IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 18, 2019, accepted July 8, 2019, date of publication July 18, 2019, date of current version August 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929681

Down Syndrome Prediction Using a Cascaded Machine Learning Framework Designed for Imbalanced and Feature-correlated Data

LING LI[1], WANYING LIU[1], HONGGUO ZHANG[2], YUTING JIANG[2], XIAONAN HU[2], AND RUIZHI LIU[2]

[1]School of Communication Engineering, Jilin University, Changchun 130012, China
[2]Center for Reproductive Medicine, Center for Prenatal Diagnosis, The First hospital of Jilin University, Changchun 130021, China

Corresponding author: Ling Li (liling2002@jlu.edu.cn)

This work was supported in part by the Jilin Province Nature Foundation under Grant 20170101140JC, in part by the Special Project of Industrial Technology Research and Development in Jilin Province under Grant 2019C052-6 and in part by the Industrial Project of Science and Technology Development Plan of Jilin Province under Grant 20190302073GX.

ABSTRACT Down syndrome (DS) caused by the presence of part or all of a third copy of chromosome 21 is the most common form of aneuploidy. The prenatal screening for DS is a key component of antenatal care and is recommended to be universally offered to women irrespective of age or background. The objective of this paper is to introduce a noninvasive and accurate diagnosis procedure for DS and to minimize social and financial cost of prenatal diagnosis. Recently, machine learning has received considerable attention in predictive analytics for medical problems. However, there is few its applications on DS prediction reported due to the difficulty of dealing with highly imbalanced and feature-correlated screening data. In this paper, we propose a cascaded machine learning framework designed for DS prediction based on three complementary stages: 1) pre-judgment with isolation forest technique, 2) model ensemble by voting strategy, and 3) final judgment using logistic regression approach. The experimental results show that the performance of this framework on maternal serum screening data set, when evaluated with different evaluation parameters, is superior to those of some machine learning methods. The best suggested combination of input features for DS screening is the group of alpha-fetoprotein, human chorionic gonadotropin, unconjugated estriol, and maternal age. In addition, our method has the potential to generate further accurate prediction for imbalanced and feature-correlated data, thereby providing a novel and effective approach for certain diseases analysis.

INDEX TERMS Bioinformatics, down syndrome prediction, imbalanced learning, cascaded framework, ensemble learning, noninvasive prenatal diagnosis.

I. INTRODUCTION

Down syndrome (DS), also called trisomy 21, is the most common chromosomal abnormality that occurs in 14.7 per 10,000 live births in China [1]. DS is typically associated with physical growth delays, characteristic facial features, and mild to moderate intellectual disability [2]. The lifetime economic burden of each person born with DS in China was estimated to be US$47,000 in 2003 [3]. No cure treatment is available for DS until now. Hence, screening for DS is a key component of antenatal care and is recommended to be universally offered to women irrespective of age or background. Since 1990s, maternal serum screening (MSS) has been performed in China for non-invasive detection of DS risk and other abnormalities [4]. Statistical mixture models constructed in accordance with the related concentration data of maternal serum markers have been utilized in MSS [5], [6]. The common screening programs are the double test (alpha-fetoprotein [AFP] and human chorionic gonadotropin [HCG or free-β-HCG]), triple test (AFP + HCG [or free-β-HCG] + unconjugated estriol [uE3]), and quadruple test (AFP + HCG [or free-β-HCG] + unconjugated estriol [uE3] + inhibin A). The detection rate of the abovementioned methods ranges from 60% to 80% in different healthcare institutions, and the false positive rate (FPR) is often limited to 3%–13% [7], [8].

Before 2012, pregnant women in China with MSS outcomes rated as ''high risk'' were suggested to undergo

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

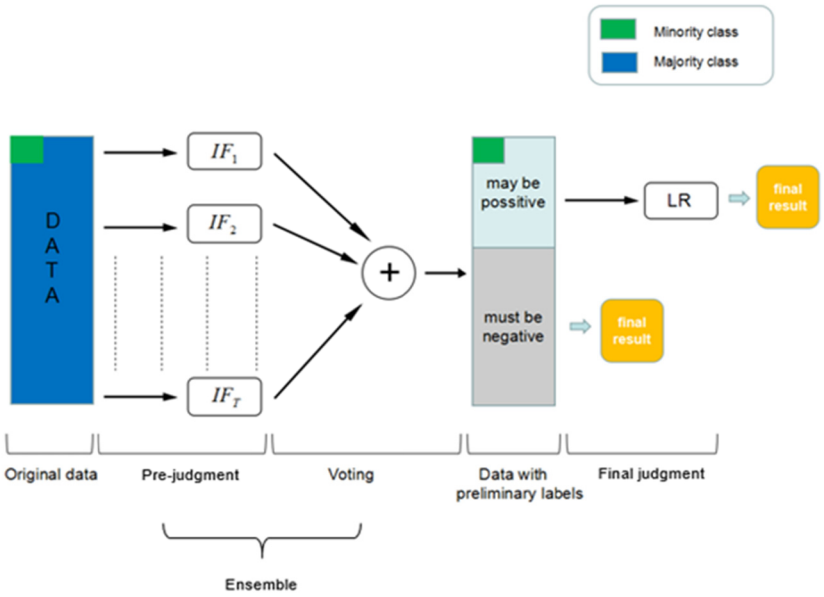


FIGURE 1. Schematic representation of CVIFLR.

amniocentesis or chorionic villi sampling (CVS) tests, which are the ''gold standards'' for the diagnosis of chromosomal abnormality. However, such methods are invasive and have some infection risk, fetal damage during examination, and miscarriage rates of approximately 0.4% for amniocentesis and 1.1% for CVS [9]. Non-invasive methods for prenatal screening with a high accuracy are urgent to protect numerous women from invasive procedures. In the last years, non-invasive prenatal DNA test (NIPT) has gained particular attention in scientific community of this domain. NIPT is a new type of genetic test that screens for birth defects and inherited diseases [10]. This test examines a small amount of baby's DNA naturally found in the blood of pregnant women. NIPT results often offered to the pregnant patients with high risk of MSS for further screening are accurate but time-consuming and costly.

Recently, machine learning methods have been received considerable attention and have been widely used in cancer diagnosing [11]–[13] and other common diseases diagnosis [14], [15]. They have also been applied to understand complex disease progresses [16] and generate disease-specific medications from biomedical literature and clinical data repository [17]. However, few applications of machine learning on DS screening have been reported due to the highly imbalanced and feature-correlated data. In 2016, Neocleous *et al*. present the trained artificial neural networks with under-sampling strategy (under-sampled ANN) on the MSS dataset provided by the Fetal Medicine Foundation to predict chromosomal abnormality [18]. However, the given model in [18] could not be promoted globally because of large differences in the concentrations of serum markers related to DS in different regions and races. Considering the incidence of DS (occurring in 14.7 per 10,000 live births), the data collection for developing a data-hungry ANN model is an extensive work. Lightweight machine learning methods trained with a small number of DS samples are highly needed and necessary for the models specially designed for people of a certain ethnic or region available.

