applied sciences

MDPI

Article

A Cardiovascular Disease Risk Score Model Based on High Contribution Characteristics

Mengxiao Peng [1], Fan Hou [1], Zhixiang Cheng [1], Tongtong Shen [1], Kaixian Liu [1], Cai Zhao [1,*] and Wen Zheng [1,2,*]

[1] Institute of Public-Safety and Big Data, College of Data Science, Taiyuan University of Technology, University Street, Yuci District, Jinzhong 030600, China
[2] Center for Big Data Research in Health, Changzhi Medical College, East Jiefang Street, Changzhi 046000, China
* Correspondence: zhaocai@tyut.edu.cn (C.Z.); zhengwen@tyut.edu.cn (W.Z.)

Abstract: Cardiovascular disease (CVD) risk prediction shows great significance for disease diagnosis and treatment, especially early intervention for CVD, which has a direct impact on preventing and reducing adverse outcomes. In this paper, we collected clinical indicators and outcomes of 14,832 patients with cardiovascular disease in Shanxi, China, and proposed a cardiovascular disease risk prediction model, XGBH, based on key contributing characteristics to perform risk scoring of patients' clinical outcomes. The XGBH risk prediction model had high accuracy, with a significant improvement compared to the baseline risk score (AUC = 0.80 vs. AUC = 0.65). At the same time, we found that with the addition of conventional biometric variables, the accuracy of the model's CVD risk prediction would also be improved. Finally, we designed a simpler model to quantify disease risk based on only three questions answered by the patient,with only a modest reduction in accuracy (AUC = 0.79), and providing a valid risk assessment for CVD. Overall, our models may allow early-stage intervention in high-risk patients, as well as a cost-effective screening approach. Further prospective studies and studies in other populations are needed to assess the actual clinical effect of XGBH risk prediction models.

Keywords: cardiovascular disease; machine learning; risk score

Citation: Peng, M.; Hou, F.; Cheng, Z.; Shen, T.; Liu, K.; Zhao, C.; Zheng, W. A Cardiovascular Disease Risk Score Model Based on High Contribution Characteristics. *Appl. Sci.* **2023**, *13*, 893. https://doi.org/10.3390/app13020893

Academic Editors: Qi-Huang Zheng and Zhonghua Sun

Received: 30 October 2022
Revised: 30 December 2022
Accepted: 4 January 2023
Published: 9 January 2023

1. Introduction

CVD is a series of diseases involving the circulatory system, including angina pectoris, myocardial infarction, coronary heart disease, heart failure, arrhythmia and else, which is generally related to atherosclerosis [1]. According to the 2019 Global Burden of Disease Report statistics [2], the number of CVD patients has steadily increased, reaching 523 million in 2019, of which 18.6 million died, accounting for one-third of the total deaths [3]. Studies have found that cardiovascular disease is often associated with a number of factors, including age, smoking, high blood pressure, diabetes, blood lipid levels, chronic kidney disease, alcohol consumption, insufficient physical activity, unreasonable diet, family history and so on [4]. Therefore, we need to explore the relationship between risk factors and diseases, and use data analysis as the theoretical support to find its inherent laws to achieve accurate prediction of disease occurrence.

Because of the high fatality rate of cardiovascular disease, many institutions have carried out prospective studies of cardiovascular disease. Typical representa tives include the method recommended by the American Heart Association/American College of Cardiology (ACC/AHA) [5], Coronary risk assessment study in European system [6], UK disease risk study based on QResearch database [7,8], cardiovascular disease cohort study in many provinces and cities in China [9], etc. These institutions have launched risk assessment tools for cardiovascular disease, coronary heart disease, stroke, heart failure and other diseases. However, existing CVD risk assessment models have an implicit assumption that each risk factor has a linear relationship with the probability of CVD [10]. Such an

assumption may oversimplify the relationship because it includes a large number of Risk factors with nonlinear interactions [11]. Because of their restrictive modelling assumptions and limited number of predictors, the existing algorithms usually fail to predict CVD risk correctly [12], especially for certain subgroups [13].

Recent years have seen remarkable advances in the application of machine learning (ML) in healthcare and medical research, thanks to high-performance computers [14]. The machine learning model can establish a complex nonlinear relationship between risk factors and diseases by minimizing the error between the predicted results and the real results [15]. This kind of machine learning algorithm has good prediction ability and can also detect potentially risk variables, such as early screening imaging tools or medical therapy (e.g., anti-platelet therapy) [16]. In the field of cardiovascular disease prediction, Typical of this is Sabrina Mezzatesta et al. [17] who used non-linear SVC to predict CVD risk of patients in the US and Italy, P.Unnikrishnan et al. [18] used 8 indicators to establish a risk assessment model based on Support Vector Machines (SVM) to predict the sensitivity and specificity of CVD [19]. Although they have obtained good model accuracy, the number of indicators chosen can cause more burden to the patient when actually applied to the clinical setting. Existing models can make predictions but do not provide an accurate score of risk of disease. Therefore, this paper aims to establish a CVD risk scoring model with less features and high accuracy.

