Statistics and Application 统计学与应用, 2023, 12(4), 1111-1120
Published Online August 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.124114

Hans汉斯

基于心脏病数据的两种机器学习预测模型比较研究

代　倩，黄希芬

云南师范大学数学学院，云南　昆明

收稿日期：2023年7月26日；录用日期：2023年8月23日；发布日期：2023年8月30日

摘　要

现如今中国正面临的两大主要压力是人口老龄化进程加快和代谢危险因素流行，心血管疾病的发病率和患病率一直保持上升状态，并成为了我国居民死亡的首要原因。与此同时，医学与统计相结合，建立出具有一定预测效果的模型，可以帮助更有效地治疗和控制病情，使心脏病风险预测模型成为公共卫生安全的重要工具。本文首先对心脏病数据集进行预处理，再通过混合采样的方法获得平衡数据，依次构建随机森林和全连接神经网络模型，对它们分别进行比较研究，阐述了随机森林算法在预测心脏病患病情况时有显著优势。建立恰当的模型之后可以有效地对患者进行方便快捷的心脏病预测，有效提高临床诊断的准确率，帮助心脏病患者尽早进行医疗干预获得健康。

关键词

心脏病预测模型，不平衡数据，随机森林

Comparative Study of Two Machine Learning Prediction Models Based on Heart Disease Data

Qian Dai, Xifen Huang

School of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Jul. 26th, 2023; accepted: Aug. 23rd, 2023; published: Aug. 30th, 2023

Abstract

The two major pressures facing China today are the accelerated aging of the population and the

文章引用: 代倩, 黄希芬. 基于心脏病数据的两种机器学习预测模型比较研究[J]. 统计学与应用, 2023, 12(4): 1111-1120. DOI: 10.12677/sa.2023.124114

prevalence of metabolic risk factors. The incidence and prevalence of cardiovascular diseases have been on the rise, and have become the leading cause of death in China. At the same time, the combination of medicine and statistics can establish a model with certain predictive effect, which can help to treat and control the disease more effectively, making the heart disease risk prediction model an important tool for public health security. In this paper, the heart disease data set is pre-processed first, and then balanced data is obtained by mixed sampling method. The random forest and fully connected neural network models are constructed successively, and the comparison between them is carried out, and the significant advantages of the random forest algorithm in predicting the incidence of heart disease are expounded. After establishing a proper model, patients can effectively predict heart disease conveniently and quickly, effectively improve the accuracy of clinical diagnosis, and help patients with heart disease to get healthy as soon as possible through medical intervention.

Keywords

Heart Disease Prediction Model, Imbalanced Data, Random Forest

1. 引言

1.1. 研究背景及意义

1.1.1. 研究背景

随着工业化、城镇化、人口老龄化进程加快以及生态环境、生活行为方式等发生改变，慢性非传染性疾病已成为居民的主要死亡原因和疾病负担[1]。心脑血管疾病、癌症、慢性呼吸系统疾病、糖尿病等慢性病导致的负担占总疾病负担的 70%以上，成为阻碍提高健康预期寿命的重要原因之一[2]。

心脏病是一种常见的循环系统疾病，它涉及心脏、血管和调节血液循环的神经体液组织。会严重影响患者的劳动能力，是内科疾病中的常见病。心脏病和脑血管病的特点是患病率高、致残率高、复发率高和死亡率高，给社会和经济发展带来了沉重的负担[3]。据统计，我国现有的高血压患者达 2.7 亿、脑卒中患者 1300 万、冠心病患者 1100 万。高血压、血脂异常、糖尿病等是心脑血管疾病的主要危险因素，而肥胖、吸烟、缺乏运动、不健康的饮食习惯等也会增加患病的风险[4]。中国 18 岁以上居民中，约四分之一的人患有高血压，接近一半的人血脂异常，并且这些比例还在上升。对这些危险因素进行干预，不仅可以预防或延缓心脑血管疾病的发生，还可以与药物治疗配合，预防其复发。

