http://dx.doi.org/10.52810/JIR.2024.003

REVIEW

机器学习模型在心血管疾病中的应用

蒋子悠[1,*]

[1] 北京工商大学，计算机与人工智能学院，北京 100048

学术编辑： 马慧鋆；**收稿日期：** 2024-03-12；**录用日期：** 2024-04-17；**发布日期：** 2024-05-07

***通讯作者：** ✉蒋子悠，2307010206@st.btbu.edu.cn

文章引用

蒋子悠. 机器学习模型在心血管疾病中的应用. 智能机器人, 2024, **1**(1): 26–38.

Citation

Jiang, Z. (2024). Machine Learning Models in Cardiovascular Diseases. Journal of Intelligent Robots, 1(1), 26–38.

摘　要

随着当今社会带给人们的高强度工作生活压力，心血管疾病问题的日益严峻，发病率逐年增加，全球对此类疾病的关注与日俱增。传统的预测方法虽有一定预测能力，但是特异性较低，而机器学习和深度学习技术在为心血管疾病的高效预测和设计提供了新的解决方案。本文综述了机器学习和深度学习在心血管疾病预测中的应用，从心血管疾病问题现状引出对其预测的重要性，介绍了其遭遇的挑战，以及预测模型的优势性能评估。尽管面临诸多挑战，机器学习模型在预测心血管疾病研究中的应用仍具有巨大潜力，有望为降低心血管疾病发病率提供新的支持策略。

关键词: 人工智能，深度学习，机器学习，心血管疾病

Machine Learning Models in Cardiovascular Diseases

Ziyou Jiang[1,*]

[1]School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

Academic Editor: Huijun Ma; **Submitted:** 2024-03-12; **Accepted:** 2024-04-17; **Published:** 2024-05-07

***Correspondence Author:** ✉Ziyou Jiang, 2307010206@st.btbu.edu.cn

Abstract

With the high intensity of work and life pressure brought by today's society, the problem of cardiovascular dis-

ease is becoming more and more serious, the incidence rate is increasing year by year, and the global attention to such diseases is increasing day by day. Although traditional prediction methods have some predictive ability, but the specificity is low, while machine learning and deep learning technology in the efficient prediction and design of cardiovascular disease provides a new solution. This paper reviews the application of machine learning and deep learning in CVD prediction, introducing the importance of prediction of CVDs from the current state of the problem, the challenges encountered, and the evaluation of the superior performance of prediction models. Despite the challenges, the application of machine learning models in predictive cardiovascular disease research has great potential and is expected to provide new support strategies for reducing the incidence of cardiovascular disease.

Keywords: Artificial intelligence, deep learning, machine learning, cardiovascular disease

1 引言

心血管疾病是全球导致死亡的首要原因，在全球范围内是主要的健康威胁之一 [1]。特别是在，中国现约有 3. 3 亿人口患有心血管疾病，其死亡率占居民疾病总死亡率的 40%，其中农村心血管疾病死亡率持续高于城市，且目前心血管疾病的患病人数和致死率在我国仍呈上升趋势 [2]，因此预测诊疗心血管疾病至关重要。心血管疾病具有隐匿性高，发病率和死亡率高等特点 [3]，故准确预测心血管疾病发病风险对于当今社会有重要的现实意义。

在传统的预测方法中 [34, 35]，过往只基于数量有限的风险因素，例如基础疾病，吸烟史，肥胖程度和年龄等，虽做出了一定程度上的预测，但明显对于每个病患的特异性和针对性仍需提高，难以做出精准针对于个人的预测。而机器学习方法 [36, 37] 利用计算机系统以算法和统计模型 [38–42]，依靠模式和推理来执行特定的任务，而不使用明确指令的一种方法进而实现预测或决策任务，能为疾病预测的准确性，便捷性，高效性的提升提供了帮助，也可以应用于各种工业任务中 [43–50]。

在下文提到的可用于心血管疾病心血管疾病的机器学习模型之间存在着密切的关联和区别。传统的机器学习模型 [51–57] 如支持向量机、随机森林、Logistic 回归和决策树，通常依赖于手工设计的特征进行训练和预测。其中，随机森林是由多个决策树构成的集成学习方法，通过组合多个决策树来提高预测性能和鲁棒性。与传统模型不同，卷积神经网络（CNN）是一种深度学习模型 [58–70]，具有自动学习特征表示的能力。CNN 通过多层卷积和池化操作来提取图像等数据的特征，并通过全连接层进行分类。相比于传统模型，CNN 在处理复杂数据（如图像、语音）时表现出更好的性能。尽管这些模型在应用和性能上有所不同，但它们在机器学习领域都扮演着重要角色。随机森林的集成方法可以提高预测准确性和鲁棒性，而 CNN 则适用于处理复杂数据并自动学习特征表示。在实际应用中，根据数据类型、问题复杂度和性能需求等因素，选择合适的模型进行建模和预测是至关重要的。

2 机器学习历史

机器学习的历史可以追溯到人工神经网络的研究起源。在 1943 年，Warren McCulloch 和 Walter Pitts 提出了一种神经网络的层次结构模型 [4]，这个模型模仿了生物神经系统的多层次神经元结构，用于数据识别、学习和处理。该模型由输入层、隐藏层和输出层组成，并通过训练来调整连接权重，以便对复杂的数据模式进行拟合和预测。这个模型奠定了神经网络计算理论的基础，并为机器学习的发展提供了重要的起点。

在 1957 年，教授 Frank Rosenblatt 引入了 Perceptron（感知器）的概念 [5]，并首次使用算法精确地定义了自组织和自学习的神经网络的数学模型。他设计了世界上第一个计算机神经网络，这个机器学习算法成为了神经

网络模型发展的基础性工作。在 1982 年，Hopfield 发布了一篇具有深远影响的关于神经网络模型的文章 [6]。他在文章中构建了一个能量函数，并将这一概念融入到 Hopfield 网络中。同时，通过对动力系统性质的研究，他成功地实现了使用 Hopfield 网络进行最优化求解。这一成果极大地推动了神经网络的应用研究，并为未来的发展开辟了更多可能性，包括支持向量机和逻辑回归等方法。这些模型在理论上相对简单，训练方法也容易掌握。它们能够从给定的训练样本数据及其对应结果中学习到内在模式规则，以完成对象识别、任务分类和简单结果预测等初级智能工作。相比于其他传统的基于固定规则和单一标准的方法，这些基于统计规律的浅层学习方法具有很多优势，取得了不少成功的应用，同时也加深了人们对浅层学习的理解。然而，其相关问题也逐渐暴露出来，例如学习能力不强，只能提取初级特征等。