In this paper, a cascaded framework of voting isolation forests and logistic regression (CVIFLR) is proposed using highly imbalanced and feature-correlated data. All data come from the singleton pregnancy cases undergoing second trimester screening in the triple test in Jilin Province of China. To show the effectiveness of the proposed approach, we perform an experimental comparison with the under-sampled ANN and state-of-the-art imbalance learning methods [19]–[23]. Our results show that CVIFLR outperforms other methods, and the best combination of the input features for DS screening are PAPP-A MoM, β-hCG MoM, uE3 MoM, and MA.

The remainder of the paper is organized as follows: Section II describes the proposed CVIFLR framework. Section III presents the data analysis, its features along with the proposed framework, and the experimental set-up. Section IV summarizes the performance evaluation and cross-validation results, and analyzes the results and the characteristics of different methods. Finally, Section V provides the conclusion and the possible feature work.

II. THE PROPOSED FRAMEWORK

CVIFLR is a method for addressing imbalanced and feature-correlated data for DS prediction. To realize DS prediction with high accuracy and achieve large coverage of the available input data, CVIFLR has three complementary stages shown in Fig. 1: 1) pre-judgment, 2) voting, and 3) final judgment.

A. PRE-JUDGMENT

Isolation forests (IFs) are used at the first stage of CVIFLR to determine the suspected DS samples identified as anomalies. IF is an anomaly detection method proposed in [24] and it is an ensemble of isolation trees

$$IF = \{t_1, \ldots, t_T\} \tag{1}$$

For each tree t_i, the number of steps required to isolate a sample is represented by h(x). The average number of iterations required to isolate a sample in an isolation forest is then expressed as

$$E(h(x)) = \frac{1}{T}\sum\nolimits_{t\in IF} h_t(x) \tag{2}$$

The main idea is that only a few of iterations are required to isolate an anomaly. The number of iterations required to isolate an observation x is affected by that of the samples ψ assigned to the root node. To account this figure, a normalized anomaly score $s(x, \psi)$ is defined as

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}} \tag{3}$$

where

$$c(\psi) = \begin{cases} 2H(\psi - 1) - \dfrac{2(\psi - 1)}{M} & for \psi > 2 \\ 1 & for \psi = 2 \\ 0 & otherwise \end{cases} \tag{4}$$

and M denotes the number of samples in the data and the harmonic number is

$$H(i) = \ln(i) + 0.5772156649 \tag{5}$$

If $s(x, \psi) \geq S_{th}$, the sample x is regarded as an anomaly. Otherwise, x is regarded as a normal instance. On the basis of (6) and (7), S_{th} is the threshold fitted by the input parameter of contamination and the number of samples M:

$$M \times contamination \approx \sum\nolimits_{i=i}^{M} flag_i \tag{6}$$

$$flag_i = \begin{cases} = 1 & if\ s_i(x, \psi) \geq S_{th} \\ = 0 & else \end{cases} \tag{7}$$

According to [24], [25], the parameter contamination represents the amount of contamination of the dataset, i.e. the proportion of outliers in the data set. When we default a bigger contamination in training, which means $\sum_{i=i}^{M} \mathrm{flag_i}$ contains more '1' samples, we can get a smaller S_{th}. Therefore, more samples in test set will be labeled as the anomalies by IF. However, these anomalies consist of both real DS samples and mistaken ones, which should be further distinguished.

B. MODEL ENSEMBLE BY VOTING

Ensembles often improve prediction accuracy and robustness in learning machines [26], [27]. In the second stage of CVIFLR, an ensemble strategy is used to refine the samples flowing to the next stage. We train each IF with a partition of the negative data in the train set. Then, the VIF is defined by a set of IFs as

$$VIF = \{IF_1, \ldots, IF_n\} \tag{8}$$

On the basis of (3), for each IF, $s(x, \psi)$ is easy to compute by (3). Similarly, if $s(ix, \psi)j \geq S_{th}$, x is regarded as an anomaly. The voting matrix V is then defined as

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1M} \\ v_{21} & v_{22} & \cdots & v_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nM} \end{bmatrix} \tag{9}$$

where v_{ij} is

$$v_{ij} = \begin{cases} = 1 & if\ s_{ij}(x, \psi) \geq S_{th}^i \\ = 0 & else \end{cases} \tag{10}$$

S_{th}^i is the S_{th} of IF_i, and $s_{ij}(x, \psi)$ denotes the anomaly score of sample x_j calculated by base model IF_i. The comprehensive poll of x_j is then expressed as

$$Vote_{sum}(x_j) = \sum\nolimits_{i=1}^{n} v_{ij} \tag{11}$$

By comparing $Vote_{sum}(x_j)$ and the threshold $Vote_{th}$, we redefine the decision function as follows:

(a) If $Vote_{sum}(x_j) < Vote_{th}$, the sample is labeled as "must be negative";

(b) If $Vote_{sum}(x_j) \geq Vote_{th}$, the sample is regarded as "may be positive".

The $Vote_{th}$ is the threshold in the new decision function. It is a hyper-parameter determined by experiments. Choosing proper $Vote_{th}$ and the contamination, we can ensure that all the samples labeled as "must be negative" are real negative ones. However, the "may be positive" samples will be still suspected and fed to the next stage for final judgment.

On the basis of (6), (7) and (11), the larger the contamination we default, the smaller S_{th} will be learned, and a larger $Vote_{sum}$ for each sample will be calculated. If $Vote_{th}$ remains the same, more negative samples will be regarded as "may be positive". As a result, the data flowing to the next stage will be more imbalanced which produces the disadvantage for training the model in third stage. However, if the predefined contamination is too small to recognize all the positive samples as "may be positive", the final stage has no chance to identify the positive samples regarded as "must be negative" by the previous stage. Therefore, the detection rate of the whole framework will decrease. In this study, under the premise that all DS samples are classified into "may be positive", we select a small contamination (0.15) and large $Vote_{th}$ (54), so that the data flowing to the final stage are less imbalanced.

As IF defines the "anomaly" as "more likely to be separated", the "may be positive" data detected by VIF are not only less imbalanced but also more dispersed in space than raw data. Therefore the classification effect in the final stage is significantly improved.

algorithm 1 CVIFLR

Input:

$-P$: set of positive examples(examples with DS)

$-N$: set of negative examples(examples without chromosome anomalies)

$-n$: number of partitions-1

$-C$: predefined contamination of each IF

$-Vote_{th}$: hyper-parameter

Output:

$CVIFLR = \{IF_1, IF_2, ..., IF_n, LR_{VIF}\}$:a set of Isolation Forest models and a LR model

Output on test example x:

The label of 0 (negative) or 1 (positive)

begain algorithm

$(I)Initialazation\ and\ partitioning\ of\ N$:

$\{N_1, N_2, ..., N_n, N_{n+1}\} := Do.partition(N, n+1)$

$i =: 1$

while $i < n+1$ **do**

$(II)Isolation\ Forest\ training$:

$IF_i = IF_C(N_i)$

$i =: i+1$

end while

$(III)Training\ set\ for\ Logistic\ Regression\ (represented\ by\ T')$:

$T = N_{n+1} \cup P$

$x_j \in T$

while $i < n+1$ **do**

$IF_i(T)$

end while

$Vote_{sum}(x_j) = \sum_{i=1}^{n} V_{IF_i}(x_j)$

if $Vote_{sum}(x_j) \geqslant Vote_{th}$ **then**

$x_j \in T'$

end if

$(IV)Logistic\ Regression\ training$:

$LR_{VIF} = LR(T')$

end algorithm

FIGURE 2. Pseudocode of CVIFL.