1.1.2. 心脏病研究意义

心脏病是一种致死率高的危险疾病，需要尽早发现。但很多心脏病患者在早期没有胸闷、气短、乏力、心绞痛等典型症状，这就是“沉默的心脏病”。这导致很多病人无法及时发现自己有心脏病，错过了最佳治疗时机。因此，通过了解身体各项指标和日常生活习惯，来判断是否有心脏病对降低心脏病的死亡率非常重要。

机器学习是一门新兴的学科，它通过周围的环境学习来模拟人类的智能。随机森林是一种基于决策树的集成学习算法，它可以减少过拟合的风险，提高模型的泛化能力。因为每个决策树只使用了部分样本和特征，所以每个决策树都是独立的，并且随机森林可以通过多个决策树的投票或平均值来降低单个

决策树的错误率。全连接神经网络有着比浅层体系结构更强的数据处理能力，可以利用出色的服务器性能对海量大数据进行特征挖掘与处理。本文主要利用随机森林和全连接神经网络两种模型对心脏病数据进行训练学习，经过对比得出最优分类算法模型，有利于构建适用于心血管疾病的预测模型，从而辅助医师对心脏病患者提供诊断评估依据。

总而言之，利用随机森林模型对是否患有心脏病提供较为准确的预测，能为医生诊断提供依据，将机器学习与医学相结合，精准诊断并治疗，提高医疗资源的利用效率，从而降低心脏病的死亡率。该方法经推广后可用于其他疾病的预测，对于控制各类疾病发生具有重要积极意义。

1.2. 文献综述

目前有许多针对心脏病的研究模型，然而大部分仅针对热门特征变量进行建模，并未考虑一些新型的影响因素，且使用国内患者数据的影响因素比较陈旧，这会降低模型的预测效能。其次这些研究都是基于类别平衡的数据集，但真实的临床数据多为类别不均衡数据，基于这种数据构建的机器学习模型性能差距较大，且无法直接给出模型基于哪些因素进行预测，这将无法满足医疗领域要求模型可解释的需求。因此本文根据疾病特点使用经典模型进行建模对比，且使用某地区患者数据，使其更好地服务于心脏病的预测。

2011 年，G. Subbalakshmi 等[5]将年龄、血压等因素作为预测患心脏病的指标，用贝叶斯分类器作为内核函数，建立支持决策的心脏病预测系统(DSHDPS)。Alickovic E.等[6]采用自回归模型提取特征，并通过 k 邻近算法、多层感知器和支持向量机等技术来判别心率是否正常。Dimopoulos A. C.等[7]将传统心血管疾病评分系统分别与决策树、随机森林和 KNN 算法结合进行对比分析，发现机器学习适合风险预测研究。Gokulnath 等[8]提出了 SVM 的优化函数，将其应用于遗传算法中，提升心脏病预测的效果。Khourdifi 等[9]利用粒子群优化算法和蚁群优化算法对人工神经网络进行优化，结果证明了所提出的混合方法在处理心脏病分类数据中的有效性和鲁棒性。Valarmathi 等[10]使用三种不同的超参数优化算法对随机森林和极端梯度提升算法进行参数调整与测试，研究发现随机森林通过随机搜索方式进行参数调优的心脏病预测结果更好。

2. 数据来源与处理

2.1. 数据来源

本文采用来自 CDC 公开的 2021 年行为风险因素检测系统(BRFSS)的原始心脏病数据集，其数据总量为 438,693 条，304 个特征变量。分类标签为是否患有心脏病，1 代表患有心脏病病，2 代表未患心脏病，在所有原始数据中共有 22,831 条记录为患有心脏病。其中部分特征都存在明显的关联性，且对心脏病的影响较小。

2.2. 数据预处理

在训练机器学习模型之前，特征选择是一个很重要的预处理过程，它能从原始特征集中选择最有用的特征子集，以提高模型的预测性能和泛化能力。当我们遇到维数灾难问题，如果可以挑选出重要的特征变量，然后再进行后续的学习过程，那么维数灾难问题可以减轻。并且剔除不相关的特征变量不仅可以降低学习任务的难度，使模型更易理解，还可以减少过拟合的风险，帮助我们创建一个稳定性良好、泛化性能更强的模型。本文对心脏病数据集的预处理主要采取以下步骤：