在 2006 年，随着科技社会的发展和计算机硬件技术的快速进步，计算能力的提升为机器学习的发展提供了更高的上限。Geoffrey Hinton 和 Ruslan Salakhutdinov[7] 提出了深度学习模型，该模型的主要观点包括：具有多个隐藏层的人工神经网络具有良好的特征学习能力；通过逐层初始化来克服训练的难度，实现网络的整体调优。这一模型的提出标志着深度神经网络机器学习新时代的开始。机器学习为计算机系统提供了利用数据和经验的自学能力。基于一系列算法，机器学习使得计算机能够识别模式、做出决策并对未来事件进行预测，而无需人为编写具体的指令。

心血管疾病预测是一项复杂的任务，其特点包括多因素关联、早期预测的重要性、大量数据的可用性以及疾病发展的动态变化。这些因素之间存在复杂的相互作用，需要实时或近实时的监测和预测。故此在下文中提到了相应可解决这些问题的机器学习模型。

机器学习方法在处理这些特点方面具有以下优势：首先，机器学习算法可以处理大量的复杂数据，并从中提取有用的信息。这有助于处理多因素关联的心血管疾病预测问题。其次，机器学习算法可以自动从数据中学习有用的特征，而不需要人工进行特征工程。这在处理高维数据和动态变化的心血管疾病预测问题时非常有用。此外，虽然传统的统计方法在模型的解释性方面有一定的优势，但机器学习方法也可以通过特征重要性、模型解释等方式提供一定的解释性。这有助于理解模型的预测结果并进行进一步的研究。最后，机器学习算法可以通过不断的学习和优化来提高预测的准确性和可靠性。这对于处理动态变化的心血管疾病预测问题非常有帮助。总之，由于其强大的数据处理能力、自动学习特征、模型可解释性和持续改进的能力，使用机器学习方法对于心血管疾病预测是有效的。

3 基于机器学习模型的心血管疾病预测方法

3.1 支持向量机

支持向量机（support vector machines, 下文将之简称为 SVM）是一种用于监督学习的算法 [8]，它不依赖于预先设定的模型，而是利用数据本身来生成预测 (如图 1 所示)。作为数据驱动的方法，SVM 特别擅长分类任务。SVM 通过寻找一个最优的决策边界，即最大化两类数据点之间的间隔的超平面，来实现分类。这个超平面不仅将输入数据映射到更高维度的空间，还生成一个 n 维向量，并尽可能扩大两个类别之间的距离，以实现最佳的类别分离。在处理非线性可分的数据时，SVM 使用核技巧将原始的特征空间映射到一个更高维度的空间，使得原本非线性可分的数据在新的空间中变得线性可分。这种映射能力使得 SVM 在处理复杂的分类问题时具有很高的灵活性和强大的泛化能力。SVM 的另一个重要特性是它只关注那些位于决策边界附近的数据点，这些数据点被称为支持向量。这使得 SVM 在处理大规模数据集时具有较高的效率，因为它不必处理所有的数据点，只需关注那些对构建决策边界至关重要的支持向量。SVM 尤其适用于小样本数据集，能有效处理高维数据问题，因此在疾病监测和生物信息学领域的分类问题中表现出色 [9, 10]，并得到广泛应用。

Shi 等构建了颅内小动脉瘤破裂风险的预测模型 [11]，基于 504 例颅内小动脉瘤患者（395 例破裂动脉瘤、109

例未破裂动脉瘤）数据，采用 SVM，Logistic 回归等方法构建预测模型，实验结果辨明 SVM 在内部验证集中表现均最好。Joloudari 等 [12] 提出了一种新的高效 CAD(Computer Aided Design, 简称 CAD) 诊断方法，称为 GSVMA，通过选择关键特征来帮助 CAD 的有效诊断和预测。GSVMA 方法有两个关键障碍。第一种方法是遗传优化，它选择最重要的特征。第二种是带有方差分析核的 SVM 算法，用于对输入数据集进行分类，所提出的 GSVMA 方法在 55 个特征中的 31 个特征上表现最佳，包括准确度（89.45%）、F 测量（80.49%）、特异性（100%）、灵敏度（81.22%）和曲线下面积（AUC）（100%）。结果表示,SVM 预测精准度高。Aljarah 等 [13] 最近实施了 GOA 和 SVM 的混合，以最大限度地提高 SVM 分类的准确性。混合 GOA-SVM 在 18 个数据集上进行了测试。将实验结果与 GA、PSO、GWO、CS、萤火虫算法（FF）、蝙蝠算法和多元宇宙优化器进行了对比。尽管在 GOA-SVM 方面已经做了很多努力，但它有一些缺点，即被困在局部最优中。

SVM 在心血管疾病预测中展示了显著的优势，包括其处理高维数据的强大能力，这对于分析如基因表达谱这样的复杂生物信息学数据至关重要。SVM 的核技巧使其能够解决非线性问题，从而捕捉数据中的复杂关系。其泛化能力得益于最大化边界间距的设计，使得模型即使在未知数据上也能保持准确的预测。在训练过程中，SVM 能识别出最具影响力的支持向量，有助于挖掘与疾病密切相关的关键基因。此外，通过不同的核函数，SVM 可以灵活适应多种数据和问题，同时其健壮性减少了对个别数据点的依赖。然而，SVM 也存在一些局限性。它对数据中的噪声和异常值较为敏感，这可能影响其性能。参数调整是 SVM 面临的一个挑战，需要经验和技巧来找到最佳的参数组合。对于大规模数据集，SVM 的训练可能耗时且占用大量内存。尽管 SVM 能揭示重要的支持向量，但其决策边界的复杂性可能导致模型解释性不如简单算法。对于多类别问题，需要采取特定策略来解决，这可能增加模型的复杂性。综上所述，虽然 SVM 是一种功能强大的机器学习工具，尤其适合处理包括心血管疾病预测在内的复杂分类问题，但要充分发挥其潜能，就需要进行细致的数据预处理、特征选择和参数调整。

