Identifying Metabolic Syndrome Easily and Cost Effectively Using Non-Invasive Methods with Machine Learning Models

Wei Xu1,*, Zikai Zhang2,*, Kerong Hu1,*, Ping Fang1, Ran Li1, Dehong Kong1, Miao Xuan1, Yang Yue3, Dunmin She4,5, Ying Xue1,6

1Department of Endocrinology and Metabolism, Tongji Hospital, School of Medicine, Tongji University, Shanghai, People’s Republic of China; 2Department of Oncology, Tongji Hospital, School of Medicine, Tongji University, Shanghai, People’s Republic of China; 3School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia; 4Clinical Medical College, Yangzhou University, Yangzhou, Jiangsu, People’s Republic of China; 5Department of Endocrinology, Northern Jiangsu People’s Hospital Affiliated to Yangzhou University, Yangzhou, Jiangsu, People’s Republic of China

*These authors contributed equally to this work

Correspondence: Ying Xue, Department of Endocrinology and Metabolism, Tongji Hospital, School of Medicine, Tongji University, No. 389, Xincun Road, Shanghai, People’s Republic of China, Tel +86-21-66111061, Email doxy6622@163.com; Dunmin She, Clinical Medical College, Yangzhou University, No. 98, Nantong West Road, Yangzhou, Jiangsu, People’s Republic of China, Email sdm1979@126.com

Purpose: The objective of this study was to employ machine learning (ML) models utilizing non-invasive factors to achieve early and low-cost identification of MetS in a large physical examination population.

Patients and Methods: The study enrolled 9171 participants who underwent physical examinations at Northern Jiangsu People’s Hospital in 2009 and 2019, to determine MetS based on criteria established by the Chinese Diabetes Society. Non-invasive characteristics such as gender, age, body mass index (BMI), systolic blood pressure (SBP), and diastolic blood pressure (DBP) were collected and used as input variables to train and evaluate ML models for MetS identification. Several ML models were used for MetS identification, including logistic regression (LR), k-nearest neighbors algorithm (k-NN), naive bayesian (NB), decision tree (DT), random forest (RF), artificial neural network (ANN), and support vector machine (SVM).

Results: Our ML models all showed good performance in the 10-fold cross-validation except for the SVM model. In the external validation, the NB model exhibited the best performance with an AUC of 0.976, accuracy of 0.923, sensitivity of 98.32%, and specificity of 91.32%.

Conclusion: This study proposed a new non-invasive method for early and low-cost identification of MetS by using ML models. This approach has the potential to serve as a highly sensitive, convenient, and cost-effective tool for large-scale MetS screening.

Keywords: metabolic syndrome, machine learning methods, non-invasive method, naive Bayesian

Introduction

Metabolic syndrome (MetS) refers to a set of combined metabolic disorders, including obesity, hypertension (HBP), hyperglycemia, and dyslipidemia such as elevated triglycerides (TG), and reduced high-density lipoprotein cholesterol (HDL-C), which is one of the vital risk factors for accelerating progression of cardiovascular disease (CVD), chronic kidney disease (CKD) and type 2 diabetes (T2DM). Several criteria defining MetS in adults have been established by expert panels from organizations such as the National Cholesterol Education Program’s Adult Treatment Panel III (NCEP: ATP III), the World Health Organization (WHO), the European Group for the Study of Insulin Resistance, the International Diabetes Federation (IDF) and the Chinese Diabetes Society. However, there is still no globally accepted definition of MetS.

MetS is a growing global health concern, with an estimated number of affected individuals exceeding one billion and continuing to rise. In a study conducted in the United States, a weighted prevalence of MetS was found to be 34.7%
among 17,048 adults, increasing to 48.6% among those aged 60 years or older.\textsuperscript{12} In another survey of 8,814 individuals in the United States, the prevalence of MetS was over 40% among those aged 60–69 years.\textsuperscript{13} The prevalence of MetS in the Chinese population is relatively lower compared to populations in developed countries. In a survey of 2,975 subjects in China, the prevalence of MetS was found to be 12.6%.\textsuperscript{14} A cross-sectional survey conducted in 2001 in 31 Chinese provinces and cities found the standardized prevalence of MetS was 13.7% (9.8% for men and 17.8% for women).\textsuperscript{15} It is crucial to find ways to early diagnose patients at high risk for MetS, allowing earlier intervention and treatment of MetS to decrease their comorbidity and improve their quality of life.

