Hans Journal of Data Mining 数据挖掘, 2022, 12(3), 220-234
Published Online July 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.123023

Hans汉斯

基于Smote-XGBoost算法的心脏病预测模型研究

管锦寒[1]，杨　健[1*]，陈俊钰[1]，李　璐[2]

[1]山西财经大学信息学院，山西　太原
[2]安徽医科大学基础医学院，安徽　合肥

收稿日期：2022年5月22日；录用日期：2022年6月23日；发布日期：2022年6月30日

摘　要

该模型首先采用合成少数类过采样技术编辑的最近邻来平衡训练数据分布，然后通过集成学习算法XGBoost预测心脏病。为了验证模型效果，本文采用心脏病患者真实医疗数据，利用专家咨询法提取特征，并通过混淆矩阵进行模型评估。与4类基线算法相比，所提模型在AUC、Accuracy、Recall和F-Score指标的评测下均表现良好。实验结果显示，所提模型能够为心脏病预测提供更精准、更智能的辅助参考，同时可以在一定程度上提高诊断的效率和心脏病预测的准确率。

关键词

心脏病预测，Smote-Enn算法，XGBoost算法，混淆矩阵

A Study of Heart Disease Prediction Model Based on Smote-XGBoost Algorithm

Jinhan Guan[1], Jian Yang[1*], Junyu Chen[1], Lu Li[2]

[1]School of Information, Shanxi University of Finance and Economics, Taiyuan Shanxi
[2]School of Basic Medical Sciences, Anhui Medical University, Hefei Anhui

Received: May 22nd, 2022; accepted: Jun. 23rd, 2022; published: Jun. 30th, 2022

Abstract

The proposed model uses nearest neighbors edited by synthetic minority class oversampling techniques to balance the training data distribution, and then predicts heart disease by ensemble

*通讯作者。

文章引用: 管锦寒, 杨健, 陈俊钰, 李璐. 基于 Smote-XGBoost 算法的心脏病预测模型研究[J]. 数据挖掘, 2022, 12(3): 220-234. DOI: 10.12677/hjdm.2022.123023

learning algorithm XGBoost. To detect the prediction reliability, a real medical dataset of heart disease patients are used, features are extracted using expert consultation method, and the model is evaluated by confusion matrix. Compared with the four types of baseline algorithms, the proposed model performs well in terms of AUC, Accuracy, Recall and F-Score metrics. The experimental results show that the proposed model can provide a more accurate and intelligent auxiliary reference for heart disease prediction, and it can also improve the efficiency of diagnosis and the accuracy of heart disease prediction to some extent.

Keywords

Heart Disease Prediction, Smote-Enn Algorithm, XGBoost Algorithm, Confusion Matrix

1. 引言

传统心血管疾病的诊断主要依靠医生经验，并结合患者的各项生理检测数据，家族遗传病史，生活习惯，以及医学影像，最终形成诊断结果。但医生的资质和实践经历，以及患者所在地区的整体医疗水平都是形成最终诊断结果的不可控因素，且在一定程度上会出现误诊漏诊现象。随着医疗信息化的不断发展，医疗信息系统中存储了海量的患者病理信息，若对存储的病患数据进行分析，且在发病之前进行必要的预测和诊断，则可以显著降低发病率和死亡率。而机器学习算法在处理复杂且高度非线性的分类和预测问题中一直具有显著优势。不同的机器学习模型已广泛用于许多疾病分类和预测问题，例如在检测心血管疾病的早期症状时，利用机器学习算法分析患者的电子病历，通过记录患者的症状、身体特征、临床实验室测试值，从而进行生物统计分析，可以发现医生无法检测到的模式和相关性。同时机器学习算法也应用在糖尿病、帕金森氏症、高血压和埃博拉病毒等疾病的预测。

集成学习[1]是很多机器学习算法的构建基础，其实现方法是集成多个弱分类器的优势，达到以多弱敌一强的效果，进而提高模型的预测有效性和鲁棒性。目前，诸如随机森林、GBDT 之类的集成学习被广泛使用。其中，XGBoost 是 GBDT 的一种高效实现。与其他集成模型相比，通过引入正则项和列采样来提高模型的鲁棒性，并且当每棵树选择分割点时，采用并行化策略来提高模型的运行速度。此外，XGBoost 只需要较少的训练时间，可以在一定程度上克服计算速度和准确性方面的局限，从而在预测模型的应用上适配性更高。当前，在库存预测和疾病预诊等几大领域，XGBoost 算法已取得十分有效的成果，并且许多应用研究成果可以达到和专业人士相媲美的程度。

本研究以某医院真实心脏病患者数据为研究对象，提出一种基于 SMOTE-ENN 和 XGBoost 算法的预测模型，根据患者病理信息，预测其是否发生 MACCE。此外，与其他四种常用模型(NB，LR，SVM 和 RF)进行对比研究，结果显示本文所提模型在预测效果方面具有显著优势。

2. 研究现状

近些年来，国内外的学者从不同的角度对心脏病预测进行了研究。总体来说，机器学习算法在心脏病诊断方面的研究，主要有两个切入点：1) 数据集的选取和处理；2) 算法的选择和优化。

从数据集选取角度，很多学者都直接采用了开源数据集，例如 Nahato 等和 Dwivedi 等[2] [3]都是以 Statlog 数据集为研究对象，进而建立不同的心脏病预测模型。Statlog 数据集包含 13 个特征，总计 270

个样本数据，其中 150 个样本被标记未患病，120 个样本标记患病，并且该数据集中不含有缺失值。Wiharto 等和 Krishnan 等[4] [5]使用 Cleveland 数据集构造模型。其中收录患病程度由低到高四种不同类型的患者数据以及健康类型数据。原始数据集包含 303 个病患信息，以及 76 个特征，但大部分研究均使用其中 14 个特征子集，其输出值包括健康和其他四类心脏病类型。此外，Sellami 等和 Wang 等[6] [7]对 MIT-BIH 数据集进行了研究，该数图 1 据集是第一个普遍用于评估心律失常检测的标准测试数据集，包括超过 10 万次心跳数据，总共有 18 种标签值，且由 1 种正常心律 + 17 种非正常心律组成。