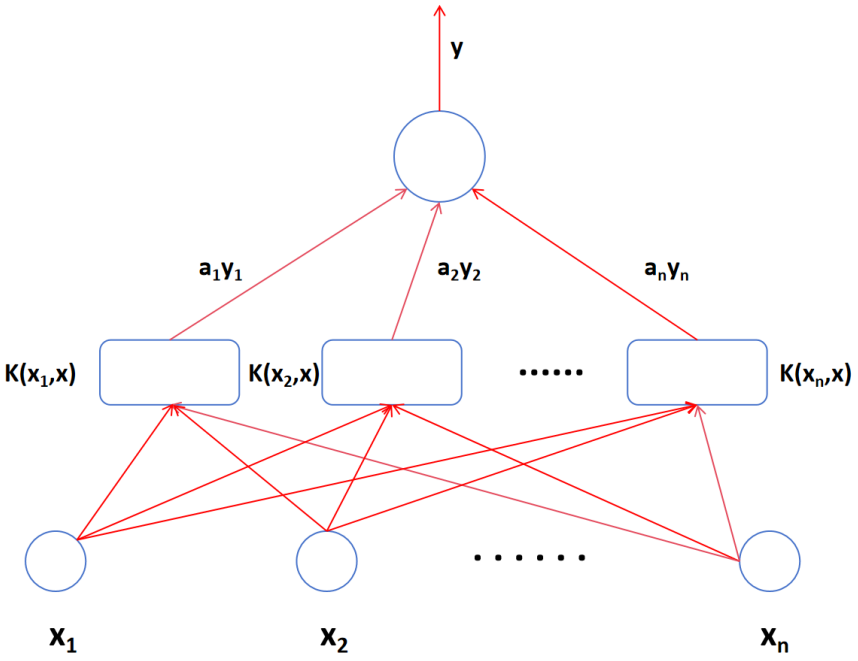

图 1. 支持向量机示意图

3.2 随机森林

随机森林是一种监督学习中的集成算法，它由多个随机生成的决策树构成，这些决策树共同作用以形成一个强大的分类器 [14](如图 2 所示)。该算法通过组合多个弱分类器（即单个决策树），显著提升了分类性能。为了

确保高效性，随机森林要求决策树之间保持独立性且可并行创建 [15]。它通过自主抽样法从原始数据中抽取多个样本，并利用决策树作为基分类器对这些样本进行训练。最终，众多决策树的结果通过投票机制汇总，以得出最终的分类或预测结果。随机森林可以结合多种基分类器，包括 CART 决策树 [16]、基于核函数的极限学习机（KELM）[17]、BP 神经网络 [18]、SVM 等。其引入的随机性和集成方法使得模型具有较低的过拟合风险和良好的抗噪声能力，因此，在医学 [19]、管理学、经济学等多个领域得到了广泛应用 [20]。

Li等 [21] 于 2014 年从浙江省 101056 人中选取 29930 名心血管疾病高危受试者，Logistic 回归分析显示，近 30 个指标与心血管疾病相关，包括性别、年龄、家庭收入、吸烟、饮酒、肥胖、腰围过大、胆固醇异常、低密度脂蛋白异常、空腹血糖异常等。他们采用多种方法构建预测模型，使用多变量回归模型作为性能评估的基准（曲线下面积，AUC = 0.7143）。结果表明，随机森林优于其他方法，AUC 为 0.787，比基准有显著提高。石胜源等 [22] 根据重要性评分, 对特征进行降序排序，并运行 (SWSFS) 过程，结果显示，在变量数为 8 时具有最优的分类准确性，所以将重要性评分排在前 8 位的特征纳入随机森林模型进行分析，所选变量为收缩压、胆固醇、性别、葡萄糖、舒张压、年龄、体重、吸烟情况。实验可以看出随机森林模型在 4 个指标 (精确度, 精度, 召回率,F1-score) 上都优于其它 3 种方法 (逻辑回归,K 近邻,SVC), 结果表明随机森林本身精度比大多数单个算法要好并具有一定的抗过拟合能力和抗噪声能力, 对比其他算法具有一定的优势。

随机森林是一种集成学习算法，它通过组合多个决策树来做出最终的预测。该模型具有许多优势，使其在各种类型的预测任务中表现出色。首先，随机森林模型通常能够提供较高的预测准确率，有时甚至可以与神经网络相媲美，而且往往比逻辑回归等传统方法更高。其次，该模型对错误数据和离群点具有较好的鲁棒性，这意味着即使数据集中存在一些异常值，模型的预测性能也不会受到太大影响。此外，由于随机森林由多个决策树组成，因此可以有效减少单一决策树可能出现的过度拟合问题。随机森林算法同样适合处理大规模数据集，因为其训练过程可以并行化，从而提高计算效率。这对于处理大数据问题非常有益。此外，随机森林还可以评估各个特征对于最终预测的重要性，这有助于理解模型的决策过程。

然而，随机森林也存在一些局限性。首先，该模型中存在许多超参数，如树的数量、树的深度等，这些参数的选择对模型的性能有很大影响，需要进行细致的调节优化。其次，虽然随机森林整体上是一个黑箱模型，但与单个决策树相比，它的解释性相对较差。此外，在训练过程中，随机森林可能需要较多的计算资源，尤其是在树的数量很多时。最后，传统的随机森林采用平均多数投票规则，可能无法区分强弱分类器，这在一定程度上限制了模型的性能。总而言之，随机森林是一个强大且多功能的预测模型，适用于各种类型的预测任务。然而，为了达到最佳性能，需要对模型进行适当的调优和参数选择。在实际应用中，结合模型的优势和局限性，选择合适的算法和调优方法是关键。通过合理的调优和参数选择，可以充分发挥随机森林的优势，提高预测性能，同时克服其局限性。

