An Augmented Model with Inferred Blood Features for the Self-diagnosis of Metabolic Syndrome

Tianshu Zhou[1,2] Ying Zhang[1,2] Chengkai Wu[1] Chao Shen[3] Jingsong Li[1,2] Zhong Liu[3]

[1] Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, People's Republic of China
[2] Connected Healthcare Big Data Research Center, Zhejiang Lab, Hangzhou, People's Republic of China
[3] Health Management Center, The First Affiliated Hospital, Medical School of Zhejiang University, Hangzhou, People's Republic of China

Methods Inf Med 2020;59:18–30.

Address for correspondence Jingsong Li, PhD, Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, People's Republic of China (e-mail: ljs@zju.edu.cn).

Zhong Liu, MD, PhD, Health Management Center, The First Affiliated Hospital, Medical School of Zhejiang University, Hangzhou 310006, People's Republic of China (e-mail: liuzhongzheyi@zju.edu.cn).

Abstract

Background and Objectives The penetration rate of physical examinations in China is substantially lower than that in developed countries. Therefore, an auxiliary approach that does not depend on hospital health checks for the diagnosis of metabolic syndrome (MetS) is needed.

Methods In this study, we proposed an augmented method with inferred blood features that uses self-care inputs available at home for the auxiliary diagnosis of MetS. The dataset used for modeling contained data on 91,420 individuals who had at least 2 consecutive years of health checks. We trained three separate models using a regularized gradient-boosted decision tree. The first model used only home-based features; additional blood test data (including triglyceride [TG] data, fasting blood glucose data, and high-density lipoprotein cholesterol [HDL-C] data) were included in the second model. However, in the augmented approach, the blood test data were manipulated using multivariate imputation by chained equations prior to inclusion in the third model. The performance of the three models for MetS auxiliary diagnosis was then quantitatively compared.

Results The results showed that the third model exhibited the highest classification accuracy for MetS in comparison with the other two models (area under the curve [AUC]: 3rd vs. 2nd vs. 1st = 0.971 vs. 0.950 vs. 0.905, $p < 0.001$). We further revealed that with full sets of the three measurements from earlier blood test data, the classification accuracy of MetS can be further improved (AUC: without vs. with = 0.971 vs. 0.993). However, the magnitude of improvement was not statistically significant at the 1% level of significance ($p = 0.014$).

Conclusion Our findings demonstrate the feasibility of the third model for MetS homecare applications and lend novel insights into innovative research on the health management of MetS. Further validation and implementation of our proposed model might improve quality of life and ultimately benefit the general population.

Keywords
- self-care
- machine learning
- systolic blood pressure
- diastolic blood pressure
- metabolic syndrome

received
December 4, 2018
accepted after revision
March 15, 2020

DOI https://doi.org/10.1055/s-0040-1710382.
ISSN 0026-1270.

Introduction

Metabolic syndrome (MetS) is a multimetabolic disorder that can cause hyperglycemia, hypertension, hyperlipidemia, atherosclerosis, thrombosis, and inflammation.[1,2] The most serious clinical consequences of MetS include type II diabetes and cardiovascular disease, which is the leading cause of death in China.[3,4] MetS is also closely associated with multiple types of cancer.[5–10] Due to economic development and lifestyle changes in the Chinese population, the prevalence of MetS is increasing each year, and the age range of affected patients has shown a significant decreasing trend compared with that of previous years.[11,12]

Regular medical examination has provided an avenue for discovery and risk reduction in MetS.[13] However, a low national coverage rate of medical examinations (32.75% in 2016) has been reported in China,[14] and there are serious problems regarding insufficient and unevenly distributed medical resources.[15] Since the increasing prevalence of MetS places increasing pressure on economic development and social health protection policies, the improvement and diversification of MetS examinations should be explored.

Several studies were performed to lower the threshold of self-management for MetS. Ichikawa et al and Shimoda et al[16–18] proposed prediction models for identifying health guidance candidates. This method identifies a health guidance candidate using available electronic health records, including demographic information (sex, age, height, and weight), and examination results (blood pressure and levels of blood test indexes). However, the method, which relies on clinical data, is not highly applicable to China's national conditions given people's reduced awareness of medical examinations. In such scenarios, a home-based approach to monitoring MetS is needed.

Blood test data (e.g., triglyceride [TG] data) are necessary for the diagnosis of MetS. However, blood test data are unavailable at home, which makes self-diagnosis of MetS difficult. There is a correlation between certain blood test measurements and home-based measurements.[19] By using a large amount of health check-up data, the relationship between blood test data and home-measurable data may be helpful for predicting appropriate blood test data and can thus contribute to MetS auxiliary diagnosis.

We proposed a novel augmented method with inferred blood features based on a large amount of health check-up data for MetS self-care to compensate for the lack of blood test data in MetS self-care. Our model ensures timely, convenient, and accurate MetS risk assessment in the context of low national medical examination penetration rates, which could be significant for improving the quality of national health.

Methods

Data Source

Data were collected from the health inspection database of the First Affiliated Hospital, Medical School of Zhejiang University. The database contained data from 295,241 physical examinations performed between January 2011 and September 2017. The data include the sex, age, history, medication history, lifestyle records, height, weight, body mass index, systolic blood pressure (SBP), and diastolic blood pressure (DBP) of patients, as well as the following blood test data: fasting blood glucose (FBG), TG levels, and high-density lipoprotein cholesterol (HDL-C). The scope of the health check targets includes public institutions, government agencies, private companies, etc., without restrictions on sex or age. The baseline data were generated from the first health check of this population, which excluded patients with baseline coronary heart disease, type I diabetes, and familial hyperlipidemia.

A total of 96,506 people had at least 2 consecutive years of complete health check data. Among them, 15,984 (16.6%) had MetS the first year, and 17,060 (18.7%) had MetS the following year. Of these individuals, 5,086 (who converted to non-MetS the following year) were not included in the study. Therefore, 91,420 individuals were included, among whom 10,898 had MetS at baseline and 6,162 converted from non-MetS to MetS.

►Table 1 lists the demographic and clinical characteristics of the study population at baseline and the following year. Among them, 47,098 were males, accounting for 51.5% of the total sample. The average age of the individuals was 43.7 years. Compared with the first year, all features deteriorated slightly on average the following year (body mass index [BMI], systolic blood pressure [SBP], diastolic blood pressure [DBP], FBG, and TG were higher; HDL-C was lower).

We set up positive samples for people who were diagnosed with MetS the following year. ►Fig. 1 shows the distribution of the densities of the various indicators in the positive and negative samples. We observed significant differences between the positive samples and the negative

Table 1 Characteristics of the datasets

Variable	$n = 91{,}420$	
	Previous year	Subsequent year
Mean age (SD)		43.7 (14.0)
Percentage of male participants (total number)		51.5 (47,098)
Percentage of non-MetS patients (total number)	88.1% (80,522)	81.3% (74,360)
Percentage of MetS patients (total number)	11.9% (10,898)	18.7% (17,060)
Mean BMI (SD)	22.8 (3.1)	22.9 (3.2)
Mean SBP (SD)	121.1 (16.6)	122.3 (16.8)
Mean DBP (SD)	74.0 (10.4)	74.4 (10.7)
Mean FBG (SD)	4.92 (0.90)	4.97 (0.96)
Mean TG (SD)	1.297 (0.879)	1.358 (0.923)
Mean HDL-C (SD)	1.335 (0.344)	1.323 (0.347)

Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; MetS, metabolic syndrome; SBP, systolic blood pressure; SD, standard deviation; TG, triglyceride.