In this work, we propose XGBH model, which is based on a small number of features to achieve high-precision prediction. The key step of XGBH is to introduce histogram algorithm, which can effectively reduce the number of features in the sample and save memory space. To supplement previous studies, we simulated the performance of the real world on 14,832 retrospective data on cardiovascular patients and the kaggle competition cardiovascular disease dataset. At the same time, the XGBH model is compared with four previous machine learning models, and the superiority of the model is proved. We also analyze the importance of the characteristics of the data set, select features with high impact to build a CVD risk prediction model with with less characteristics and high accuracy. Finally, we introduce the scoring card model into XGBH, which can predict the probability of patients and quantify the risk of patients, so as to effectively evaluate the risk of cardiovascular disease.

2. Materials and Methods

2.1. Data Description

The dataset used in this study contains data from the real Baiqiuen Hospital in Shanxi, China, which is stored in the data centre. The data provided by the hospital includes EHR, medical imaging data and genetic data. Secure access mechanisms have been created to protect the privacy and security of our patients. We used a three-year dataset from 2017 to 2020. Our data focus on inpatient data, including 1913 inpatients with a total of 14,832 medical records, including 8179 disease samples and 6653 health samples. Inpatient data consists mainly of structured and unstructured text data. Structured data includes laboratory data and basic information about the patient, such as age, gender and lifestyle. The unstructured text data includes the patient's account of the condition, the doctor's interrogation notes and the diagnosis.

The study protocol was approved by the Shanxi Bethune Hospital (Shanxi Academy of Medical Sciences) Medical Ethics Committee (approval number: YXLL-2022-094), and the methods used in this study were conducted in accordance with the approved guidelines. Participants were informed of the objectives and methods of the study, informed consent was obtained from the participants or their guardians by written signature or thumbprint, and they could withdraw from the study at any time without giving any reason.

In the past, European populations have mainly been considered for studies using machine learning for CVD risk prediction, and a sample of 55,168 cases of the kaggle competition CVD dataset was added to the dataset in order to make the model more widely available. A total of 70,000 samples, and each sample has 12-dimensional features,

including four objective features (Age, Height, Weight, Gender), four inspection features (Ap_hi, Ap_lo, Chol, Gluc), three subjective features (Smoke, Aclo, Active), and one target feature(Cardio), which contains 35,021 health samples, accounting for 49.97% of the total, 34,979 disease samples, accounting for 50.03% of the total, and the ratio of the number of patients with disease to not suffering from disease is close to 1:1. The target class cardio is equal to 1 when the patient has cardiovascular disease and 0 if the patient is healthy. objective characteristics include height and weight characteristics, so a new characteristic of BMI was constructed. The detailed characteristic information is shown in Table 1:

Table 1. Characteristics of the dataset.

Variable Type	Feature	Value Type	Feature Meaning
Objective	Age	int (days)	count in days
	Height	int (cm)	count in centimeters
	Weight	float (kg)	count in kilograms
	BMI	float (kg/m^2)	Body mass index
	Gender	categorical code	1: women, 2: men
Examination	Ap_hi	int (mmHg)	Systolic blood pressure
	Ap_lo	int (mmHg)	Diastolic blood pressure
	Chol	categorical code	Cholesterol 1: normal, 2: above normal, 3: well above normal
	Gluc	categorical code	Glucagons 1: normal, 2: above normal, 3: well above normal
Subjective	Smoke	binary	whether patient smokes or not
	Aclo	binary	Alcohol intake
	Active	binary	Physical activity
Target variable	Cardio	binary	Presence or absence of cardiovascular disease

BMI = Height/(Weight * Weight).

2.2. XGBoost Histogram Model

By comparing four machine learning models:logistic regression [20], linear support vector machine [21], random forest [22], XGBoost [23], we validated the algorithm based on the Kaggle competition cardiovascular disease dataset and chose XGBoost, which had the highest accuracy, as the base model. As XGBoost uses a pre-ranking method to handle node splitting, although the splitting points calculated in this way are more accurate . However, the training time in use is long and the memory usage is large. In this paper, the XGBH model is proposed as a fast high-performance gradient enhancement framework, a tree-based learning algorithm [24]. XGBH introduces a histogram algorithm based on XGBoost, by which the introduction of this algorithm can reduce the number of features in the sample and save memory space.

The basic idea of the Histogram algorithm is to discretize the data by partitioning the continuous feature values into k boxes. The feature histogram is constructed using the k discrete boxes. Instead of traversing all the sample points to find the segmentation points, the algorithm looks between the boxes, speeding up the process and reducing memory. Instead of losing accuracy it will have the effect of regularisation and improve the generalisation ability of the algorithm. The specific implementation process is as Algorithm 1:

Algorithm 1 Find Best Split By Histogram

Require: Training data X,Current Model $T_{C-1}(X)$
Ensure: First order gradient G,second order gradient H
  for all Leaf p in T_{C-1}(X) **do**
    for all f in X.Features **do**
      H = new Histogram()
      for i in (0,num_of_row) **do**
        H[f.bins[i]].g +=g_i; H[f.bins[i]].n+=1
      end for
      for i in (0,len(H) **do**
        S_L+= H[i].g; n_L+=H[i].n
        S_R=S_P-S_L;n_R=n_P-n_L
        $\triangle$ loss=${S_L}^2/n_L$+ ${S_R}^2/n_R$+${S_P}^2/n_P$
      end for
      if $\triangle$ loss > loss(p_m,f_m,v_m) **then**
        (p_m,f_m,v_m)= (p,f,h[i].value)
      end if
    end for
  end for

From the algorithm: the histogram optimisation algorithm needs to pre-transform the feature values into bin values before training, make a segmentation function on the values of the features, divide the values of all samples on that feature into a certain segment (bin), and finally discrete the values of the features.