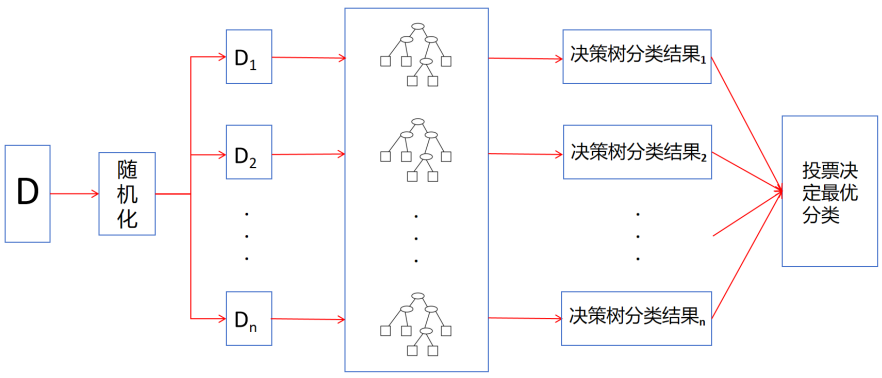

图 2. 随机森林示意图

3.3 Logistic 回归

Logistic 回归是一种用于二元分类问题的统计技术 [23]，广泛应用于数据挖掘、疾病风险预测、市场营销等领域。与线性回归不同，Logistic 回归模型的输出是类别性的，通常用 0 和 1 表示，例如一个邮件是否为垃圾邮件或者一个病人是否患有某种病状。该模型通过逻辑函数（Sigmoid 函数）将线性组合的输入转化为介于 0 和 1 之间的概率值 (如图 3 所示)，从而可以解释为事件发生的概率 [24]。模型参数通常通过最大化似然函数来估计，使用诸如梯度下降或牛顿法等优化算法求解。Logistic 回归在形式上简洁且易于实现，不需要线性关系假设，对数据的分布要求相对宽松，并且输出具有直观的概率解释。不过，它默认自变量间相互独立，不适用于处理复杂的非线性关系，尽管存在局限性，Logistic 回归因其高效性和广泛的适用性，仍是现实社会中分类问题中常用的重要工具。Feng 等围绕急性缺血性卒中（arterial ischemic stroke, AIS）患者术后发生出血性转化的独立预测因素开展研究 [25]。基于 90 例前循环大动脉闭塞所致 AIS 患者, 建立多元 Logistic 回归模型证实, 较差的侧支循环及较高的血小板/淋巴细胞比率与出血转化显著相关。Thanuja Nishadi A S 等 [26] 人提出了在具有 4238 条记录的 Framingham 数据集上对心脏病进行分类的逻辑回归模型,logistic 回归的准确率为 86.66%。Montu Saw 等 [27] 人提出了逻辑回归模型来对心脏病进行分类。该研究使用 Framingham 数据集和物流回归，准确率为 87.02%。

Logistic 回归训练速度快, 这主要得益于其计算量仅与特征的数目相关, 使得它能够快速拟合大规模数据集。其次，解释性强也是 Logistic 回归的一大优势，由于模型的形式简单易懂，我们可以直接从特征的权重看出不同特征对结果的影响，这对于模型的解释和理解非常有帮助。再者，Logistic 回归特别适合处理二分类问题，无需对输入特征进行缩放，这使得它在许多实际问题中非常实用。此外，内存资源占用小，只需存储各个维度的特征值，节省了内存资源，对于资源有限的环境来说，这是一个重要的优势 [28]，直接对分类可能性建模，无需事先假设数据分布，减少了假设不准确带来的问题，使得它在各种情况下都能得到应用。最后，以概率形式输出结果，适用于需要利用概率辅助决策的任务 [29]，例如信用评分、疾病诊断等。

然而，Logistic 回归也存在一些局限性。首先，由于其决策面是线性的，它无法解决非线性问题，这限制了其在复杂问题上的应用。其次，对多重共线性敏感，在存在多重共线性的数据中，模型的性能可能会受到影响，这可能导致模型的解释性和预测能力下降。再者，难处理数据不平衡，在数据不平衡的情况下，Logistic 回归的表现可能不佳，这可能导致模型对于少数类的预测性能较差。此外，由于模型形式的简单性，可能无法充分拟合数据的复杂分布，导致准确率不高，这在复杂问题上可能是一个重要的限制。最后，Logistic 回归本身无法自动筛选特征，有时需要借助其他方法如 GBDT 来选取重要特征，这增加了模型建立的复杂性。

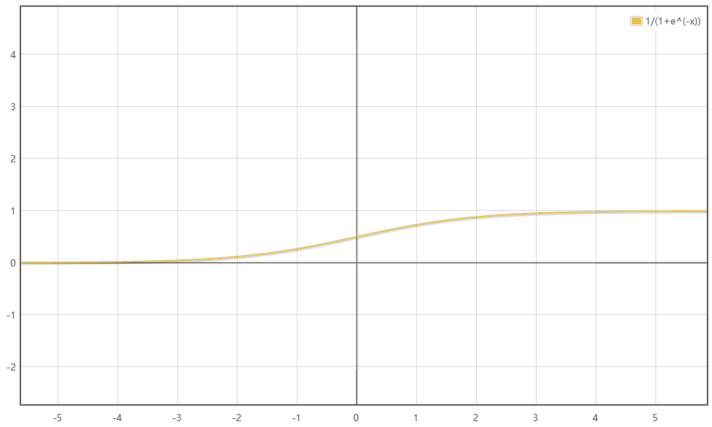

图 3. Logistic 回归函数曲线图

3.4 决策树

决策树主要用于分类和回归问题，它通过对数据集的特征进行递归划分，形成一个树形结构的模型 (如图 4 所示)。决策树的每个内部节点表示一个特征属性上的判断，每个分支代表一个判断结果，而每个叶子节点代表一种类别（对于分类问题）或者一个数值（对于回归问题）。决策树模型拥有明晰的建立过程 (如图 5 所示), 易于理解和解释，数据准备工作相对较少，能够处理数值型和类别型数据，且对数据的微小变化非常敏感，以及可以处理多输出问题。然而，它也容易过拟合（指模型在训练集上表现优异，但在验证集或测试集上表现不佳的现象，可能成因为训练数据量相对较少或者噪声较多），对于类别较多的数据，错误可能会增加。Hostettler 等利用决策树分析法构建了一个用于预测动脉瘤性蛛网膜下腔出血患者预后的模型 [30]。该模型以 548 名患者的数据集为基础，并实现了 71.1% 的预测准确度。