Artificial intelligence (AI) is transforming the way healthcare is being delivered. Due to recent advances in machine learning (ML), digital data collecting and natural language processing, AI is being extensively applied in fields previously considered to be the domain of human experts, such as disease diagnosis.\textsuperscript{16,17} ML is an application of AI and computer science that looks for patterns in data and can learn from prior experience to improve the accuracy of algorithms and make inferences from the examples provided. Due to the flexibility in handling enormous amounts of data and multiple variables, various ML models have been used to predict disease risk using real-world data and health outcomes. Predictive models are typically constructed by training datasets with known outcomes and then applying the trained models to new data to evaluate their performance.\textsuperscript{18} Recently, ML models such as artificial neural network (ANN), decision tree (DT), support vector machine (SVM), and random forest (RF) have been successfully applied to identify MetS.\textsuperscript{19–22}

To date, the vast majority of previous studies still collected biochemical or biophysical parameters for MS diagnosis through invasive procedures.\textsuperscript{23–25} Large-scale screening for MetS in all patients might be expensive and time-consuming. In addition, blood drawing may be a painful and inconvenient procedure for those without MetS. Therefore, it has become essential to develop a rapid, simple, non-invasive and cost-effective method to screen for MetS in high-risk populations. Consequently, this study aimed to develop a new non-invasive method with ML models for easy and low-cost identification of MetS.

**Methods**

**Study Population**

A total of 9,171 participants who received physical examinations in Northern Jiangsu People’s Hospital affiliated with Yangzhou University in 2009 and 2019 (n = 4,847 in the 2009 group and n = 4,324 in the 2019 group) were enrolled. Individuals aged 20 years or older were included in the study. In contrast, individuals under 20 years of age or underweight (BMI < 18.5 kg/m\(^2\)) were excluded from the study. Ethical approval was already provided for the study by the Ethical Committee of Northern Jiangsu People’s Hospital affiliated with Yangzhou University. And the study was carried out in accordance with recommendations of the Declaration of Helsinki.

**Data Source and Collection**

The following data were collected: age, sex, body height (BH, m), body weight (BW, kg), systolic blood pressure (SBP, mmHg), diastolic blood pressure (DBP, mmHg) and blood biochemical tests. BH and BW were measured with subjects wearing light clothes and barefoot. BP was measured three times by trained nurses according to standardized protocols. Body mass index (BMI) was calculated by using the following formula: \( \text{BMI (kg/m}^2\text{)} = \frac{\text{weight (kg)}}{\text{height square (m}^2\text{)}} \).

Blood samples were obtained from an antecubital vein after 8h of fasting. The samples were immediately transported to the clinical laboratory of Northern Jiangsu People’s Hospital affiliated with Yangzhou University (Yangzhou, Jiangsu, China) for processing. Blood biochemical parameters, including serum uric acid [UA reference range, 143–339 μmol/L (2.4–5.7 mg/dL)], fasting blood glucose [FBG, 3.9–6.1 mmol/L (2.4–5.7 mg/dL)], triglyceride [TG, < 1.7 mmol/L (<150.5 mg/dL)], total cholesterol [TC < 5.17 mmol/L (200 mg/dL)], HDL-C [1.29–1.55 mmol/L (50–60 mg/dL)] and low-density lipoprotein cholesterol [LDL-C < 3.37 mmol/L (128 mg/dL)], were measured using an automatic biochemical analyzer (Cobas 8000; Roche, Switzerland). Table 1 and Table 2 present the non-invasive factors for 20 subjects randomly selected from the training dataset and the external validation dataset, respectively.
Definition of MetS

MetS was defined according to the Chinese guideline for MetS from Chinese Diabetes Society. The subjects were diagnosed with MetS if they fulfilled three or more of the following criteria: Overweight and/or obesity (BMI ≥ 25 kg/m²); hyperglycemia [FPG ≥ 6.1 mmol/L (110 mg/dL) and/or 2h PG ≥ 7.8 mmol/L (140 mg/dl)]; HBP (SBP ≥ 140 mmHg or DBP ≥ 90 mmHg); dyslipidemia: [fasting plasma TG ≥ 1.7 mmol/L (150 mg/dL), and/or fasting HDL-C<0.9 mmol/L (35 mg/dl) (male) or <1.0 mmol/L (39 mg/dl) (female)].