从算法选择角度，Saxena K 等人[8]使用 SVM，CMAR，贝叶斯分类器和 C4.5 算法，提出了一个有效框架，可以根据病人的健康情况来预测他们的风险因素，该框架经评估被认为能够准确的预测与冠状动脉疾病有关的风险水平；Beyene C 等人[9]采用 SVM，KNN，NB，ANN 算法，并结合各种特征选择方法，实现了快速诊断，最大限度的降低了病人的等待时间，同时也降低了挽救病人的成本；Soni J 等人[10]使用不同的目标属性和算法相结合来进行有效的心脏病预测，并对比了决策树，朴素贝叶斯，KNN，神经网络等算法在预测中的准确性，且在研究中发现决策树和贝叶斯分类在应用遗传算法减少实际数据大小以获得最佳子数据后，其准确性进一步提高；王凤利[11]使用在 DS 证据理论基础上的优化神经网络心脏病预测模型，其算法模型相较与普通的神经网络模型在预测方面更加有效，同时算法的鲁棒性也很好；蔡勋玮[12]提出了一种集成 SVM 和 DS 证据理论的有效模型，同时将该模型与逻辑回归和普通的 SVM 算法在心脏病预测方面的表现进行比较，实验结果表明该模型预测优势显著，并对于一些疾病早期信号较为敏感；李孝虔[13]使用特征工程，结合卷积神经网络提出了一种心脏病预测模型，该方法的预测准确度能够达到 89.89%，对于心脏病的预测有一定的借鉴作用。

综上所述，大多数研究都采用了 UCI 开源的心脏病数据集，该数据集分布较为平衡，但实际的临床数据往往是不平衡的，本文采用真实的回访数据，通过 Smote-Enn 算法进行平衡处理，提高数据集质量，使得训练的模型也更加真实可靠。此外，本文使用 XGBoost 模型进行心脏病预测，其核心思想是“以弱敌强”，通过提升树来提高模型预测表现。

3. 模型构建方法

构建基于 SMOTE-ENN 和 XGBoost 算法的心脏病预测模型的实现过程包括数据预处理、模型构建、性能评估等过程，具体实现流程如图 1 所示。

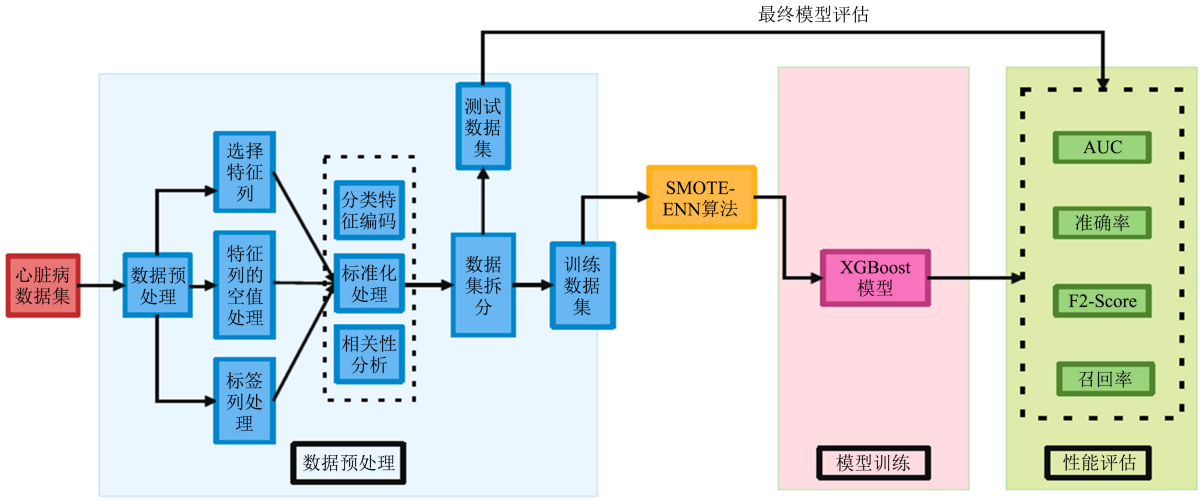

Figure 1. Flow chart of the proposed model implementation
图 1. 模型的实现流程

3.1. 数据集介绍

本文所采用的数据集来源于某医院病患真实心脏病回访数据(HDD)。该数据集包含 4356 条样本实例和 69 个特征维度。其中，MACCE 是标签列，表示主要不良心脑血管事件，0 表示不发生，1 表示发生。

3.2. 数据预处理

3.2.1. 结合专家咨询剔除不适合做特征的列

① 医生认为非必要的特征(总计 38 个，如 ID 等)

② 医生认为相关的特征，根据建议保留其中 1 项(共剔除 8 个)

3.2.2. 含空值的特征处理

① 对于类型变量的处理，采取新加一类来表示空值。

② 对于数值变量的处理，以 DMI 特征维基准，该特征是医生建议纳入的度量特征中含空值项最多的特征，共缺失 1082 项。对于其余特征，若空值数大于 1082，则直接剔除该特征(共 9 列)，若空值数小于 1082，则保留该特征，仅删除含空值的行。

由此，可以得到包括 14 个特征列和 1 个标签列(MACCE)在内总计 3528 条样本数据。数据集特征描述如表 1 所示。

Table 1. Characterization of the HDD dataset
表 1. HDD 数据集特征描述

序号	特征名	描述	类型	数据范围
1	Sex	性别	类别型	1 = 男，2 = 女
2	Age	年龄	数值型	[20, 88]
3	DM	有无糖尿病	类别型	0 = 无糖尿病，1 = 有糖尿病
4	HT	高血压史	类别类	0 = 无高血压史，1 = 有高血压史
5	CVD_history	脑血管病史	类别型	0 = 缺血性脑血管病, 1 = 出血性脑血管病
6	Afl_af	房颤扑	类别型	0 = 不患房颤扑，1 = 患房颤扑
7	LVEF	左心射血分数	数值型	[18, 88]
8	WBC	白细胞	数值型	[2.5, 32.1]
9	NE	中性	数值型	[13.9, 95.8]
10	SCV_number	病变支数	数值型	[0, 3]
11	REV_type	重建类型	类别型	1 = PCI 技术，2 = CABG 技术
12	LM_lesion	有无冠状动脉左主干病变	类别型	0 = 无，1 = 有
13	ASA	住院是否使用 ASA	类别型	0 = 不使用，1 = 使用
14	ACEI	住院是否使用 ACEI	类别型	0 = 不使用，1 = 使用
15	MACCE	是否发生主要不良心脑血管事件	因变量	0 = 不发生，1 = 发生

3.3. 标准化处理

为了提高数据的适用性，本文采用公式(1)中的最大最小规范法[14]来降低在心脏病预测过程中的数值复杂性，从而提高模型的准确性。

$$D_{\text{norm}} = \frac{D^h - D_{\min}}{D_{\max} - D_{\min}} \times \left[\text{new}_{\max} - \text{new}_{\min}\right] + \text{new}_{\min} \tag{1}$$