决策树模型是一种广泛使用的机器学习算法，它的结构类似于流程图，非常直观，使得非专家也能轻松理解其工作原理。其次，由于决策树对每个特征单独处理，因此不受特征尺度的影响，无需进行数据缩放。此外，它能够很好地处理包含二元特征和连续特征的离散型数据。同时，决策树还可以作为集成学习的基模型，如随机森林和梯度提升树等高级树模型都是通过组合多个决策树来提高预测性能的。相较于其他复杂的机器学习模型，决策树在小数据集上的分类和回归问题上表现尤为出色。

然而，决策树也存在一些缺点。首先，它容易在训练集上过度特化，导致泛化能力较弱，这通常需要通过剪枝等方法来缓解过拟合问题。其次，由于每条路径都对应一条 if-then 规则，决策树对局部数据非常敏感，容易陷入局部最优解。最后，数据的小变化可能会导致生成完全不同的树，从而影响最终的预测结果，这使得决策树的稳定性较差。综上所述，在实际应用中，尽管决策树存在一些缺点，但其简单易懂且对数据预处理要求不高的特点使其被广泛使用。为了克服容易过拟合的问题，通常会采用集成学习方法如随机森林来替代单一决策树模型，以提高模型的稳定性和泛化能力。

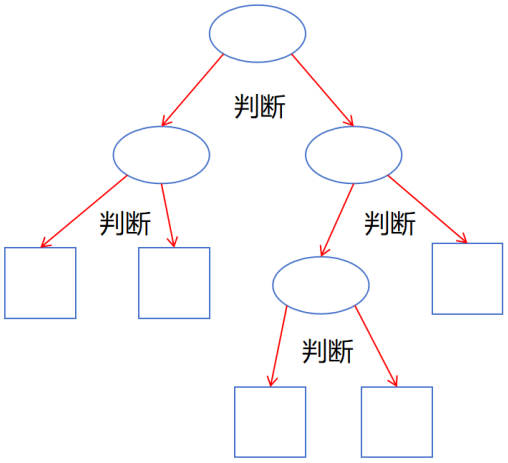

图 4. 决策树示意图

3.5 卷积神经网络（CNN）

卷积神经网络（Convolutional Neural Networks, CNNs）是一种深度学习模型，具有分层结构 (如图 6 所示)，它在图像识别、视频分析以及众多计算机视觉任务中表现卓越。CNN 的核心特点是利用卷积层来自动并有效地提取空间层级特征，这使得网络能够捕捉到图像中的局部模式，如边缘、纹理等，以及更高层次的抽象概念。一个典型的 CNN 包含若干种类型的层：1) 卷积层（Convolutional Layer）：使用一组可学习的滤波器来扫描

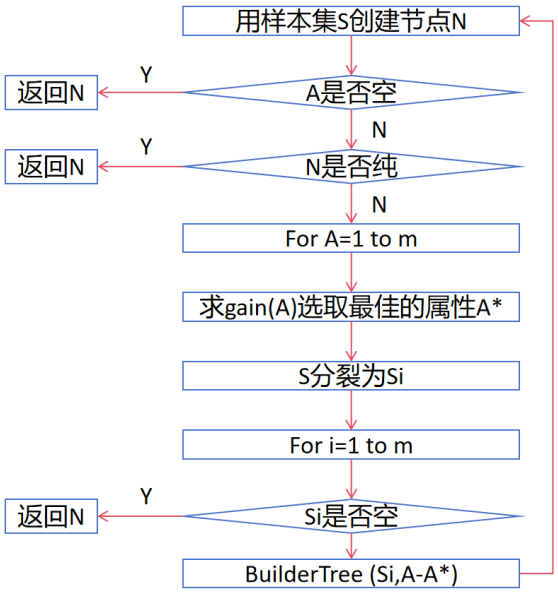

图 5. 建立决策树的流程图

输入数据，生成特征图（feature maps）。2) 激活层（Activation Layer）：引入非线性激活函数，比如 ReLU，以增加模型的表达能力。3) 池化层（Pooling Layer）：通过降低特征图的维度来减少参数数量和计算量，同时提高模型对小变形的鲁棒性。4) 全连接层（Fully Connected Layer）：在网络的末端，将高维特征映射到样本的标签空间。CNN 通过反向传播算法进行训练，该算法根据损失函数的梯度来更新网络中的权重。随着训练的进行，CNN 能够逐渐学习到复杂的特征表示，从而提高对未见数据的泛化能力。在心血管疾病预测领域，CNN 可以用于分析医学影像数据，如心脏 MRI 或 CT 扫描，以检测心脏结构和功能的异常。此外，通过处理心电图（ECG）信号，CNN 有助于识别心律失常和其他心脏病症。在这些应用中，CNN 通常需要针对特定问题进行微调，这可能包括调整网络架构、选择合适的激活函数和损失函数，以及实施特定的数据增强策略。

卷积神经网络（CNN）作为一种深度学习模型，它们能够通过卷积层自动学习图像的空间层次结构，这使得在处理图像数据时非常有效。其次，CNN 的卷积层采用局部连接和参数共享的机制，大大减少了模型的参数数量，降低了过拟合的风险，并提高了计算效率 [31]。此外，由于参数共享，CNN 对输入数据的平移具有不变性，无论特征出现在输入数据的哪个位置，CNN 都能识别它。CNN 还能自动提取复杂的特征，随着网络深度的增加，可以学习到更加抽象的高级特征。最后，CNN 特别适合处理高维度的数据，如图像和视频，因为它们可以有效地处理像素之间的空间关系。然而，CNN 也存在一些局限性 [32]。首先，尽管 CNN 在很多任务上表现出色，但它们的决策过程缺乏可解释性，因此有时被称为“黑盒”模型。其次，CNN 通常需要大量的标注数据来训练，这在某些领域可能难以获得。此外，CNN 的训练和推理过程可能需要大量的计算资源，尤其是对于大型网络和大规模数据集。再者，CNN 的性能很大程度上取决于超参数的选择，如网络结构、卷积核大小、步长等，这些超参数的调整往往需要大量的实验和经验。最后，CNN 通常假设输入数据具有固定的大小和格式，这可能导致在处理不同规模或格式的数据时需要额外的预处理步骤。

3.6 循环神经网络（RNN）

循环神经网络（Recurrent Neural Networks, RNN）是一种专为处理序列数据设计的深度学习模型 (结构如图 7 所示)，其核心特性在于能够利用内部状态来存储和处理时间序列中的信息。这种网络特别适合于分析和预测那些随时间展开的数据，如股票价格、气象变化、语音识别或文本生成等场景。在 RNN 的架构中，信息可以在