C. FINAL JUDGEMENT

A logistic regression (LR) [28] model is used as the final stage of CVIFLR to identify true and false positive instances from the suspected samples. The LR classifier is defined as

$$\hat{y} = g(Z) = \frac{1}{1+e^{-Z}} \tag{12}$$

where

$$\begin{aligned} Z &= W^T x + b = w_0 u_0 + w_1 u_1 + \cdots + w_L u_L + b \\ &= \sum_{k=0}^{L} w_k u_k + b \end{aligned} \tag{13}$$

and $x = (u_1, u_2, \ldots, u_L)$ is the suspected sample vector of features, and u_k denotes the value of the feature k. Then, the cost function is defined as

$$J(W, b) = -\frac{1}{M'} \sum_{i=1}^{M'} \left[y^{(i)} log\hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) \right] \tag{14}$$

where $y^{(i)} \in (0, 1)$ represents two different classes of DS ($y^{(i)} = 1$) and euploid ($y^{(i)} = 0$). M' represents the number of samples flowing to the last stage. On the basis of (14), $w_1, w_2, \cdots, w_L, b$ in (13) are calculated using the gradient descent.

D. PSEUDOCODE AND DETAILS OF THE ALGORITHM

Figure 2 shows the pseudocode of the CVIFLR algorithm. P represents the set of positive examples, i.e., the set of available cases of DS, and N is the set of negative examples, i.e., the euploid cases that we assume to outnumber greatly. Precisely, $P = \{(x, y)\, x \in F, y = 1\}$ and $N = \{(x, y)\, x \in F, y = 0$ where y $\in (1, 0)$ indicates the two different classes of DS ($y = 1$) and euploid ($y = 0$), and F is a set of real-valued vectors representing the features associated with DS examples. The working procedure of CVIFLR is as follows.

First, the algorithm initializes working parameters. Then, it subdivides the negative examples in $n + 1$ partitions $\{N_1, N_2, \ldots, N_n, N_{n+1}\}$, such that $\bigcup_{i=1}^{n+1} N_i = N$ and $\bigcap_{i=1}^{n+1} N_i = \emptyset$ where N is divided into n+1 partitions, and the first n partitions are used to train n base IFs, the last partition constitutes the original training set T for LR such that $T = N_{n+1} \bigcup P$.

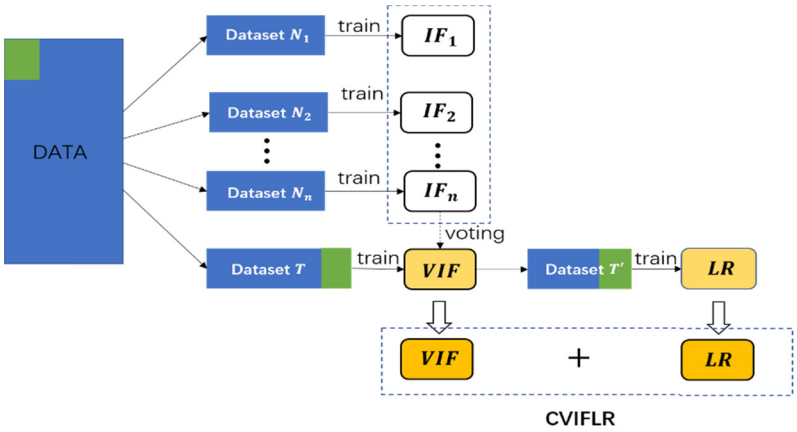


FIGURE 3. Training strategy of CVIFLR.

The first while loop iterates a set of steps on the n partitions of the data to train different IF with a predefined contamination in each iteration. The second while loop predicts the examples in the dataset T by using the IF models obtained from the last while loop. The real training set T' for the LR model is a subset of T. As described in the pseudocode, $Vote_{sum}\,(x)$ should be calculated for each example x in T and compared with the predefined Vote_{th} to decide whether x is a member of T'. $Vote_{sum}$ for a given variant x_j is computed by summing the votes provided by different base IFs as

$$Vote_{sum}\left(x_j\right) = \sum_{i=1}^{n} V_{IF_i}(x_j) \tag{15}$$

If $\text{Vote}_{\text{sum}}\left(\text{x}_\text{j}\right) \geq \text{Vote}_{\text{th}}$, $\text{x}_\text{j} \in \text{T}'$, otherwise $\text{x}_\text{j} \notin \text{T}'$. The parameter contamination C and Vote_{th} should be adjusted to ensure that all positive examples in T are selected as the numbers of T', such that $T' = P \bigcup N_{n+1}^{sub}$, where $N_{n+1}^{sub} \in N_{n+1}$. Finally, by using the training set T', the algorithm trains the LR model, which is the last stage of CVIFLR. Fig. 3 displays the overall training strategies of CVIFLR.

III. DATA AND EXPERIMENTAL SET-UP

A. DATA

The MSS dataset used in this study is provided by the Center for Reproductive Medicine, Center for Prenatal Diagnosis in the First Hospital of Jilin University. The present study is supported by the Ethics Committee of this hospital (no. 2016-419, dated 10$^{\text{th}}$ Dec. 2016). Written informed consent is provided by each participant and signed.

The dataset contains 100,244 negative and 108 positive cases, resulting in a near imbalance of 1:928. All cases have been diagnosed by the invasive tests or paid a return visit to check the screening results.

More concretely, each example in the dataset is represented by a vector of 22 features. Some features are of apparently great importance, because they are taken during pregnancy, for instance, the biochemical pregnancy-associated plasma protein-A (PAPP-A), β- human chorionic gonadotropin (β-hCG), unconjugated estriol (uE3), and ultrasonographic markers such as crown-rump length (CRL), biparital dianctoc (BPD), nuchal translucency (NT). Some features are related to the historical and physiological data of the pregnant women, such as nationality, weight, maternal age, menstrual cycle, history of abnormal pregnancies, whether or not the menstruation is regular, whether or not the vaginal bleed, the vagina bleeding time, smoking or drug habits, the history of insulin-dependent diabetes, and way of conception are. Additionally, the values of biochemical markers are normalized with their multiples of the medians (PAPP-A MoM, β-hCG MoM, uE3 MoM), which is an effective data normalization method in medical data [29], [30]. Another two features are the fetal chromosome karyotype and the result of telephone follow-up, both of which are the important basis to ascertain the real label of the samples. The former is available only for the high-risk pregnant women valued by the statistical mixture models, and the latter is available for the samples that have a successful delivery.