D_{norm} 即为进行归一化处理后的 HDD 数据集，其范围位于区间[0, 1]，D^h 为原始 HDD 数据集，$D_{\min}$ 为最小数值，$D_{\max}$ 为最大数值，$\text{new}_{\max}$ 和 $\text{new}_{\min}$ 是变换后 HDD 数据集取值范围，通常 $\text{new}_{\max} = 1$，$\text{new}_{\min} = 0$。

3.4. 相关性分析

特征之间的相关性会影响机器学习模型的性能。例如，在统计建模中，使用最小二乘法求解线性回归模型的充分必要条件是特征之间没有相关性，否则模型就会出现偏移，因此，需要先计算特征之间的皮尔逊相关系数(Pcc)，进而确定两特征值之间的相关程度，然后再建模。皮尔逊相关系数的计算方式如公式(2)所示：

$$r_{xy} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{2}$$

其中，$\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$ 表示 x 的平均值，$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} y_i$ 表示 y 的平均值，r_{xy} 表示 x 和 y 之间的系数，其范围在[−1, 1]之间变化。

若 $r_{xy} = 1$，则 x 和 y 完全正相关。

若 $r_{xy} = 0$，则 x 和 y 之间的线性关系不明显。

若 $r_{xy} = -1$，则 x 和 y 完全负相关。

值得注意的是，若两个特征根据 Pcc 证明彼此线性相关，则意味着可以忽略其中之一来优化数据集。如图 2 所示，在 HDD 数据集中计算出所有特征的 Pcc 都非常小(最大小于 0.5)，因此可以假定所选择的特征对 MACCE 的影响是相互独立的。

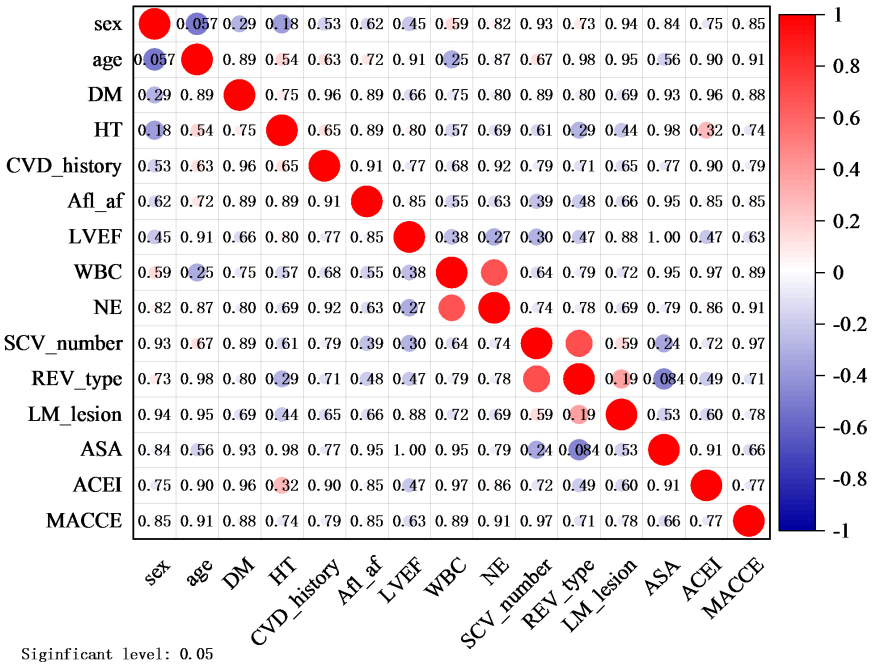

Figure 2. Correlation matrix of HDD data set
图 2. HDD 数据集相关性矩阵

3.5. 基于 Smote-Enn 算法的不平衡数据处理

本文对处理后 HDD 数据集的标签列的分布情况进行了统计。由下表 2 可知，该数据集的标签类别是极不平衡的，少数类与多数类比值达到了 1:8，因此，需要对数据集进行平衡处理，使得每类的数据量基本一致。

Table 2. Predicted column distribution of HDD dataset
表 2. HDD 数据集预测列分布

HDD 数据集	0	1	总计
数量	3133	395	3528
占比(%)	88.80%	11.20%	100%

通常采用欠采样，过采样和混合采样三种方法来解决机器学习中不平衡数据问题。图 3 列出了这三种数据平衡方法的示例。过采样方法通过主动生成占比小的一类的数据样本来平衡训练数据；欠采样则通过消除占比大的一类的数据样本来实现该目标；混合采样是在过采样处理的基础上采用欠采样来平衡数据。

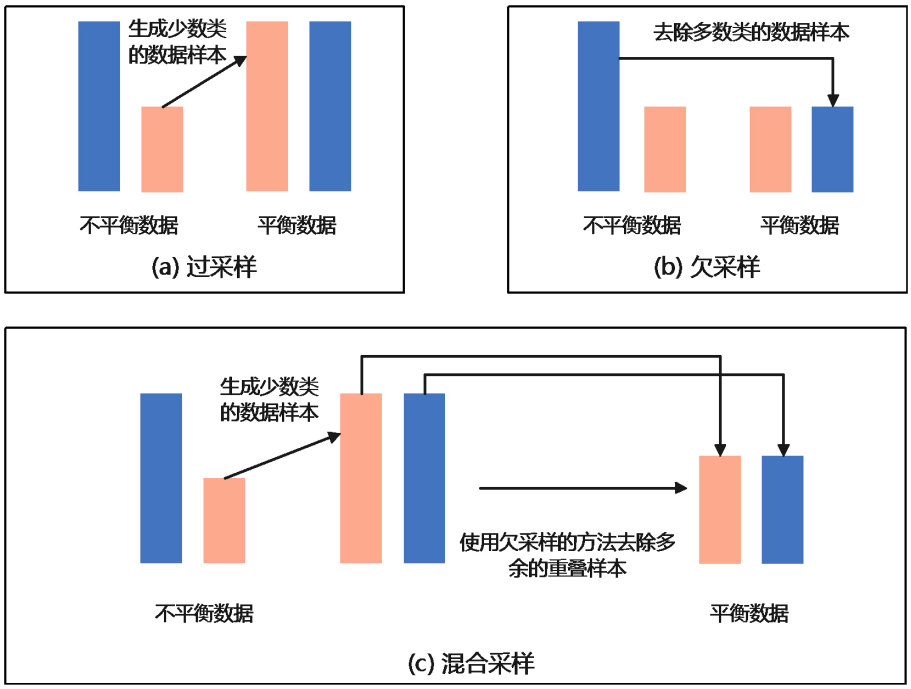

Figure 3. Three methods of balancing data
图 3. 处理不平衡数据的三种方法

本文采用 SMOTE-ENN 混合采样方法[15]来平衡 HDD 训练集。通常，SMOTE 是对占比小的一类数据进行过采样，在处理异常数据集，即处理失衡状态的数据时，其通过结合扩充少样本和削减多样本的方式，使数据集最终获得易于算法应用处理的平衡状态，当数据集整体达到平衡时，算法停止，然后使用“最近邻”(ENN)消除两个类之间的重叠样本，与此同时保持平衡分布。在实际操作中，首先应用 SMOTE