Where H[f.bins[i]].g is the sum of the gradients of the samples in each bin, H[f.bins[i]].n is the number of samples in each bin, S_L, S_R, S_P represents the gradient sum on the left side of the current bin, the gradient sum on the right side, and the total gradient sum, n_L, n_R, and n_P represent the number of samples on the left side, the right side and the total number of samples.

2.3. Feature Importance

The problem with these new methods is that they are "black boxes" where the basis of the prediction is unknown [25]. Ideally, the model should provide operational advice for prevention, with the explanation of what needs to be improved to change a poor state or the identification of early risks determining the usefulness of the model. The existing method requires the patient to test more indicators in clinical practice, which increases the burden and inconvenience to the patient due to the high number of indicators used. Interpretable analysis of machine learning models can capture the most influential features, from which simpler predictive models can be constructed. Therefore, we need to use machine learning models to evaluate the importance of features in the dataset. The basic idea of feature importance assessment is to calculate the degree of decline in model performance scores by randomly ranking a particular feature, with the more fluctuating values playing a more important role. The specific method is as follows (Algorithm 2):

Algorithm 2 Feature Importance

1: Input: trained mode $\hat{f}$, feature matrix X, target vector Y, and error function Ł$\left(Y, \hat{Y}\right)$
2: Calculate the original prediction error: $e_{orig}\left(\hat{f}\right) = L\left(Y, \hat{f}(X)\right)$
3: For each feature $j = 1, 2, 3...p$, by randomly arranging the j feature,perturbed characteristic matrix X_perm_j
4: Calculate the new error: $e_{perm}\left(\hat{f}\right) = L\left(Y, \hat{f}\left(X_{perm_j}\right)\right)$
5: Calculate the importance parameters: $FI_j = e_{perm}\left(\hat{f}\right)/e_{orig}\left(\hat{f}\right)$
6: Sort by size FI_j

$\hat{Y}$ represents the target vector predicted after training. Finally, the importance of each feature can be obtained through the ordered FI_j which is helpful for our subsequent experiments.

2.4. Scorecard Model

Although feature importance assessment of machine learning models allows a simple model to be built using a small number of high importance features. However the model can only determine if a patient has cardiovascular disease and the risk of disease cannot be accurately predicted. A means of measuring the risk probability in the form of scores is the scorecard in the risk control scenario [26], which is a prediction of the probability of default, overdue and other behaviors within a certain period of time in the future. In this paper, the scorecard is applied to medical treatment, which can predict the probability of disease and quantify the disease risk of patients, so as to carry out effective risk assessment of cardiovascular disease. The scorecard works as follows:

Suppose that the prevalence probability of the sample is p, and $1 - p$ is the normal probability of the sample. The formula for calculating the *Odds* is shown in Formula (1):

$$Odds = \frac{p}{1-p} \tag{1}$$

Equation (2) is derived from (1):

$$p = \frac{Odds}{1+Odds} \tag{2}$$

The expression of the score card is shown in Equation (3):

$$score = A - Blog(Odds) \tag{3}$$

Point to double odds (*PDO*) means the score increased when *Odds* doubled. Setting the score of a particular point with *Odds* of θ_0 to $score_0$ and the score of a point with *Odds* of $2\theta_0$ to $score_0 + PDO$. Taking the above formula into account gives (4) and (5)

$$score_0 = A + Bln(\theta_0) \tag{4}$$

$$score_0 + PDO = A + Bln(2\theta_0) \tag{5}$$

where A and B are constants. Solving the above equation will give you the values of A and B.

$$B = \frac{PDO}{\ln 2} \tag{6}$$

$$A = score_0 - Bln(\theta_0) \tag{7}$$

Suppose that when the $score_1$ is, the *Odds* value is x_1, and when the *Odds* value is $2x_1$, the score is $score_2$, which satisfies Formula (8).

$$score_2 = A - \frac{PDO}{\ln 2} \cdot \log(2x_1) = score_1 + PDO \tag{8}$$

Then the total scorecard formula is

$$score_{total} = A + Bln(Odds) = A + B(\theta^T x) = A + B(w_0 + w_1 x_1 + ... + w_n x_n) \tag{9}$$

w is the regression coefficient of the logistic regression model, x_1 is the value of each characteristic transformed into WOE, WOE essentially means that when a characteristic variable takes on a certain value, that variable acts as an effect of the independent variable on the proportion of default cases. In this way the average of the scores on the individual characteristics is obtained.

Therefore, given a particular *Odds* corresponding to the Score value and *PDO*, the calculation can be substituted into Formulas (7) and (6) to find A and B. Through the above calculation process, the scorecard calculation can be converted into a problem of finding the $log(Odds)$ of a user being ill $log(Odds)$.

2.5. Statistical Analysis

Quantitative variables that obey the normal distribution are described by mean ± standard deviation, and quantitative variables that do not obey the normal distribution are described by median ± interquartile range. The categorical variables are described as quantity and proportion. Statistical differences were determined using the t-test with Welch correction or the Mann-Whitney U test, the Wilcoxon signed-rank test, or the Kruskal-Wallis test. The analysis was performed using RStudio version 7.2 (RStudio, New York, NY, USA). The baseline characteristics of the data are detailed in Table S1.

