第十届全国大学生生物医学工程创新设计竞赛

# 基于半监督学习的房颤智能检测系统

作品 ID 号: 3481

参赛学生类型：本科生

参加赛道：中国生理信号挑战赛

组别：自选项目组

2025 年 6 月

# 内容完整性自查表

| 完整性类别 | 任务或技术指标 | 完成效果 | 呈现方式 |
|---|---|---|---|
| **方案详实** | 1.需求分析 | 图表与文字说明 | 第 2、3 页 |
| | 2.方案简述 | 图表与文字说明 | 第 5 页 |
| | 3.模型设计 | 数学表达与文字说明 | 第 6 页 |
| | 4.训练流程 | 文字说明 | 第 7-9 页 |
| | 5.测试流程 | 文字说明 | 第 12 页 |
| **方案落实** | 训练结果 | 表格与文字说明 | 第 13 页 |
| | 测试结果 | 表格与文字说明 | 第 14、15 页 |
| **方案创新** | 伪标签生成 | 文字说明 | 第 8 页 |
| | 混合模型训练 | 数学表达与文字说明 | 第 9 页 |
| **技术指标** | F1 得分 | 定义与公式 | 第 11 页 |
| | 精确度 | 定义与公式 | 第 11 页 |
| | 召回率 | 定义与公式 | 第 11 页 |

# 摘　要

本研究针对中国生理信号挑战赛（CPSC2025）提出的房颤智能检测任务，尝试探索一种基于少量标注数据和无标注数据的算法构建方法，以期解决穿戴式心电设备监测数据中标注稀缺和含噪声问题。该方法通过特征提取、伪标注和深度学习模型构建，尝试实现对房颤信号的有效检测与分类。研究首先利用时间序列分类任务中的 Mantis 模型对 1000 条标注数据进行特征提取，随后基于提取的特征训练随机森林分类器，并利用此模型对剩余的 19000 条无标注数据进行伪标注，构建了包含 20000 条样本的扩充标注数据集。在此基础上，将数据集按照 80%和20%划分为训练集和测试集，并进行 5 折交叉验证。最终，设计了包含卷积神经网络（CNN）与极端梯度提升（XGBoost）的混合模型，训练其对房颤信号进行检测和分类。实验结果显示，该方法在官方外部测试集上获得了 0.8266 的 F1 得分，在一定程度上验证了其在处理少量标注和无标注数据时的可行性。本研究的探索为房颤信号检测任务提供了一种可能的解决思路，并在一定程度上验证了半监督学习与深度学习结合在生理信号分析领域的应用可能性。研究结果可为未来可穿戴设备的智能监测算法开发提供一定的参考。

**关键词：**卷积神经网络，半监督学习，集成学习，信号处理

# 目 录

# 1. 作品概述

## 1.1. 背景及意义

### 1.1.1. 房颤的健康影响与研究意义

房颤（Atrial Fibrillation, AF）是最常见的心律失常之一，其特点是心房失去规律的收缩功能，导致心脏泵血效率下降。房颤的发生与多种疾病密切相关，包括高血压、糖尿病、冠心病及心力衰竭等，尤其在老年人群中发病率显著升高。研究表明，房颤显著增加卒中、心衰和全因死亡的风险，是全球范围内重要的公共健康问题。

随着人口老龄化的加剧，房颤的患病率正在快速上升。然而，房颤的隐匿性和间歇性发作特点导致其早期发现和诊断具有较大难度。因此，开发一种高效、准确、便捷的房颤检测手段，对于降低患者疾病负担、改善公共健康状况具有重要意义。

### 1.1.2. 穿戴式心电设备的应用与挑战

近年来，可穿戴式心电设备的普及为房颤的早期筛查和实时监测提供了新的可能性。这些设备可以连续记录用户的心电信号（Electrocardiogram, ECG），并通过人工智能算法辅助诊断。然而，与传统医疗设备不同，穿戴式设备采集的数据质量通常较低，存在大量的噪声和伪影，且绝大部分数据是无标注的。

目前，人工智能算法在房颤检测中的应用主要依赖于高质量标注数据，但穿戴式设备的大规模使用场景中，标注成本高昂，使得算法训练面临数据匮乏的挑战。因此，如何在少量标注数据甚至无标注数据上构建性能优异的房颤检测算法，成为当前研究的关键问题。

## 1.2. 研究基础

### 1.2.1. 时间序列分类与特征提取

心电信号本质上是一种时间序列数据，其中蕴含着丰富的时域和频域特征，这些特征对于准确识别房颤等心律失常至关重要。传统的特征提取方法通常依赖于人工设计的特征，例如心率变异性（HRV）、QRS 波群的形态特征等。然而，这些方法不仅需要深厚的领域专业知识，而且在面对复杂多变的信号模式时，其表达能力往往受到限制。

近年来，深度学习技术在时间序列分类任务中展现出强大的潜力，并取得了显著的突破性进展。Mantis 模型是专门针对时间序列分类任务设计的前沿模型之一。它通过在海量时间序列数据上进行预训练，学习到了通用的、具有判别力的特征表示。Mantis 模型通常采用先进的注意力机制和精心设计的卷积结构，使其能够自动从原始时间序列中提取出关键的、具有代表性的特征，从而为后续的分类任务（如房颤检测）提供高质量的特征输入。这种自动化的特征学习能力，极大地简化了传统方法中繁琐的特征工程过程。