Table 1 Non-Invasive Factors of 10 Randomly Selected Subjects from Training Dataset

<table>
<thead>
<tr>
<th>Samples</th>
<th>Age, (Years)</th>
<th>Gender</th>
<th>BMI (kg/m²)</th>
<th>SBP, (mmHg)</th>
<th>DBP, (mmHg)</th>
<th>MetS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>Male</td>
<td>24.57</td>
<td>138</td>
<td>86</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>Female</td>
<td>25.31</td>
<td>130</td>
<td>85</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>Male</td>
<td>25.06</td>
<td>132</td>
<td>88</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>Female</td>
<td>22.53</td>
<td>116</td>
<td>74</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>Female</td>
<td>21.45</td>
<td>136</td>
<td>90</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>Male</td>
<td>19.79</td>
<td>142</td>
<td>82</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>Female</td>
<td>18.07</td>
<td>104</td>
<td>64</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>Female</td>
<td>22.03</td>
<td>116</td>
<td>78</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>Male</td>
<td>23.67</td>
<td>160</td>
<td>110</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>62</td>
<td>Male</td>
<td>30.30</td>
<td>145</td>
<td>100</td>
<td>No</td>
</tr>
</tbody>
</table>

Abbreviations: BH, body height; BW, body weight; SBP, systolic blood pressure; DBP, diastolic blood pressure; MetS, metabolic syndrome.

Table 2 Non-Invasive Factors of 10 Randomly Selected Subjects from External Validation Dataset

<table>
<thead>
<tr>
<th>Samples</th>
<th>Age, (Years)</th>
<th>Gender</th>
<th>BMI (kg/m²)</th>
<th>SBP, (mmHg)</th>
<th>DBP, (mmHg)</th>
<th>MetS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>Male</td>
<td>25.83</td>
<td>149</td>
<td>88</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>Male</td>
<td>26.53</td>
<td>161</td>
<td>92</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>Female</td>
<td>24.34</td>
<td>135</td>
<td>81</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>Male</td>
<td>24.77</td>
<td>140</td>
<td>84</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>Male</td>
<td>26.83</td>
<td>159</td>
<td>91</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>41</td>
<td>Female</td>
<td>22.92</td>
<td>123</td>
<td>72</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>Male</td>
<td>23.03</td>
<td>127</td>
<td>75</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>Female</td>
<td>24.95</td>
<td>136</td>
<td>82</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>Male</td>
<td>23.03</td>
<td>127</td>
<td>75</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>Female</td>
<td>20.19</td>
<td>104</td>
<td>52</td>
<td>No</td>
</tr>
</tbody>
</table>

Abbreviations: BH, body height; BW, body weight; SBP, systolic blood pressure; DBP, diastolic blood pressure; MetS, metabolic syndrome.
ML Technique
Overview
In this study, we used ML methods including logistic regression (LR), k-nearest neighbors algorithm (k-NN), Naive
Bayesian (NB), DT, RF, ANN, and SVM for the reason that these methods are common ML methods and have been
successfully used for medical disease diagnosis. In our study, we set 2009 health examination data as our training dataset
and 2019 medical examination data as our external validation dataset.

LR
LR algorithm is a widely used supervised ML technique. LR model uses independent variables to predict the occurrence
of binary outcomes following the distribution of the general linear model (GLM). LR model linearizes the relationship
between the dependent and independent variables using the log-transformation function. As an evolution of linear
regression, LR is widely used to estimate parameters in a model by maximum likelihood estimation (MLE). MLE is
a popular method for determining the parameters that maximize the likelihood of observing the data given the model and
its parameters. In this study, LR models were performed on the training dataset. The LR model assumed that the
probability of the dependent variable (y, presence of MetS) taking a value of 1 (MetS occurrence) was related to the
independent variables \(x_1, x_2, \ldots, x_n\) through the logistic function:

\[
P(y = 1|x_1, \ldots, x_n) = \frac{1}{1 + \exp(-z)}
\]