技术从少数类样本中随机生成新样本来增加少数类别的数量。然后，使用 ENN 去除多余的重叠样本。在实施 SMOTE-ENN 之后，少数类的总数增加，并且 HDD 训练集的少数类的更新百分比也变得更加均衡，达到 46.5%。表 3 为 SMOTE-ENN 的伪代码。

Table 3. Pseudocode for SMOTE-ENN
表 3. SMOTE-ENN 伪代码

算法 1 SMOTE-ENN 伪代码
Input　　数据集 D
Output 平衡数据集 BD
1: foreach 获得数据集 D 少数类 *mp* 的数据样本 do
2:　计算少数类数据样本的 K 近邻 Kmp_i
3:　构造新的合成数据样本 $mp_{new} = mp_i + \left(\widehat{mp_i} - mp_i\right) + \delta$
4:　将生成的 mp_{new} 加入到数据集 D 的 mp_i 中
5: end for
6: foreach 获得数据集 D 中的数据样本 p　do
7:　if p_i <> K 邻近的多数类 then
8:　　将 p_i 从数据集 D 中移除
9:　end if
10: end for
11: return 平衡数据集 BD

3.6. XGBoost 算法

集成学习是联合多个分类器的优势来提高模型的学习能力。目前，由 Chen T 等人[16]提出的 XGBoost 算法是集成度最高且最快的决策树算法。该算法的基础分类器利用 CART 构建，并且由多个相关的决策树共同决定，上一个决策树的训练和预测结果会成为下一个决策树的输入样本的影响因素。因此，XGBoost 是一种高度灵活且用途广泛的工具，可以解决大多数回归和分类以及用户创建的目标函数。其模型结构如图 4 所示。

XGBoost 通过执行目标函数的二阶泰勒展开式来提高计算速度和准确性。损失函数为：

$$\int_{obj}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_l^{t-1} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) + C f_t \in F \tag{3}$$

t = 1, 2, 3 表示决策树的数量，是样本量，y_i 是目标函数值，$\hat{y}_l^{t-1}$ 是第个样本的树的预测值，f 是决策树的集合，$f_t\left(x_i\right)$ 是第 t 颗树的叶子节点分数，$\Omega\left(f_t\right)$ 是用于规避训练模型复杂度的正则化函数，以避免过度拟合，C 是常数项。

损失函数的二阶泰勒展开式为：

$$f_{obj}^{(t)} = \sum_i^n \left[L\left(y_i, \hat{y}_l^{t-1} + f_t\left(x_i\right)\right) + \frac{1}{2} L^n\left(y_i, \hat{y}_l^{t-1} + f_t^2\left(x_i\right)\right) + \Omega\left(f_t\right) \right] \tag{4}$$

$$\Omega\left(f_t\right) = \frac{1}{2}\lambda \sum_j^T \left\| w_j \right\|_2 + \gamma T \tag{5}$$

其中，w_j 是第 j 棵树的权重，λ 则是 w_i 的惩罚系数，而 γ 是叶子节点的惩罚系数，用于控制树的复杂性。

XGBoost 的优势在于其创建树的可靠目标函数，同时，它提供了几个有效参数，包括树的最大深度 T，树数(迭代数) l，惩罚系数 λ，γ，以及学习率，最小权重值。本文将采用 Parzen 估计树(TPE)来优化 XGBoost 超参数空间。TPE 优化算法具有良好的性能，并且可以自动调整参数，使得效果最优。

为了避免过拟合，XGBoost 应用如下两种策略。缩减，表示算法完成一次学习所需时间的长短，在进行完一次迭代后，所有叶子节点都会乘以该系数，从而减少单颗树对最终结果产生的影响，为后面的树留出更多空间。列抽样，主要借鉴的了随机森林的做法，不仅能降低过拟合，还能减少计算。