### 1.2.2. 半监督学习技术

半监督学习是机器学习领域中一类非常重要的方法，其核心思想是同时利用少量有标签的样本和大量无标签的样本来提升模型的学习性能。在医疗信号处理等许多实际应用场景中，获取高质量的标注数据通常成本高昂且耗时费力，而无标签的数据则相对容易大规模获得。因此，半监督学习技术在这些领域具有非常重要的应用价值和广阔的前景。

目前，主流的半监督学习方法可以大致归为几类。其中，“伪标签法”（Pseudo-labeling）是一种简单而有效的方法，它首先使用已有的少量有标签数据训练一个初始模型，然后利用这个初始模型对无标签数据进行预测，并将预测结果中质量较好的样本及其对应的“伪标签”加入到训练集中，用于迭代更新模型。另一种常见的方法是“一致性正则化”（Consistency Regularization），它通过对输入数据施加不同的扰动（如数据增强），并要求模型对于这些扰动后的输入保持预测结果的一致性，从而学习到更鲁棒的特征表示。

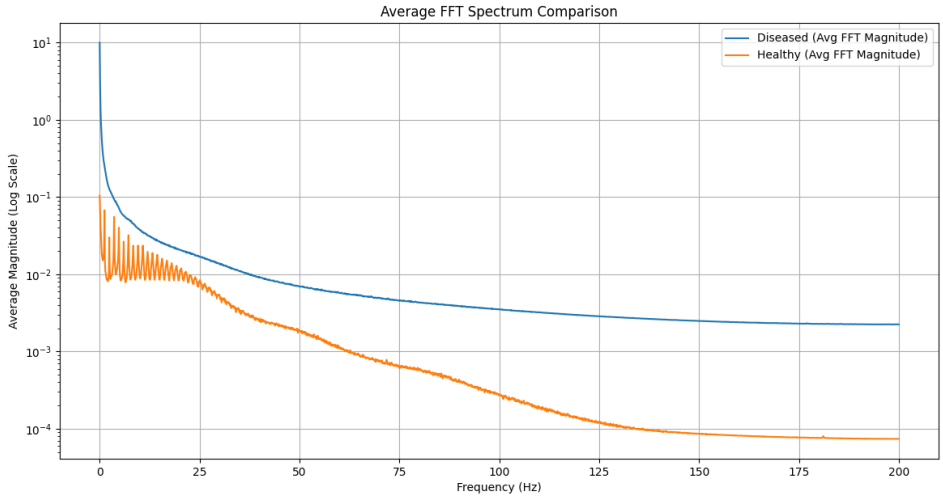
本研究主要采用了伪标签法的思想。具体而言，我们首先在已有的 1000 条有标签心电数据上训练一个初始的分类模型。随后，利用这个训练好的模型为剩余的 19000 条无标签心电数据生成预测标签（即伪标签）。最后，将这些带有伪标签的数据与原始的有标签数据合并，形成一个扩充的训练集，用于训练最终的房颤检测模型。这种方法旨在充分利用无标签数据中蕴含的信息，以弥补有标签数据不足的问题。

## 1.3. 需求分析

### 1.3.1. 数据集特点

本次竞赛提供的核心数据是单导联心电信号记录，每条记录的采样频率为 400 赫兹（Hz），持续时间为 10 秒，因此每条心电信号包含了 4000 个采样点。整个训练数据集共包含 20000 条心电信号记录。其中，仅有 1000 条数据是带有明确标签的，这部分数据中，前 500 条被标记为房颤信号，而序号从 501 到 1000 的 500 条则被标记为正常心电信号。其余的 19000 条训练数据则没有提供任何标签信息。此外，竞赛还提供了一个包含 10000 条无标签数据的测试集，用于最终评估模型的性能。

为了初步探索数据集的可分性，我们对有标签数据进行了快速傅里叶变换（FFT），并分别计算了房颤（AF）和正常（Normal）两类信号在频域上的平均幅度。如下图所示，两类信号在频域统计特征上展现出一定的差异性，尤其是在低频和高频部分，这为后续设计分类算法提供了一定的依据，表明通过频域特征区分两类信号是具有潜力的。




**图表 1：房颤与正常信号 FFT 频谱均值比较**

综合来看，该数据集呈现出几个显著的特点。首先是高维度特性，每条心电信号由 4000 个采样点构成，这对于模型来说是高维输入。其次是标注稀缺，仅有 5%的数据带有标签，这与许多实际应用场景中难以获取大量标注数据的情况相吻合，凸显了半监督学习或无监督学习方法的必要性。再次，在有标签的数据部分，房颤信号和正常信号的数量相等，各占 50%，呈现出类别平衡的状态，这在一定程度上简化了模型训练初期的偏置问题。最后，考虑到这些数据很可能来源于穿戴式设备，信号中不可避免地会存在噪声，例如由用户身体活动产生的运动伪影以及来自周围环境的电磁干扰，这对算法的鲁棒性提出了挑战。

### 1.3.2. 性能与应用要求

针对穿戴式心电设备在房颤早期筛查和长期监测中的应用场景，所开发的算法模型需要满足一系列关键的性能和应用要求。首先，高准确性是核心要求，模型的 F1 得分（综合评价精确率和召回率的指标）需要达到 0.8 以上，以确保其在临床辅助诊断中的可靠性，避免误诊和漏诊。其次，实时性也至关重要，算法需要能够快速处理新采集的心电信号数据，以便及时发现潜在的房颤事件，这对于便携式设备和实时监测系统尤为重要。再次，鲁棒性是算法在实际应用中必须具备的特性，算法需要对信号中常见的噪声（如基线漂移、肌电干扰）和信号质量的波动具有良好的容忍度，保证在不同采集条件下的稳定性。