Moreover, the unpredictable correlations of the features exist in the data. For instance, according to [31], an inverse trend exists in the median MoM levels of the serum markers in relation with maternal weight.

B. EXPERIMENTAL SET-UP

1) PERFORMANCE EVALUATION

For model performance testing, tenfold cross-valuation (tenfold CV) is performed to partition the dataset into 10 folds and assure that each fold contains a similar number of the positives and negatives. We use different performance measurements, mainly the precision and recall curve (PRC), the receiver operating characteristic curve (ROC), the area under the PRC (AUPRC), and the area under the ROC (AUROC). With imbalanced data, the AUPRC is more informative than the AUROC [32], [33]. However, in the field of medicine, the ROC curve is more meaningful, because the sensitivity, also called the true positive rate (TPR), and 1-specificity, also called the false positive rate (FPR), have been widely regarded as the indicator of the effectiveness of prenatal screening.

TABLE 1. Groups of input features that are used as inputs to different methods compared in this paper.

Features	Combination of input features (IDs)							
	3	4a	4b	4c	5a	5b	5c	6
WOP	No	No	Yes	No	Yes	No	Yes	Yes
MA	No	Yes	No	No	Yes	Yes	No	Yes
GW	No	No	No	Yes	No	Yes	Yes	Yes
AFP MoM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
β -hCG MoM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
uE3 MoM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

2) COMPARISON WITH STATE-OF-THE-ART METHODS

We compared the CVIFLR with under-sampled ANN [19] and state-of-the-art methods for the imbalance learning, namely, BalanceBagging [20], BalanceCascade [22], SMOTEENN [21], and SMOTETomek [23]. All methods are designed for the imbalance data. We know that BalanceBagging and BalanceCascade are the ensemble leaning methods based on under-sampling and lost function strategies. Under-sampled ANN uses the K-nearest neighbor to reduce the population of the majority class. SMOTEENN and SMOTETomek utilize combined strategies targeting the optimum proportion of over-sampling and under-sampling to solve the class imbalance problem.

Moreover, we compare the effect of using different classifiers in the final stage. Besides LR, we also take common classifiers, Quadratic Discriminant Analysis (QDA), Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), and Random Forests (RF) in the comparison. All of these classifiers have adjustable classification performance with different decision thresholds and are more applicable to practical medical scenarios. For example, we can choose a smaller threshold to improve TPR at the cost of increased FPR, which is meaningful to reduce the miss rate for screening diseases.

IV. FEATURE SELECTION

According to [24], [25], IF is proposed for the data with continuous value features. Adding features with the discrete value will increase the height of isolation trees, but the detection rate of the whole model usually has no improvement. In this paper, only continuous value features are considerable as the input features of CVIFLR. Unfortunately, some of them are not available for most of samples both in minority and majority classes according to the will of participant. In consideration of this fact, we preliminarily choose 6 most common features including AFP MoM, β-hCG MoM, uE3 MoM, weight, maternal age, and gestational weeks of pregnant woman as the alternative input features. Among them, **AFP** MoM, β-hCG MoM, uE3 MoM have always been used as the key features for DS screening both in medicine statistical mixture models and machine learning methods [18], [34]–[36]. Then we select the optimal group of the features required for DS examinations according to the classification effect of VIFLR.

Table 1 presents the different combinations of the features used in the experiments. In the first row, we show the ID of every combination that corresponds to the figures in Section IV. WOP, MA, and GW stand for the weight, maternal age, and gestational weeks of pregnant woman, respectively. The word **Yes** indicates that the specific feature is used in the respective group. Similarly, the word **No** indicates a feature that is not used. All features in Table 1 are available and continuous for every case in the dataset used in this study.

V. RESULTS AND DISCUSSION

In this section, we present the results of the experimental comparison between CVIFLR and state-of-the-art imbalance learning methods for DS prediction in accordance with the experimental set-up described in the previous section. First, we compare these methods by using different metrics and show the tenfold CV results of PRC and ROC in Figs. 4 and 5. Then, by using the TPR and FPR metrics which are widely regarded as the indicators of the effectiveness of prenatal screening, we determine the best combination of the input features from those listed in Table 1. Finally, we compare different classifiers in the third stage of the proposed framework and demonstrate the results of the comparison in Fig. 7 and 8.

A. CVIFLR OUTPERFORMS STATE-OF-THE-ART METHODS

In Fig. 4, for all recall levels, CVIFLR (blue curve) reaches a higher precision and its AUPRC is larger than those of other state-of-the-art methods. These results are also confirmed by ROC curves in Fig. 5. Although the differences among CVIFLR, BalanceBagging and BalanceCascade as measured by the AUROC are not very large, CVIFLR (blue curve) always has larger TPR values especially at a small level of FPR, which is meaningful for the DS screening in medicine.

