**第十九届“挑战杯”全国大学生课外学术科技作品竞赛**

**“人工智能+”挑战赛**

**项目策划书**

作品名称：面向中医脉诊智能化的多源脉搏信号建模与脉象分

类模型设计

项目赛道：“人工智能+”挑战赛

项目领域： 中医专病临床智能辅助诊疗大模型

团队负责人：李天羽

团队成员：金耀仑、高岩、马文龙、梁承涵、刘培林、蔡逸

作品形式（可多选）：

□视频形式

☑图文形式

□其他



**一、背景与目标**

本课题面向“中医专病诊疗大模型”，强调对**中医典籍、诊疗指南、名老中医医案与电子病历**的深度结构化，形成覆盖“辨证论治—方剂推荐—疗效评估—健康管理”的智能决策引擎。在此整体框架中，我们聚焦“**脉象识别**”这一关键模态：通过将双通道脉搏波形与医理知识以**文本化表示**嵌入到中文大模型中，以参数高效微调（LoRA）实现对“特征—证候—标签“映射的精准对齐，使“AI+中医”的知识传承与创新落在可复现、可部署的工程路径上。

**二、脉象数据集**

本研究共纳入 **217 名受试者** 的**双通道指尖脉搏**数据，每例连续采集 **45 s**、采样率 **200 Hz**，并采用 **8 bit（0–28）** 的幅值量化以保证对细微波形变化的分辨能力。每个样本除原始脉搏波形外，同时记录**脉象类别标签**与**性别、身高、体重、年龄**等生理信息，作为后续建模的辅助变量。数据采集在**浙江中医药大学附属第二医院**开展，由**朱勤副院长**在专家门诊场景中**同步完成脉诊与标注**，以临床实践为依据给出脉象类别，确保标签的一致性与可信度。朱勤医生具有**40 年临床与科研经验**，为本数据集的专业性与可靠性提供了重要保障。

为便于复现与后续分析，数据集具有如下特征：**（i）双通道同步**采集，有利于表征跨指通道的互补信息；**（ii）高时域采样**（200 Hz，45 s），覆盖充足心动周期，利于稳健的单周期提取与时频分析；**（iii）带有人口学/体格信息的**标签化样本，便于开展多变量建模与亚组分析。实际评测与模型选择阶段，建议以**受试者为单位**进行分层分组划分，避免同一受试者数据在训练与测试间泄漏。

图1 数据集采集

**三、特征提取设计**

**3.1 预处理与单周期构建**

脉搏波在采集过程中易受环境与生理多源干扰，常见问题包括**基线漂移**与**宽频噪声**。为保证后续分类的可辨性与稳定性，我们构建了“**归一化—去噪—基线校正—周期分割—周期归一**”的一体化流水线：

首先对原始信号做幅值归一化，统一量纲以利于跨样本比较；随后采用**卡尔曼带通滤波**（自适应增益，将有效频段限制在 0.5–20 Hz）与**FIR 低通滤波**串联，前者抑制低频漂移与部分生理噪声，后者进一步平滑高频成分，从而最大化保留脉搏有效信息。为校正基线漂移，我们在波谷点上进行**三次样条（Cubic Spline）拟合**，以拟合曲线作为基线并从原始信号中扣除，实现形态对齐。随后通过**峰—谷配对**完成周期定位与分割，并在周期级别执行归一/反归一操作以消除节律差异。为提高特征稳健性，我们将同一样本的多个有效周期进行**相位对齐与平均**，得到代表性“单周期”用于特征构建。

**3.2 特征构建与定义**

在单周期层面，我们提取时域、频域与医理相关形态指标，共计约 **62 个**数值特征：

**时域**：主搏幅度、重搏波幅度、中峡（切迹）幅度，以及关键波峰/波谷间的时间—幅度差、持续时长、上升/下降斜率、分段面积等；

**频域**：1–5 Hz **五个子带**的谱能量，以表征心血管节律与血流动力成分；

**形态学指数 K**：反映血管弹性、外周阻力与波速等综合影响，定义为

其中 Pm​ 为单周期内的平均压力（以归一化脉搏波幅近似），Ps、Pd 分别对应该周期的收缩峰与舒张谷幅值；

**血流动力学指标—心指数 CI**：度量单位体表面积的心输出能力，

其中 CO 为心输出量（由脉搏波相关量化特征推估），BSA 由个体身高体重计算得到。

每个样本包含**食指与中指**两路脉搏数据，我们分别构建两指的特征矩阵，并融合**病人生理信息**（如年龄、身高体重等）作为辅助变量。特征级联后统一执行**标准化**，生成最终样本向量。考虑到高维特征的冗余与共线性，我们采用**极度随机树（Extra Trees）进行重要性筛选，去除冗余维度，再以PCA** 进行正交降维，累计方差贡献率控制在 **≥95%**，以提升模型的泛化与稳健性。