1) 变量筛选：依据一些关于影响心脏病风险因素的国内外研究，从临床诊断经验中提取患者的病史和症状，结合心脏病相关的参考文献，选择出高频率且在医学角度与心脏病相关性较强的因素作为自变

量。删除特征关联性较为明显的变量因素。最后我们从 304 个特征变量中选取了 30 个特征变量作为解释变量进行研究分析见表 1。

2) 缺失值处理：统计数据包括实验数据和调查数据，在此类调查数据获取的过程中，客观因素的限制或人为原因造成数据缺失的情况是非常常见的。由于缺失情况对变量特征值损失率过高且数量较大，例如，空值，调查对象回答“不知道”或拒绝回答的，因此我们将这部分数据直接剔除。

3) 离散化：离散化是指对连续的数值变量进行分段处理，可以分为等频和等长两种离散化方法。等频和等长离散化方法不仅可以提高算法的运行速度，而且更有，利于提升模型的预测精度。在此数据集中，如 Sex (受访者年龄)的值为[18, 65+)范围内的连续变量，按照等长离散化对年龄数据进行分段处理，得到 6 类分类数据。

4) 哑编码：当变量不是定量特征时，无法对模型进行训练，因此引入哑变量将不能够定量处理的变量量化，然后得到可用于模型训练的特征。对于有 n 个分类属性的自变量，通常需要选取 1 个分类作为参照，因此可以产生 $n-1$ 个哑变量。如 BMI (体重指数)分为过轻、正常、过重和肥胖四类多分类有序变量，将其赋值为 1、2、3、4，通过数字的大小关系来体现程度之间一定的等级关系。

5) 规范化处理：将现有数据进行规范化处理，转化为便于数据分析处理的特征形式。本文旨在分析预测是否患有心脏病，因此分类标签数值化处理为 1 = 未患有心脏病，0 = 为患有心脏病。

Table 1. Raw data variable interpretation
表 1. 原始数据变量解释

变量名称	变量解释	分类变量赋值
Heart_Disease	是否被告知患有心脏病	患有心脏病 = 0；未患有心脏病 = 1
Sex	受访者性别	男性 = 1；女性 = 2
Age	受访者年龄	18~24 = 1；25~34 = 2；35~44 = 3；45~54 = 4；55~64 = 5；65+ = 6
Height	受访者身高	121~140 = 1；141~160 = 2；161~180 = 3；181~200 = 4；201~220 = 5
BMI	体重指数	过轻 = 1；正常 = 2；过重 = 3；肥胖 = 4
Marriage	婚姻状况	已婚 = 1；离异 = 2；寡妇或鳏夫 = 3；分居 = 4；未婚 = 5；未婚夫妇中的一员 = 6
Edu	受教育程度	高中未毕业 = 1；高中毕业 = 2；上过大学或技术学校 = 3；大专或技校毕业 = 4
Income1	收入类别	<15,000$ = 1；15,000$~25,000$ = 2；25,000$~35,000$ = 3；35000$~50,000$ = 4；50,000$~100,000$ = 5；100,000$~200,000$ = 6；>200,000$ = 7
Urbstat	城市/农村状况	城市 = 1；农村 = 2；
French	食用炸土豆的频率	从不 = 0；每天一次 = 1；每周一次 = 2；每月不到一次 = 3；每月/年一次 = 4
Fruit	每天食用的水果总量	0~200 = 1；201~400 = 2；401~600 = 3；601~800 = 4；801~2000 = 5；2000+ = 6