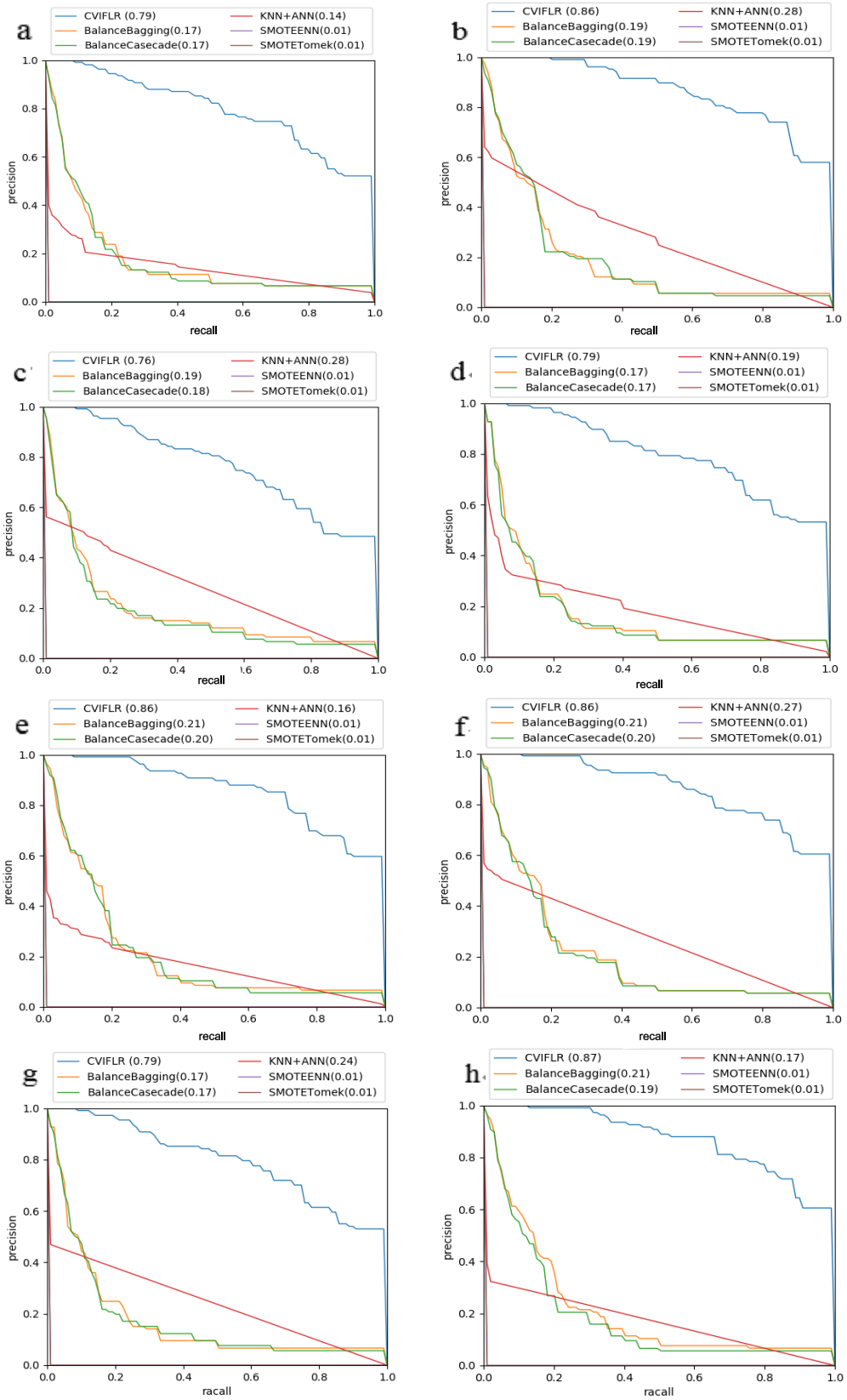


FIGURE 4. Comparison of PRCs of methods. Different combinations of input futures are used in each subgraph. Subgraph a–h correspond to feature combinations (IDs) of 3 ,4a,4b,4c,5a,5b,5c,6. Numbers in parentheses represent AUPRC values.

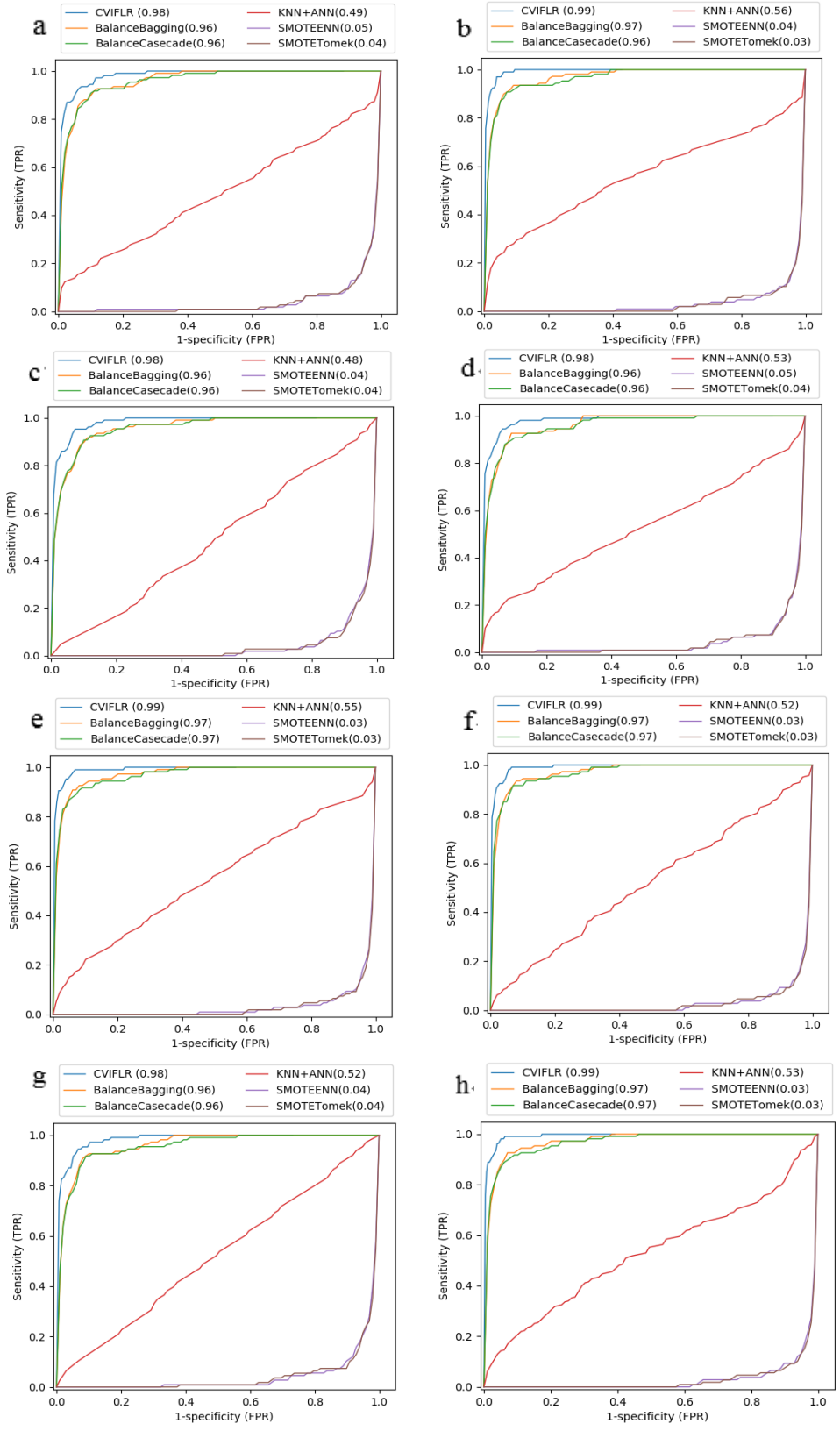


FIGURE 5. Comparison of ROCs of methods. Different combinations of input futures are used in each subgraph. Subgraphs a–h correspond to feature combinations (IDs) of 3, 4a, 4b, 4c, 5a, 5b, 5c, and 6. Numbers in parentheses represent AUROC values.