由于本研究需要对多种脉象进行特征提取，而其中一部分脉象的下降沿非常平滑，导致降中峡和次搏峰的位置提取困难。所以本研究对传统的通过极大极小值提取降中峡和次搏峰的方法进行了改进，采用了平均周期波形一阶导数和二阶导数相结合的方法，能更好把握曲线的运动趋势。由图可见，对下降沿近似一条直线的细脉，算法也可以准确定位到降中峡和次搏峰。

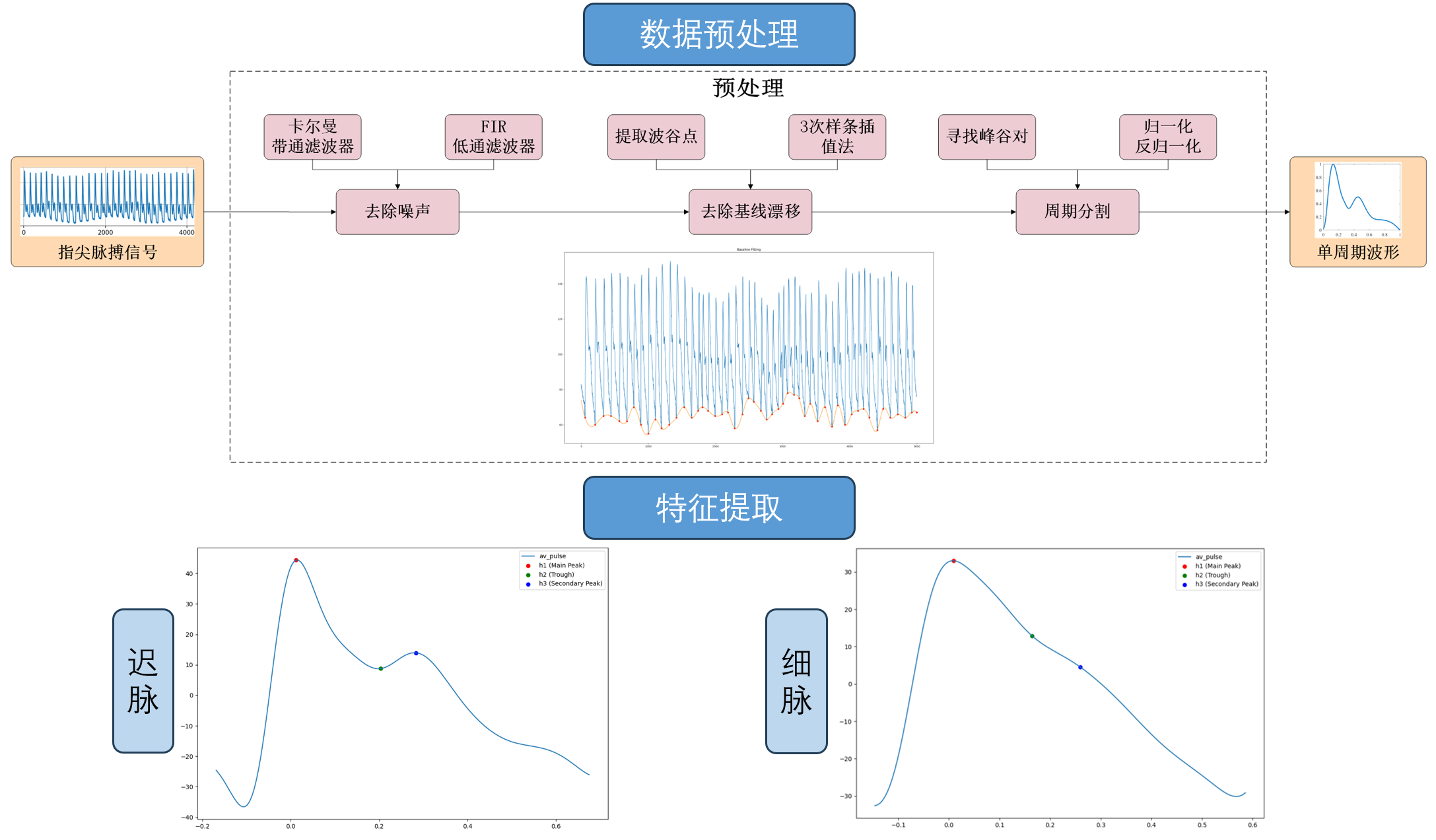


图2 脉搏数据预处理与特征提取

**四、模型训练设计方案**

**4.1 总体思路与技术路线**

我们将脉象分类视作一种**结构化指令学习**问题：对双通道脉搏信号的关键信息进行**文本化表示**，嵌入标准化的医学判别提示（prompt），再以**大语言模型（LLM）进行监督微调（SFT）**，将“特征→诊断标签”的映射显式对齐在一个受控的输出空间内。训练阶段仅更新极少量的低秩适配器**（LoRA）**参数，从而在保持基座通用能力的同时，实现对中医专病脉象知识的高效注入与固化。