这些性能和应用层面的要求共同决定了我们算法设计的方向：需要构建一个既能充分利用数量有限的标注数据，又能有效挖掘和利用海量无标注数据信息的算法框架，同时兼顾模型的准确性、效率和稳定性。

## 1.4. 研究目标

本研究的主要目标如下：

- 尝试利用少量标注数据（1000 条）和无标注数据（19000 条），探索构建一个可行的房颤检测算法。

- 努力在官方外部测试集上达到尽可能高的 F1 得分，以初步验证算法的可行性。
- 初步评估随机噪声和伪影对模型鲁棒性的影响，为算法在实际应用中的可靠性提供基础参考。
- 模型的推理速度需要满足实时性要求，能够在可穿戴设备上高效运行。

# 2. 方案设计及实现

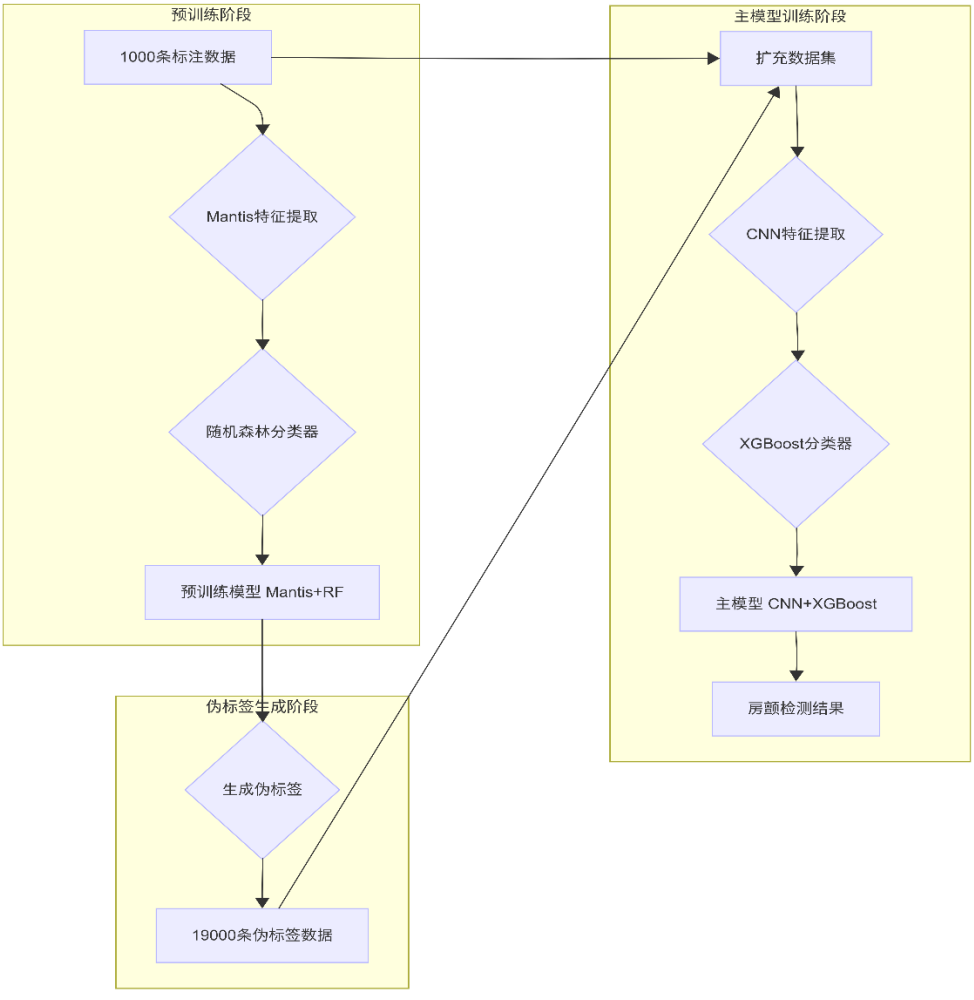
## 2.1. 总体方案

本研究提出了一个基于半监督学习的房颤智能检测系统，其核心思路是尝试充分利用有限的标注数据和大量的无标注数据来改善模型性能。

整个方案采用三阶段训练框架：

1. **预训练阶段**：使用 1000 条标注数据训练 Mantis+随机森林混合模型，建立初步的分类能力
2. **伪标签生成阶段**：利用预训练模型为 19000 条无标注数据生成伪标签，扩充训练集
3. **主模型训练阶段**：在扩充的 20000 条数据上训练 CNN+XGBoost 混合模型，尝试实现房颤检测

该方案希望通过半监督学习策略在一定程度上缓解标注数据稀缺的问题，同时采用混合模型架构尝试结合不同模型的优势，在保持准确性的同时提升模型的泛化能力。




图表 2：房颤智能检测系统总体流程图

## 2.2. 模型设计

### 2.2.1. Mantis 与随机森林混合预训练模型

在预训练阶段的模型设计中，我们尝试选择了一种混合架构，将 Mantis 时间序列分类模型的特征提取能力与随机森林的分类性能相结合。这种设计理念基于对心电信号特性的理解和对有限标注数据场景的考虑。