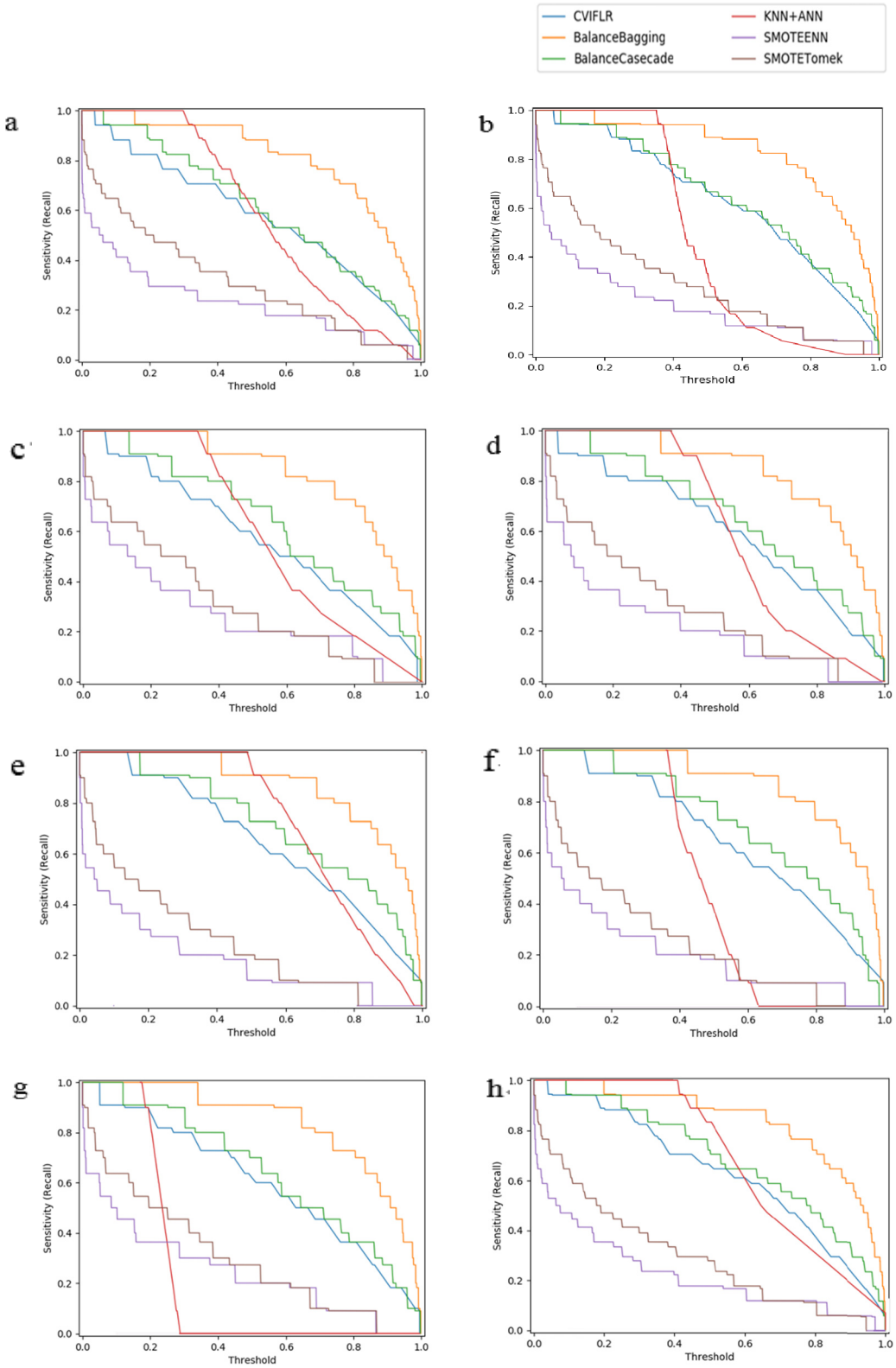


FIGURE 6. Sensitivity (Recall) comparison of methods vs. varying normalized score threshold.

The experimental results show that our proposed method is well-suited for the DS prediction in imbalanced and feature-correlated experimental settings. In this method, the cascaded framework of pre-judgment, voting, and final judgment allows the review for suspected samples to achieve high prediction accuracy. It only requires a few or no minority class

data to train the IF at the first stage, which is for the success of the imbalanced data learning.

On the basis of AUROC in Fig. 5, the ensemble learning algorithms, BalanceBagging and BalanceCascade, are the second-best methods for DS prediction. However, with AUPRC, the under-sampled ANN ranks second in Fig. 4. Moreover, with the independent metrics, we find that the combined methods, SMOTEENN and SMOTETomek, show the poorest performance both in Fig. 4 and Fig. 5.

The results show that the combined methods are not feasible for the DS prediction, since the valuable cases for minority class are difficult to create through oversampling techniques when having unpredictable correlation of the features. Although, the ensemble learning methods may allow us to overcome slight imbalances by sampling techniques, on the MSS dataset in which the positives and negatives nearly result in an imbalance of 1:1000, the performances are not satisfactory. Finally, since the under-sampled ANN is a data-hungry learning method, it is not suitable for highly unbalanced datasets which has a small number of minority class cases, such as MSS dataset.

To show the thresholds used in AUPRC and AUROC in Fig.4 and Fig.5, we plot the sensitivity as a function of the thresholds predicted by CVIFLR and the other methods in Fig. 6. As the sensitivity (the Y-axis of ROC) is equal to the recall (the X-axis of PRC), we can find the threshold of each point in ROC and PRC in Fig. 6.

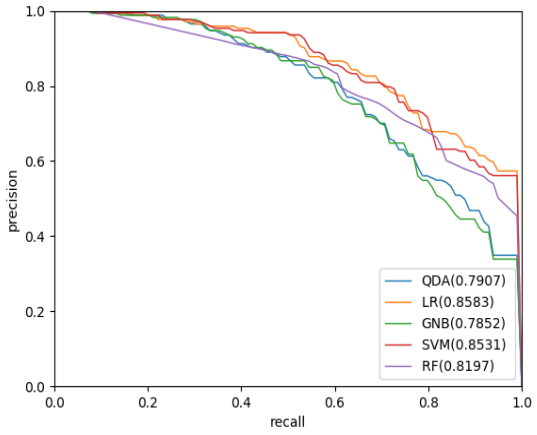


FIGURE 7. Comparison of ROCs of different classifier in final stage.

B. LR OUTPERFORMS OTHER CLASSIFIERS IN THE FINAL STAGE OF THE PROPOSED FRAMEWORK

Figures 7 and 8 show that LR outperforms other classifiers in the final stage of the proposed framework. In Fig.7, for all recall levels, LR (yellow curve) reaches a high precision and its AUPRC is larger than those of other classifiers. These results are also confirmed by ROC curves. In Fig.8, we can see that LR (yellow curve) always performs with a higher TPR at the same FPR, especially when FPR is small, which is more suitable for the DS screening. Additionally, LR is fairly efficient in terms of time and memory requirement, which is another factor for the final decision classifier in this paper.

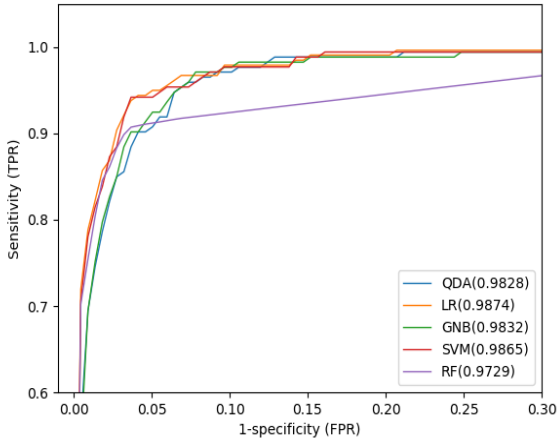


FIGURE 8. Comparison of PRCs of different classifier in final stage.

