EL	90.3%	71.4%	55.6%	0.63

VI. CONCLUSION

In this paper, we study the PPH predictive diagnosis problem by resorting the machine learning techniques. Two main contributions are (1) the collection of large clinical dataset, and (2) the well-designed ensemble learning method. Our PPH and DIC dataset involves 3842 and 212 records, respectively. The ensemble learning is designed to integrate four basis methods, i.e. random forest, extreme gradient boosting, gradient boosting decision tree and SVM for PPH prediction. With the trained prediction diagnostic model, the accurate results have been obtained. As shown, the accuracy of correct PPH diagnosis would achieves 96.7%; the total disseminated intravascular coagulation (DIC) prediction accuracy would surpass 90%. As a result, the proposed model based on machine learning enables to predict successfully the risk of PPH, and assess the critical level of PPH patient. We anticipate our study results have the great potential to the reduction the future mortality of pregnant women. In the future, we may further extend the dataset to further adapt our proposed method to improve the generalizability.

APPENDIX

In this analysis, we have used the grid-search method to obtain the proper configuration of various ML methods and the designed EL method. To be specific, we have adopted the *sklearn.ensemble* software tool for RF and GBDT, whilst we used the *xgboost* software tool for XGB. For the other SVM methods, the *sklearn* software tool was also used. According to our study, the used hyperparameters of RF, GBDT, XGB and SVM for PPH are listed in the following Table 8.

TABLE 8. Hyperparameters for the PPH dataset.

Base learner	Hyperparameters
RF	*n_estimators=20, criterion='gini', max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features='sqrt'*
GBDT	*n_estimators=200, criterion='friedman_mse', learning_rate=0.2, loss='deviance', subsample=0.8, min_samples_split=2, min_samples_leaf=1*
XGB	*n_estimators=200, learning_rate=0.2, min_child_weight=1, max_depth=10, gamma=0, subsample=0.8, scale_pos_weight=1*
SVM	*C=0.1, kernel='linear', decision_function_shape='ovr'*

TABLE 9. Hyperparameters for the DIC dataset.

Base learner	Hyperparameters
RF	*n_estimators=30, criterion='gini', max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features='sqrt'*
GBDT	*n_estimators=20, criterion='friedman_mse', learning_rate=0.25, loss='deviance', subsample=0.9, min_samples_split=2, min_samples_leaf=1*
XGB	*n_estimators=30, learning_rate=0.13, min_child_weight=1, max_depth=10, gamma=0, subsample=1, scale_pos_weight=150*

In the same manner, we have determined the Hyperparameters of RF, GBDT and XGB for DIC dataset, which are listed in the Table 9.

REFERENCES

[1] R. Chávez-Santiago, M. Szydełko, and A. Kliks, "5G: The convergence of wireless communications," *Wireless Pers. Commun.*, vol. 83, no. 3, pp. 1617–1642, Mar. 2015.

[2] B. Li, S. Li, A. Nallanathan, and C. Zhao, "Deep sensing for future spectrum and location awareness 5G communications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1331–1344, Jul. 2015.

[3] B. Li, Z. Zhou, and W. X. Zou, "On the efficient beam-forming training for 60GHz wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 504–515, Feb. 2013.

[4] S. Li, L. Da Xu, and S. Zhao, "5G Internet of Things: A survey," *J. Ind. Inf. Integr.*, vol. 10, pp. 1–9, Jun. 2018.

[5] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C.-H. Youn, "5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 16–23, Apr. 2018.

[6] J. Mwangama, B. Malila, T. Douglas, and M. Rangaka, "What can 5G do for healthcare in Africa?" *Nature Electron.*, vol. 3, no. 1, pp. 7–9, Jan. 2020.

[7] B. Gu, "5G key technologies and their impact on the Internet of Things," *Wireless Internet Technol.*, vol. 7, pp. 30–31, May 2019.