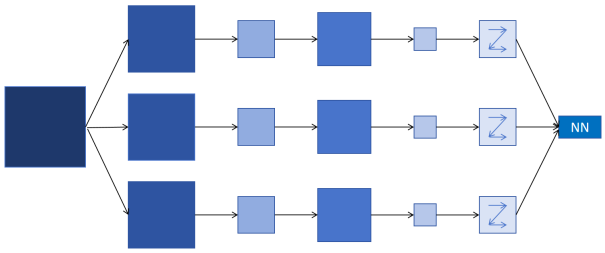

图 6. 简化卷积神经网络结构示意图

序列的不同时间步之间传递，允许网络捕捉到时间上的动态关系。理论上，RNN 可以处理任意长度的序列，因为它们可以使用其内部的“记忆”来持续追踪之前的输入信息。尽管 RNN 具有处理时间依赖关系的潜力，它们也面临着一些挑战，最著名的是梯度消失和梯度爆炸问题，这影响了它们学习长期依赖能力的稳定性。为了应对这些挑战，研究者们引入了改进型的 RNN 结构，例如长短时记忆网络（LSTM）和门控循环单元（GRU），这些变体通过引入门机制来控制信息的流动，有效缓解了原始 RNN 的局限性。

在心血管疾病预测领域，RNN 可用于分析患者的历史健康记录，包括血压、胆固醇水平、心电图 (ECG) 数据等随时间变化的指标。通过对这些时间序列数据的建模和分析，RNN 有助于揭示潜在的风险因素和疾病发展趋势，为临床决策提供辅助信息。循环神经网络（RNN）是深度学习中处理序列数据的重要工具，其最大的优势在于能够捕捉时间序列中各个时刻之间的依赖关系。这种能力源自于 RNN 内部的记忆机制，它允许信息在序列的不同时间步之间传递，使得网络可以展现出对历史信息的记忆。这一特性使 RNN 非常适合于自然语言处理、语音识别、股票走势预测等领域。然而，RNN 也存在一些缺点 [33]。最显著的是梯度消失和梯度爆炸问题，这些问题会影响网络学习长期依赖关系的能力，尤其是在处理较长的序列时。此外，由于其循环结构，RNN 可能在计算资源消耗上较高，特别是在需要回溯大量时间步的场景下。相比于其他类型的神经网络，RNN 的设计和调试也更为复杂，因为它们需要考虑时间维度和循环连接的稳定性。因此，在实际应用中，选择合适的网络架构和训练策略对于发挥 RNN 的优势至关重要。

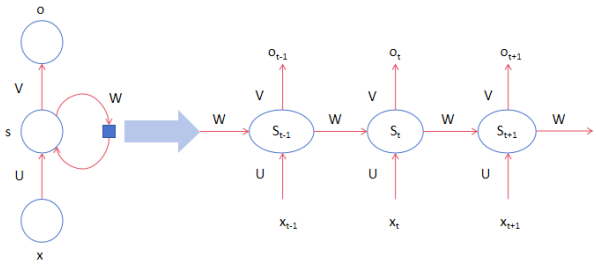

图 7. 循环神经网络结构示意图

4 结论

机器学习技术在医疗健康领域中的应用日益增多，通过分析和学习患者的临床数据、检测结果以及监测记录等信息，能够预测疾病的发生风险。这种预测能力对医生而言是一个重要的辅助工具，它能够帮助医生提前识别患者可能面临的健康问题，并采取预防措施或早期治疗策略，从而降低患病风险，减少医疗成本，对于疾病的防控和健康管理具有显著的意义。此外，利用机器学习算法分析大量的医疗数据，可以帮助研究人员发现一些传统方法难以识别的、但对疾病影响较大的关键指标和潜在特征。这些新发现的特征和指标为疾病的诊断和治疗提供了新的科学依据，有助于改善现有的诊疗方案，甚至可能引领医学领域的新突破。

然而，由于不同的机器学习算法有着各自的特点、优势和局限性，它们适用于预测和分析不同类型的数据和疾

病。例如，一些算法可能在处理结构化数据方面表现出色，而其他算法则更擅长分析非结构化数据。因此，在实际应用中，选择合适的机器学习算法至关重要。这通常需要结合疾病类型、可用数据的性质以及所需预测精度等因素进行综合考量。只有选择了与特定疾病和数据类型相匹配的算法，才能确保预测结果的准确性和可靠性，从而更好地支持医疗决策。

创新说明

这篇文章主要探讨了如何应用机器学习和深度学习技术于心血管疾病的预测，并详细分析了各种模型如支持向量机、随机森林、Logistic 回归、决策树等模型，还考虑了深度学习模型如卷积神经网络（CNN）与循环神经网络（RNN）的应用，展示了这些模型在处理复杂数据（如图像、语音）时的优越性能。

针对心血管疾病的高发病率和高死亡率问题，文章提出了采用先进的机器学习和深度学习技术进行预测和诊断，这为传统的心血管疾病预测方法提供了全新的解决方案。

文章深入探讨了各个模型在心血管疾病预测中的优势和挑战，以及如何通过特征选择、参数调整等方式提高模型的预测性能，这对于实际的疾病预测任务具有重要的指导意义。本文认为尽管存在一定挑战，但在合理选择模型和调整参数的前提下，机器学习模型仍然在心血管疾病的预测中具有巨大的潜力和应用价值。

参考文献

[1] Sacco RL, Roth GA, Reddy KS, et al. The heart of 25 by 25:Achieving the goal of reducing global and regional premature deaths from cardiovascular diseases and stroke: A modeling study from the American Heart Association and World Heart Federation. Circulation, 2016, 133(23): e674-e690.

[2] 中国心血管健康与疾病报告编写组.《中国心血管健康与疾病报告 2019》概要［J］. 中国循环杂志，2020，35（9）: 833-854.

[3] 周航宇，姚兴伟.1990～2019 年不良饮食危险因素对中国 ≥55 岁人群心血管疾病负担的影响 [J]. 中国循环杂志,2023,38(12):1279-1284.