TABLE 2. FPR at different TPR levels with suggested input feature groups.

Features groups (IDs)							
3	4a	4b	4c	5a	5b	5c	6
TPR=85%							
2.1%	1.3%	3.1%	2.9%	1.3%	1.4%	6.1%	1.3%
TPR=90%							
5.3%	2.2%	5.9%	5.1%	2.6%	1.9%	12.1%	2.6%
TPR=95%							
10.7%	4.0%	7.0%	7.3%	5.2%	4.7%	27.3%	4.3%
TPR=100%							
26.4%	10.3%	21.9%	27.8%	28.0%	19.9%	39.0%	18.0%

C. BEST SCOMBINATION OF INPUT FEATURES

In Table 2, we give the FPRs at different TPR levels with the suggested groups of the input features listed in Table 1. In medical perspective, if the given TPR is the same, the smaller the FPR, the better the effect of prenatal screening. The experimental results show that the best combination of the features for CVIFLR is PAPP-A MoM, β-hCG MoM, uE3 MoM, and MA, which always results in the smallest FPR at different TPR levels

VI. CONCLUSION

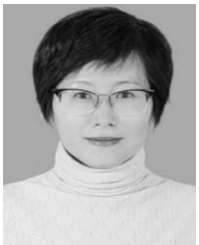
In this paper, we propose a framework called CVIFLR to solve the prediction problem for the DS, motivated by the increasing role of ensemble machine learning algorithms in the predictive analytics. The proposed method, despite having highly imbalanced data in the case study, produced an AUROC of 0.99 on the testing data. It is also found to be superior in the overall performance, when an extensive comparison is made with the state-of-the-art methods using different classification metrics. The experimental results show that the best suggested combination of input features is PAPP-A MoM, β-hCG MoM, uE3 MoM, and MA. With the suggested features, CVIFLR produced a TPR of 95% at the FPR of 4%.

The proposed framework can be used as the decision support system to predict DS. Furthermore, it offers

a potential classification method for the imbalanced and feature-correlated data, which may be helpful for biologists and physicians to screen or diagnose rare diseases.

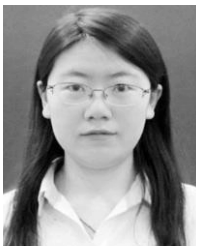
REFERENCES

[1] *Report on Birth Defects Prevention and Treatment of PRC*, (in Chinese), Ministry Health People's Republic China, Beijing, China, 2012.
[2] M. E. Weijerman and J. P. de Winter, "Clinical practice: The care of children with down syndrome," *Eur. J. Pediatr.*, vol. 169, no. 12, pp. 1445–1452, Dec. 2010.
[3] Y. Chen, X. Qian, J. Zhang, J. Li, A. Chu, and S. O. Schweitzer, "Preliminary study into the economic burden of down syndrome in China," *Birth Defects Res. A, Clin. Mol. Teratol.*, vol. 82, no. 1, pp. 25–33, Jan. 2008.
[4] K. Spencer, C. E. Spencer, M. Power, C. Dawson, and K. H. Nicolaides, "Screening for chromosomal abnormalities in the first trimester using ultrasound and maternal serum biochemistry in a one-stop clinic: A review of three years prospective experience," *BJOG, Int. J. Obstetrics Gynaecol.*, vol. 110, no. 3, pp. 281–286, Mar. 2003.
[5] K. H. Nicolaides, "First-trimester screening for chromosomal abnormalities," *Seminars Perinatol.*, vol. 29, no. 4, pp. 190–194, Aug. 2005.
[6] N. Maiz, C. Valencia, K. O. Kagan, D. Wright, and K. H. Nicolaides, "Ductus venosus Doppler in screening for trisomies 21, 18 and 13 and turner syndrome at 11–13 weeks of gestation," *Ultrasound Obstetrics Gynecol.*, vol. 33, no. 5, pp. 512–517, May 2009.
[7] X.-M. Bian and Q.-W. Qi, "Prenatal screening and diagnosis of chromosomal abnormalities has a long way to go," (in Chinese), *Chin J. Practical Gynecol. Obstetrics*, vol. 26, no. 12, pp. 889–891, Dec. 2010.
[8] S. De-Shun and Q. Li-Min, "Comparison of the efficacy of triple screening and quadruple screening in maternal serum," (in Chinese), *Guide China Med.*, vol. 15, no. 22, pp. 21–23, Aug. 2017.
[9] C. Enzensberger, C. Pulvermacher, J. Degenhardt, A. Kawacki, U. Germer, U. Gembruch, and R. Axt-Fliedner, "Fetal loss rate and associated risk factors after amniocentesis, chorionic villus sampling and fetal blood sampling," *Ultraschall Der Medizin Eur. J. Ultrasound*, vol. 33, no. 7, pp. E75–E79, Dec. 2012.
[10] G. Ashoor, A. Syngelaki, M. Wagner, C. Birdir, and K. H. Nicolaides, "Chromosome-selective sequencing of maternal plasma cell–free DNA for first-trimester detection of trisomy 21 and trisomy 18," *Amer. J. Obstetrics Gynecol.*, vol. 206, no. 4, pp. 322.e1–322.e5, Apr. 2012.
[11] X. Yuan, L. Xie, and M. Abouelenien, "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data," *Pattern Recognit.*, vol. 77, pp. 160–172, May 2018.
[12] H. Yang and Y.-P. P. Chen, "Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information," *Expert Syst. Appl.*, vol. 42, nos. 15–16, pp. 6168–6176, Sep. 2015.
[13] A. Mouelhi, H. Rmili, J. B. Ali, M. Sayadi, R. Doghri, and K. Mrad, "Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images," *Comput. Methods Programs Biomed.*, vol. 165, pp. 37–51, Oct. 2018.
[14] K. Buchan, M. Filannino, and Ö. Uzuner, "Automatic prediction of coronary artery disease from clinical narratives," *J. Biomed. Inf.*, vol. 72, pp. 23–32, Aug. 2017.
[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
[16] M. L. P. Bueno, A. Hommersom, P. J. F. Lucas, M. Lappenschaar, and J. G. E. Janzing, "Understanding disease processes by partitioned dynamic Bayesian networks," *J. Biomed. Inform.*, vol. 61, pp. 283–297, Jun. 2016.
[17] L. Wang, P. J. Haug, and G. D. Fiol, "Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository," *J. Biomed. Inform.*, vol. 69, pp. 259–266, May 2017.
[18] A. C. Neocleous, K. H. Nicolaides, and C. N. Schizas, "Intelligent non-invasive diagnosis of aneuploidy: Raw values and highly imbalanced dataset," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 5, pp. 1271–1279, Sep. 2017.
[19] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.
[20] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[22] G. E. Batista, A. L. Bazan, and M. C. Monard, "Balancing training data for automated annotation of keywords: A case study," in *Proc. WOB*, São Carlos, Brazil, Dec. 2003, pp. 10–18.
[23] G. Louppe and P. Geurts, "Ensembles on random patches," in *Proc. ECML PKDD*, 2012, pp. 346–361.
[24] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 413–422.
[25] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, p. 3, Mar. 2012.
[26] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.*, Jun. 2000, pp. 1–15.
[27] R. Matteo and G. Valentini, "Ensemble methods," in *Advances in Machine Learning and Data Mining for Astronomy*, vol. 22. Boca Raton, FL, USA: CRC Press, 2012, pp. 563–593.
[28] H.-F. Yu, F.-L. Huang, and C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," *Mach. Learn.*, vol. 85, nos. 1–2, pp. 41–75, 2011.
[29] K. O. Kagan, D. Wright, A. Baker, D. Sahota, and K. H. Nicolaides, "Screening for trisomy 21 by maternal age, fetal nuchal translucency thickness, free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A," *Ultrasound Obstetrics Gynecol.*, vol. 31, no. 6, pp. 618–624, Jun. 2010.
[30] K. O. Kagan, D. Wright, K. Spencer, F. S. Molina, and K. H. Nicolaides, "First-trimester screening for trisomy 21 by free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A: Impact of maternal and pregnancy characteristics," *Ultrasound Obstetrics Gynecol.*, vol. 31, no. 5, pp. 493–502, May 2010.
[31] J. J. Hsu, T. T. Hsieh, and Y. K. Soong, "Influence of maternal age and weight on second-trimester serum alpha-fetoprotein, total and free beta human chorionic gonadotropin levels," *Changgeng Yi Xue Za Zhi*, vol. 20, no. 3, pp. 181–186, Sep. 1997.
[32] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. 0118432.
[33] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, New York, Ny, USA, ACM, Jun. 2006, pp. 233–240.
[34] J. P. Bestwick, W. J. Huttly, and N. J. Wald, "The estimation of median nuchal translucency values between 10 and 14 weeks of pregnancy," *J. Med. Screening*, vol. 21, no. 2, pp. 110–112, Jun. 2014.
[35] T. Ghi, T. Arcangeli, F. Ravennati, G. Salsi, E. Montaguti, G. Pacella, E. Maroni, M. C. Pittalis, E. Pompilii, G. Pilu, and N. Rizzo, "Prenatal diagnosis versus first-trimester screening of trisomy 21 among pregnant women aged 35 or more," *J. Maternal-Fetal Neonatal Med.*, vol. 28, no. 6, pp. 674–678, May 2015.
[36] A. Kaul, C. Singh, R. Gupta, N. Arora, and A. Gupta, "Observational study comparing the performance of first-trimester screening protocols for detecting trisomy 21 in a North Indian population," *Int. J. Gynecol. Obstetrics*, vol. 137, no. 1, pp. 14–19, Apr. 2017.