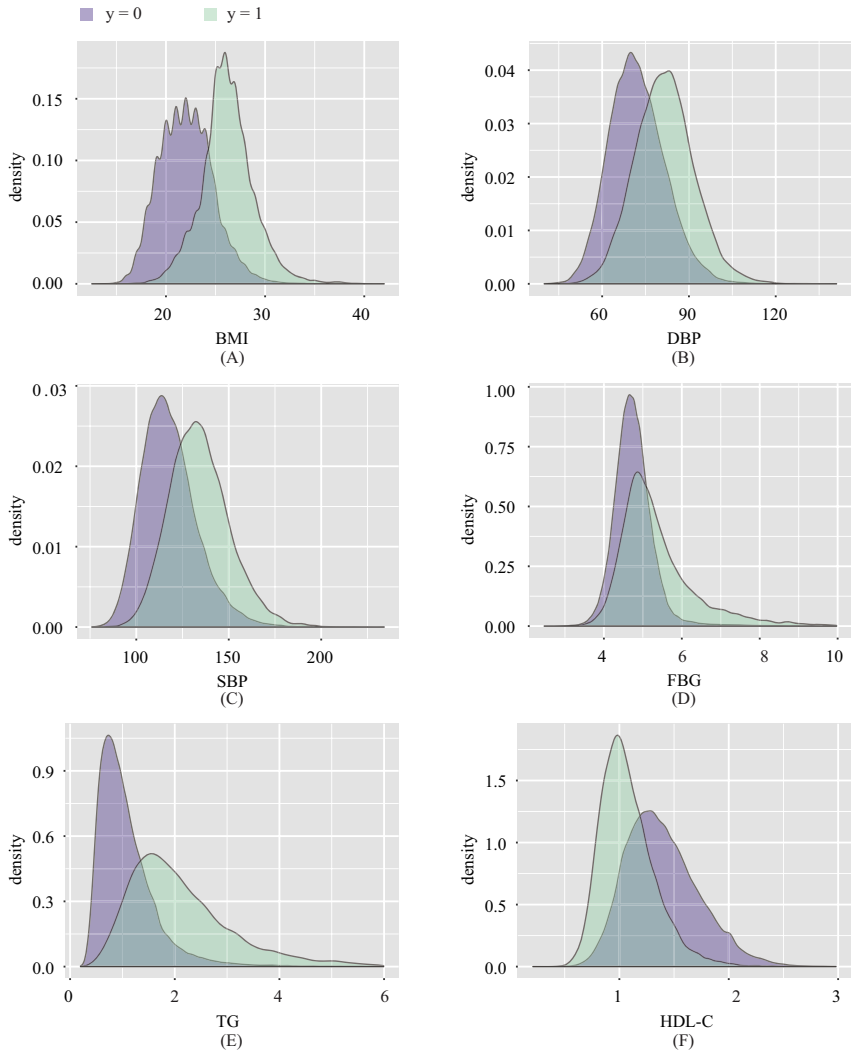

Fig. 1 (A–F) Density map for BMI, DBP, SBP, FBG, TG and HDL-C for positive and negative MetS samples. BMI, body mass index; DBP, diastolic blood pressure; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; MetS, metabolic syndrome; SBP, systolic blood pressure; TG, triglyceride.

samples with respect to the distribution of the inspection indicators, which demonstrated that the health check indicators of the dataset in the baseline year were significantly related to future diagnosis of MetS.

In this study, the diagnosis of MetS was based on the new MetS definition criteria that were jointly developed by the 2009 guidelines of the International Diabetes Federation and the American Heart Association/National Heart, Lung and Blood Institute.[20] The criteria are primarily employed to assess the risk of obesity (BMI or waist circumference) and cardiovascular risk (SBP, DBP, FBG, TG, and HDL-C). We used BMI as an assessment of obesity risk. The latest study published in *Metabolism* reported that the use of BMI to evaluate MetS risk is equivalent to the use of waist circumference. In other words, the use of BMI to evaluate MetS risk has greater clinical potential than the use of waist circumference.[21]

Based on the analysis of the population data, several methods were utilized for data processing. ►**Fig. 2** displays a map of the research workflow.

Data Preprocessing

Standardization

To improve the comparability among the feature indexes and the convergence speed and data processing performance, we applied the *Z*-score standardization method to normalize each continuous variable.[22] This method normalizes the data based on the mean and standard deviation (SD) of the raw data. *Z*-scores have a mean of zero and a SD of one; they are informative when the empirical distribution is close to a normal distribution. In such cases, *Z*-scores may be used to compare relative locations of values from distributions with different means or SDs.

Sample Balance

A serious class imbalance was observed in the dataset, and the number of patients who had MetS the following year (18.7% of the total) was significantly smaller than the untransformed population. To prevent deviations in the results and improve the results, we employed the synthetic minority oversampling