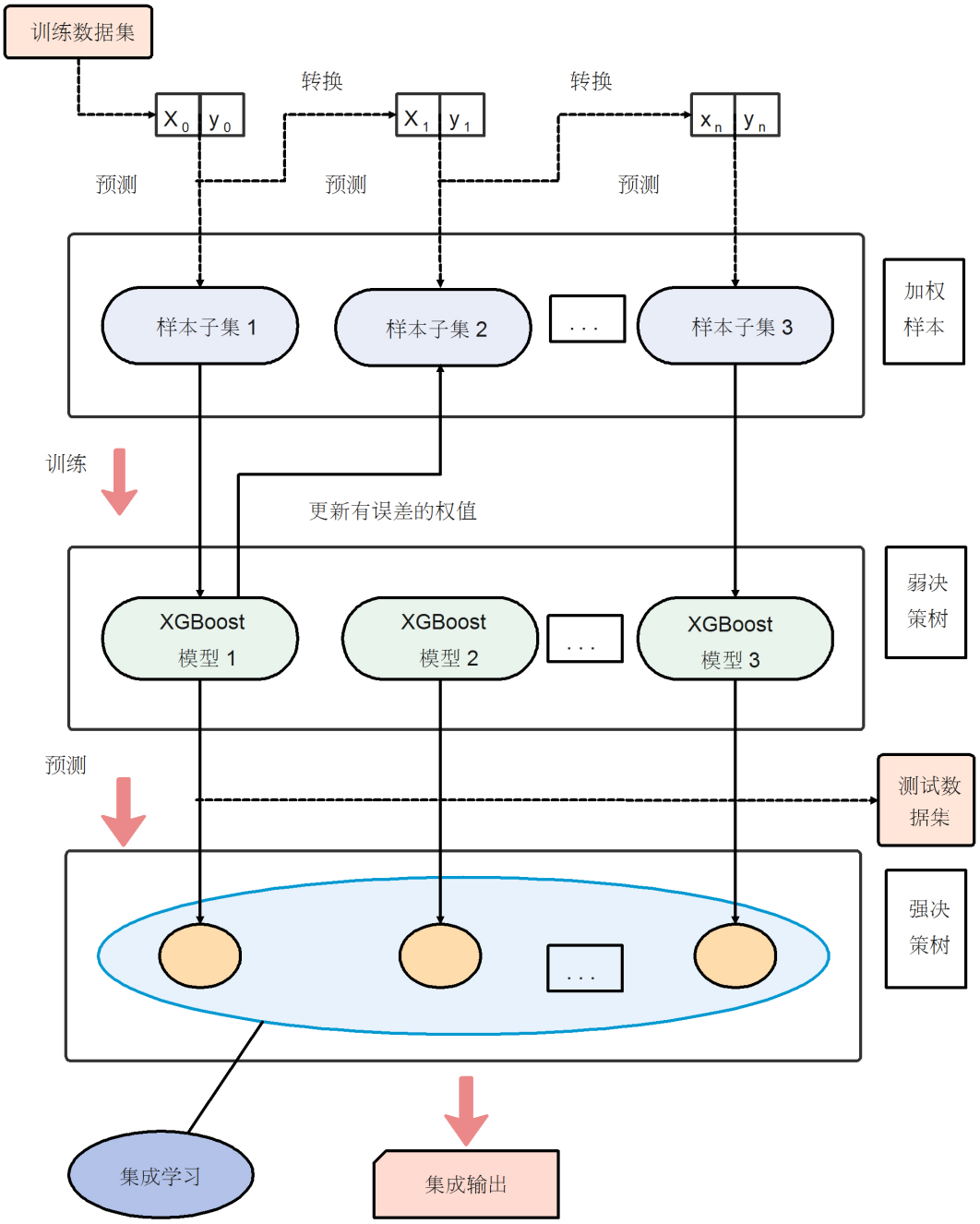

Figure 4. Structure of XGBoost algorithm
图 4. XGBoost 模型结构

4. 实验设置与评估指标

本文对 XGBoost 超参数进行了设置，具体如表 4 所示。

Table 4. Parameter setting of XGBoost
表 4. XGBoost 参数设置

参数	最大深度 T	学习率	惩罚系数 λ	惩罚系数 γ	树数(迭代数) l	最小权重值
值	2	0.15	0.12	0.16	300	0

本文引入四种基线模型与所提方法进行对比，实验中采用 5 折交叉验证来分别评测各模型在不同评估指标的预测表现，具体实现流程如下图 5 所示。

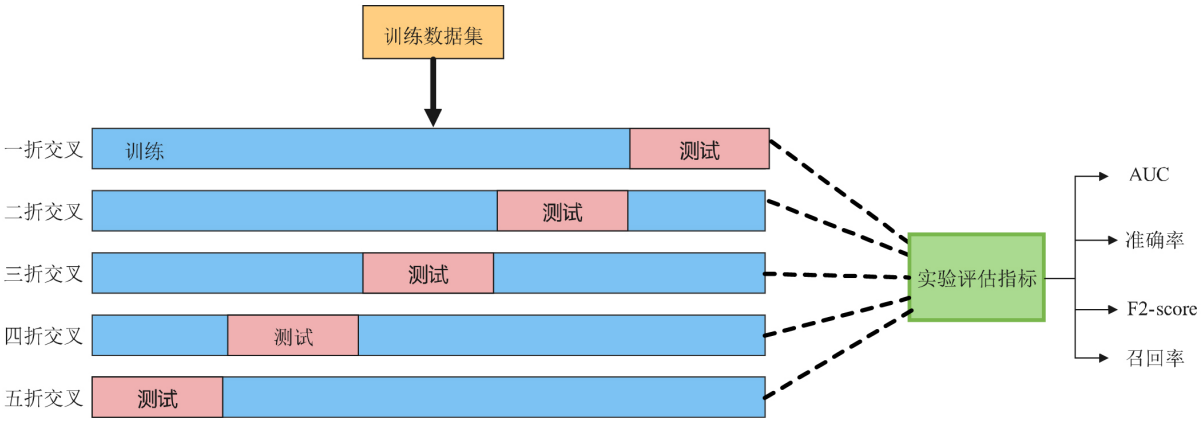

Figure 5. Experimental evaluation process
图 5. 实验评估流程

4.1. 基线算法

本文将 XGboost 模型与以下 4 种基线模型进行比较。

4.1.1. 随机森林

随机森林[17]是一种 Bagging 类型的算法，算法模型中包含了许多决策树，模型中的每棵决策树都随机的采样数据集中的小部分进行训练，每棵决策树的输出相似但并不相同，最后将每棵树的输出结果综合形成最佳结果。

对于给定数据集， $X=\{x_1,x_2,x_3,\cdots,x_n\}$，响应数据集 $Y=\{y_1,y_2,y_3,\cdots,y_n\}$，它重复了 $b=1$ 到 B 上的 Bagging。最终预测的样本集 x' 是由每棵决策树对 x 数据集的预测结果 $\sum_{b=1}^{B} fb(x')$ 的平均数来决定的。

$$j=\frac{1}{B}\sum_{b=1}^{B} fb(x') \tag{6}$$

其中，预测不确定性是通过其标准差来实现的。

$$\sigma=\sqrt{\frac{\sum_{b=1}^{B}\left(fb(x'-\hat{f}\right)^2}{B-1}} \tag{7}$$

4.1.2. 支持向量机

支持向量机[18]在保证分类准确性的同时，还尽可能的加大类别之间的区别，使得类别更容易区分。支持向量机是一种线性分类器，即应该寻找到合适的参数来描述分类的直线，即超平面，超平面可以由以下线性方程来描述：

$$f(x)=w^{T}x+b \tag{8}$$

其中，w 表示维度的系数向量，b 表示偏移量。

最终的最优化问题是：

$$\mathrm{Min}_{w,b,\xi_i}\frac{1}{2}w^{2}+C\sum_{i=1}^{n}\xi_i \tag{9}$$

$$s.t.\ \ y_i\left(w^{T}x_i+b\right)\geq 1-\xi_i,\ \ \xi_i\geq 0,\ \ \forall i\in\{1,2,\cdots,m\} \tag{10}$$