Evaluation metrics play an important role in determining how well a trained model performs. The performance of the XGBH model in prediction of CVD and the dataset was quantitatively evaluated using the testing Area Under Curve(AUC), Recall, Precision, F1-score. In other research fields, precision is also known as positive predictive value (PPV), and recall is also known as sensitivity. These metrics would help identify where the model is unable to predict correctly. Some terms help us to calculate these metrics. They include True Positive (TP), which represents that the positives are correctly identified aspositive. True Negative (TN) means that the negatives are correctly identifed as negative False Positive (FP) denotes that the negatives are wrongly identifed as positive, and False Negative (FN) where positives are wrongly identified as negative. These formulations related to the evaluation metrics are defined in Formulas (10)–(12):

The discrimination ability of the model was evaluated by using receiver operator characteristic (ROC) curve analysis. The AUC > 0.5 indicated better predictive values, the closer the AUC to 1, the better the model performance. The area under the ROC is the AUC, which is created by plotting the true positive rate versus the false positive rate at different thresholds.

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{Precision} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

The KS (Kolmogorov-Smirnov), which measures the difference between the cumulative distribution of normal and default samples, is used in this paper to assess the ability of the model to discriminate between CVD risks. The KS value gives an indicator of the model's discriminatory ability expressed in quantitative terms. The greater the cumulative difference between the normal and sick samples, the greater the KS value, then the greater the model's ability to discriminate between CVD risks.Greater than 0.3 indicates that the model has strong discriminative ability. The formula for calculating KS values is given in Formula (13):

$$\text{KS} = max(\text{TPR} - \text{FPR}) \tag{13}$$

3. Results

3.1. Model Evaluation

In this paper, the XGBH model is compared with four machine learning models of logistic regression, linear classification support vector machine, random forest and eXtreme Gradient Boosting(XGBoost). 80% of the kaggle competition dataset and 80% of the Shanxi Bethune Hospital dataset were randomly selected as the training set and 20% as the test set to cross-validate the performance of the model, and adjust the classification model according to the parameters of the classification algorithm.

Table 2 provides the results of the quantitative evaluation of five machine learning models, logistic regression, linear categorical support vector machine, random forest, XGBoost and XGBH, in terms of AUC, Recall, Precision and F1-score. Firstly, as shown in Table 2, the XGBH prediction model is superior to the other four baseline models in terms of AUC, Recall, Precision and F1-score, regardless of BMI characteristics. The AUC and Precision reached 0.8059 and 0.7578 respectively, indicating that the XGBH model not only performed best but also had higher accuracy in predicting the risk of CVD. Then we tried to add the new feature BMI to the prediction model. As shown in Table 2, after adding the BMI feature, the Precision had a small decrease but the rest of the metrics increased, showing higher predictability than before. There is also an improvement in runtime, with an XGBoost runtime of 4.263 s and an XGBH runtime of 3.742 s.

Table 2. Characteristics of the dataset.

Dataset	Model	AUC	Recall	Precision	F1 Score
Without BMI	LinearSVC	0.651 (0.643–0.659)	0.601	0.664	0.631
	Logistic Regression	0.697 (0.689–0.704)	0.657	0.709	0.682
	Random Forest	0.713 (0.706–0.720)	0.697	0.717	0.707
	XGBoost	0.802 (0.795–0.809)	0.686	0.755	0.719
	XGBH	**0.806 (0.799–0.813)**	**0.703**	**0.758**	**0.729**
With BMI	LinearSVC	0.656 (0.648–0.664)	0.613	0.666	0.638
	Logistic Regression	0.712 (0.705–0.720)	0.675	0.726	0.699
	Random Forest	0.715 (0.707–0.722)	0.703	0.717	0.710
	XGBoost	0.803 (0.796–0.810)	0.687	0.753	0.718
	XGBH	**0.807 (0.800–0.814)**	**0.704**	**0.757**	**0.730**

3.2. Feature Importance Analysis

To estimate the contribution of each feature to the prediction, we analysed the feature importance of each prediction model using the PermutationImportance method [27]. The specific method is to randomly arrange the values of a feature column in the data set, trained the model using the disordered feature values. Feature importance is identified by looking at the extent to which the feature values affect the performance of the model. The PermutationImportance method is used to calculate the feature importance of the target variable, and then the feature importance weights are ranked and a table of feature weights is drawn as shown in Table 3. As can be seen from Table 3, the ranking of the importance of the features of the models with higher prediction accuracy is much closer. Among them, the important features of the two models with the best prediction accuracy: XGBoost and XGBH are consistent, namely systolic blood pressure (Ap_hi), cholesterol (Chol), Age, diastolic blood pressure (Ap_lo), and body mass index (BMI). And systolic blood pressure (Ap_hi) showed the greatest feature weighting of all four models, which indicated that it was the most predictive feature.

In order to reduce the burden of patients, the previous feature importance analysis drives us to establish a simpler prediction model based on a small number of high importance features instead of the model based on all features. According to the order of feature weights of the XGBH model in Table 3, taking CVD as the target feature. The XGBH model proposed in this paper was used to make predictions based on different numbers of features and to plot ROC curves, as shown in Figure 1. As can be seen from the figure, the AUC of the model is 0.6353 [0.6262, 0.6444] when predictions are made using the most influential feature Ap_hi. Then using the top three features in terms of influence (ap_hi, chol, age), the AUC of the model reached 0.7999 [0.7926, 0.8072], and finally using the top five features in terms of influence (ap_hi, chol, age, ap_lo, BMI) for prediction, the AUC of the model reached 0.8030 [0.7957, 0.8102], with the AUC increasing from 0.6353 to 0.803 as the number of features increased.When increasing from three to five features, the AUC rise only increased from 0.7999 to 0.803, indicating that ap_lo, BMI provided a smaller

contribution to the increase in model accuracy. Taken together, we can perform an accurate CVD risk assessment using a questionnaire with only three questions (1. blood pressure 2. Is cholesterol normal? 3. What age?).