Continued

Vegetable	每天食用的蔬菜总量	0~200 = 1；201~400 = 2；401~600 = 3； 601~800 = 4；801~2000 = 5；2000+ = 6
Smoke	吸烟的频率	从不 = 0；有时候 = 1；每天 = 2
Alcohol	每天平均饮酒量	从不饮酒 = 0；1~20 = 1；21~40 = 2；41~60 = 3
Exercise	每月是否参加体育锻炼	患有 = 1；没有 = 2
Walk	走路或爬楼梯是否有困难	患有 = 1；没有 = 2
Hblood	是否被告知患有高血压	患有 = 1；没有 = 2
HCholesterol	是否被告知患有高胆固醇	患有 = 1；没有 = 2
Angina	是否患有心绞痛或冠心病	患有 = 1；没有 = 2
Stroke	是否被告知患有中风	患有 = 1；没有 = 2
Asthma	是否被告知患有哮喘	患有 = 1；没有 = 2
Skin Cancer	是否被告知患有皮肤癌	患有 = 1；没有 = 2
Other Cancer	是否被告知患有其他癌症	患有 = 1；没有 = 2
Bronchitis	是否被告知患有慢性支气管炎	患有 = 1；没有 = 2
Depression	是否被告知患有抑郁症	患有 = 1；没有 = 2
Kidney	是否被告知患有肾脏疾病	患有 = 1；没有 = 2
Diabetes	是否被告知患有糖尿病	患有 = 1；没有 = 2
Arthritis	是否被告知患有关节炎	患有 = 1；没有 = 2
Aids	是否被告知患有 HIV	患有 = 1；没有 = 2
Body Health	身体健康状况	身体健康状况不佳的天数为零 = 1； 1~13 天身体健康状况不佳 = 2； 14~30 天身体健康状况不佳 = 3
Mental Health	心理健康状况	心理健康状况不佳的天数为零 = 1； 1~13 天心理健康状况不佳 = 2； 14~30 天心理健康状况不佳 = 3

2.3. 患病人数总体分布情况

首先查看数据标签的比例，标签特征为是否患有心脏病，黄色代表患有心脏病的记录共有 2994 条，比例为 6%，蓝色代表未患有心脏病的记录共有 48,305 条，比例为 94%。标签特征相差较大，处于不平衡状态。若直接对此数据进行建模分析是没有意义的，例如：利用机器学习方法建模分类，如果数据集中不均衡的比例超过了 4:1，分类器就会倾向于样本较多的类别，如果训练集中有 90%的样本都属于同一种类，那么分类器会将所有的样本都判断为属于该类，尽管最后的分类准确度很高，但在这种情况下分类器是无效的。所以我们考虑混合采样的方法对数据进行预处理。采样得到的数据中患有心脏病比未患有心脏病约等于 1:1 见图 1。