4.1.3. 朴素贝叶斯

朴素贝叶斯[19]是一种在贝叶斯定理基础上的分类算法。分类器中各个不同的特征可以独立的影响最终输出结果，且每个特征权重占比相同。每个数据实例 D 都被分配到后续概率最高的类别中。该模型是通过高斯函数训练的，其先验概率为

$$P\left(X_f\right)=\mathrm{priority}\in(0:1) \tag{11}$$

$$P\left(X_{f1},X_{f2},\cdots,X_{fn}\,|\,C\right)=\prod_{i=1}^{n}P\left(X_f\,|\,C\right) \tag{12}$$

$$P\left(X_{f1},X_{f2},\cdots,X_{fn}\,|\,c\right)=\frac{P\left(C_i\,|\,X_f\right)}{P\left(C_i\right)} \tag{13}$$

最后，根据关联的概率对测试数据进行分类

$$C_{nb}=\arg\max P\left(C_k\right)\prod_{i=1}^{n}P\left(X_{fi}|C_k\right),\ \text{for}\ k=1,2 \tag{14}$$

4.1.4. 逻辑回归

逻辑回归[20]是一种基于线性回归的分类算法，在二分类中的应用非常广泛。逻辑回归首先构建评估指标，在此基础上构建一个大致分布 $Y=a1X+B$，并使用平滑函数来减小极端值对于整体的影响，从而使整体分布更加集中。常见的函数有 sigmoid 函数，其公式如下：

$$p=\mathrm{sigmoid}\left(y\right)=\frac{1}{1+\mathrm{e}^{-y}} \tag{15}$$

通过交叉熵函数

$$L=-\sum\left[y_{\mathrm{true}}\log\left(p\right)+\left(1-y_{\mathrm{true}}\right)\log\left(1-p\right)\right] \tag{16}$$

作为评估目标，从而更新逻辑回归中的参数，使得到的分布更加准确。

4.2. 评估指标与实验结果

4.2.1. 混淆矩阵

混淆矩阵是一种可视化工具，用来分析每个模型在各个类别上的分类情况由四部分组成，即真阳性(TP)，真阴性(TN)，假阳性(FP)和假阴性(FN)。具体如图 6 所示。

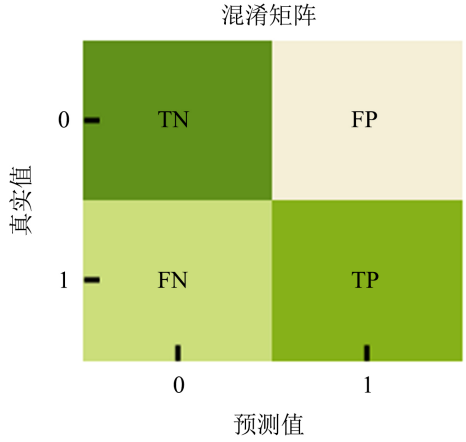

Figure 6. Confusion Matrix
图 6. 混淆矩阵

其中，TP 为正确预测患有心脏病的心脏病患者；TN 为正确预测为健康的健康人；FP 为健康人错误地预测患有心脏病；FN 为心脏病患者被错误地预测为健康。

4.2.2. 评价指标

本文选用了如下 4 种性能指标对算法的预测有效性进行评测。

① ROC 曲线下面积(AUC)

ROC 曲线下面积(AUC)：通过绘制真实正率(TRP)即敏感度或召回率与 False Positive Rate 的关系曲线创建 AUC。

② 准确率(Accuracy)

准确率是评估样本预测正确的比例，即真实值和预测值相同时所占样本总数的比率。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{17}$$

③ F-Score

由于精确率和召回率这两个指标之间相互影响，且呈现出负相关关系。F^{β} 分数就是将二者合并为一个指标，β的不同取值，代表了不同的偏重方向，在本文中，β值取 2，表示衡量中更加看重 Recall 指标。

$$F2-\text{Score} = \left(1+2^2\right)\cdot\frac{\text{Precision}\cdot\text{Recall}}{2^2\,\text{Precision}+\text{Recall}} \tag{18}$$

④ 召回率(Recall)

召回率(Recall)，是指在所有真实为正类(TP + FN)中，被判定为正类(TP)所占的比例，即有样本中的正类被正确预测的比率。召回率越高，表示实际患有心脏病的患者预测出来的概率更高。

$$\text{rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

4.2.3. 实验结果

本节对比了 RF，SVM，NB，LR 和 XGBoost 模型获得的混淆矩阵和评估结果。实验中，上述机器学习模型在测试集上做出了 705 项测试。在实际的 705 条记录中，样本中有 92 条记录的 MACCE 为“发生(1)”，613 条记录 MACCE 为“未发生(0)”。如表 5 所示，本文使用混淆矩阵来表示各类模型对 MACCE 的预测结果。

Table 5. Confusion matrix for different algorithms
表 5. 不同算法的混淆矩阵

分类器	混淆矩阵	描述
RF	预测(否) 预测(是) 真实(否) TN = 527 FP = 86 613 真实(是) FN = 20 TP = 72 92 547 158	在 705 条记录中，随机森林预测为“是”158 次，为“否”547 次 **True Negatives (TN)**：正确预测 527 个受试者为“否”，实际上也未发生 MACCE。 **True Positives (TP)**: 正确预测 72 个受试者为“是”，实际上会发生 MACCE。 **False Negatives (FN)**：20 个受试者被错误地预测为“否”，而实际上会发生 MACCE。 **False Positives (FP)**：86 个受试者被错误地预测为“是”，而实际上未发生 MACCE。
SVM	预测(否) 预测(是) 真实(否) TN = 501 FP = 112 613 真实(是) FN = 23 TP = 69 92 524 181	在 705 条记录中，支持向量机预测为“是”181 次，为“否”524 次 **True Negatives (TN)**：正确预测 501 个受试者为“否”，实际上也未发生 MACCE。 **True Positives (TP)**: 正确预测 69 个受试者为“是”，实际上会发生 MACCE。 **False Negatives (FN)**：23 个受试者被错误地预测为“否”，而实际上会发生 MACCE。 **False Positives (FP)**：112 个受试者被错误地预测为“是”，而实际上未发生 MACCE。
NB	预测(否) 预测(是) 真实(否) TN = 518 FP = 95 613 真实(是) FN = 21 TP = 71 92 539 166	在 705 条记录中，朴素贝叶斯预测为“是”166 次，为“否”539 次 **True Negatives (TN)**：正确预测 518 个受试者为“否”，实际上也未发生 MACCE。 **True Positives (TP)**: 正确预测 71 个受试者为“是”，实际上会发生 MACCE。 **False Negatives (FN)**：21 个受试者被错误地预测为“否”，而实际上会发生 MACCE。 **False Positives (FP)**：95 个受试者被错误地预测为“是”，而实际上未发生 MACCE。
LR	预测(否) 预测(是) 真实(否) TN = 511 FP = 92 613 真实(是) FN = 23 TP = 69 92 524 161	在 705 条记录中，逻辑回归预测为“是”161 次，为“否”524 次 **True Negatives (TN)**：正确预测 511 个受试者为“否”，实际上未发生 MACCE。 **True Positives (TP)**: 正确预测 69 个受试者为“是”，实际上发生 MACCE。 **False Negatives (FN)**：23 个受试者被错误地预测为“否”，而实际上发生 MACCE。 **False Positives (FP)**：92 个受试者被错误地预测为“是”，而实际上未发生 MACCE。
XGBoost	预测(否) 预测(是) 真实(否) TN = 537 FP = 76 613 真实(是) FN = 14 TP = 78 92 551 154	在 705 条记录中，XGBoost 预测为“是”154 次，为“否”551 次 **True Negatives (TN)**：正确预测 537 个受试者为“否”，实际未发生 MACCE。 **True Positives (TP)**: 正确预测 78 个受试者为“是”，实际发生 MACCE。 **False Negatives (FN)**：14 个受试者被错误地预测为“否”，而实际发生 MACCE。 **False Positives (FP)**：76 个受试者被错误地预测为“是”，而实际未发生 MACCE。