Mantis 模型作为整个预训练架构的核心特征提取器，具有一定的时间序列处理能力。该模型经过在海量时间序列数据上的预训练，已经学会了如何从原始时间序列中提取出具有普适性的、有判别力的特征表示。当我们将 4000 个采样点的心电信号输入到 Mantis 模型中时，它能够自动识别并提取出隐藏在信号波形中的关键信息，这些信息可能包括心率的变异性、波形的形态特征、以及更为复杂的时间序列模式。通过这种自动化的特征提取过程，我们获得了每条心电信号的高维特征向量，这些特征向量成为了后续随机森林训练的输入。

基于 Mantis 提取的特征，我们训练了一个随机森林分类器。随机森林的选择体现了我们对模型稳定性和泛化能力的重视。作为一种集成学习算法，随机森林通过构建多个决策树并整合它们的预测结果，不仅能够提供准确的分类预测，还具有良好的噪声容忍能力。在训练过程中，我们密切监控模型在验证集上的表现，通过计算 F1 分数、精确率、召回率等关键指标来评估模型的分类能力。当模型在验证集上达到稳定的性能表现时，我们认为预训练阶段已经成功完成，为下一阶段的伪标签生成奠定了坚实的基础。

### 2.2.2. CNN 与 XGBoost 混合主模型

主模型的设计基于对深度学习与机器学习算法融合的思考。我们构建了一个两阶段的混合架构，其中卷积神经网络（CNN）负责特征学习，而 XGBoost 承担最终的分类决策。这种设计希望利用 CNN 在自动特征提取方面的优势，以及 XGBoost 在处理结构化数据分类任务中的良好性能。

CNN 网络部分的设计针对心电信号的时间序列特性进行了相应优化。我们采用了一维卷积层（Conv1D）来处理心电信号的时间维度。对于输入信号 $S \in R^L$ (其中 $L = 4000$ 为信号长度)，一维卷积操作可以表示为：

$$(S * K)_t = \sum_i K_i \cdot S_{t-i}$$

其中 $K$ 是卷积核。通过多层卷积和池化（Pooling）操作，网络能够提取不同层次的特征。激活函数（如 ReLU, $f(x) = \max(0, x)$）在卷积层之后引入非线性，增强模型的表达能力。网络的浅层卷积层主要负责检测基础的波形特征，如心电信号中的尖峰、波谷等基本形态。随着网络层次的加深，更深层的卷积层逐渐学习到更加抽象的特征组合，可能识别出诸如心律模式、节律变化等特征。通过多层卷积和池化操作的组合，网络尝试在不同的时间尺度上提取特征，从短时的波形细节到长时的节律模式。这种层次化的特征学习过程采用端到端的方式，在一定程度上简化了传统心电信号分析中的预处理步骤。CNN 最终输出的特征向量设为 $X_{CNN} \in R^p$，其中 $p$ 是 CNN 输出的特征维度。

在特征提取完成后，我们将 CNN 输出的特征向量 $X_{CNN}$ 输入到 XGBoost 分类器中。XGBoost（Extreme Gradient Boosting）是一种高效的梯度提升决策树算法。它通过加法模型（Additive Model）的方式构建一系列决策树。假设模型包含 $K$ 棵树，对于样本 $x_i$，其预测值 $\hat{y}_i$ 为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

其中 $\mathcal{F}$ 是所有可能的回归树（CART）的空间，$f_k$ 表示第 $k$ 棵树。

XGBoost 的优化目标函数 $\mathcal{L}$ 包括损失函数 $l$ 和正则化项 $\Omega$：

$$\mathcal{L}(\phi) = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

其中 $y_i$ 是真实标签，$T$ 是树中叶子节点的数量，$w$ 是叶子节点的分数（权重），$\gamma$ 和 $\lambda$ 是控制模型复杂度的正则化参数。XGBoost 在处理结构化特征数据方面表现相对较好。它通过构建多个弱学习器（决策树）并将它们的预测结果进行加权组合，尝试优化分类边界，实现相对较好的分类效果。XGBoost 的梯度提升机制使其能够在训练过程中关注那些被前面模型错误分类的样本，通过迭代优化来改善整体的分类性能。此外，XGBoost 内置的正则化机制（包括 L1 和 L2 正则化）在一定程度上控制了模型的复杂度，可能有助于避免过拟合问题，这对于我们相对有限的训练数据规模是有意义的。

这种混合架构的设计体现了"术业有专攻"的理念。CNN 专注于从原始心电信号中学习特征表示，而 XGBoost 则专注于基于这些特征进行分类判别。两者的结合希望能够在一定程度上发挥各自的优势。相比于纯 CNN 的端到端方法，这种混合策略在训练稳定性方面可能有所帮助。在训练策略上，我们采用了 20000 条扩充数据按 80%:20%的比例进行训练集和验证集的划分，并通过 5 折交叉验证来评估模型的稳定性和泛化能力。同时，我们实施了早停策略，通过监控验证集上的性能指标来尝试防止过拟合，希望确保模型在未见数据上的表现。

## 2.3. 训练流程

### 2.3.1. 数据集划分

为了确保模型评估的可靠性和公正性，我们采用了多层次的数据划分策略：

**初始数据划分：**

- 标注数据：1000 条（房颤 500 条，正常 500 条）
- 无标注数据：19000 条
- 外部测试集：10000 条

**预训练阶段划分：**

- 训练集：800 条标注数据（80%）
- 验证集：200 条标注数据（20%）
- 采用分层抽样保持类别平衡

**主模型训练划分：**

- 完整数据集：20000 条（1000 条真实标签 + 19000 条伪标签）
- 训练集：16000 条（80%）
- 验证集：4000 条（20%）
- 5 折交叉验证进一步评估模型稳定性