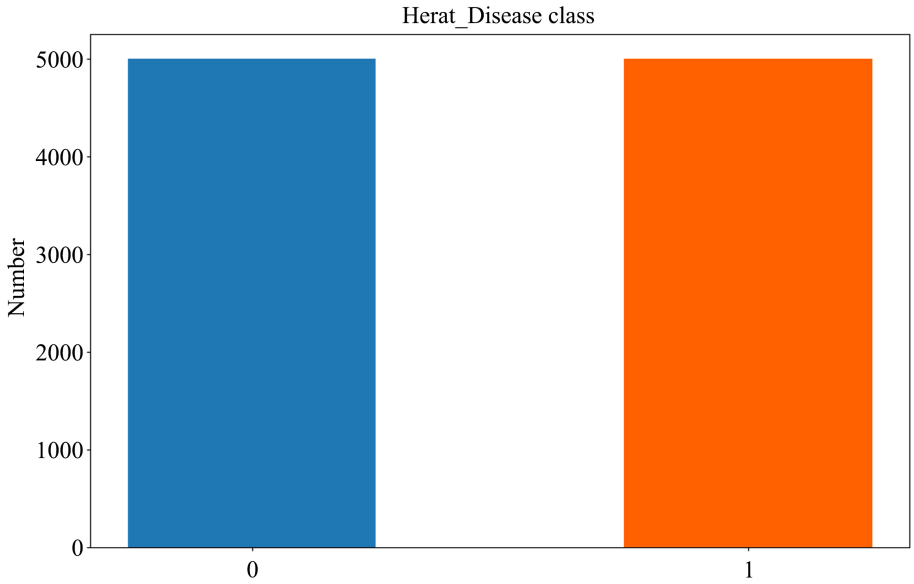

Figure 1. Distribution of heart disease after sampling
图 1. 采样后心脏病人数分布图

3. 模型构建

本研究通过信息增益分析训练模型的重要特征信息，模型中数据特征重要性降序排序结果见图 2。通过对比不同特征的信息增益值可知，这些重要特征是对模型输出具有不同的影响。其中年龄、高血压、高胆固醇和肥胖是心脏病的关键危险因素。不同的生活方式也会导致人们患有心脏病，包括：食用油炸食品、蔬菜的频率、吸烟、缺乏运动。食用高饱和脂肪和高胆固醇的饮食也会增加心脏病和相关疾病的患病风险，如：心绞痛(如动脉粥样硬化)。此外，部分现有研究中较少涉及的外在因素，如：收入、教育、婚姻等也会在一定程度上增加患心脏病和心脏病发作的风险。

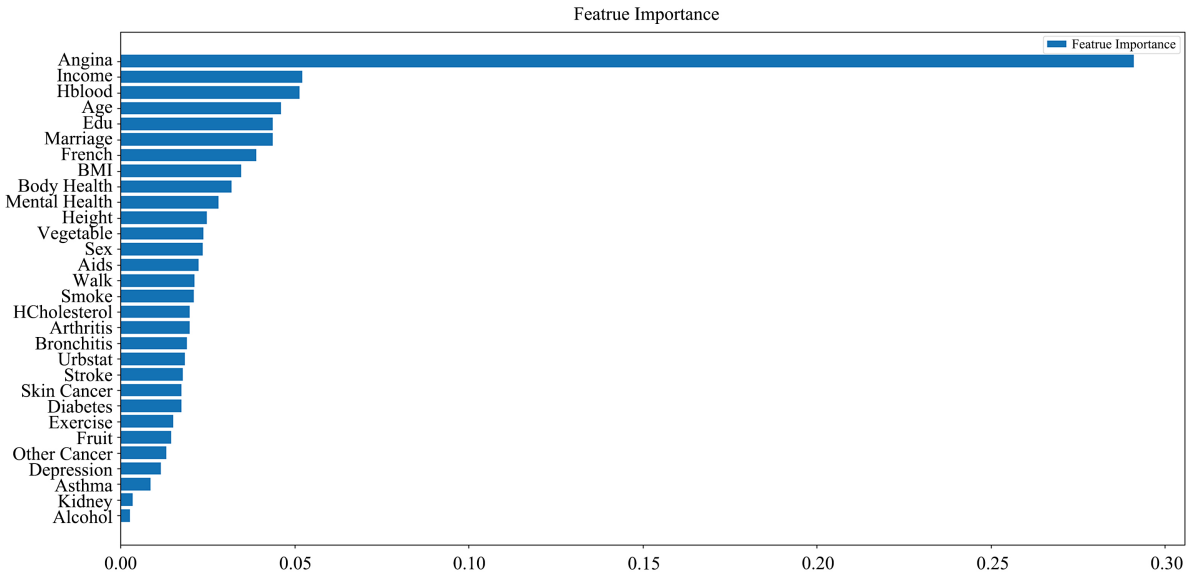

Figure 2. Data feature importance ranking
图 2. 数据特征重要性排序

3.1. 随机森林(Random Forest)

在随机森林模型中，我们将分别将数据的 70%、30%作为训练集和测试集，建立模型进行预测分析。此方法使用 Python 中的开源 Random Forest 包来建立建模。此模型需录入年龄、食用炸土豆的频率和体重特征等不同属性的解释变量数据进行抽样，建立大量决策树预测模型[11]。为了获得最佳的变量组合形式，用 OOB 样本在已训练好的决策树上运行，计算袋外预测误差 e_1，然后固定其他不变，依次计算决策树第 500 个特征的特征值，得到袋子外误差 e_{i2}。其中随机森林 OOB 得分为 0.8854，OOB 错误率为 0.1146。