Where

\[z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n\]

was a linear combination of the independent variables. The values of \(b_0, b_1, b_2\) and \(b_n\) were the coefficients of the model, which represented the effect of each independent variable. \(x_1, x_2\) and \(x_n\) in our
model represented age, gender, BMI, SBP, and DBP.

k-NN
K-NN is a non-parametric ML method developed by Evelyn Fix and Joseph Hodges in 1951. The k-NN algorithm is
a passive learning method based on partial classification. K-NN is a simple and intuitive method that is often used as
a baseline model for comparing more complex models. When the Euclidean distance between a data and its neighbors is
d, then it has a weight of \(1/d\), so the closer the distance, the greater the influence. The input data is classified based on its
proximity to other data points in the feature space. The K-NN algorithm in our ML models was summarized in the
following steps:

1. We chose the value of K equal to 5 and selected the Euclidean distance.
2. For each new data point, calculate the distances to all other data points in the training set.
3. The nearest K data points were selected based on the Euclidean distance.
4. The output label was determined by the majority vote of the K-nearest neighbors.

NB
NB is a probabilistic ML algorithm based on Bayes’ theorem. Bayes probability comes from two parts: the prior
probability and the posterior probability. NB is well-suited to handle uncertain or incomplete data. It can also model
complex relationships between variables and can deal with both discrete and continuous variables. Therefore, applica-
tions of NB include diagnosis and risk management in healthcare. When we use our experience to predict the likelihood
of future events, we get the prior probability \(P(H)\), where H represents a hypothesis. And when the new evidence is
updated, we get the posterior probability \(P(H|D)\), where D represents the observed data. Thus, we generated the following
formula to compute the posterior probability based on the prior probability \(P(H)\) and the likelihood of the data. The formula was:

\[
P(H) = P(D|H) \times \frac{P(H)}{P(D)}
\]

The NB algorithm in our ML models involved three steps:

1. Model specification: Input variables in our model and assign the prior probabilities between them.
2. Parameter estimation: estimate the probabilities that define the relationships between them.
3. Causal inference: use the model to answer queries about the posterior probabilities as well as to make predictions.
DT
DT is a type of ML model that uses a tree structure to make predictions or to classify data. DT consists of nodes and branches. Nodes represent data features, while branches represent possible values of these features. With the parent database as the root node, the types of data in the database are discriminated by the branches of the tree structure, and the overall probability obtained using the recursive approach is the conditional probability of the tree structure. It works by recursively partitioning the data into sub-datasets. At each step, the model selects the feature that provides the maximum information gain. The formula for the information gain is: Information gain = Entropy (root) – (weighted average) × Entropy (children). Where Entropy (root) is the entropy of the root node and Entropy (children) is the entropy of the children node. The weighted average is based on the proportion of data points belonging to each children node. There are several algorithms for split criterion and missing value imputation methods. In our model, we chose the ID4.5 algorithm. The main weakness of DT is over-fitting, particularly if the tree is too deep or there is too much noise in the training dataset. To overcome this problem in ML models, it is necessary to prune and remove some nodes or branches.\(^\text{26}\)

RF
RF uses several DTs and discriminates samples by the ensemble of these DTs, which increases the robustness of the model. For a single DT model, the performance is highly sensitive to noise. By using bagging techniques in the RF model, noise decreases its impact on the average performance of many trees as far as these trees are not strongly correlated.

ANN
ANN is a nonlinear ML method based on the human brain.\(^\text{27}\) ANN consists of a multi-layer neuronal network, each layer of which has one input and one output. Each layer of neurons takes the output of the corresponding neuron of the previous layer as its input and sends information messages to the corresponding neuron of the next layer as its output. The mathematical formula for a single neuron was: \(y = f(\sum w_i x_i + b)\), where \(y\) was the output of the neuron, \(x_i\) were the features of our model, \(w_i\) were the weights associated with each feature, \(b\) was the bias term and \(f()\) was the activation function which transforms the weighted sum of the inputs into the output of the neuron. The weights and bias terms are learned during the training and the activation function can be chosen based on the problem being solved. The ANN process is performed by comparing partial cases with each other and correcting the weights of each layer.