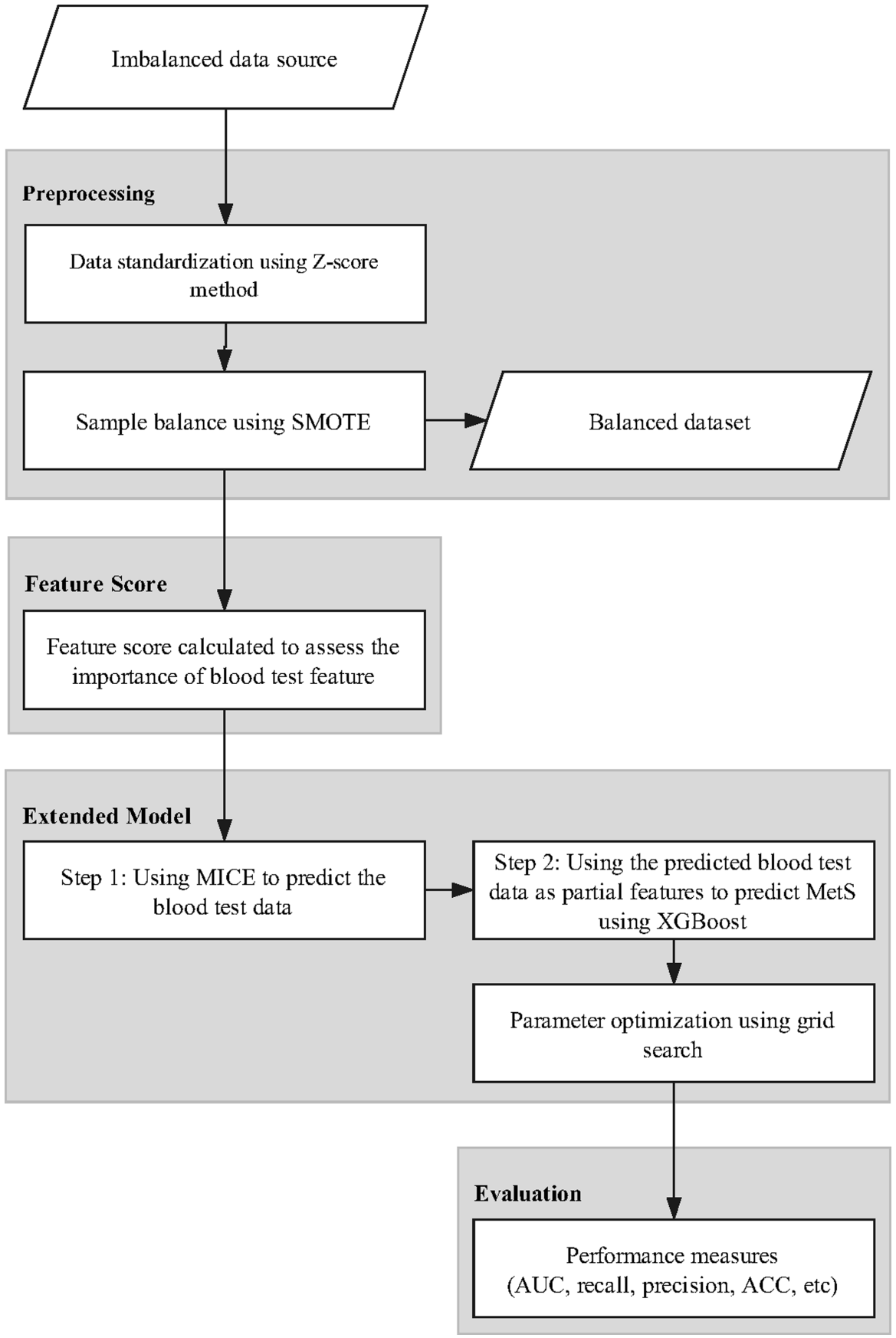

Fig. 2 Map of the research process. AUC, area under the curve; ACC, accuracy; MICE, multivariate imputation by chained equations; SMOTE, synthetic minority oversampling technique.

technique (SMOTE) to solve the problem of unbalanced categorical data.[23]

The SMOTE is an improved oversampling technique that is based on a random oversampling algorithm. The main idea is to use the similarity among the few existing classes of samples in the feature space to create artificial data. The basic principle is to use Eq. (1) to linearly interpolate between the closely spaced samples of the minority class to generate a new minority sample. For each sample from the minority class (x), five samples from the minority class with the smallest Euclidean distance from the original sample were identified (nearest neighbors), and one of them was randomly chosen (x^{NN}). Because the data constructed by the algorithm is a new sample that does not exist in the original dataset, the risk of overfitting to the minority-class data is minimized.[24]

$$x^{SMOTE} = x + u \times (x^{NN} - x) \quad (1)$$

where u was randomly chosen from $U(0,1)$. u was the same for all variables but differed for each SMOTE sample; this guarantees that the SMOTE sample lies on the line joining the two original samples used to generate it.

A randomly sampled dataset that represents one-fifth of the data was employed as a test dataset. Of the remaining data, four-fifths were subsampled as the training dataset, and the rest were used for validation. The training dataset consisted of 58,509 people, the validation dataset consisted of 14,627 people, and the test dataset consisted of 18,284 people. To improve the performance of the classifier, we used the SMOTE implementation from the DMwR package[25] of R software (version 3.4.3) to oversample the unbalanced training dataset. After SMOTE oversampling, 69,335 training samples, of which 21,838 were positive samples (31.5% of the total), were obtained.

Features

To understand the impact of existing health check indicators on the development of MetS, we generated a regularized gradient-boosted decision tree model using eXtreme Gradient Boosting (XGBoost) to estimate the importance of the model features, which indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function. The feature importance scores are then averaged across all of the decision trees within the model.

By estimating the feature importance, we obtain three indicators: Gain, Cover, and Frequency. Features are classified by Gain. Gain is the improvement in accuracy brought by a feature to the branches it is on. Cover measures the relative quantity of observations related to a feature. Frequency is a simpler way to measure the Gain. It counts only the number of times a feature is used in all generated trees.

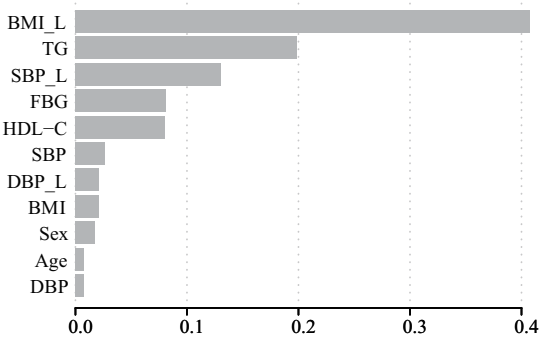

Fig. 3 Order of importance of the model features. BMI, body mass index; DBP, diastolic blood pressure; FBG, fasting blood glucose; HDL, high-density lipoprotein; MetS, metabolic syndrome; SBP, systolic blood pressure; TG, triglyceride.

Features included sex, age, BMI, SBP, DBP, TG, FBG, and HDL-C of the previous year and BMI, SBP, and DBP of the subsequent year, which were referred to as body mass index of the subsequent year (BMI_L), systolic blood pressure of the subsequent year (SBP_L), and diastolic blood pressure of the subsequent year (DBP_L). Positive samples were patients who were diagnosed with MetS in the subsequent year. The inputs were the health check indicators for 2 consecutive years of the study population. The output was the diagnosis of MetS in the subsequent year. We calculated the importance of the features to assess the extent to which these features (2-year home-based data and blood test data) affected the classification of MetS. ►Fig. 3 and ►Table 2 present the prioritization of each feature in the model.

As shown, the contribution of BMI_L to the outcome of MetS was the largest, and the third largest was SBP_L, which reflects that the features of the subsequent year are important to the

Table 2 Model feature importance

Feature	Gain	Cover	Frequency
BMI_L	0.408	0.190	0.106
TG	0.199	0.149	0.124
SBP_L	0.131	0.113	0.080
FBG	0.081	0.140	0.129
HDL-C	0.080	0.086	0.139
SBP	0.026	0.062	0.088
DBP_L	0.022	0.075	0.077
BMI	0.021	0.089	0.113
Sex	0.018	0.031	0.029
Age	0.007	0.034	0.060
DBP	0.007	0.032	0.055

Abbreviations: BMI-L, body mass index of the subsequent year; DBP, diastolic blood pressure; DBP-L, diastolic blood pressure of the subsequent year; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; SBP, systolic blood pressure; SBP-L, systolic blood pressure of the subsequent year; TG, triglyceride.

Table 3 Variables of each model

Model	Variables
HOME	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L
RBTIBE	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L, TG, HDL-C, FBG
IB	Step 1: sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L
	Step 2: sex, age, BMI, SBP, DBP, TG (inferred), HDL-C (inferred), FBG (inferred), BMI_L, SBP_L, DBP_L

Abbreviations: BMI-L, body mass index of the subsequent year; DBP, diastolic blood pressure; DBP-L, diastolic blood pressure of the subsequent year; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; SBP, systolic blood pressure; SBP-L, systolic blood pressure of the subsequent year; TG, triglyceride.

current recognition. The blood test indexes of the previous year (TG, HDL-C, and FBG) comprise the next largest contribution, which means that historical blood test data are relatively important for the outcome of MetS.

The mean and SD of the BMI and BMI_L shown in ►Table 2 are similar; however, the importance of these two features is significantly different, as shown in ►Fig. 3. We calculated the differences between the BMI_L and BMI for all individuals and performed a *t*-test with the differences and the value zero. The resulting *p*-value is 0.022; therefore, the difference between the 2-year BMI and zero is statistically significant at the 5% level of significance. This result shows that for the same individual, the BMI in year 1 does not fully reflect the BMI_L in year 2 (►Table 3).