得到随机森林模型的混淆矩阵见图 3：

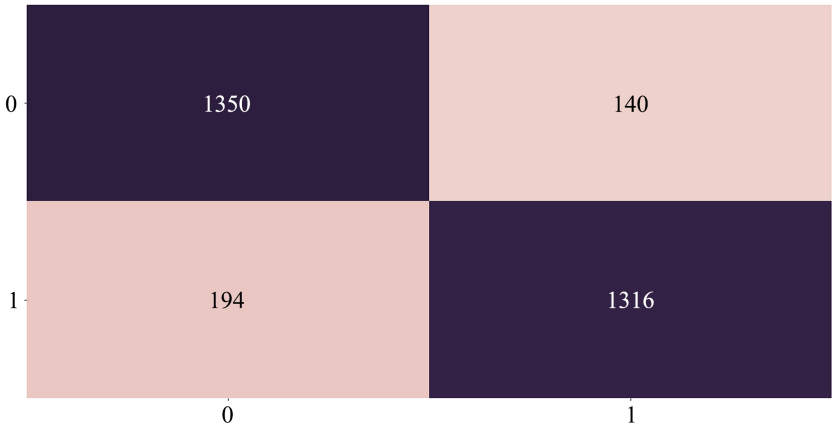

Figure 3. Random forest confusion matrix
图 3. 随机森林混淆矩阵

在随机森林模型测试集的 3000 个样本中，预测正确的共有 2666 个，预测错误的共有 334 个。其中有 140 个样本是将未患有心脏病的样本预测为患有心脏病的样本，而有 194 个样本是将实际患有心脏病的样本预测为未患有心脏病的类别。

3.2. 全连接神经网络

本文设置了 4 层隐藏层神经网络来预测是否患有心脏病，以性别、年龄、高血压、高胆固醇等 30 个变量作为输入，隐藏层节点设为 50，输出层节点为 2 的二分类模型。模型图见图 4：

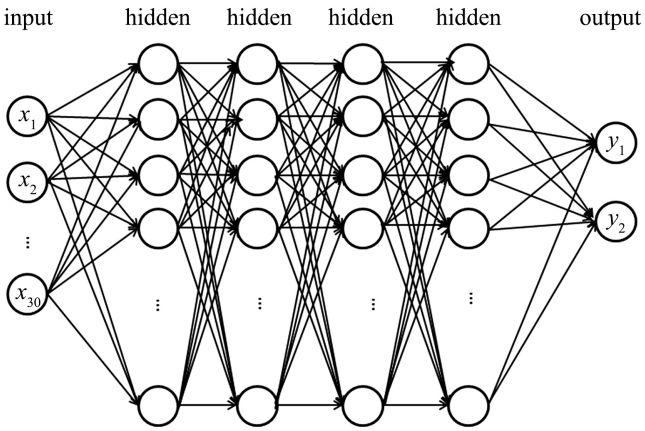

Figure 4. Fully connected neural network structure diagram
图 4. 全连接神经网络结构图

得到全连接神经网络的混淆矩阵见图 5：

Figure 5. Fully connected neural network confusion matrix
图 5. 全连接神经网络混淆矩阵

在全连接神经网络模型测试集的 3000 个样本中，预测正确的共有个 2375，预测错误的共有个 625。其中有 338 个样本是将未患有心脏病的样本预测为患有心脏病的样本，而有 287 个样本是将实际患有心脏病的样本预测为未患有心脏病的类别。