SVM
SVM is based on mapping data vectors into a high-dimensional feature space and then dividing the two classes with maximum geometric separation on a hyperplane. By maximum geometric separation, it is necessary to separate the two classes and to ensure that the two classes can produce the maximum distance in the hyperplane. The hyperplane can be represented by the equation: \(w \times x + b = 0\), where \(w\) is a vector perpendicular to the hyperplane and \(b\) is the bias of the hyperplane. If \(w \times x + b = 0\), then it belongs to one category, and if \(w \times x + b < 0\), then it belongs to the other class. Thus, the distance can be calculated as Distance = \(\frac{|w \times x + b|}{||w||}\), where \(||w||\) is the Euclidean norm of \(w\). The maximum geometric separation can be understood as maximizing \(\frac{2}{||w||}\), subject to \(y_i (w \times x_i + b) \geq 1\) for all \(i = 1, 2, \ldots, n\) where \(x_i\) is the \(i\)-th data point and \(y_i\) is the class label (+1 or −1), and \(n\) is the number of data points. Additionally, nonlinear SVM models can be used to handle non-linearly separable datasets. By using the kernel trick, these models can map the original feature space to a higher-dimensional space, where the data may become linearly separable.

Statistical Analysis
Statistical analysis was performed using Stata 15.0SE (Texas, USA). ML analysis was conducted using Orange 3.27.1 (Ljubljana, Slovenia). Continuous variables were represented as mean ± standard deviation (normally distributed). Categorical variables were expressed as numbers (percentages). To compare basic demographic characteristics between the two groups, we used independent \(t\)-tests for continuously normally distributed variables and chi-square tests for categorical variables. The features used in our ML models were age, gender, BMI, SBP, and DBP. We used a 10-fold cross-validation approach to internally validate the performance of our ML method. Receiver operating characteristics (ROC) curves were designed to illustrate the diagnostic capabilities of the ML algorithms in our study. The area under the curve (AUC), classification accuracy (CA), positive predictive value (precision), true positive rate (recall), and sensitivity and specificity were used to assess the applicability of the ML models.
**Result**

**Clinical Characteristics of Participants in the Training and Validation Datasets**

The health examination data from the 2009 group were set as the training dataset, while the medical examination data from the 2019 group were used as the external validation dataset. The participants in the training dataset and external validation dataset are shown in **Table 3**. We collected 4847 participants from the 2009 group and 4324 participants from the 2019 group. The prevalence of MetS was 7.1% in the 2009 group and 13.8% in the 2019 group. A higher percentage of subjects with MetS than those without MetS were male in both the 2009 and 2019 groups. In the 2009 and 2019 groups, patients with MetS had significantly higher age, BMI, SBP, DBP, FBG, TG, TC, LDL-C, and SUA compared to those without MetS.

**The 10-Fold Cross-Validation Performance of the ML Models in the Training Dataset**

**Table 4** presented the results of 10-fold cross validation of ML models including ANN, LR, NB, RF, k-NN, SVM, and DT on the training dataset. According to **Table 4**, the specificity of all models except SVM model was above 90%. The sensitivity of both the NB model and the SVM model was above 50%, while the sensitivity of all other models was around 20%. Among the six ML models, ANN, LR and NB had the highest AUC (0.926 vs 0.916 vs 0.904). Thus, we chose these three models for external validation.

**The External Validation Performance of the ML Models**

The external validation results of ANN, LR, and NB are shown in **Table 5**. The ROC curves of these three ML models are displayed in **Figure 1**. Both ANN and NB models had good performance in the external validation, and their AUCs and accuracies were close to each other and relatively high. The sensitivity of ANN model was lower than that of NB.