Table 3. Feature importance of each prediction.

Model	Logistic Regression	Random Forest	XGBoost	XGBH
1	Ap_hi (0.1383 ± 0.0070)	Ap_hi (0.1326 ± 0.0028)	Ap_hi (0.13750 ± 0.0065)	Ap_hi (0.1406 ± 0.0047)
2	Weght (0.1218 ± 0.0056)	Chol (0.0302 ± 0.0052)	Chol (0.0321 ± 0.0050)	Chol (0.0358 ± 0.0034)
3	BMI (0.0473 ± 0.0030)	Age (0.0239 ± 0.0058)	Age (0.0268 ± 0.0030)	Age (0.0276 ± 0.0043)
4	Height (0.0434 ± 0.0050)	Active (0.0024 ± 0.0016)	Ap_lo (0.0059 ± 0.0007)	Ap_lo (0.0063 ± 0.0007)
5	Age (0.0319 ± 0.0079)	Ap_lo (0.0023 ± 0.0026)	BMI (0.0045 ± 0.0026)	BMI (0.0036 ± 0.0023)
6	Chol (0.0012 ± 0.0008)	Smoke (0.0008 ± 0.0017)	Active (0.0020 ± 0.0011)	Active (0.0034 ± 0.0017)
7	Smoke (0 ± 0.0000)	Gender (0.0002 ± 0.0043)	Height (0.0018 ± 0.0015)	Gender (0.0016 ± 0.0010)
8	Alco (−0.0000 ± 0.0001)	Alco (−0.0000 ± 0.0008)	Gender (0.0014 ± 0.0009)	Gluc (0.0015 ± 0.0011)
9	Active (−0.0001 ± 0.0004)	Gluc (−0.0012 ± 0.0014)	Smoke (0.0012 ± 0.0009)	Smoke (0.0012 ± 0.0005)
10	Gluc (−0.0001 ± 0.0003)	BMI (−0.0092 ± 0.0022)	Weight (0.0010 ± 0.0024)	Weight (0.0010 ± 0.0007)
11	Gender (−0.0002 ± 0.0003)	Height (−0.0100 ± 0.0049)	Aclo (0.0003 ± 0.0003)	Height (0.0010 ± 0.0012)
12	Ap_lo (−0.0003 ± 0.0006)	Weight (−0.0125 ± 0.0023)	Gluc (0.0002 ± 0.0005)	Aclo (0.0004 ± 0.0006)

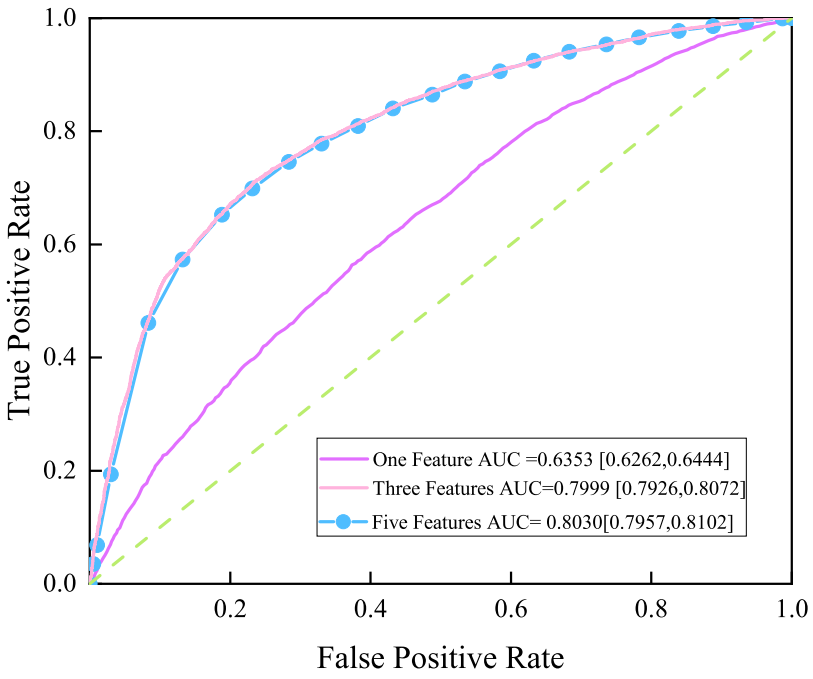

Figure 1. ROC curve.One feature(ap_hi), Three features(ap_hi, chol,age), Five features(ap_hi, chol,age, ap_lo,BMI.

3.3. Scorecard Model

The XGBH model was used to build a scorecard model for a data set containing three features. According to the principle of scorecard, the value of $score_0$ and PDO are taken as 600 and 20 respectively, then the highest score for medical scoring of the test set through the scoring card constructed by the model is 754, and the lowest score is 586. The dataset is equally divided into 20 groups according to interval prevalence and score, draw the number of sick and non sick patients in each group as shown in Figure 2. The group with high interval prevalence rate corresponds to a smaller group number, the prevalence curve is gradually reduced, and the overall number of patients is increasing.