4. 模型评价

4.1. 模型结果

通过运用随机森林和全连接神经网络算法建立模型，两种模型的混淆矩阵结果见表 2 和表 3；各个模型 ROC 曲线汇总图见图 6；两个模型的评价指标结果见表 4。

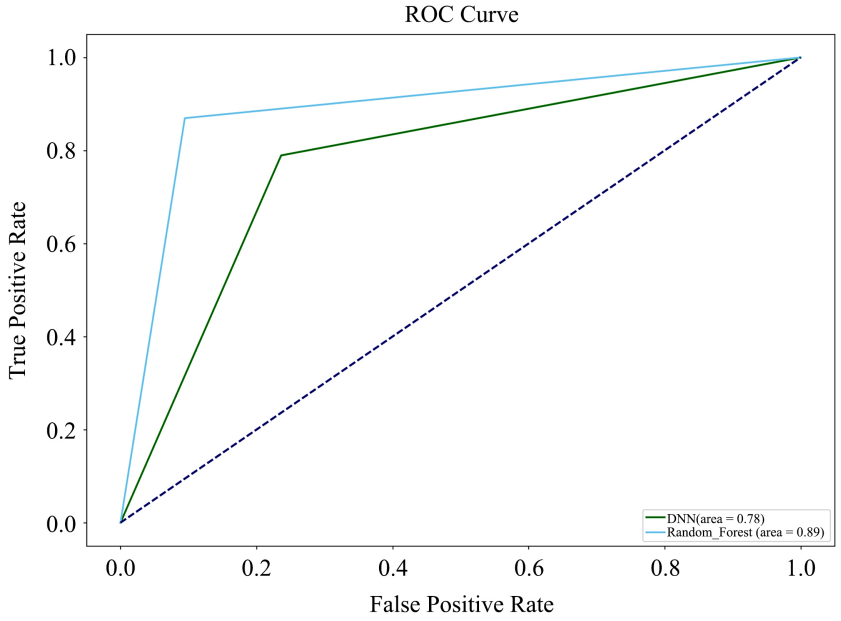

Figure 6. ROC curves of random forest and fully connected neural network models
图 6. 随机森林和全连接神经网络模型的 ROC 曲线

Table 2. Random forest confusion matrix
表 2. 随机森林混淆矩阵

	实际为患有心脏病	实际为未患有心脏病
预测为患有心脏病	1350	140
预测为不患有心脏病	194	1316

Table 3. Fully connected neural network confusion matrix
表 3. 全连接神经网络混淆矩阵

	实际为患有心脏病	实际为未患有心脏病
预测为患有心脏病	1152	338
预测为不患有心脏病	287	1223

Table 4. Comparison of model classification performance index
表 4. 模型分类性能指标对比

	准确率	精确率	召回率	AUC
随机森林	0.89	0.91	0.87	0.8878
全连接神经网络	0.85	0.77	0.80	0.7815

ROC 曲线的横轴 FPR 表示假阳率，纵轴 TPR 表示真阳率，ROC 曲线越靠近左上角，模型的性能越好[12]，因为此时 TPR 高、FPR 低，随机森林模型能够很好地区分正例和负例。

4.2. 结果分析

心脏病患者的数量和影响心脏病的因素逐渐增多，如何选择重要特征变量发现患有心脏病的患者是预测模型的核心内容。提前发现心脏病患者，及时介入治疗，有利于减低心脏病带来的死亡风险。敏感度和召回率两个指标共同反应了模型是否正确预测患者患有心脏病的能力，根据表 4 可得出两种机器学习模型中基于随机森林模型相较于全连接神经网络模型的分类性能更加优越。随着大数据时代的发展，大部分医学数据难以实现较为高效且快速的数据处理，故传统模型会面临淘汰。在较为前沿的机器学习算法所构建的模型中，随机森林在精确度、召回率、AUC (Area Under Curve)以及准确率 4 个方面都具有较好的效果，AUC 的取值范围在 0.5 到 1 之间，越接近 1 表示模型的性能越好，说明随机森林模型在数据量较大、指标分类较为复杂的医学邻域有较为广阔的前景。同时，相较于全连接神经网络，随机森林不依赖高维变量的组合优化，针对存在较大异质性的数据的情况，我们可以加大决策树来消除具有极端情况的异质性，从而表现出良好的适用性和泛化性。