**Table 3** General Characteristics of Subjects in the Training and Validation Datasets According to the Development of Metabolic Syndrome

<table>
<thead>
<tr>
<th>Variables</th>
<th>Training Dataset (2009 Group)</th>
<th>External Validation Dataset (2019 Group)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-MetS</td>
<td>MetS</td>
</tr>
<tr>
<td>N=4501 (92.9)</td>
<td>N=346 (7.1)</td>
<td>N=3728 (86.2)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>2867 (63.70)</td>
<td>298 (86.13)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>1634 (36.30)</td>
<td>48 (13.87)</td>
</tr>
<tr>
<td>Age, (years)</td>
<td>45.97 ± 13.74</td>
<td>54.73 ± 12.81</td>
</tr>
<tr>
<td>BMI (kg/m2)</td>
<td>23.35 ± 2.87</td>
<td>27.39 ± 2.27</td>
</tr>
<tr>
<td>SBP (mmHg)</td>
<td>120.35 ± 21.05</td>
<td>141.13 ± 15.98</td>
</tr>
<tr>
<td>DBP (mmHg)</td>
<td>79.90 ± 10.46</td>
<td>93.34 ± 9.67</td>
</tr>
<tr>
<td>FBG (mg/dL)</td>
<td>94.87 ± 18.42</td>
<td>125.76 ± 36.77</td>
</tr>
<tr>
<td>TG (mg/dL)</td>
<td>134.46 ± 106.74</td>
<td>266.23 ± 246.29</td>
</tr>
<tr>
<td>TC (mg/dL)</td>
<td>176.11 ± 33.70</td>
<td>187.95 ± 37.55</td>
</tr>
<tr>
<td>HDL-C (mg/dL)</td>
<td>48.57 ± 12.28</td>
<td>48.57 ± 9.96</td>
</tr>
<tr>
<td>LDL-C (mg/dL)</td>
<td>95.49 ± 25.51</td>
<td>99.09 ± 26.48</td>
</tr>
<tr>
<td>SUA (mg/dL)</td>
<td>5.22 ± 1.33</td>
<td>6.19 ± 1.36</td>
</tr>
</tbody>
</table>

**Notes**: Continuous variables were represented as mean ± standard deviation and categorical data were represented as a number (percentage). P-values of normal-distribution date derived from Student's t-test; and P-values of categorical data derived from chi-squared test.

**Abbreviations**: MetS, metabolic syndrome; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TG, triglycerides; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; SUA, serum uric acid.
model (71.98% vs 98.32%), and the specificity of NB model was lower than that of ANN model (91.32% vs 99.38%). LR did not perform well in the external test dataset, as the sensitivity of the model was only 3.02%. Since this is a model for screening tests, high sensitivity is more important because high sensitivity minimizes the possibility of missed diagnoses and ensures the potential patients can receive further testing and treatment early. Therefore, we believed that the NB model had the best performance in our study.

**Discussion**

The purpose of this study is to develop simple methods for early and low-cost identification of MetS using ML models based on feature information obtained by noninvasive means. In this study, non-invasive and easy-to-obtain features including gender, age, BMI, SBP, and DBP were considered as input variables for easy and low-cost identification, and

---

**Table 4** The 10-Fold Cross-Validation Performance of the Machine Learning Models in the Training Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>True Positive</th>
<th>False Positive</th>
<th>True Negative</th>
<th>False Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Neural Network</td>
<td>58</td>
<td>60</td>
<td>4441</td>
<td>288</td>
<td>16.76%</td>
<td>98.67%</td>
<td>0.926</td>
<td>0.928</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>54</td>
<td>61</td>
<td>4440</td>
<td>292</td>
<td>15.61%</td>
<td>98.64%</td>
<td>0.916</td>
<td>0.927</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>212</td>
<td>421</td>
<td>4080</td>
<td>134</td>
<td>61.27%</td>
<td>90.65%</td>
<td>0.904</td>
<td>0.885</td>
</tr>
<tr>
<td>Random Forest</td>
<td>97</td>
<td>127</td>
<td>4374</td>
<td>249</td>
<td>28.03%</td>
<td>97.18%</td>
<td>0.896</td>
<td>0.922</td>
</tr>
<tr>
<td>K-NN</td>
<td>50</td>
<td>75</td>
<td>4426</td>
<td>296</td>
<td>14.45%</td>
<td>98.33%</td>
<td>0.811</td>
<td>0.924</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>195</td>
<td>895</td>
<td>3606</td>
<td>151</td>
<td>56.36%</td>
<td>80.12%</td>
<td>0.652</td>
<td>0.652</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>103</td>
<td>158</td>
<td>4343</td>
<td>243</td>
<td>29.77%</td>
<td>96.49%</td>
<td>0.591</td>
<td>0.921</td>
</tr>
</tbody>
</table>

**Abbreviations:** AUC, the area under the curve; CA, classification accuracy; Precision, positive predictive value; Recall, true positive rate.