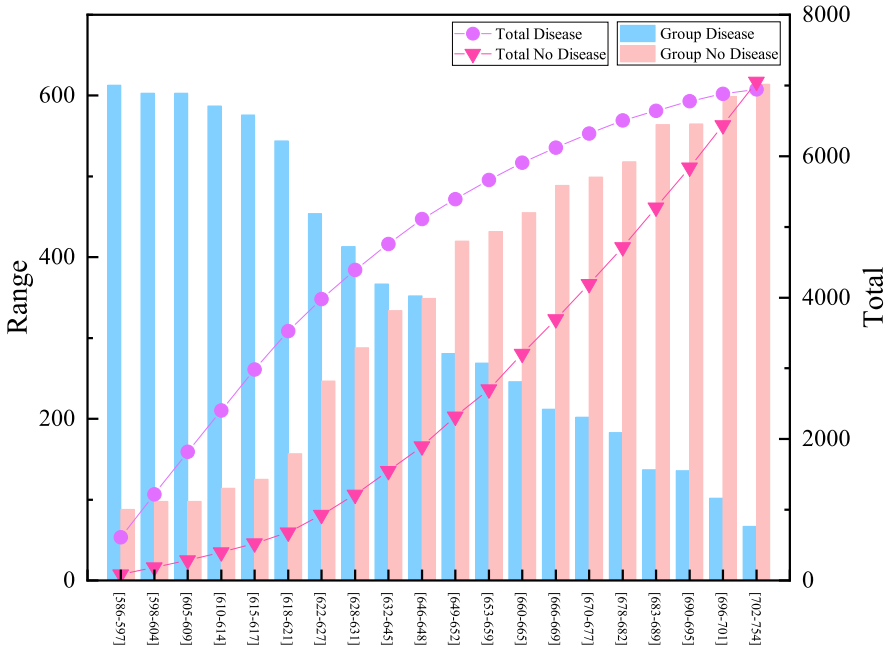

Figure 2. Scoring results and the number of patients.

KS evaluates the risk discrimination ability of the model by measuring the difference between the cumulative distributions of good and bad samples. Greater than 0.3 indicates that the model has strong discriminative ability. The performance graph of the scorecard model is plotted according to the value of KS as shown in Figure 3. It can be seen from the table that the maximum KS value in group 9 indicates that this group has the best effect on distinguishing normal samples from diseased samples. The passing rate of about 95% is used as the criterion for whether or not to be sick, and the fractions correspond to [683, 689]. If the sample is lower than 683, it is judged to be sick, and further medical tests such as electrocardiogram and cardiac color ultrasound are required to achieve effective prevention and treatment for cardiovascular disease.

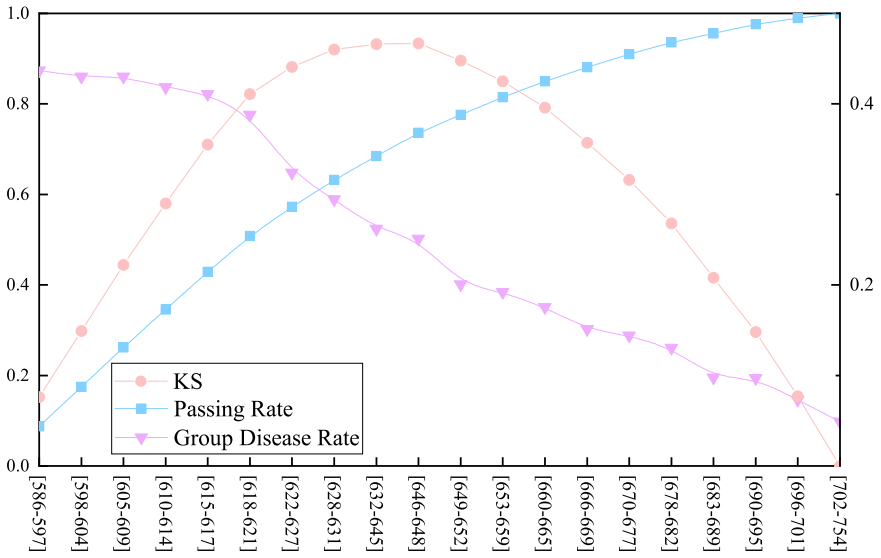

Figure 3. Performance graph of the scorecard model. The blue line represents the passing rate within the group, the orange line represents the change in the KS value, and the purple line represents the probability of disease within the group.

4. Discussion

In this study we present an XGBH model to predict the possibility of CVD occurrence by applying the model to the treatment data of 70,000 patients in Europe and Asia. Among the four baseline algorithms of logistic regression, linear classification support vector machine, random forest and XGBoost, the XGBH model has the highest accuracy, AUC and precision reached 0.8059 and 0.7578 respectively. The strength of our study is that with increasing health awareness, most people conduct health screening every 1–2 years, which contributes to the availability of patient treatment data. As the prediction model is only based on the retrospective data of patients, ML algorithm can be used for simpler and more effective CVD prediction. Compared with the traditional CVD prediction model, this method avoids the additional cost and burden of baseline data collection.

In addition, several previous studies have analysed cardiovascular patient data to assess the risk of various cardiovascular events, but they have mainly used statistical analyses [28,29] and have only been done on European patients. In contrast, in this study we used a hospital dataset containing data from the real Baiqiuen Hospital in Shanxi, China, including 1913 inpatients with a total of 14,832 medical records, allowing the model to be applied more widely. As well, the importance of missing values or non-response is not usually assessed in the development of conventional CVD risk prediction tools. This study suggests that the addition of conventional biological characteristic variables, such as BMI in particular, will also have an improved accuracy for CVD risk prediction.