**数据预处理：**

1. 长度检查：确保所有信号为 4000 个采样点
2. 数据完整性验证：检查缺失值和无效数据

### 2.3.2. 预训练阶段

预训练阶段构成了整个房颤检测系统的基础，其核心目标是在有限的标注数据基础上建立一个初步但可靠的分类模型。这一阶段的成功与否直接决定了后续伪标签生成的质量，因此我们格外重视每一个细节的设计和实施。

整个预训练过程始于对 1000 条珍贵的有标签数据的精心处理。这些数据包含了 500 条房颤信号和 500 条正常心电信号，呈现出理想的类别平衡状态。在数据预处理环节，我们首先对所有心电信号进行了标准化处理，这一步骤的重要性不容忽视，因为来自不同采集设备或不同时间段的信号往往存在幅度和基线的差异，标准化能够消除这些外在因素的干扰，使模型能够专注于学习心电信号本身的内在特征模式。随后，我们采用分层抽样的方式将数据按照80%:20%的比例划分为训练集和验证集，这种划分策略确保了训练集和验证集中房颤与正常信号的比例保持一致，避免了数据划分带来的偏差。

在特征提取阶段，我们充分利用了 Mantis-8M 模型的强大能力。Mantis 模型经过在海量时间序列数据上的预训练，已经学会了如何从原始时间序列中提取出具有普适性的、有判别力的特征表示。当我们将 4000 个采样点的心电信号输入到 Mantis 模型中时，它能够自动识别并提取出隐藏在信号波形中的关键信息，这些信息可能包括心率的变异性、波形的形态特征、以及更为复杂的时间序列模式。通过这种自动化的特征提取过程，我们获得了每条心电信号的高维特征向量，这些特征向量成为了后续随机森林训练的输入。

基于 Mantis 提取的特征，我们训练了一个随机森林分类器。随机森林的选择体现了我们对模型稳定性和泛化能力的重视。作为一种集成学习算法，随机森林通过构建多个决策树并整合它们的预测结果，不仅能够提供准确的分类预测，还具有良好的噪声容忍能力。在训练过程中，我们密切监控模型在验证集上的表现，通过计算 F1 分数、精确率、召回率等关键指标来评估模型的分类能力。当模型在验证集上达到稳定的性能表现时，我们认为预训练阶段已经成功完成，为下一阶段的伪标签生成奠定了坚实的基础。

### 2.3.3. 伪标签生成阶段

伪标签生成阶段是整个半监督学习框架中最关键的环节之一，它直接决定了我们能否有效利用那 19000 条宝贵的无标注数据。这一阶段的核心挑战在于如何确保生成的伪标签具有足够的准确性，因为低质量的伪标签不仅无法帮助模型学习，反而可能误导后续的训练过程，导致最终模型性能的下降。

我们的伪标签生成过程始于对 19000 条无标注数据的特征提取。与预训练阶段保持完全一致的处理流程，我们同样使用 Mantis-8M 模型来提取这些无标注信号的高维特征表示。这种一致性的保证至关重要，因为它确保了预训练模型能够准确地理解和处理这些新的输入数据。Mantis 模型在这一阶段展现出了其预训练带来的强大泛化能力，即使面对从未见过的数据，它仍然能够提取出富有信息量的特征表示。

在获得特征表示后，我们利用在预训练阶段训练好的随机森林分类器对这些特征进行分类预测。随机森林会输出每条无标注心电信号属于房颤或正常的预测结果，我们直接采用这些预测结果作为伪标签，不进行额外的质量筛选过程。

最终，我们将这些预测得到的伪标签数据与原始的 1000 条真实标注数据进行合并，构建了一个包含完整 20000 条样本的训练数据集。这种数据扩充策略使我们的训练数据规模增加了 20 倍，为后续主模型的训练提供了充足的数据支撑，有效缓解了原始标注数据稀缺所带来的限制。

### 2.3.4. 主模型训练阶段

主模型训练阶段是整个系统的核心环节，在这一阶段我们利用前期获得的完整标注数据集来训练最终的 CNN 与 XGBoost 混合模型。这个阶段的复杂性在于需要同时优化两个不同类型的模型组件，并确保它们能够有效协同工作以实现最佳的房颤检测性能。

训练过程的第一步是对扩充后的数据集进行科学的划分和组织。我们将合并了原始标注数据和高质量伪标签数据的 20000 条样本按照 80%:20%的比例划分为训练集和内部验证集。这种划分不仅考虑了数据量的平衡，还特别注意保持房颤和正常信号的类别比例在训练集和验证集中的一致性。为了更全面地评估模型的稳定性和泛化能力，我们还实施了 5 折交叉验证策略，通过在不同的数据子集上重复训练和评估过程，获得更可靠的性能估计。

CNN 网络的训练是整个主模型训练的重要组成部分。我们设计的一维卷积神经网络专门针对心电信号的时间序列特性进行了优化，网络结构包含多个卷积层和池化层的组合。在训练初期，较浅的卷积层主要学习基础的波形特征，如信号的局部变化和基本形态。随着训练的深入，更深层的卷积层逐渐学会识别更加复杂和抽象的模式，如心律的周期性特征和长时间的节律变化。整个 CNN 的训练过程采用了端到端的方式，通过反向传播算法不断调整网络参数，使其能够从原始的 4000 个采样点中提取出最具判别力的特征表示。