**Table 5** The External Validation Performance of the Machine Learning Models

<table>
<thead>
<tr>
<th>Model</th>
<th>True Positive</th>
<th>False Positive</th>
<th>True Negative</th>
<th>False Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Neural Network</td>
<td>429</td>
<td>23</td>
<td>3705</td>
<td>167</td>
<td>71.98%</td>
<td>99.38%</td>
<td>0.989</td>
<td>0.956</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>586</td>
<td>321</td>
<td>3407</td>
<td>10</td>
<td>98.32%</td>
<td>91.32%</td>
<td>0.976</td>
<td>0.923</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>18</td>
<td>0</td>
<td>3728</td>
<td>578</td>
<td>3.02%</td>
<td>100%</td>
<td>0.979</td>
<td>0.866</td>
</tr>
</tbody>
</table>

**Abbreviations:** AUC, the area under the curve; CA, classification accuracy; Precision, positive predictive value; Recall, true positive rate.

---

Figure 1 Area under the receiver-operating characteristic curves of three machine learning algorithms on the external validation set. (A) Artificial Neural Network; (B) Naive Bayesian; (C) Logistic Regression.
were used to train and test various ML models. And these parameters have been reported to be closely related to the occurrence and development of Mets.\textsuperscript{28} Our study attempted to detect MetS using various ML models, including LR, k-NN, NB, DT, RF, ANN, and SVM. Our results of 10-fold cross-validation showed that ANN, LR, and NB were the three best-performing ML models in terms of AUC, accuracy, sensitivity, and specificity among all evaluated models. Of these, NB performed the best in external validation with an AUC of 0.976, accuracy of 0.923, sensitivity of 98.32\%, and specificity of 91.32\%, indicating that the NB model had the ability to detect 98.32\% of participants with MetS and 91.32\% of participants with non-MetS. The high sensitivity means that the test can accurately detect as many true positives as possible and reduce the likelihood of false negatives. On the other hand, specificity is also an important consideration, as high specificity reduces the occurrence of false-positive results, avoiding unnecessary follow-up testing and anxiety. As a screening test model, the focus of this study is on early detecting of potential cases or disease risks for take preventing measures or further intervention steps. Therefore, high sensitivity is more important in this scenario.

NB and ANN are algorithms that can be used for nonlinear statistical modeling.\textsuperscript{29,30} In contrast, LR is a simple technique that uses linear combinations of variables, which means that the model is ineffective for modeling complex interactions with severe nonlinearity. NB and ANN have theoretical advantage over LR for effectively capturing nonlinearities between factors and outcomes. Therefore, NB and ANN may be a better choice than LR-based models when the relationships between the variables are nonlinear, complex, and multidimensional, such as in the biological process and development of disease.\textsuperscript{36} One advantage of NB over alternative ML approaches such as ANN and SVM is that we are allowed to understand the decision process of the NB model, whereas alternative ML methods are “black box” algorithms.\textsuperscript{29–31} In our study, the sensitivity and specificity of the NB model exceeded 90\%, which was superior to the ANN model with a sensitivity of 71.98\% for its screening test model, suggesting a more pronounced advantage of NB in ML models for diagnosing MetS based on non-invasive data from a physically examined population.

Previous studies have utilized ML models to diagnose MetS, but the results vary widely. Vrbaski et al used four ML methods (linear regression, ANN, DT, and RF) to predict MetS in the Serbian population, and found that RF had the best performance.\textsuperscript{32} Worachartcheewan et al reported that DT correctly classified MetS and non-MetS in the Thai population by more than 99\% and was the best ML method, surpassing ANN and SVM.\textsuperscript{21,33,34} Lin et al developed ANN and LR models to detect MetS using data from psychiatric patients treated with SGA in Taiwan, China. They found that, compared to LR model, ANN could better identify MetS with higher accuracy, sensitivity, and specificity.\textsuperscript{35} We believe that the differences in results in the previous literature mentioned above may be due to differences in population (including race, disease, etc.). Therefore, it is particularly important to develop ML models for the diagnosis of MetS based on data from the mainland Chinese population.