基于混淆矩阵，对各个机器学习算法在各评估指标上的评测如下。

从图 7(a)中可以看出，在 AUC 指标上的表现，XGBoost 算法最高，为 86.32%，随机森林次之为 83.28%，而支持向量机(SVM)最小，为 76.85%。

准确率是评估机器学习算法性能的最重要指标。由图 7(b)可知，XGBoost 达到了 87.23%的最高准确率。随机森林(RF)的准确度为 84.96%。逻辑回归(LR)和支持向量机(SVM)获得了相同的最低的准确度，为 82.27%。朴素贝叶斯(NB)为 83.55%。准确率关注的是分类器预测的准不准的问题，但作为疾病预测更应关注的患者是否发生 MACCE，分类器能不能检测出来。因此，本研究对各模型结果的召回率(Recall)进行了评测。由图 7(c)可以看出，XGBoost 在五个分类模型中的召回率最高，意味着真正发生 MACCE 能够评估出的概率为 84.78%。随机森林和朴素贝叶斯的召回率分别为 78.26%和 77.17%，逻辑回归和支持向量机获得了相同的召回率，为 75%。

本研究采用 F2-Score 进行评测，突出召回率在指标中的占比。如图 7(d)，XGBoost 模型的 F2-score 最高为 70.95%，随机森林次之，为 68.44%，逻辑回归和朴素贝叶斯两模型的 F2-Score 相当，分别为 65.22%和 66.48%，支持向量机最终获得了 62.84%。

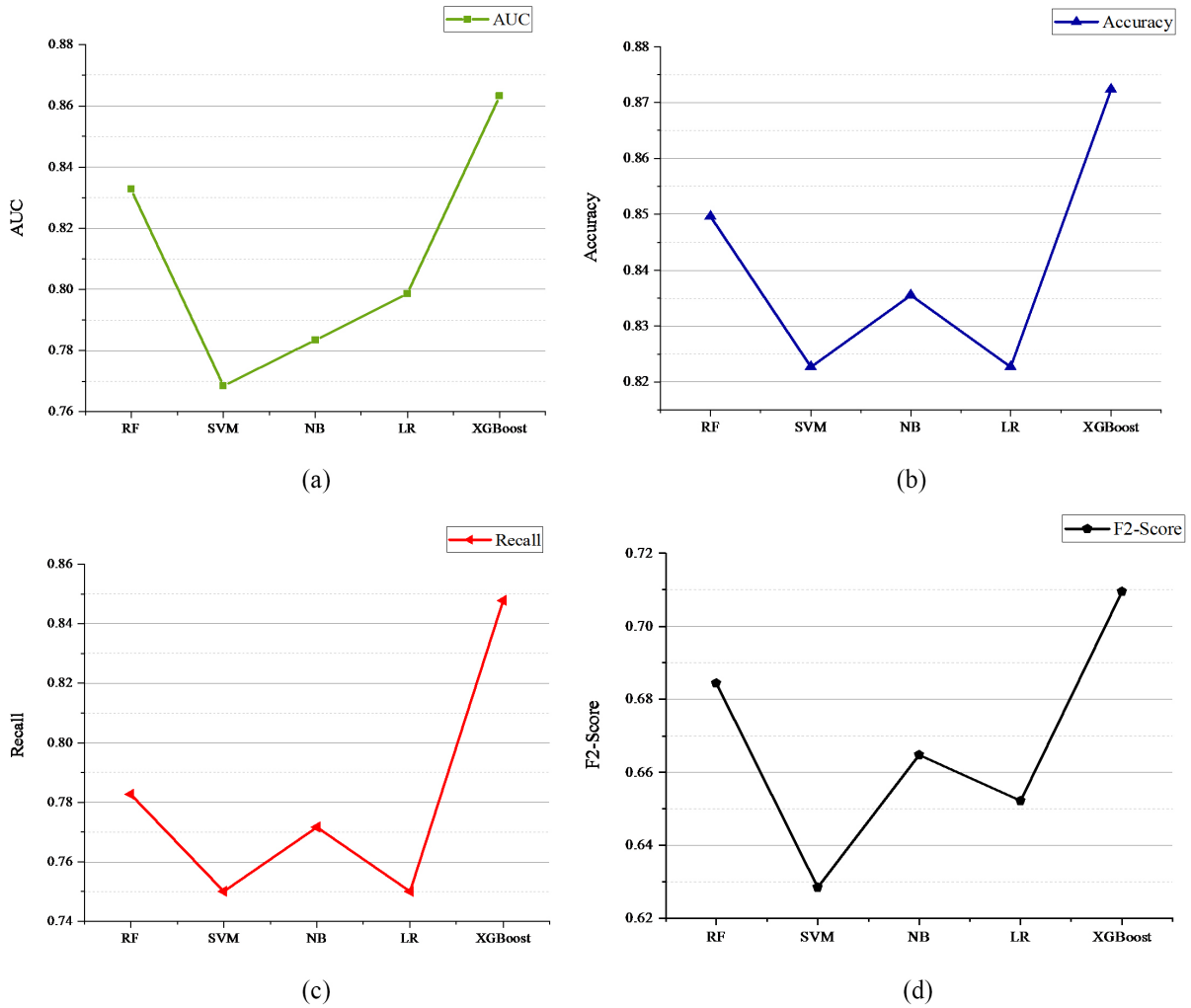

Figure 7. Evaluation results of each classification algorithm under different indicators
图 7. 各分类算法在不同指标下的评估结果