Clinical Feature Augmented Model

It is possible to evaluate the risk of MetS in the following year by using only home-based data; however, blood test data contribute to MetS diagnosis (►Fig. 3). The substantial importance of the data from the three blood tests has become an important basis for us to interpolate the data from the three blood tests. Therefore, in our proposed method (►Fig. 4C), inferred blood test data could be helpful as supplementary data.

The goal is to evaluate MetS risk in the following year based on the absence of a health check-up (i.e., no blood test data). We aim to use a large amount of health check-up data to obtain a model that can predict blood test data by learning the relationship between home-based data and blood test data to provide additional effective features for the final classification model. The augmented model consists of two steps. In the first step, blood test features are inferred by the multivariate imputation by chained equations (MICE).[26] In the second step, the results obtained from the first step are combined with the home-based data for the modeling of MetS by using the regularized gradient-boosted decision tree algorithm.[27]

MICE

MICE is a practical approach to creating imputed datasets based on a set of imputation models, with one model for each variable with missing values. MICE is an increasingly popular method of performing multiple imputations. Here, we outlined the MICE algorithm for a set of variables, x_1, ..., x_k, some or all of which have missing values. Initially, all missing values are filled in at random. The first variable (say x_1) with missing values is regressed on all other variables x_2, ..., x_k The estimation is restricted to individuals with observed x_1 Missing values in x_1 are replaced by simulated draws from the posterior predictive distribution of x_1, an important step known as *proper imputation*. Next, x_2 with missing values is regressed on all other variables $x_1, x_3, ..., x_k$ and using the imputed values of x_1. Again, missing values of x_2 are replaced by draws from the posterior predictive distribution of x_2. The process is repeated in turn; one such round is called a *cycle*. The procedure is repeated for several cycles to produce a single imputed data point to stabilize the results, and the whole procedure is repeated independently m times to give m imputed data points. MICE has the ability to handle different variable types (continuous, binary, unordered categorical, and ordered categorical) as each variable is imputed using its own imputation model.[28] Compared with *k*-nearest neighbors interpolation and recursive partitioning and regression tree interpolation, the MICE interpolation method has better flexibility and higher precision. We applied the MICE package in R to perform interpolation.

Regularized Gradient-Boosted Decision Tree

The regularized gradient-boosted decision tree algorithm is an algorithm implemented by XGBoost.[29] Compared with the traditional gradient boosting decision tree algorithm, the regularized gradient-boosted decision tree method adds a regularization term helping to smooth the final learned weights to reduce the risk of overfitting. The regularized objective tends to choose a model that employs simple and predictive functions. The objective function consists of a loss function and complexity, which limits the number of leaves and prevents overfitting to some extent; the function is defined as

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)$$

where $\Omega(f) = \tilde{a}T + \frac{1}{2}\ddot{e}||\grave{u}||^2$.

Here, i is the sample id, k is the tree id (number of rounds), $l(\hat{y}_i, y_i)$ represents the prediction error of the ith sample, $\Sigma_k \Omega(f_k)$ penalizes the complexity of the tree, T is the number of leaf nodes, and ω is the value of the node. When the regularization parameter is set to zero, the objective will fall back to the traditional gradient tree boosting.

The tree ensemble model in Eq. (2) includes functions as parameters and is trained in an additive manner. For each iteration, the training objective function of a tree can be written as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

where $\hat{y}_i^{(t)}$ is the prediction of the ith instance at the $t-1$ iteration, which is employed to fit the residual $f(x)$. The

objective function is approximated by Taylor's second-order expansion as follows:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (4)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are the first- and second-order gradient statistics on the loss function.

In addition to the regularized objective, shrinkage and column subsampling are used to further prevent overfitting.

Model Generation

As shown in ►Fig. 4C, during the training phase of the augmented model with inferred blood features (abbreviated as IB), the original training data were subsampled into 10 equal parts. Each time, nine complete parts were used to interpolate the blood test data for the one remaining missing part. After 10 imputations, all datasets had inferred blood test values. MICE was used to impute blood test data from home-based data, and the inferred results were provided to the regularized gradient-boosted decision tree algorithm as additional features. In the testing phase, only home-based features were used in the test dataset, and blood test data were inferred by the same method. The prediction of the blood test data using MICE in the test dataset utilized a priori knowledge of the large training dataset.

We compared the performance of IB with those of two other models. One model (abbreviated as HOME) was given only the home-based features (►Fig. 4A), and the other (abbreviated as RBTIBE) was trained with extra blood test data (►Fig. 4B). The augmented model and the two other models were compared.

In addition to the HOME model, another baseline model could be the one that includes only features from year 1. However, the main goal of this work is to provide a continuous