Our study also has several limitations. Firstly, there are no published cholesterol thresholds, therefore we are unable to accurately assess the effect of specific cholesterol values on cardiovascular disease. Secondly, the outliers ap_ho and ap_hi that appeared in the original dataset were not processed because the model accuracy was reduced after trying to delete the outliers. Thirdly, the feasibility and acceptability of the new three-question risk assessment model proposed in this paper has not been further investigated in clinical practice. And the current study uses a range of machine learning algorithms, which suggests that the importance of different risk factors changes in interesting ways depending on the modelling technique. The models based on decision tree are very similar to each other, and the performance of gradient hoist is better than that of random forest. Neural networks and logistic regression place more emphasis on categorical variables and CVD-related medical conditions, clustering patients with similar characteristics into groups. This may facilitate further exploration of various predictive risk factors and the development of new risk prediction methods and algorithms in the future.

Our work has shown that accurate CVD prediction can be achieved based on a small number of characteristics of European and Asian patients. These results may have many implications for the subsequent treatment of patients. Our predictive models can form the basis of the initial CVD diagnostic screening process to prevent the development of CVD and its associated adverse health outcomes. Future prospective studies and research with other populations are needed to assess the clinical impact of the model.

5. Conclusions

With the recent introduction of ACC/AHA and similar guidelines internationally, CVD risk prediction has become increasingly important in clinical decision making. And the emergence of machine learning methods offers an exciting prospect for improved and more personalized cardiovascular disease risk assessment. Globally, there are still a large number of individuals at risk of CVD who have not been identified by previous tools, while some individuals who are not at risk receive unnecessary prophylactic treatment. For example, about half of all myocardial infarction (MI) and stroke will occur in people who are not expected to be at risk of cardiovascular disease. Machine learning (ML) offers an alternative to standard predictive modelling that can address current limitations. This may help drive the development of personalized medicine as it allows for better risk management for individual patients.

By introducing the histogram idea, the XGBH model in this paper reduces the number of samples and features without loss of accuracy, saves memory space, and shows better prediction performance. Our proposed model obtained the best results in all four evaluation metrics compared to the four models of logistic regression, linear classification support vector machine, random forest and XGBoost, with AUC, Precision reaching 0.8059 and 0.7578 respectively. The dataset used in this study contains data from the real Baiqiuen Hospital in Shanxi, China, including 1913 inpatients with a total of 14,832 medical records. To make the model more widely available, a portion of the kaggle competition cardiovascular disease dataset was added to the dataset, with a total of 70,000 samples. Then we analysed the five most influential features on the model (ap_hi, chol, age, ap_lo, BMI) by feature importance analysis, using different numbers of features for prediction, and the final AUC of the model reached 0.7999 [95% CI, 0.7926, 0.8072], making only three features (1. systolic blood pressure 2. Whether cholesterol is normal 3. Age) is needed to make a more accurate risk assessment of CVD. Finally we have developed a cardiovascular disease risk scale based on the XGBH model, which is able to quantify the patient's risk of developing the disease and thus provide a valid risk assessment for cardiovascular disease. In conclusion, the method proposed in this paper is superior to the existing XGBoost model in terms of accuracy and is more applicable to the prediction of CVD risk in a wider range of patients. Only three indicators of the patient are required for accurate prediction. Finally, the disease risk of patients is quantified according to the scoring model, and the risk of developing the disease is objectively evaluated for early warning and prevention.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app13020893/s1, Table S1: Baseline characteristics of patients.

Author Contributions: M.P. contributed to the methodology, software, and drafted the manuscript. F.H. and Z.C. contributed to drafted the Review and Editing, T.S. and K.L. contributed to the interpretation of the data and critically revised the manuscript. C.Z. and W.Z. responsible for overseeing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China,Grant No. 11702289, Key core technology and generic technology research and development project of Shanxi Province, No. 2020XXX013 and National Key Research and Development Project.

Institutional Review Board Statement: Ethics approval was granted by the Shanxi Bethune Hospital (Shanxi Academy of Medical Sciences) Medical Ethics Committee (approval number: YXLL-2022-094).

Informed Consent Statement: Informed consent was obtained from all individual participants included in the study.

Data Availability Statement: Some of the data analyzed during the this study are included in the Supplementary Information File and the full data are available upon reasonable request by contacting the corresponding author.

Conflicts of Interest: The authors declared no potential conflict of interest with respect to the research, authorship and publication of this article.