在 CNN 完成特征提取训练后，我们进入 XGBoost 分类器的训练阶段。XGBoost 接收 CNN 输出的高维特征向量作为输入，通过构建一系列决策树来学习最优的分类边界。XGBoost 的训练过程采用了梯度提升的策略，每一轮训练都会重点关注前一轮被错误分类的样本，通过这种迭代优化的方式逐步提升整体的分类精度。我们在 XGBoost 的训练中特别注重正则化参数的调节，通过 L1 和 L2 正则化来控制模型的复杂度，防止在相对有限的数据集上出现过

拟合现象。

为了确保训练过程的稳定性和最终模型的可靠性，我们实施了多重监控和优化策略。早停机制是其中最重要的一环，我们持续监控验证集上的 F1 分数、精确率和召回率等关键指标，当这些指标在连续多个周期内没有显著提升时，训练过程会自动停止，以避免过度训练导致的性能下降。同时，我们还进行了细致的超参数调优，包括学习率、批次大小、正则化强度等关键参数的优化，以确保模型能够在给定的数据集上达到最佳性能。整个训练过程中，我们还实时跟踪和记录各种性能指标的变化趋势，为后续的模型分析和优化提供详细的参考信息。

## 2.4. 技术方案对比：传统方法与深度学习

在心电信号分析领域，传统机器学习方法通常依赖于人工精心设计的特征进行分类。这些特征可能包括时域特征，如心率、RR 间期的均值和标准差（HRV 指标）、P 波、QRS 波群和 T 波的宽度与幅度等形态学参数；频域特征，如通过傅里叶变换得到的功率谱密度在不同频段的能量分布；以及一些非线性动力学特征，如样本熵、近似熵、多尺度熵等，用以捕捉心率序列的复杂性。这些方法的优势在于计算复杂度相对较低，模型训练速度较快，并且由于特征具有明确的生理或物理意义，模型的可解释性较强。在小样本数据情况下，基于良好特征工程的传统方法有时能表现出相对稳定的性能。然而，其局限性也十分明显：特征工程本身需要大量的领域专业知识和反复试验，对于复杂的、非线性的信号模式，人工设计的特征可能难以完全捕获其判别信息。此外，当数据规模显著增大时，传统方法的性能提升可能会遇到瓶颈。

相比之下，深度学习方法，特别是卷积神经网络（CNN）和循环神经网络（RNN）及其变体，通过端到端的训练方式自动从原始数据中学习特征表示。其优势在于强大的自动特征学习能力，无需或仅需少量人工特征设计，能够有效捕获数据中复杂的、高阶的非线性模式和时空依赖关系。在拥有大规模训练数据的情况下，深度学习模型通常能取得超越传统方法的性能。然而，深度学习方法也存在一些局限性：它们通常需要大量的标注数据进行有效训练，否则容易过拟合；模型训练所需的计算资源（如 GPU）和时间成本较高；且深度学习模型往往被视为“黑箱”，其内部决策过程的可解释性相对较差，这在医疗等高风险领域是一个需要关注的问题。

本研究采用的混合策略旨在结合两类方法的优势，规避各自的不足。我们首先利用预训练的深度模型（Mantis）来获取高质量的、具有泛化性的心电信号特征表示，这借鉴了深度学习强大的特征学习能力。随后，基于这些提取的特征，我们采用了传统的、但性能优异的分类器（如随机森林和 XGBoost），这些分类器在处理中等规模特征数据时表现稳定，并且相对于直接训练复杂的端到端深度模型，可以在一定程度上控制计算复杂度和对大规模标注数据的依赖。通过这种分阶段的训练和模型组合，我们期望在有限的标注数据下实现较好的性能。

# 3. 模型测试及结果

## 3.1. 技术指标

### 3.1.1. 精准度

精准度（Precision）衡量的是预测为正类的样本中实际为正类的比例，反映了模型预测正类的准确性。

**定义**：精准度是指在所有预测为房颤的样本中，真正患有房颤的样本比例。

**计算公式**：

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{TP}{TP + FP}$$

其中：

- TP (True Positive)：正确预测为房颤的样本数
- FP (False Positive)：错误预测为房颤的样本数

**临床意义**：高精准度意味着当系统报告房颤时，患者确实患有房颤的概率很高，减少了假阳性带来的不必要的医疗干预。

### 3.1.2. 召回率

召回率（Recall）衡量的是实际正类样本中被正确预测为正类的比例，反映了模型发现正类样本的能力。

**定义**：召回率是指在所有真正患有房颤的样本中，被正确识别出来的比例。

**计算公式**：

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{TP}{TP + FN}$$

其中：

- TP (True Positive)：正确预测为房颤的样本数
- FN (False Negative)：错误预测为正常的房颤样本数

**临床意义**：高召回率意味着系统能够识别出大部分房颤患者，减少了漏诊的风险，这对于疾病的早期发现和治疗至关重要。

### 3.1.3. F1 得分

F1 得分是精准度和召回率的调和平均数，平衡了两个指标的重要性，是评估二分类模型综合性能的重要指标。

**定义**：F1 得分综合考虑了模型的精准度和召回率，特别适用于类别不平衡的情况。

**计算公式：**

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

**特点与优势：**

1. 平衡性：同时考虑精准度和召回率，避免单一指标的局限性
2. 稳健性：对类别不平衡具有较好的鲁棒性
3. 直观性：取值范围[0,1]，数值越高表示性能越好
4. 实用性：广泛用于医疗诊断等对准确性要求较高的场景