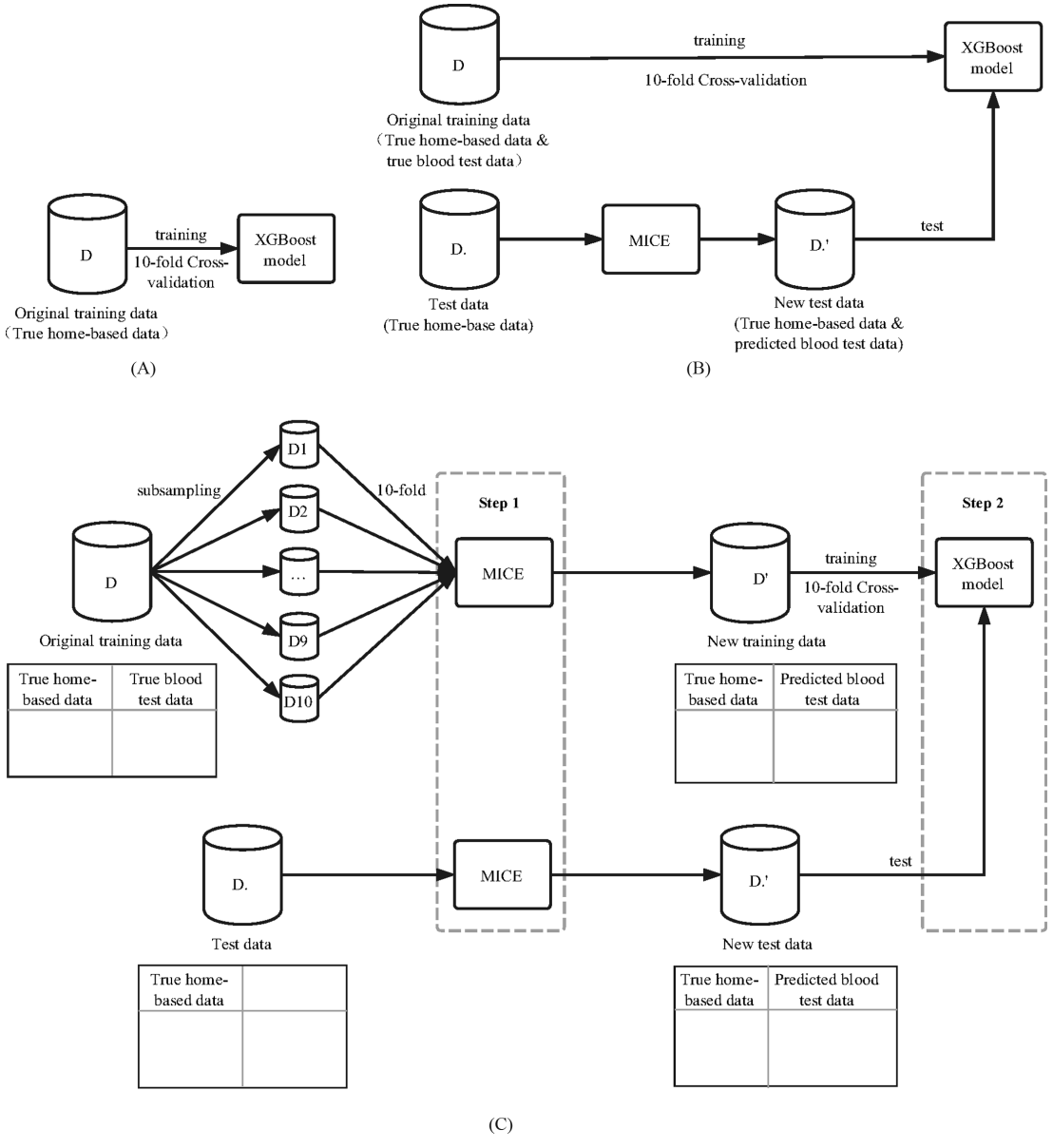

Fig. 4 (**A–C**) Block diagram for the three models.

self-assisted diagnosis of MetS, so we did not use previous data to predict the risk of developing MetS in the future. Therefore, we chose HOME as the baseline model, which continuously uses the latest physiological data (BMI, SBP, and DBP) as the input. Furthermore, the IB model proposed here also requires the latest physiological data to infer the blood features for modeling; therefore, for a consistency comparison, the HOME model was selected as the baseline model.

HOME contained home-based variables (sex, age, BMI, SBP, DBP, BMI_L, SBP_L, and DBP_L) and used only home-based variables to directly train the regularized gradient-boosted decision tree model to achieve a MetS auxiliary diagnosis, while RBTIBE was trained using additional true blood test data (TG, HDL-C, FBG); accordingly, the three features missing from the test dataset were interpolated using MICE. IB consisted of two steps: step 1 used home-based variables (sex, age, BMI, SBP, DBP, BMI_L, SBP_L, and DBP_L) to predict blood test data (TG, HDL-C, FBG), and step 2 merged the home-based variables and the inferred values from the first step for MetS modeling. Compared with HOME, the difference was that the training data had additional blood test data; compared with RBTIBE, the difference was that the inferred blood test features were used for regularized gradient-boosted decision tree training rather than real blood test features. Tenfold cross-validation was applied in the boosting part of the three models for parameter adjustment and selection, which ensured the reliability of area under the curve (AUC) and limited overfitting to some extent.

Our purpose is to achieve a better model for MetS self-care. Blood test data are unavailable at home; therefore, the three models were generated to compare their performance using a test dataset that contains only home-based data.

In the regularized gradient-boosted decision tree model, parameter optimization was performed using a grid search. The parameters were general parameters, booster parameters, and task parameters. We chose a relatively high learning speed (0.3) and the optimal number of trees based on the selected learning rate. We prioritized tree-specific parameters (max_depth, min_child_weight, gamma, subsample, colsample_bytree) for decided learning rate and number of trees. Tune regularization parameters (lambda, α) were optimized to help reduce model complexity and enhance performance. Then, we lowered the learning rate and decided the optimal parameters. ►Table 4 lists the final classifier parameter values.

Table 4 Parameters of the regularized gradient boosted decision tree model

Parameter	Model		
	HOME	RBTIBE	IB
nrounds	100	100	100
booster	gbtree	gbtree	gbtree
objective	reg:logistic	reg:logistic	reg:logistic
eta	0.1	0.1	0.1
gamma	0.6	0.5	0.4
max_depth	6	6	6
max_delta_step	0	0	0
min_child_weight	1	1	1
subsample	0.9	0.9	0.8
colsample_bytree	0.5	0.7	0.8

Evaluation Metrics

The AUC, sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), precision (positive predictive value, PPV), negative predictive value (NPV), accuracy (ACC), F1 score, and the area under the precision-recall curve (AUPRC) were used to evaluate the predictive performance of the three models. In predictive analytics, the number of false positives, false negatives, true positives, and true negatives in a confusion matrix are written relatively as FP, FN, TP, and TN, respectively. The calculation formulas of the evaluation metrics are as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} = 1 - \mathrm{FNR} \quad (5)$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} = 1 - \mathrm{FPR} \quad (6)$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (7)$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \quad (8)$$

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{P+N} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN++FP+FN}} \quad (9)$$

$$F1 = 2 * \frac{\mathrm{PPV*TPR}}{\mathrm{PPV+TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}} \quad (10).$$

The receiver operating characteristic (ROC) curve[30] is plotted with the TPR as the ordinate and the false positive rate as the abscissa, which is often used to evaluate the merits of a binary classifier. The precision—recall (PR) graph[31] takes precision as the ordinate and recall as the abscissa, which visually shows the recall and precision of the learner on the sample.

Results

Model Comparison

In the first step of the augmented model, our goal was to obtain the lowest root mean square error (RMSE) for each of the predicted metrics using the strategy. ►Table 5 lists the RMSE and mean absolute percentage error of MICE.

Table 5 Interpolation effect of the MICE method

Interpolation accuracy	TG	HDL-C	FBG
RMSE	0.0693	0.0734	0.0763
MAPE	0.0173	0.0267	0.0276

Abbreviations: FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; MAPE, mean absolute percentage error; MICE, multivariate imputation by chained equations; RMSE, root mean square error; TG, triglyceride.

Table 6 Performance of the three models

Model	AUC	95%CI	Sensitivity	Specificity	Precision	NPV	F1	ACC	AUPRC	*p*-Value of AUC
HOME	0.905	0.902–0.907	0.702	0.897	0.609	0.929	0.652	0.860	0.703	<0.001
RBTIBE	0.950	0.949–0.951	0.809	0.922	0.705	0.955	0.753	0.901	0.842	Na
IB	0.971	0.970–0.971	0.856	0.935	0.751	0.966	0.800	0.920	0.917	<0.001

Abbreviations: AUC, area under the curve; ACC, accuracy; AUPRC, area under the precision-recall curve; CI, confidence interval; IB, inferred blood features; NPV, negative predictive value.

Calculations of the AUC, sensitivity (TPR), specificity (TNR), precision (PPV), NPV, ACC, F1 score, and AUPRC of the three models are shown in ►Table 6. All metrics were computed at the same threshold of 0.425. We obtained the performance of the test dataset in the model by ROC curve (►Fig. 5A) and PR graph (►Fig. 5B).

The AUC value of the test dataset reflects the total discriminative power of the classifier.[32] As shown in ►Table 6, the AUCs of HOME and RBTIBE are 0.905 (95%CI: 0.902–0.907) and 0.950 (95%CI: 0.949–0.951), respectively. The total performance of RBTIBE is greater than that of HOME ($p < 0.001$), which indicates that RBTIBE has higher reliability and accuracy.