[8] American College of Obstetricians and Gynecologists, "ACOG practice bulletin: Clinical management guidelines for obstetrician-gynecologists Number 76, October 2006: Postpartum hemorrhage," *Obstetrics Gynecol.*, vol. 108, no. 4, pp. 1039–1047, 2006.

[9] L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, and A. B. Moller, "Global causes of maternal death: A WHO systematic analysis," *Lancet Global Health*, vol. 2, no. 6, pp. 323–333, 2014.

[10] American College of Obstetricians and Gynecologists, "ACOG practice bulletin no. 205: Vaginal birth after cesarean delivery," *Obstetrics Gynecol.*, vol. 133, no. 2, pp. 110–127, 2019.

[11] C. Chen, X. Liu, D. Chen, S. Huang, X. Yan, H. Liu, Q. Chang, and Z. Liang, "A risk model to predict severe postpartum hemorrhage in patients with placenta previa: A single-center retrospective study," *Ann. Palliative Med.*, vol. 8, no. 5, pp. 611–621, Nov. 2019.

[12] C. M. Koopmans, K. Van der Tuuk, H. Groen, J. P. R. Doornbos, and I. M. de Graaf, "Prediction of postpartum hemorrhage in women with gestational hypertension or mild preeclampsia at term," *Acta Obstetricia et Gynecol. Scandinavica*, vol. 93, no. 4, pp. 399–407, Apr. 2014.

[13] W. Sittiparn and T. Siwadune, "Risk score for prediction of postpartum hemorrhages in normal labor at Chonburi Hospital," *J. Med. Assoc. Thailand*, vol. 100, no. 4, pp. 382–388, 2017.

[14] S. Jurarut, L. Somnimit, and P. Chadakarn, "A risk score for predicting postpartum hemorrhage in association with cesarean delivery," *Thai J. Obstetrics Gynaecol.*, vol. 23, no. 1, pp. 3–11, 2015.

[15] C. Neary, S. Naheed, D. McLernon, and M. Black, "Predicting risk of postpartum haemorrhage: A systematic review," *BJOG: Int. J. Obstetrics Gynaecol.*, vol. 128, no. 1, pp. 46–53, Jan. 2021.

[16] N. Prata, S. Hamza, S. Bell, D. Karasek, F. Vahidnia, and M. Holston, "Inability to predict postpartum hemorrhage: Insights from Egyptian intervention data," *BMC Pregnancy Childbirth*, vol. 11, no. 1, Nov. 2011.

[17] E. Sheiner, L. Sarid, A. Levy, D. S. Seidman, and M. Hallak, "Obstetric risk factors and outcome of pregnancies complicated with early postpartum hemorrhage: A population-based study," *J. Maternal-Fetal Neonatal Med.*, vol. 18, no. 3, pp. 149–154, Sep. 2005.

[18] A. J. Dilla, J. H. Waters, and M. H. Yazer, "Clinical validation of risk stratification criteria for peripartum hemorrhage," *Obstetrics Gynecol.*, vol. 122, no. 1, pp. 120–126, Jul. 2013.

[19] M. S. Kramer, C. Berg, H. Abenhaim, M. Dahhou, J. Rouleau, A. Mehrabadi, and K. S. Joseph, "Incidence, risk factors, and temporal trends in severe postpartum hemorrhage," *Amer. J. Obstetrics Gynecol.*, vol. 209, no. 5, p. 449, 2013.

[20] A. Onan, "On the performance of ensemble learning for automated diagnosis of breast cancer," in *Advances in Intelligent Systems and Computing*. 2015, pp. 119–129.

[21] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction," *BMC Med. Res. Methodol.*, vol. 19, p. 64, Mar. 2019.

[22] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informat. Med. Unlocked*, vol. 20, 2020, Art. no. 100402.

[23] Y. Doomah, S. Xu, L. Cao, S. Liang, G. Allornuvor, and X. Ying, "A fuzzy expert system to predict the risk of postpartum hemorrhage," *Acta Inf. Medica*, vol. 27, no. 5, p. 318, 2019.