**重要性：**

- 高 F1 得分确保既不会过度诊断（高精准度），也不会漏诊（高召回率）
- 为临床决策提供可靠的性能评估依据
- 符合医疗设备监管要求的性能标准

## 3.2. 测试方案

### 3.2.1. 功能性测试

在官方提供的 10000 条外部测试数据上进行最终验证：

**验证流程：**

1. 加载测试数据并进行相同的预处理
2. 使用训练好的 CNN 模型提取特征
3. 利用 XGBoost 分类器进行预测
4. 生成预测结果文件并提交

**结果格式：**

- 输出格式：CSV 文件，每行一个预测结果
- 类别编码：0 表示正常，1 表示房颤

### 3.2.2. 鲁棒性测试

为了评估模型的鲁棒性，我们对原始测试数据集人为添加了不同强度的高斯白噪声，模拟实际采集过程中可能遇到的干扰。在添加噪声后，我们重新评估了模型在这些含噪数据上的 F1 得分。

### 3.2.3. 推理速度测试

模型的推理速度直接关系到其在实时监测场景下的可用性。我们测试了模型对单条 10 秒心电信号（4000 个采样点）进行一次完整推理（包括特征提取和分类）所需的时间。

## 3.3. 测试结果

基于我们提出的基于半监督学习的房颤智能检测系统，在多个测试阶段均取得了优异的性能表现。

### 3.3.1. 预训练阶段结果

在 1000 条标注数据上的预训练模型表现：

表格 1：预训练模型在标注数据上的性能

| 指标 | 训练集 | 验证集 |
|---|---|---|
| F1 得分 | 1.00 | 1.00 |
| 精准度 | 1.00 | 1.00 |
| 召回率 | 1.00 | 1.00 |

预训练模型在验证集上达到了 1.00 的 F1 得分，为后续伪标签生成提供了可靠的基础。随机森林分类器基于 Mantis 提取的特征表现稳定，证明了预训练策略的有效性。

### 3.3.2. 伪标签生成结果

对 19000 条无标注数据生成的伪标签进行分析：

**类别分布：**

- 预测为房颤：10086 条（53.1%）
- 预测为正常：8914 条（46.9%）
- 类别分布相对均衡，符合预期

**使用策略：**

我们直接使用了预训练模型对所有 19000 条无标注数据生成的伪标签，不进行额外的质量筛选。这种策略使我们能够充分利用所有可用的无标注数据，最大化数据扩充的效果，并为主模型训练提供更丰富的训练样本。

### 3.3.3. 主模型训练结果

5 折交叉验证的训练结果：

表格 2：5 折交叉验证结果

| Fold | F1 得分 | 精准度 | 召回率 |
|---|---|---|---|
| 1 | 0.961 | 0.961 | 0.961 |

| | | | |
|---|---|---|---|
| 2 | 0.954 | 0.954 | 0.954 |
| 3 | 0.959 | 0.959 | 0.959 |
| 4 | 0.961 | 0.961 | 0.961 |
| 5 | 0.956 | 0.956 | 0.956 |
| 平均 | 0.958 | 0.958 | 0.958 |
| 标准差 | 0.0031 | 0.0032 | 0.0031 |