Furthermore, the performance of each indicator of IB is better than that of RBTIBE (AUC: 0.971 vs. 0.950, $p < 0.001$; ACC: 0.920 vs. 0.901; F1: 0.800 vs. 0.753; AUPRC: 0.917 vs. 0.842), confirming the advantage of the blood test data imputation in the training process. In the prediction of true positives, recall was increased to a value of 0.859 (RBTIBE: 0.809), meaning that IB has a better precise positioning rate and a lower missing rate for people at high risk of MetS. Additionally, the precision in the test dataset was more satisfying, and the specificity was improved from 0.922 to 0.935, which reflects the improved correct recognition rate for patients at low risk of MetS of IB.

The performance improvement of IB compared with HOME was due to the input of additional inferred blood test features, which are important factors in the diagnosis of MetS, in both the training and testing processes. Interestingly, the only difference between RBTIBE and IB is that during the training process, IB used the inferred blood test features for the regularized gradient-boosted decision tree model instead of the actual blood test features used in RBTIBE. To our knowledge, the inferred blood test data derived from the same MICE method in both the training and testing processes in IB may have had better data consistency and lower estimation bias than, respectively using the actual blood test data in training process and inferred data in the testing process in RBTIBE, thereby optimizing the training model and improving the performance in IB.

Multiscene Model Analysis

Some of the blood test information provided could be useful for improving the performance of the augmented model if the patient has undergone a physical examination in the previous year. We developed seven extra-augmented models (IB_1–IB_7) for different scenarios to evaluate the applicability of our augmented method. ►Table 7 lists the scenarios and the corresponding models. The performances are shown in ►Table 8 and ►Fig. 6.

As shown in ►Table 8, the seven augmented models in the scenarios all demonstrate good predictive performance, and the performance of the augmented models could be further improved if more detailed blood test data could be obtained, i.e., AUC varied from 0.979 to 0.993. That is, the augmented method is also suitable when previous blood test data are provided and guarantee excellent performance in terms of home-based MetS auxiliary diagnosis. If a person can provide extra blood test results from physical examinations for self-care, the model will show even better predictive performance. However, the performance of the best model (IB_7) did not differ significantly from that of IB ($p < 0.014$).

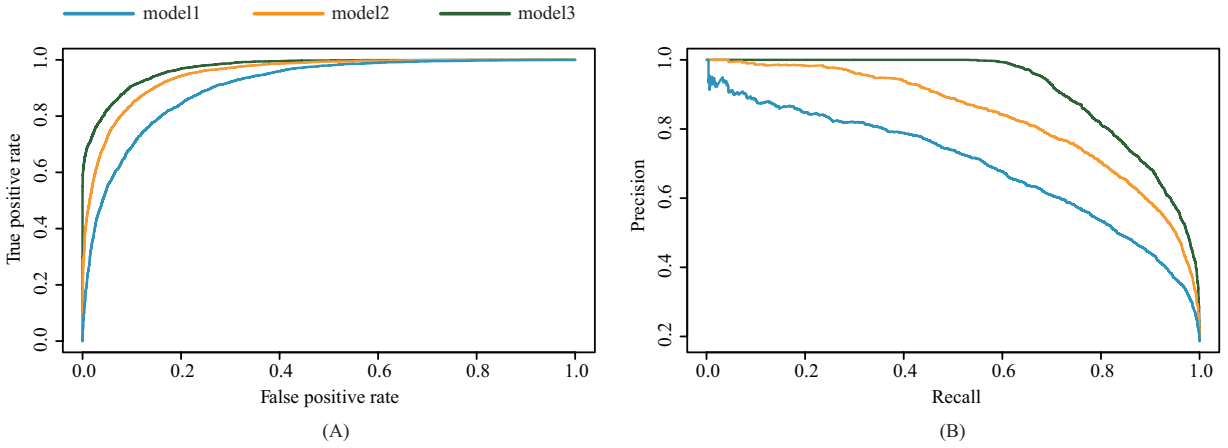

Fig. 5 (**A**) ROC curves of the three classifiers. (**B**) Precision-recall graphs of the three classifiers. ROC, receiver operating characteristic.

Table 7 Different scenarios of the augmented models

Model	Measured variables	Inferred variables in step 1
IB	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	TG, HDL-C, FBG
IB_1	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	TG, HDL-C
IB_2	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	HDL-C, FBG
IB_3	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	TG, FBG
IB_4	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	HDL-C
IB_5	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	TG
IB_6	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	FBG
IB_7	sex, age, BMI, SBP, DBP, BMI_L, SBP_L, DBP_L	none

Abbreviations: BMI, body mass index; BMI-L, body mass index of the subsequent year; DBP, diastolic blood pressure; DBP-L, diastolic blood pressure of the subsequent year; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; IB, inferred blood features; SBP, systolic blood pressure; SBP-L, systolic blood pressure of the subsequent year; TG, triglyceride.

Table 8 Performance of IB in different scenarios

Model	AUC	95% CI	Sensitivity	Specificity	Precision	NPV	F1 score	AUPRC	ACC
IB	0.971	0.970–0.971	0.856	0.935	0.751	0.966	0.800	0.917	0.920
IB_1	0.979	0.978–0.979	0.881	0.943	0.779	0.972	0.827	0.934	0.931
IB_2	0.984	0.983–0.984	0.899	0.949	0.801	0.976	0.847	0.947	0.940
IB_3	0.980	0.979–0.980	0.882	0.944	0.782	0.972	0.828	0.938	0.932
IB_4	0.987	0.987–0.987	0.910	0.953	0.815	0.979	0.860	0.958	0.945
IB_5	0.982	0.981–0.982	0.887	0.946	0.791	0.973	0.836	0.945	0.935
IB_6	0.986	0.986–0.986	0.906	0.952	0.814	0.978	0.857	0.958	0.944
IB_7	0.993	0.993–0.993	0.941	0.961	0.848	0.986	0.892	0.976	0.958

Abbreviations: AUC, area under the curve; ACC, accuracy; AUPRC, area under the precision-recall curve; IB, inferred blood features; CI, confidence interval; NPV, negative predictive value.

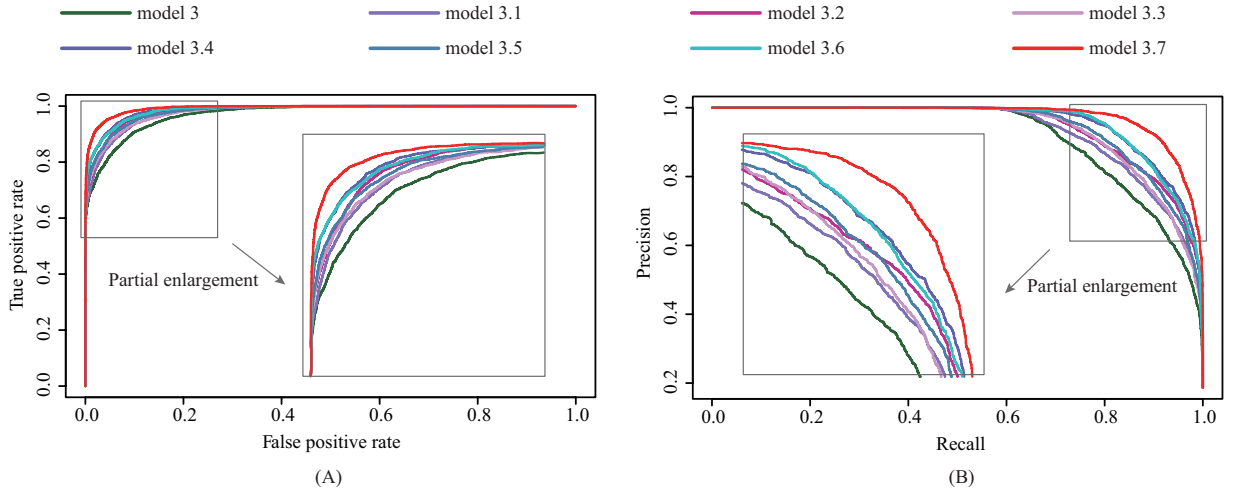

Fig. 6 (**A**) ROC curves of the classifiers. (**B**) Precision-recall graphs of the classifiers. ROC, receiver operating characteristic.