Although similar studies have been conducted in the past, our study still has strong advantages. First, our study is more representative of the prevalence characteristics of MetS in the Chinese population. There are significant regional and population variations in the epidemiology of MetS due to differences in lifestyle, demographic composition (age, sex, race), and definition of MetS. Majority of the previous studies were conducted in non-Chinese populations, and only two studies were focused on populations in Taiwan, China,\textsuperscript{35,36} but one of them enrolled patients with psychiatric disorders \textsuperscript{35} rather than a physically examined population as study subjects. Therefore, the above findings were not generalizable. Our study was conducted in the physical examination population from the eastern part of mainland China, better reflecting the characteristics of MetS prevalence in the Chinese population. Second, our study is simpler and more efficient by using information obtained by non-invasive means. Our study required input of only demographic and anthropometric characteristics, including sex, age, BMI, SBP, and DBP, whereas most other studies required the use of invasive procedures to collect blood samples to measure biochemical or biophysical parameters (e.g., TG, HDL-C, FBG).\textsuperscript{20,37–40} Kakudi et al employed genetically optimised Bayesian ARTMAP model to diagnose MetS using biochemical and biophysical parameters and found that the model had AUCs of 86.42\%, 87.04\%, 91.08\%, and 89.24\% for the young female, middle-aged female, young male, and middle-aged male individuals, respectively.\textsuperscript{41} None of the above models in that study performed as well as NB model in our study. Worachartcheewan et al reported that their DT model could correctly classify more than 99\% of MetS and non-MetS in the Thai population using 15 blood chemical parameters.\textsuperscript{34} In our study, approximately 98\% of MetS cases and 91\% of non-MetS were successfully predicted by the NB model, which was sufficient to screen for MetS and ensure that potential patients can receive...
further testing and treatment as early as possible. Thus, our study provided a faster, simpler, and more cost-effective method for the early diagnosis of MetS than previous studies. Third, our study considered the temporal dimension and analyzed data from two time points, 2009 as well as 2019. Using the 2009 data as the training set and the 2019 data as the external validation set, it was found that NB achieved the best results in both datasets, with no time constraint in diagnosing MetS. Compared to traditional cross-sectional studies limited to one time point (e.g., Lin et al), our study is more generalizable.

Our study also has several limitations. First, all data were obtained from clients who underwent health screening at the one center. Further analysis of population data from other regions of China is needed to better ensure the generalizability of our model. Second, our study is a cross-sectional study, which prevents us from making causal inferences and requires further follow-up studies to assess. Third, due to the limited information in our physical examination database, we were unable to analyze other important MetS parameters such as educational status, history of alcohol consumption, smoking history, medication use, and past medical history.

Conclusion
In this study, we developed seven ML methods, including LR, k-NN, NB, DT, RF, ANN, and SVM, for easy and low-cost identification of MetS using non-invasive factors (gender, age, BMI, SBP, and DBP). Our performance evaluation of these models showed that NB was the most effective model for diagnosing MetS. The greatest advantage of this ML model is its ability to predict MetS using non-invasive, easily accessible metrics. Our ML model has a number of potential practical applications, such as developing a formula based on a trained model, creating a mobile or web-based application, or integrating the model into an existing clinical decision support system. Moreover, we will further improve and validate our model by using an external longitudinal cohort study dataset.

Data Sharing Statement
The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics Statement
All procedures performed in studies were in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval for this study was approved by the Ethics Committee of Northern Jiangsu People’s Hospital Affiliated with Yangzhou University.

Acknowledgments
We are grateful to all interviewers, nurses, and physicians who participated in the study.

Author Contributions
All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis, and interpretation, or in all these areas; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding
This study was supported by the Fundamental Research Funds for the Central Universities (22120220625), National Natural Science Foundation of China (81974105), Cultivation project for the National Natural Science Foundation of Shanghai Tongji Hospital.

Disclosure
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
References


Dovepress

Diabetes, Metabolic Syndrome and Obesity 2023:16

https://doi.org/10.2147/DMSO.S413829

Diabetes, Metabolic Syndrome and Obesity 2023:16

https://doi.org/10.2147/DMSO.S413829