LING LI was born in Qiqihar City, Heilongjiang Province, China, in 1965. She received the B.S. degree in computer applications from Changchun Institute of Posts and Telecommunications, Changchun, Jilin province, in 1987. She received the M.S. degree in computer applications from Jilin University of Technology, Changchun, Jilin, China, in 1994.

Since 2000, she has been an Associate Professor with the Communication Engineering Department, Jilin University. She is the author of more than 20 articles, and more than five inventions. Her research interests include communication network protocol and architecture and application of next generation network technology, application development and research of cloud computing, construction of large data platform, machine learning and application of big data algorithm in communication network, medicine, and other fields.



WANYING LIU was born in Songyuan City, Jilin Province, China, in 1994. She received the B.S. degree in communication engineering from Jilin University, in 2017. She is currently pursuing the M.S. degree in communication and information system, Jilin University, Changchun, Jilin, China.

Her research interests include development of machine learning and data mining algorithms in medicine and bioinformatics.

HONGGUO ZHANG was born in Xinxiang City, Henan Province, China, in 1977. He received the B.S. degree in biology from Xinyang Normal University, Henan Province, in 2000, and the M.S. and Ph.D. degrees in cell biology from Northeast Normal University, China, in 2006 and 2010, respectively.

Since 2010, he has been a Research Assistant in First Hospital, Jilin University. He is the author of one book, and more than 70 articles. His research interests include biomedical engineering technology, reproductive medicine, and the technology of prenatal diagnosis.

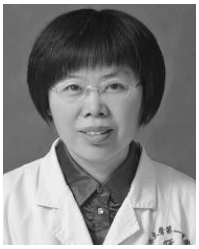
He was a recipient of Science and Technology Progress Award in Jilin Province, and Natural Science Academic Achievement Award in Jilin Province.

YUTING JIANG was born in Changchun City, Jilin Province, China, in 1986. She received the B.S. degree in biology from Wuhan Institute of technology, Hubei Province, in 2009, and the M.S. and Ph.D. degrees in cell biology from Jilin University, China, in 2013. She is currently pursuing the Ph.D. degree in cell biology at the First Hospital of Jilin University, Changchun, Jilin, China.

From 2013 to 2018, she was with Assisted Reproduction Laboratory. She is the author of three articles (first author). Her research interests include embryonic development disorder, genetic factor for male infertility, and prenatal diagnosis.

XIAONAN HU received the B.S. degree in bioscience from Changchun Normal University, Changchun, China, in 2012. She is currently pursuing the M.S. degree in cytobiology from the First Hospital of Jilin University, Changchun, Jilin, China.

From 2016 to 2017, she was a Research Assistant with the Reproductive Medicine Laboratory, First Hospital of Jilin University. Since 2017, she has been an Assistant Secretary. She participated in the research and writing of several articles and a research project. Her research interests include cytobiology, reproductive medicine, medical Genetics, prevention and control of birth defects, prenatal screening of Down syndrome, and genetic factors of male infertility.



RUIZHI LIU was born in Jilin, China, in 1965. She received the B.S. degree in medicine from Jilin Medical University, in 1988, and the M.S. and Ph.D. degrees in pathophysiology from the Norman Bethune University of Medical Science (NBUMS), in 1998.

Since 2006, she has been an Associate Professor with the Department of Obstetrics and Gynecology, Clinical Hospital of Jilin University. She is the author of four books, and more than 150 articles. She set up the Center of Reproductive Medicine and Center of Prenatal Diagnosis in the First Hospital of Jilin University, and is currently appointed as Chief. Her research interests include chromosomal polymorphisms, spermatogenesis failure, azoospermia factor, microdeletions, non-obstructive azoospermia-related gene mutations, and prenatal genetic diagnosis.

Dr. Liu was selected to be a member of the Chinese Society of Genetic Medicine (GSC) and Chinese Society of Reproductive Medicine (CSRM).