Discussion

The main purpose of our proposed model is to provide a ubiquitous self-diagnosis approach to MetS for self-care in the context of low physical examination awareness of individuals in China. Therefore, in the application scenarios, the model must support the smallest amount of input data that can be acquired at home. However, the recall of HOME (0.702) with home-based inputs did not satisfy the availability for MetS self-diagnosis and management. Thus, we took the blood test data into consideration to enrich features (RBTIBE) and utilized the MICE method to impute the blood test data instead of the raw data (IB); thus, our study provides new ideas for innovative research in health management.

Among the three models, the performance of RBTIBE was much better than that of HOME, which implies the importance of blood test data for the auxiliary diagnosis of MetS. Furthermore, we developed an augmented model (IB) that uses a large amount of physical examination data to predict the blood test items instead of using real blood test data. Concretely, the MICE method was used to learn the relationship between blood test data and home-based data within the context physical examination data, and the output was used in the second step to develop a better predictive performance model. As shown in ►Table 6, the AUC, ACC, F1 score, and AUPRC of IB were better than those of RBTIBE, which confirmed the advantage of the blood test data imputation in the training process. The superior results in IB showed that our model, which is constructed from existing health check-up data, may have the ability to provide MetS self-diagnosis and promote health management, verifying the availability of the augmented method and the feasibility of MetS self-diagnosis. In addition, the recall of IB was 0.856, which embodies the model's good ability to recognize MetS patients in the second year. The ability of the augmented model to identify the at-risk MetS population is acceptable, especially for the minority who developed MetS in the second year.

Since prevention and treatment of MetS have become a global issue,[20] several algorithmic approaches have already been applied to various aspects of MetS care, including the findings of associated risk factors,[33–38] prediction of complications,[39,40] and large-scale factors such as managing health care systems.[41,42] In particular, several studies have focused on the early prediction or diagnosis of MetS and demonstrated its clinical significance,[16,17,43] and several efforts have been made to improve the performance of models.[18,44] Several effective machine learning methods proposed by Akihiro Shimoda and Daisuke Ichikawa could immediately obtain an accurate diagnosis of MetS and determine the candidates for health guidance by using an individual's historical medical examination data.[16–18] A primary motivation for our study, however, is that despite these efforts, a home-based auxiliary diagnosis method for MetS would be more versatile and more valuable in China because of the low rate of participation in physical examinations. Moreover, for the test dataset with only home-based data (missing important blood test data), our goal was to use the augmented method with inferred blood features to obtain an effective model with good performance.

The implementation of our method could guarantee that the self-diagnosis of MetS is not limited by time or place and ensures effective self-care. Compared with the MetS models in a recent study,[16] the convenience of a MetS auxiliary diagnosis at home can increase the frequency and performance of MetS self-examination, which could ameliorate China's national health check-up conditions. A variety of studies attempted to achieve more effective self-management to improve health.[45–47] The ubiquitous auxiliary diagnostic approach could substantially improve the national health level based on the following: (1) the precise prediction plays an important role in the enhancement of people's health awareness, which helps people have a clear understanding of their health condition and engage in better self-care behaviors, such as targeted treatments and avoiding blind medication. An increase in disease awareness is helpful in reducing the risk of disease; (2) our method enhances the awareness of MetS and encourages high-risk patients to go to the hospital for further examinations; and (3) considering the population with physical examination habits, our model helps to ensure their healthy self-management in daily life.

Limitations

However, this research has some limitations.

Due to the lack of some information (such as surgical records, medications, procedures, etc.), we deleted some of the original dirty data to focus on the identification of the occurrence and development of MetS but not MetS improvement. Future studies with more angular data could integrate more effective features and information to improve the generalization ability of the model.

The level of economic growth explains the geographic variation in the prevalence of MetS. Due to differences in GDP levels in different regions, the prevalence of MetS varies from region to region. The validation results of our model are applicable only to people in developed areas such as Hangzhou, Zhejiang province, China, and can hardly be generalized to populations in other regions. In the future, we will focus on collecting more types of population data for modeling and expanding the generalizability of the model.

Our data are from the hospital database in the First Affiliated Hospital, Medical School of Zhejiang University, which indicates that the majority of the population resides in the same region. This represents the limited scope of our model. In future research, additional data from multiple hospitals in different regions that represent different economic levels and medical services should be explored to detect the relationships among different data sources and establish models for a broader range of people.

In addition, the verification results of our model have not yet been proven. Some long-term follow-up studies may be implemented to verify the validity of the risk assessment model and refine and improve the model.

Conclusion

In this study, we proposed a novel augmented method to provide useful complementary variables for home-based MetS auxiliary diagnosis. This method provides novel ideas for promoting innovative research on health management in MetS. Further validation and widespread application of the augmented models could be beneficial to the achievement of self-care among individuals at risk of MetS and other chronic diseases through early prevention and intervention, which substantially improve the physical fitness of such individuals and benefit the general population.

Note

Human and/or animal subjects were not included in the project.

Funding

This work was supported by the National Natural Science Foundation of China (No. 81771936), the National Key Research and Development Program of China (No. 2018YFC0116901), the Fundamental Research Funds for the Central Universities (No. 2020FZZX002-08) and the Major Scientific Project of Zhejiang Lab (No. 2018DG0ZX01).

Conflict of Interest

None declared.

Acknowledgment

The authors would like to thank American Journal Experts (AJE) for their English language editing assistance during the preparation of this manuscript.