在大数据新发展格局下，数据类型多样，来源复杂，范围扩大，需要新型生物统计理论方法加以支撑，医疗统计在疾病预测，危险因素分析等方面起着至关重要的作用。随机森林模型的建立应不断与临床需求相适应，选出较为重要的影响因素纳入学习模型进行不断迭代，使得模型更加优化和完善。灵活地将统计应用到医学事业中是时代的大势所趋，更是实现健康中国的重中之重，把人民健康放在首位，大幅提高健康水平，努力全方位、全周期保障人民健康，为全面建设社会主义现代化国家、全面推进中华民族伟大复兴打下坚实的健康基础。

参考文献

[1] 刘彦华. 2023 中国现代生命发展指数 76.0 国人最担忧的三大疾病: 癌症、心脏病、眼病[J]. 小康, 2023(10): 46-48.

[2] 陈育德, 杨辉. 贯彻“十四五”国民健康规划, 确保实现健康预期寿命目标[J]. 中国全科医学, 2023, 26(4): 391-394+408.

[3] Cai, Y., Cui, X., Su, B.B. and Wu, S.Y. (2022) Changes in Mortality Rates of Major Chronic Diseases among Populations Aged over 60 Years and Their Contributions to Life Expectancy Increase, China, 2005-2020. *China CDC Weekly*, **4**, 866-870. https://doi.org/10.46234/ccdcw2022.179

[4] Muthiah, V.,A G M,Varieur, J T, et al. (2022) The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. *Journal of the American College of Cardiology*, **80**, 2361-2371. https://doi.org/10.1016/j.jacc.2022.11.005

[5] Subbalakshmi, G., Ramesh, K. and Chinna Rao, M. (2011) Decision Support in Heart Disease Prediction System Using Naive Bayes. *Indian Journal of Computer Science and Engineering*, **2**, 170-176.

[6] Alickovic, E. and Subasi, A. (2015) Effect of Multiscale PCA De-Noising in ECG Beat Classification for Diagnosis of Cardiovascular Diseases. *Circuits*, *Systems*, *and Signal Processing*, **34**, 513-533. https://doi.org/10.1007/s00034-014-9864-8

[7] Dimopoulos, A,C., Nikolaidou, M., Caballero, F.F., Engchuan, W., Sanchez-Niubo, A., Arndt, H., Ayuso-Mateos, J.L., Haro, J.M., Chatterji, S., Georgousopoulou, E.N., Pitsavos, C. and Panagiotakos, D.B. (2018) Machine Learning Methodologies versus Cardiovascular Risk Scores, in Predicting Disease Risk. *BMC Medical Research Methodology*, **18**, Article Number: 179. https://doi.org/10.1186/s12874-018-0644-1

[8] Gokulnath, C.B., and Shantharajah, S.P. (2019) An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease. *Cluster Computing*, **22**, 14777-14787. https://doi.org/10.1007/s10586-018-2416-4

[9] Khourdifi, Y. and Bahaj, M. (2019) Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Systems*, **12**, 242-252. https://doi.org/10.22266/ijies2019.0228.24

[10] Valarmathi, R. and Sheela, T. (2021) Heart Disease Prediction Using Hyper Parameter Optimization (HPO) Tuning. *Biomedical Signal Processing and Control*, **70**, 103033. https://doi.org/10.1016/j.bspc.2021.103033

[11] 王浩淼, 曹若菲, 林金欣, 等. 基于脑出血患者院前指标的多种机器学习预测模型构建及比较研究[C]//中国统计教育学会, 教育部高等学校统计学类专业教学指导委员会, 全国应用统计专业学位研究生教育指导委员会. 2021 年(第七届)全国大学生统计建模大赛获奖论文集(一). 2021: 394-439.

[12] 唐善成, 陈明, 王瀚博, 等. 采用变分自编码器的无监督压敏电阻表面缺陷检测[J]. 计算机集成制造系统, 2022, 28(5): 1337-1351. https://doi.org/10.13196/j.cims.2022.05.006