[24] K. Betts, S. Kisely, and R. Alati, "Predicting common maternal postpartum complications: Leveraging health administrative data and machine learning," *BJOG: Int. J. Obstetrics Gynaecol.*, vol. 126, no. 6, pp. 702–709, May 2019.

[25] K. K. Venkatesh, R. A. Strauss, C. A. Grotegut, R. P. Heine, N. C. Chescheir, J. S. A. Stringer, D. M. Stamilio, K. M. Menard, and J. E. Jelovsek, "Machine learning and statistical models to predict postpartum hemorrhage," *Obstetrics Gynecol.*, vol. 135, no. 4, pp. 935–944, Apr. 2020.

[26] W. K. B. A. Owiredu, D. N. M. Osakunor, C. A. Turpin, and O. Owusu-Afriyie, "Laboratory prediction of primary postpartum haemorrhage: A comparative cohort study," *BMC Pregnancy Childbirth*, vol. 16, no. 1, Dec. 2016.

[27] Y. Ma, M. Shao, and X. Shao, "Analysis of risk factors for intraoperative hemorrhage of cesarean scar pregnancy," *Medicine*, vol. 96, no. 25, p. 7327, Jun. 2017.

[28] Y. Liu, Y. Shen, W. Zhu, J.-B. Qiu, Q. Huang, and W.-Q. Ye, "Clinical assessment indicators of postpartum hemorrhage: A systematic review," *Chin. Nursing Res.*, vol. 4, no. 4, pp. 170–177, Dec. 2017.

[29] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discovery*, vol. 2, no. 6, pp. 493–507, Nov. 2012.
[30] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobot.*, vol. 7, p. 21, Dec. 2013.
[31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
[32] L. Li, Q. Hu, X. Wu, and D. Yu, "Exploration of classification confidence in ensemble learning," *Pattern Recognit.*, vol. 47, no. 9, pp. 3120–3131, Sep. 2014.

YAWEI ZHANG received the M.S. degree in obstetrics and gynecology from the Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, in 2010. Since 2017, she has been an Attending Physician with the Department of Obstetrics, Beijing Obstetrics and Gynecology Hospital. She was an Attending Physician with the Beijing Huairou Hospital and worked on obstetrics and gynecology for more than ten years. Her current research interests include the machine learning-based automated diagnosis, prediction of the postpartum hemorrhage, and medical data analysis.

XIN WANG received the M.S. degree in obstetrics and gynecology from Capital Medical University, Beijing, China, in 2001.

She is currently a Professor, a Chief Physician, and the Director with the Department of Obstetrics, Beijing Obstetrics and Gynecology Hospital. She visited the Cincinnati Children's Hospital as a Visiting Scholar in 2008. Her current research interests include labor management, postpartum hemorrhage, intrauterine diagnosis, and treatment of fetal diseases, such as fetal intrauterine transfusion and twin-twin transfusion syndrome.

NINGYU HAN received the M.S. degree in obstetrics and gynecology from Capital Medical University, Beijing, China, in 2020. She was an Intern with the Department of Obstetrics, Beijing Obstetrics and Gynecology Hospital. During the graduate study, she majored in management of postpartum hemorrhage and its complications. Her current research interests include the machine learning-based medical data statistics, management of labor, and postpartum hemorrhage.

RONG ZHAO received the M.S. and Ph.D. degrees in obstetrics and gynecology from Capital Medical University, Beijing, China, in 2008 and 2018, respectively. From 2008 to 2014, she was a Resident Doctor of Obstetric and Gynecology. Since 2014, she has been an Attending Physician with the Department of Obstetrics, Beijing Obstetrics and Gynecology Hospital. Her current research interests include machine learning-based prediction modeling, subclinical hypothyroidism during pregnancy, gestational diabetes mellitus, uterine fibroids during pregnancy, and prenatal diagnosis of achondroplasia.