图 8 显示了有监督的机器学习算法在 AUC，准确性，F2-Score，召回率方面的总体性能评估。从实验结果中可以看出，与所有其他机器学习算法相比，XGBoost 模型在所有评估指标上均取得了更好的结果，而 SVM 表现最差。此外，在大多数评估结果中，LR 和 NB 的性能相似。

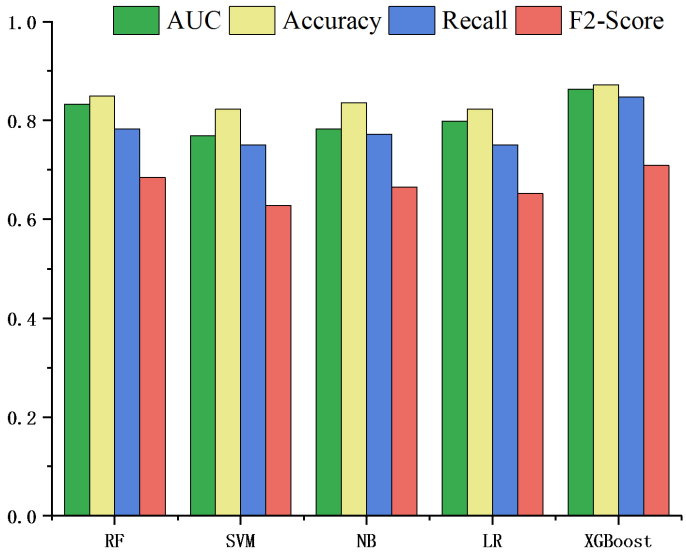

Figure 8. Algorithm performance evaluation
图 8. 算法性能评估

5. 结论与展望

在此项工作中，首先采用专家咨询进行数据预处理和特征选择，然后基于 Smote-ENN 算法处理失衡数据。最后使用随机森林、支持向量机、朴素贝叶斯、逻辑回归四种基线模型与 XGBoost 算法进行比较。实验结果表明本文所提模型在四种评测指标上的结果均表现良好，且预测准确率达到 87.23%。本文仅采用专家建议进行特征提取，在后续工作中将采用自动化的特征提取方法，以发现一些潜在的重要特征，从而提高心脏病的预测效果。

基金项目

本文系教育部人文社会科学青年基金项目(项目编号：21YJCZH197)和山西省高等学校科技创新项目(项目编号：2020L0252)的研究成果之一。

参考文献

[1] 徐继伟, 杨云. 集成学习方法: 研究综述[J]. 云南大学学报(自然科学版), 2018, 40(6): 1082-1092.

[2] Nahato, K.B., Harichandran, K.N. and Arputharaj, K. (2015) Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network. *Computational and Mathematical Methods in Medicine*, 2015, Article ID: 460189. https://doi.org/10.1155/2015/460189

[3] Dwivedi, K. (2018) Performance Evaluation of Different Machine Learning Techniques for Prediction of Heart Disease. *Neural Computing and Applications*, **29**, 685-693. https://doi.org/10.1007/s00521-016-2604-1

[4] Wiharto, W., Kusnanto, H. and Herianto, H. (2016) Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. *Healthcare Informatics Research*, **22**, 30-38. https://doi.org/10.4258/hir.2016.22.1.30

[5] Surenthiran, K., Pritheega, M. and Roslina, I. (2021) Hybrid Deep Learning Model Using Recurrent Neural Network and Gated Recurrent Unit for Heart Disease Prediction. *International Journal of Electrical & Computer Engineering*, **11**, 5467-5476. https://doi.org/10.11591/ijece.v11i6.pp5467-5476

[6] Sellami, A. and Hwang, H. (2019) A Robust Deep Convolutional Neural Network with Batch-Weighted Loss for Heartbeat Classification. *Expert Systems with Applications*, **122**, 75-84. https://doi.org/10.1016/j.eswa.2018.12.037

[7] Wang, Y., Sun, L. and Subramani, S. (2021) Cab: Classifying Arrhythmias Based on Imbalanced Sensor Data. *KSII Transactions on Internet and Information Systems*, **15**, 2304-2320. https://doi.org/10.3837/tiis.2021.07.001

[8] Purushottam, Saxena, K. and Sharma, R. (2016) Efficient Heart Disease Prediction System. *Procedia Computer Science*, **85**, 962-969. https://doi.org/10.1016/j.procs.2016.05.288

[9] Beyene, C. and Kamat, P. (2018) Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques. *International Journal of Pure and Applied Mathematics*, **118**, 165-174. https://doi.org/10.5120/2237-2860

[10] Soni, J., Ansari, U., Sharma, D. and Soni, S. (2011) Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, **17**, 43-48.

[11] 王凤利. 基于 BP 神经网络和 DS 证据理论的疾病预测模型研究[D]: [硕士学位论文]. 太原: 太原理工大学, 2016.

[12] 蔡勋玮. SVM 结合 DS 证据理论的心血管病预测方法研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2018.

[13] 李孝虔. 基于卷积神经网络的心脏病预测方法研究[D]: [硕士学位论文]. 哈尔滨: 东北林业大学, 2019.

[14] Jain, Y.K. and Bhandare, S.K. (2011) Min Max Normalization Based Data Perturbation Method for Privacy Protection. *International Journal of Computer & Communication Technology*, **2**, 45-50.

[15] 尚旭. 不平衡数据集的混合采样方法[J]. 数字技术与应用, 2016(12): 68-71.

[16] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[17] Speiser, J.L., Miller, M.E., Tooze, J. and Ip, E. (2019) A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, **134**, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

[18] Tharwat, A. (2019) Parameter Investigation of Support Vector Machine Classifier with Kernel Functions. *Knowledge and Information Systems*, **61**, 1269-1302. https://doi.org/10.1007/s10115-019-01335-4

[19] Chen, S., Webb, G.I., Liu, L. and Ma, X. (2020) A Novel Selective Naïve Bayes Algorithm. *Knowledge-Based Systems*, **192**, Article ID: 105361. https://doi.org/10.1016/j.knosys.2019.105361

[20] Wang, H.Y., Zhu, R. and Ma, P. (2018) Optimal Subsampling for Large Sample Logistic Regression. *Journal of the American Statistical Association*, **113**, 829-844. https://doi.org/10.1080/01621459.2017.1292914