References

1 Alberti KGMM, Zimmet P, Shaw J. Metabolic syndrome—a new world-wide definition. A consensus statement from the International Diabetes Federation. Diabet Med 2006;23(05):469–480
2 Zhang L, Cui HY, Liu AP. Association of hypertension, diabetes, dyslipidemia, and metabolic syndrome with overweight/obesity. Zhongguo Manxingbing Yufang Yu Kongzhi. 2009;17(06):561–563. DOI: 10.16386/j.cjpccd.issn.1004-6194.2009.06.019
3 Normann J, Mueller M, Biener M, Vafaie M, Katus HA, Giannitsis E. Effect of older age on diagnostic and prognostic performance of high-sensitivity troponin T in patients presenting to an emergency department. Am Heart J 2012;164(05):698–705.e4
4 Zang YM, Rong FAN. Annihilation of the first health-killer (cardiovascular diseases) in cooperation with multi-branches of science. Chin Heart J 2006;(05):483–488
5 Aleksandrova K, Boeing H, Jenab M, et al. Metabolic syndrome and risks of colon and rectal cancer: the European prospective investigation into cancer and nutrition study. Cancer Prev Res (Phila) 2011;4(11):1873–1883
6 Baranova A, Tran TP, Birerdinc A, Younossi ZM. Systematic review: association of polycystic ovary syndrome with metabolic syndrome and non-alcoholic fatty liver disease. Aliment Pharmacol Ther 2011;33(07):801–814
7 Chen J, Kong X, Jia X, et al. Association between metabolic syndrome and chronic kidney disease in a Chinese urban population. Clin Chim Acta 2017;470:103–108
8 Wei C, Yu Z. Analysis of correlation factors of hyperuricemia with metabolic syndrome and its related diseases in elderly male. Fudan Univ J Med Sci. 2007(3):434–437. Doi: 10.3969/j.issn.1672-8467.2007.03.026
9 Lindkvist B, Almquist M, Bjørge T, et al. Prospective cohort study of metabolic risk factors and gastric adenocarcinoma risk in the metabolic syndrome and cancer project (Me-Can). Cancer Causes Control 2013;24(01):107–116
10 Raviv NV, Sakhuja S, Schlachter M, Akinyemiju T. Metabolic syndrome and in-hospital outcomes among pancreatic cancer patients. Diabetes Metab Syndr 2017;11(Suppl 2):S643–S650
11 Ogbera AO. Prevalence and gender distribution of the metabolic syndrome. Diabetol Metab Syndr 2010;2(01):1–5
12 Xi B, He D, Hu Y, Zhou D. Prevalence of metabolic syndrome and its influencing factors among the Chinese adults: the China Health and Nutrition Survey in 2009. Prev Med 2013;57(06):867–871
13 Hsieh SD, Muto T. A simple and practical index for assessing the risk of metabolic syndrome during routine health checkups [in Japanese]. Nihon Rinsho 2004;62(06):1143–1149
14 China NBoSo. China Statistical Yearbook 2017. Beijing, China: China Statistic Press; 2017
15 Jiao Y, Hu R. Analysis of current situation of residents treated and influence factors in China. Beijing, China: China Health Insurance; 2012
16 Ichikawa D, Saito T, Oyama H. Impact of predicting health-guidance candidates using massive health check-up data: a data-driven analysis. Int J Med Inform 2017;106:32–36
17 Shimoda A, Ichikawa D, Oyama H. Prediction models to identify individuals at risk of metabolic syndrome who are unlikely to participate in a health intervention program. Int J Med Inform 2018;111:90–99
18 Shimoda A, Ichikawa D, Oyama H. Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. Comput Methods Programs Biomed 2018;163:39–46
19 Ryu SH, Park H, Mieun Y. Correlation between serum uric acid, BMI, fasting blood sugar, TG and HDL in Korean health check examinees. Cancer Res 2010;70(02):655–665
20 Alberti KGMM, Eckel RH, Grundy SM, et al; International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; International Association for the Study of Obesity. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 2009;120(16):1640–1645
21 Gurka MJ, Filipp SL, Musani SK, Sims M, DeBoer MD. Use of BMI as the marker of adiposity in a metabolic syndrome severity score: derivation and validation in predicting long-term disease outcomes. Metabolism 2018;83:68–74
22 Mohamad IB, Usman D. Standardization and its effects on K-means clustering algorithm. Res J Appl Sci Eng Technol 2013;6(17):3299–3303
23 Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: improving prediction of the minority class in boosting. Paper presented at: Knowledge Discovery in Databases: Pkdd 2003, European Conference on Principles and Practice of Knowledge Discovery in Databases; September 22–26, 2003; Cavtat-Dubrovnik, Croatia
24 El-Sayed AA, Mahmood MAM, Meguid NA, Hefny HA. Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE). Paper presented at: Complex Systems; 2016; Marrakech, Morocco
25 Torgo, L. (2010). Data Mining with R, learning with case studies Chapman and Hall/CRC. Available at: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR
26 Buuren SV, Groothuisoudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2017;45(03):1–67
27 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA
28 Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(01):40–49
29 Chen T, Tong H, Benesty M, Khotilovich V, Yuan T. xgboost: extreme gradient boosting. Paper presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; San Francisco, California, USA 2016;785–794
30 Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(08):861–874
31 Flach PA, Kull M. Precision-RECALL-GAIN CURVES: PR analysis done right. Paper presented at: International Conference on Neural Information Processing Systems; Montreal, Canada 2015
32 Melo F. Area under the ROC Curve. New York, NY: Springer; 2013

33 Taghizadeh S, Alizadeh M. The role of lipids in the pathogenesis of metabolic syndrome in adolescents. Exp Clin Endocrinol Diabetes 2018;126(01):14–22
34 Amiot MJ, Riva C, Vinet A. Effects of dietary polyphenols on metabolic syndrome features in humans: a systematic review. Obes Rev 2016;17(07):573–586
35 Srikanthan K, Feyh A, Visweshwar H, Shapiro JI, Sodhi K. Systematic review of metabolic syndrome biomarkers: a panel for early detection, management, and risk stratification in the West Virginian population. Int J Med Sci 2016;13(01):25–38
36 O'Neill S, Bohl M, Gregersen S, Hermansen K, O'Driscoll L. Blood-based biomarkers for metabolic syndrome. Trends Endocrinol Metab 2016;27(06):363–374
37 Onat A, Can G, Çoban N, et al. Lipoprotein(a) level and MIF gene variant predict incident metabolic syndrome and mortality. J Investig Med 2016;64(02):392–399
38 Ridker PM, Buring JE, Cook NR, Rifai N. C-reactive protein, the metabolic syndrome, and risk of incident cardiovascular events: an 8-year follow-up of 14,719 initially healthy American women. Circulation 2003;107(03):391–397
39 Magnussen CG, Cheriyan S, Sabin MA, et al. Continuous and dichotomous metabolic syndrome definitions in youth predict adult type 2 diabetes and carotid artery intima media thickness: the cardiovascular risk in Young Finns Study. J Pediatr 2016; 171:97–103.e1–3
40 Prokopowicz Z, Malecka-Tendera E, Matusik P. Predictive value of adiposity level, metabolic syndrome, and insulin resistance for the risk of nonalcoholic fatty liver disease diagnosis in obese children. Can J Gastroenterol Hepatol 2018;2018(01):9465784
41 Boudreau DM, Malone DC, Raebel MA, et al. Health care utilization and costs by metabolic syndrome risk factors. Metab Syndr Relat Disord 2009;7(04):305–314
42 Kan YC, Chen KH, Lin HC. Developing a ubiquitous health management system with healthy diet control for metabolic syndrome healthcare in Taiwan. Comput Meth Prog Bio 2017; 144:37–48
43 Hwang LC, Bai CH, You SL, Sun CA, Chen CJ. Description and prediction of the development of metabolic syndrome: a longitudinal analysis using a Markov model approach. PLoS One 2013;8 (06):e67436
44 Hirose H, Takayama T, Hozawa S, Hibi T, Saito I. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. Comput Biol Med 2011;41(11):1051–1056
45 Medynskiy Y, Mynatt ED. Salud!: an open infrastructure for developing and deploying health self-management applications. Paper presented at: Pervasive Computing Technologies for Healthcare; Dublin, Ireland; 2011
46 Bivins R, Marland H. Weighting for health: management, measurement and self-surveillance in the modern household. Soc Hist Med 2016;29(04):757–780
47 Borda A, Gilbert C, Gray K, Prabhu D. Consumer wearable information and health self management by older adults. Stud Health Technol Inform 2018;246:42–61