核心创新在于：

* **大模型迁移 + 文本化融合**：以文本为中介，将生理信号的判别性要素与通用中文语义先验直接对接，显著缓解小样本条件下从零训练分类器的瓶颈。
* **LoRA 结构微调**：冻结绝大多数基座权重，只在关键子层注入低秩适配器，以\*\*<1% 的参数开销\*\*完成医域任务迁移，并兼顾速度、显存与可控性。
* **指令—答案对齐**：训练时将输出严格限定在预定义的脉象标签集合内，减少自由生成导致的漂移，提升可复现性与部署稳定性。

**4.2 基座模型与可替换性**

本方案以 **Qwen-7B-Chat** 作为中文语义骨干，依托其在中文理解、格式遵从与知识问答方面的稳健性承载医学要素的抽象与跨样本归纳。

**4.3 LoRA 结构与注入位置**

我们采用标准 PEFT 框架，将 LoRA 适配器注入 Transformer 的关键线性映射以实现任务特化，同时最大限度维持基座的通用性。具体而言，注意力分支的 **q/k/v/o 投影**承担了大部分对“提示—特征—标签”关系的对齐；在需要提升非线性表达时，可对前馈分支 **up/gate/down** 进行补充注入。为加速标签分布的收敛，输出层保留可更新（或作为保留模块保存）。容量选择采用小/中/高三档的快速验证策略，以宏平均 F1 与 Cohen’s κ 为主指标，避免欠拟合或过拟合。该设计本质上将“脉象知识”**局部雕刻进 LoRA 子空间**，既便于版本化管理，也便于后续的可控回退与增量扩展。

**4.4 指令模板与任务建模**

训练样本由“**双通道脉搏文本特征 → 标准化医学问句 → 有限标签集合**”构成，prompt 明确要求模型从既定集合中**仅输出一个标签**。为保证监督信号一致，训练前对脉象名称进行同义映射与标准化。受控的模板将输出空间封闭化，降低了生成式模型常见的越界与幻觉风险，并为评测与部署提供了清晰的边界条件。

**4.5 训练目标与对齐策略**

优化目标采用因果语言建模（Causal LM）损失。为使学习聚焦于诊断结论，训练时仅对**答案段**计损（对提示段施加 label-mask），从而避免模型将容量浪费在复述提示上，加速“特征—标签”的决策对齐。结合梯度检查点与混合精度，可在单卡与中等显存条件下稳定完成训练。

**4.6 流程与可复现性**

每条样本被序列化为“Prompt + Answer”的指令对，其中 Answer 为**单一脉象标签**。在 SFT 中仅更新 LoRA 适配器与必要的输出层，训练结束同时导出两类权重产物：一是**LoRA 适配器**（体积小、便于分发与热插拔），二是**合并后的单模型**（部署路径更简洁）。为控制数据泄漏与评价波动，验证与测试建议以**患者为分组**进行划分，报告 Accuracy、Macro-F1、κ、MCC 与混淆矩阵等指标，以体现医用场景下的稳健性与可解释性。

**4.7 推理与健壮性**

部署侧采用**约束解码**：在贪心策略（temperature=0）下生成，并以标签集合做后处理匹配，保证输出单一且合规。为进一步提升稳健性，可在同一提示下对所有候选标签计算**条件对数似然**并取最大者（verbalizer scoring），同时获得类概率用于 AUC 与校准分析；该评分范式与指令式训练天然一致，具备更强的可解释性与鲁棒性。

**4.8 方法优势与可扩展性**

* **参数高效与低门槛**：LoRA 使小样本与有限算力条件下的领域迁移成为可能。
* **可维护的模块化**：按病种/科室划分 LoRA 适配器，支持多任务共存与快速热插拔。
* **面向升级的兼容性**：可在不改动数据与模板的前提下，从 Qwen-7B-Chat 平滑升级至 Qwen2.5-7B 或其他中文基座。
* **合规与可审计**：封闭标签集与严格日志记录便于比赛评审与临床落地审查。