[4] WarrenS McCulloch and Walter Pitts.A logicalcalculus of the ideas immanentin nervous activity. The bulletin of mathematical biophysics,1943,5(4):115-133.

[5] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386-408.

[6] Hopfield J J.Neural Networks and Physical Systems with Emergent Collective Computational Abilities, Proc Natl Acad Sci.USA,1982,(79):2254-2558.

[7] Geoffery E Hinton,Salakhutdinov RR.Reducing the dimensionality of data with neural networks.Science,2006,313(5786):504-7.

[8] DEO R C.Machine learning in medicine[J].Circulation,2015, 132(20):1920-1930.

[9] 薛亦诚, 刘超, 杨贵淞, 等. 基于 Logistic 回归和支持向量机的早发性结直肠癌风险预测模型 [J]. 中国现代普通外科进展,2024,27(03):195-198.

[10] 吴青, 付彦琳. 支持向量机特征选择方法综述 [J]. 西安邮电大学学报,2020,25(05):16-21.DOI:10.13682/j.issn.2095-6533.2020.05.003.

[11] SHI Z,CHEN G Z,MAO L,et al.Machine learning-based prediction of small intracranial aneurysm rupture status using CTA-derived hemodynamics:a multicenter study[J].AJNR Am J Neuroradiol,2021,42(4):648-654.

[12] Joloudari JH, Azizi F, Nematollahi MA, Alizadehsani R, Hassannataj E, Mosavi AH (2021) GSVMA: a genetic support vector machine ANOVA method for CAD diagnosis. Front Cardiovasc,8.https://doi.org/10.3389/fcvm.2021.760178.

[13] Aljarah I, Al-Zoubi AM, Faris H, Hassonah MA, Mirjalili SM, Saadeh H (2017) Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm. Cogn Comput 10:478–495.

[14] ZHANG C, MA Y Q. Ensemble machine learning: methods and applications[M]. New York: Springer, 2012.

[15] 石智强. 随机森林算法的改进及其在慢性病预警模型中的应用研究 [D]. 北京: 北京工业大学, 2018.

[16] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

[17] Huang GB, Zhou H, Ding X, Zhang R. Extreme learning machine for regression and multiclass classification.IEEE Trans Syst Man Cybern B Cybern. 2012 Apr;42(2):513-29.

[18] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986).

[19] 郑楠. 非瓣膜性心房颤动患者左心房血栓或自发显影预测评分及模型的构建与验证 [D]. 河北医科大学,2023.DOI:10.27111/d.cnki.ghyku.2023.000164.

[20] 郑楠, 刘冰, 闫洪伟, 等. 非瓣膜性心房颤动患者左心房血栓或自发显影的随机森林模型构建及抗凝结局 [J]. 岭南心血管病杂志,2024,30(01):22-27+56.

[21] Yang, L., Wu, H., Jin, X.et al.Study of cardiovascular disease prediction model based on random forest in eastern China.Sci Rep10, 5245 (2020).

[22] 石胜源, 朱磊, 叶琳, 等. 基于随机森林算法的心血管疾病预测研究 [J]. 智能计算机与应用,2021,11(04):176-178+181.

[23] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons.manufacturing technology. 2024, 130(3-4): 1405-1419.

[24] McCullagh, P. (1989). Generalized Linear Models (2nd ed.). Routledge.

[25] FENG X,YE G,CAO R,et al.Identification of predictors for hemorrhagic transformation in patients with acute ischemic stroke after endovascular therapy using the decision tree model[J].Clin Interv Aging,2020,15:1611-1624.

[26] Nishadi, A.S.T. (n.d.). International journal of advanced research and publications predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab.Montu Saw.

[27] saw(Ed.),2020 International Conference on Computer Communication and Informatics (ICCCI),IEEE,Coimbatore, India(2020), pp.1-6.

[28] 岳海涛, 何婵婵, 成羽攸, 等. 基于机器学习的冠心病风险预测模型构建与比较 [J/OL]. 中国全科医学:1-11[2024-05-12].http://kns.cnki.net/kcms/detail/13.1222.R.20240418.1001.010.html.

[29] 罗幼喜, 邓楠, 胡超竹, 等. 函数型累积 Logistic 回归模型研究与应用 [J]. 华中师范大学学报 (自然科学版),2023,57(02):185-194.DOI:10.19603/j.cnki.1000-1190.2023.02.001.

[30] HOSTETTLER I C,MUROI C,RICHTER J K,et al.Decision tree analysis in subarachnoid hemorrhage:prediction of outcome parameters during the course of aneurysmal subarachnoid hemorrhage using decision tree analysis[J].J Neurosurg,2018,129(6):1499-1510.

[31] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述 [J]. 计算机应用,2016,36(09):2508-2515+2565.

[32] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报,2017,40(06):1229-1251.

[33] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述 [J]. 计算机应用,2018,38(S2):1-6+26.

[34] Zhang X, Fang F, Liu J. Weather-classification-MARS-based photovoltaic power forecasting for energy imbalance market[J]. IEEE Transactions on Industrial Electronics, 2019, 66(11): 8692-8702.

[35] Liu J, Song D, Li Q, et al. Life cycle cost modelling and economic analysis of wind power: A state of art review[J]. Energy Conversion and Management, 2023, 277: 116628.

[36] Liu Y, Fang F, Park J H. Decentralized dissipative filtering for delayed nonlinear interconnected systems based on T–S fuzzy model[J]. IEEE Transactions on Fuzzy Systems, 2018, 27(4): 790-801.

[37] Lv Y, Fang F, Yang T, et al. An early fault detection method for induced draft fans based on MSET with informative memory matrix selection[J]. ISA transactions, 2020, 102: 325-334.

[38] Jin S, Wang S, Fang F. Game theoretical analysis on capacity configuration for microgrid based on multi-agent system[J]. International Journal of Electrical Power & Energy Systems, 2021, 125: 106485.

[39] Zhang J, Feng J, Zhou Y, et al. Linear active disturbance rejection control of waste heat recovery systems with organic Rankine cycles[J]. Energies, 2012, 5(12): 5111-5125.

[40] Fang F, Zhu Z, Jin S, et al. Two-layer game theoretic microgrid capacity optimization considering uncertainty of renewable energy[J]. IEEE Systems Journal, 2020, 15(3): 4260-4271.