References

1. Yang, L.; Wu, H.; Jin, X.; Zheng, P.; Hu, S.; Xu, X.; Yan, J. Study of cardiovascular disease prediction model based on random forest in eastern China. *Hum. Nat.* **2020**, *10*, 5245. [CrossRef] [PubMed]
2. Thomas, M.R.; Lip, G.Y.H. Novel Risk Markers and Risk Assessments for Cardiovascular Disease. *Circul. Res.* **2017**, *120*, 133–149. [CrossRef] [PubMed]
3. World Health Organization. *Global Status Report on Noncommunicable Diseases*; World Health Organization: Geneva, Switzerland, 2014.
4. Yusuf, S.; Joseph, P.; Rangarajan, S.; Islam, S.; Mente, A.; Hystad, P.; Dagenais, G. Modifiable risk factors, cardiovascular disease, and mortality in 155722 individuals from 21 high-income, middle-income, and low-income countries (PURE): A prospective cohort study. *Lancet* **2020**, *395*, 795–808. [CrossRef] [PubMed]

5. D'Agostino, R.B.; Vasan, R.S.; Pencina, M.J.; Wolf, P.A.; Cobain, M.R.; Massaro, J.M.; Kannel, W.B. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **2008**, *117*, 743–753. [CrossRef] [PubMed]
6. Hoffmann, U.; Massaro, J.M.; D'Agostino, R.B.; Kathiresan, S.; Fox, C.S.; O'Donnell, C.J. Cardiovascular Event Prediction and Risk Reclassification by Coronary, Aortic, and Valvular Calcification in the Framingham Heart Study. *J. Am. Heart Assoc. Cardiovasc. Cerebrovasc. Dis.* **2016**, *5*. [CrossRef]
7. Hippisley-Cox, J.; Coupland, C.A.; Vinogradova, Y.; Robson, J.; Minhas, R.; Sheikh, A.; Brindle, P.M. Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *BMJ Br. Med. J.* **2008**, *336*, 1475 – 1482. [CrossRef]
8. Hippisley-Cox, J.; Coupland, C.A.; Brindle, P.M. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **2017**, *357*. [CrossRef]
9. Liu, J.; Zhao, D.; Wang, W. Comparison between the results from the Chinese Multi-provincial Cohort Study and those from the Framingham Heart Study. *Chin. J. Cardiol.* **2004**, *32*, 167–172.
10. Obermeyer, Z.; Emanuel, E.J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N. Engl. L. Med.* **2016**, *375 13*, 1216–1219. [CrossRef]
11. Hou, F.; Cheng, Z.; Kang, L.; Zheng, W. Prediction of Gestational Diabetes Based on LightGBM. In Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare, Taiyuan China, 23–25 October 2020. [CrossRef]
12. Siontis, G.C.M.; Tzoulaki, I.; Siontis, K.C.; Ioannidis, J.P.A. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ Br. Med. J.* **2012**, *344*. [CrossRef]
13. Alaa, A.M.; Bolton, T.; Angelantonio, E.D.; Rudd, J.H.F.; van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE* **2019**, *14*. [CrossRef]
14. Cho, S.Y.; Kim, S.H.; Kang, S.; Lee, K.J.; Choi, D.; Kang, S.; Park, S.J.; Kim, T.; Yoon, C.H.; Youn, T.J.; et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci. Rep.* **2021**, *11*. [CrossRef]
15. Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **2002**, *35 5-6*, 352–359. . [CrossRef]
16. Roman, M. Are neural networks the ultimate risk prediction models in patients at high risk of acute myocardial infarction? *Eur. J. Prev. Cardiol.* **2020**, *27*, 2045 – 2046. [CrossRef]
17. Mezzatesta, S.; Torino, C.; Meo, P.D.; Fiumara, G.; Vilasi, A. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Comput. Methods Programs Biomed.* **2019**, *177*, 9–15. [CrossRef]
18. Unnikrishnan, P.; Kumar, D.K.; Arjunan, S.P.; Kumar, H.; Mitchell, P.; Kawasaki, R. Development of Health Parameter Model for Risk Prediction of CVD Using SVM. *Comput. Math. Methods Med.* **2016**, *2016*. [CrossRef]
19. Beuret, H.; Hausler, N.; Nanchen, D.; Méan, M.; Marques-Vidal, P.; Vaucher, J. Comparison of Swiss and European risk algorithms for cardiovascular prevention in Switzerland. *Eur. J. Prev. Cardiol.* **2021**, *28 2*, 204–210. [CrossRef]
20. Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: New York, NY, USA, 2013. [CrossRef]
21. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1–27:27. [CrossRef]
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
24. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017, .
25. Kawano, K.; Otaki, Y.; Suzuki, N.; Fujimoto, S.; Iseki, K.; Moriyama, T.; Yamagata, K.; Tsuruya, K.; Narita, I.; Kondo, M.; et al. Prediction of mortality risk of health checkup participants using machine learning-based models: The J-SHC study. *Sci. Rep.* **2022**, *12*. [CrossRef]
26. Li, Z. Research and Prototype Implementation of Financial Credit Evaluation Parallel LearningModel Support Technology. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2020.
27. Official Documentation. Available online: https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html (accessed on 15 October 2021).
28. Roh, E.; Chung, H.S.; Lee, J.S.; Kim, J.A.; Lee, Y.B.; hyeon Hong, S.; Kim, N.H.; Yoo, H.J.; Seo, J.A.; Kim, S.G.; et al. Total cholesterol variability and risk of atrial fibrillation: A nationwide population-based cohort study. *PLoS ONE* **2019**, *14*, e0215687. [CrossRef] [PubMed]
29. Chung, H.S.; Lee, J.S.; Kim, J.A.; Roh, E.; Lee, Y.B.; hyeon Hong, S.; Yoo, H.J.; Baik, S.H.; Kim, N.H.; Seo, J.A.; et al. γ-Glutamyltransferase Variability and the Risk of Mortality, Myocardial Infarction, and Stroke: A Nationwide Population-Based Cohort Study. *J. Clin. Med.* **2019**, *8*, 832. [CrossRef] [PubMed]