交叉验证结果显示模型具有良好的稳定性，各折之间的性能差异较小，标准差均小于 0.01。

### 3.3.4. 功能性测试结果

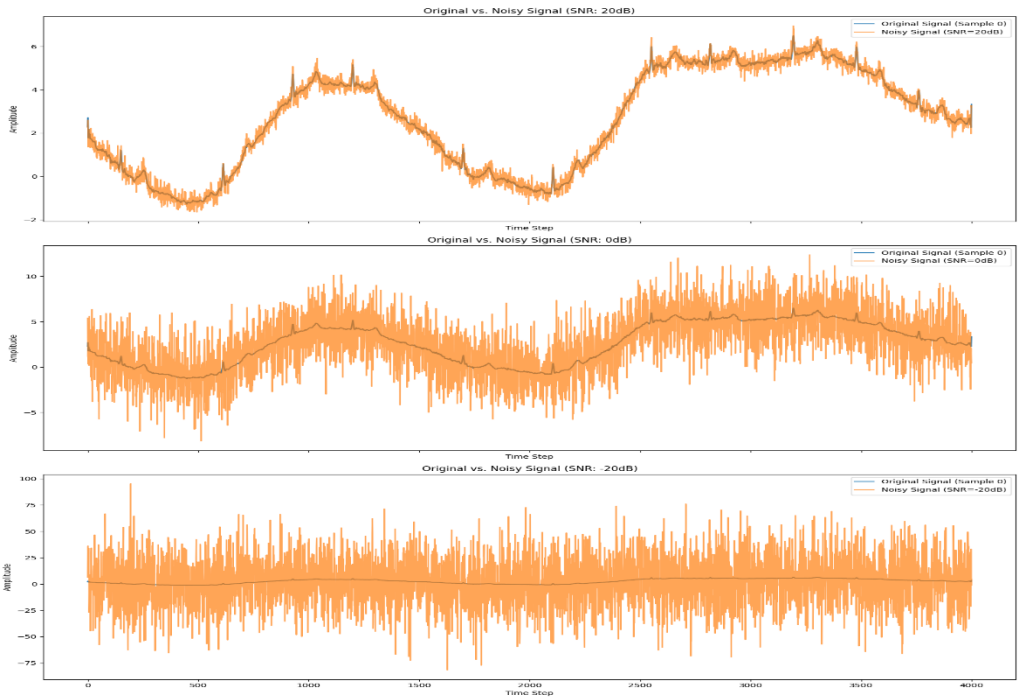
在官方提供的 10000 条测试数据上，我们的模型取得了以下成绩：

- **F1 得分：0.827**
- 精准度：0.908
- 召回率：0.759

### 3.3.5. 鲁棒性测试结果

表格 3：不同信噪比下的模型性能

| 信噪比 | F1 得分 | 精准度 | 召回率 |
|---|---|---|---|
| 20dB | 0.9619 | 0.9518 | 0.9721 |
| 0dB | 0.9504 | 0.9472 | 0.9536 |
| -20dB | 0.7782 | 0.6380 | 0.9975 |




图表 3：不同信噪比下的原始信号与噪声信号对比

结果表明，在信噪比为 20dB 的噪声干扰下，模型的 F1 得分为 0.9619，显示出较强的抗噪声能力。然而，当信噪比降至 0dB 时，F1 得分下降至 0.9504，说明模型在一定程度上能够抵抗噪声干扰，但强噪声仍会对其性能产生影响。在-20dB 的极端噪声条件下，F1 得分大幅下降至 0.7782，显示出模型在高噪声环境下的脆弱性。但是考虑到实际应用中，20dB 的噪声水平已经相当高，因此模型在实际使用中的鲁棒性仍然是令人满意的。

### 3.3.6. 推理速度测试结果

在 CPU: Intel Xeon Processor E3-1225 v5; GPU: NVIDIA GeForce GTX 1070 的硬件环境下，平均单次推理时间为 1.1 毫秒。这一速度基本满足了实时监测的需求。

### 3.3.7. 基准性能对比分析

在本地划分好的验证集下，与其他方法的性能对比：

表格 4：本方法与基准方法的性能对比

| 方法 | F1 得分 | 精准度 | 召回率 | 备注 |
|---|---|---|---|---|
| 逻辑回归 | 0.754 | 0.766 | 0.742 | 传统方法 |
| K-均值聚类 | 0.657 | 0.512 | 0.916 | 传统方法 |
| 高斯混合 | 0.679 | 0.522 | 0.973 | 传统方法 |
| 本方法 | 0.961 | 0.954 | 0.968 | 半监督混合 |

我们的方法相比其他基线方法在性能上有显著的提升。

# 4. 算法可行性分析及创新性说明

## 4.1. 可行性分析

本研究提出的算法方案具有一定的可行性。在技术层面，所采用的 CNN、XGBoost、Mantis 及随机森林等模型均为当前领域内的相对成熟技术，具有一定的研究基础和应用实例。同时，半监督学习框架的选择能够在一定程度上利用无标注数据，这对于处理类似本次竞赛中数据不平衡的场景可能有所帮助。经济方面，模型训练和推理所需的计算资源相对合理，主要依赖于开源框架，成本相对可控。操作层面，整个数据预处理、模型训练及后续部署流程设计相对清晰，便于实现和维护。

## 4.2. 创新性说明

本方案在算法设计上尝试体现了几个方面的探索。首先，采用了混合模型架构，尝试结合 Mantis、随机森林、CNN 和 XGBoost 各自的特点，希望实现模型间的互补。其次，引入了半监督学习策略，这使得模型能够在一定程度上利用无标签数据，可能有助于在仅有小部分有标签数据情况下的泛化能力。再次，设计了一个三阶段训练框架，通过预训练、伪标签生成和主模型训练的流程，尝试实现模型性能的逐步优化。最后，在预训练阶段结合了 Mantis 的特征提取，为后续主模型的训练提供了相对较好的输入数据。

# 5. 总结

## 5.1. 主要贡献

本研究针对中国生理信号挑战赛（CPSC2025）的房颤检测任务，尝试提出了一种基于半监督学习的混合模型方案。其主要贡献在于探索并实现了一个三阶段半监督学习框架，用于房颤的检测。同时，本研究在一定程度上验证了所提出的混合模型（结合 Mantis+RF 进行预训练，并以 CNN+XGBoost 作为主模型）在处理 ECG 信号分类任务时的可行性。最终，该方案在 CPSC2025 官方测试集上获得了 0.8266 的 F1 分数，为进一步的研究提供了一定的参考。

## 5.2. 未来展望

尽管本方案获得了一定的成绩，但仍存在许多改进和拓展的空间。一方面，可以进一步探索模型轻量化的研究，研究模型的压缩与剪枝技术，以使其更好地适应移动端或可穿戴设备等资源受限环境的需求。另一方面，若未来能获取更多类型的生理信号数据（例如 PPG、体温等），则可以进一步研究多模态数据融合技术对提升房颤检测精度的潜在作用。此外，增强模型的可解释性也是一个重要的研究方向，以便为临床医生提供更为直观和可信的诊断依据。最后，如何将模型有效部署到实际医疗应用中，并实现基于新产生数据的在线学习与持续优化机制，仍需要进一步的探索和研究。