[41] Lv Y, Lv X, Fang F, et al. Adaptive selective catalytic reduction model development using typical operating data in coal-fired power plants[J]. Energy, 2020, 192: 116589.

[42] Fang F, Jizhen L, Wen T. Nonlinear internal model control for the boiler-turbine coordinate systems of power unit[J]. PROCEEDINGS-CHINESE SOCIETY OF ELECTRICAL ENGINEERING, 2004, 24(4): 195-199.

[43] Wang N, Fang F, Feng M. Multi-objective optimal analysis of comfort and energy management for intelligent buildings[C]//The 26th Chinese control and decision conference (2014 CCDC). IEEE, 2014: 2783-2788.

[44] Liu J, Wang Q, Song Z, et al. Bottlenecks and countermeasures of high-penetration renewable energy development in China[J]. Engineering, 2021, 7(11): 1611-1622.

[45] Wei L, Fang F. H_∞-LQR-Based Coordinated Control for Large Coal-Fired Boiler–Turbine Generation Units[J]. IEEE Transactions on Industrial Electronics, 2016, 64(6): 5212-5221.

[46] Wang W, Liu J, Zeng D, et al. Modeling and flexible load control of combined heat and power units[J]. Applied Thermal Engineering, 2020, 166: 114624.

[47] Liu J, Zeng D, Tian L, et al. Control strategy for operating flexibility of coal-fired power plants in alternate electrical power systems[J]. Proceedings of the CSEE, 2015, 35(21): 5385-5394.

[48] Fang F, Xiong Y. Event-driven-based water level control for nuclear steam generators[J]. IEEE Transactions on Industrial electronics, 2014, 61(10): 5480-5489.

[49] Fang F, Wu X. A win–win mode: The complementary and coexistence of 5G networks and edge computing[J]. IEEE Internet of Things Journal, 2020, 8(6): 3983-4003.

[50] Fang F, Tan W, Liu J Z. Tuning of coordinated controllers for boiler-turbine units[J]. Acta Automatica Sinica, 2005, 31(2): 291-296.

[51] Lian S, Han Y, Chen X, et al. Dadu-p: A scalable accelerator for robot motion planning in a dynamic environment[C]//Proceedings of the 55th Annual Design Automation Conference. 2018: 1-6.

[52] Liu Q, Cheng L, Alves R, et al. Cluster-based flow control in hybrid software-defined wireless sensor networks[J]. Computer Networks, 2021, 187: 107788.

[53] Chang K, Wang Y, Ren H, et al. Chipgpt: How far are we from natural language hardware design[J]. arXiv preprint arXiv:2305.14019, 2023.

[54] Cheng L, Kotoulas S, Ward T E, et al. Robust and efficient large-large table outer joins on distributed infrastructures[C]//Euro-Par 2014 Parallel Processing: 20th International Conference, Porto, Portugal, August 25-29, 2014. Proceedings 20. Springer International Publishing, 2014: 258-269.

[55] Wang Y, Han Y, Zhang L, et al. ProPRAM: Exploiting the transparent logic resources in non-volatile memory for near data computing[C]//Proceedings of the 52nd Annual Design Automation Conference. 2015: 1-6.

[56] Han Y, Wang Y, Li H, et al. Data-aware DRAM refresh to squeeze the margin of retention time in hybrid memory cube[C]//2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2014: 295-300.

[57] Liu C, Pei Y, Cheng L, et al. Sampling business process event logs using graph-based ranking model[J]. Concurrency and Computation: Practice and Experience, 2021, 33(5): e5974.

[58] Cheng L, Kalapgar A, Jain A, et al. Cost-aware real-time job scheduling for hybrid cloud using deep reinforcement learning[J]. Neural Computing and Applications, 2022, 34(21): 18579-18593.

[59] Wang Y, Deng J, Fang Y, et al. Resilience-aware frequency tuning for neural-network-based approximate computing chips[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2017, 25(10): 2736-2748.

[60] Li W, Wang Y, Li H, et al. P3M: a PIM-based neural network model protection scheme for deep learning accelerator[C]//Proceedings of the 24th Asia and South Pacific Design Automation Conference. 2019: 633-638.

[61] Xu D, Zhu Z, Liu C, et al. Reliability evaluation and analysis of FPGA-based neural network acceleration system[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, 29(3): 472-484.

[62] Liang S, Liu C, Wang Y, et al. Deepburning-gl: an automated framework for generating graph neural network accelerators[C]//Proceedings of the 39th International Conference on Computer-Aided Design. 2020: 1-9.

[63] Guo J, Cheng L, Wang S. CoTV: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2023.

[64] Mao Y, Sharma V, Zheng W, et al. Elastic resource management for deep learning applications in a container cluster[J]. IEEE Transactions on Cloud Computing, 2022.

[65] Mao Y, Fu Y, Zheng W, et al. Speculative container scheduling for deep learning applications in a kubernetes cluster[J]. IEEE Systems Journal, 2021, 16(3): 3770-3781.

[66] Zheng W, Song Y, Guo Z, et al. Target-based resource allocation for deep learning applications in a multi-tenancy system[C]//2019 IEEE High Performance Extreme Computing Conference (HPEC). IEEE, 2019: 1-7.

[67] Liu Q, Cheng L, Ozcelebi T, et al. Deep reinforcement learning for IoT network dynamic clustering in edge computing[C]//2019 19th IEEE/ACM international symposium on cluster, Cloud and Grid Computing (CCGRID). IEEE, 2019: 600-603.

[68] Li J, Chen Z, Cheng L, et al. Energy data generation with wasserstein deep convolutional generative adversarial networks[J]. Energy, 2022, 257: 124694.

[69] Liu C, Chu C, Xu D, et al. HyCA: A hybrid computing architecture for fault-tolerant deep learning[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021, 41(10): 3400-3413.

[70] Xu D, Chu C, Wang Q, et al. A hybrid computing architecture for fault-tolerant deep learning accelerators[C]//2020 IEEE 38th International Conference on Computer Design (ICCD). IEEE, 2020: 478-485.

蒋子悠 现就读于北京工商大学计算机与人工智能学院软件工程工程专业。

Ziyou Jiang, Currently studying in the Software Engineering Engineering Program in the School of Computer Science and Artificial Intelligence at Beijing Technology and Business University (BTBU).