从轻量级到大模型：构建鲁棒且可解释的ECG早期心梗识别模型

王佳、张海超
西交利物浦大学
2025 年 11 月 12 日

王佳（西交利物浦大学助理教授 博士生导师）

- 北京交通大学（本科）
- 瑞典皇家理工（硕士导师：庞智博）
- 香港理工大学（博士导师：曹建农）
- 基于数据驱动强化学习的决策系统

项目与资助：

24–27，PI，用可控的机器学习研究，NSFC青年基金项目。
23–26，PI，基于反事实因果的合规推荐系统研究，江苏省高层次项目。
22–26，PI，基于复杂系统大数据的视觉缺陷检测方法与研究，AI Empowerment Tech，50万。
22–26，PI，基于大数据分析的健康照明：方法与应用，江苏天光白光电有限公司，50万。
22–25，PI，复杂系统多阶段大数据分析：方法与应用，研究发展基金，10万。

CONTENT

1. 研究背景：AI与医学交叉的临床需求

1.1 AMI诊断的“时间之困”

- 临床目标：**急性心肌梗死 (AMI) 是全球致死率最高的单病种之一**， 早期识别直接决定抢救成功率
- **核心原则：“时间就是心肌” (Time is Muscle)**
  - 从血管闭塞到心肌开始坏死，窗口期极短。
  - 临床决策的核心：**“快”** (速度) 与 **“准”** (精度)。
  - **目标：** 立即识别出需要紧急血运重建（Revascularization）的患者。
- **当前急诊胸痛分诊流程 (Triage Flow)**
  - 患者入院 –> 12导联ECG –> 抽血 (等待心肌标志物)。
  - 这个流程在识别 **STEMI**（ST段抬高型心梗）时非常高效。

1.2 NSTEMI诊断的“灰色地带”

真正的瓶颈：非ST段抬高型心梗 (NSTEMI) 患者

- **STEMI (路径清晰)**
  - **ECG表现：** 明确的ST段抬高。
  - **诊断：** 确诊闭塞。
  - **临床路径：** 立即送入导管室，进行血运重建
  - AI价值：**锦上添花**。
- **NSTEMI (灰色地带)**
  - **ECG表现：** **ST段不抬高**，表现不特异 (T波倒置, ST段压低等)。
  - **诊断：** 无法立即确认闭塞。
  - AI价值：**雪中送炭**。

1.3 AI的机遇：从“ECG判读”到“ECG虚拟冠脉造影”

临床的核心痛点：
“在ST段未抬高的患者中，隐藏着大量同样致命的急性冠脉闭塞（Occlusion），而我们无法仅凭ECG快速识别。”

- **人眼局限：** 依赖“ST段抬高”这一“强特征”，易忽略NSTEMI中微弱、高维的病理模式。
- **AI的潜力：**
  - ECG是高维时序信号，蕴含远超形态学的信息。
  - AI（特别是深度学习）擅长从12导联的复杂关联中，学习人眼无法感知的“微弱特征”（Subtle Patterns）。

1.4 我们的研究目标：构建“ECG虚拟冠脉造影”

利用2800例ECG及“金标准”造影数据，训练一个端到端模型，实现：

1. **阶段一: 分诊**
   - 是否需要做冠脉造影
2. **阶段二: ECG细粒度对齐到冠脉造影**
   - 闭塞定位 (Localization): 哪根血管（LAD/LCx/RCA）？
   - 狭窄定量 (Quantification): 狭窄率是多少（是否 > 70%）？

2. 数据不平衡：聚类增量学习

2.1 挑战：临床数据的“天然鸿沟”

- **现状:**
  - **多数类 (Majority):** 大量 NSTEMI–非闭塞 患者 (如狭窄<50%)。
  - **少数类 (Minority):** 少量我们真正要找的 NSTEMI–急性闭塞 患者 (如狭窄>70%)。

- **AI建模的“经典陷阱”:**
  - 模型会“躺平” (Get Lazy)：AI为了追求90%的“总准确率”，会**放弃学习**那10%的少数类。
  - **临床后果：** 产生一个“看似准确”但**对所有高危心梗患者全部漏诊**的“废物模型”。

2.2 常规解法：“人工合成”病例 (SMOTE)

- **方法：** 通过算法“复制”或“合成”更多的高危闭塞病例，强行让数据1:1平衡。
- **致命缺陷：**
  1. **模糊边界：** 如果制造的样本在“临界”病例周围，反而让模型更难区分结果。
  2. **引入噪声：** 合成的ECG波形可能不符合生理学逻辑，成为“脏数据”。

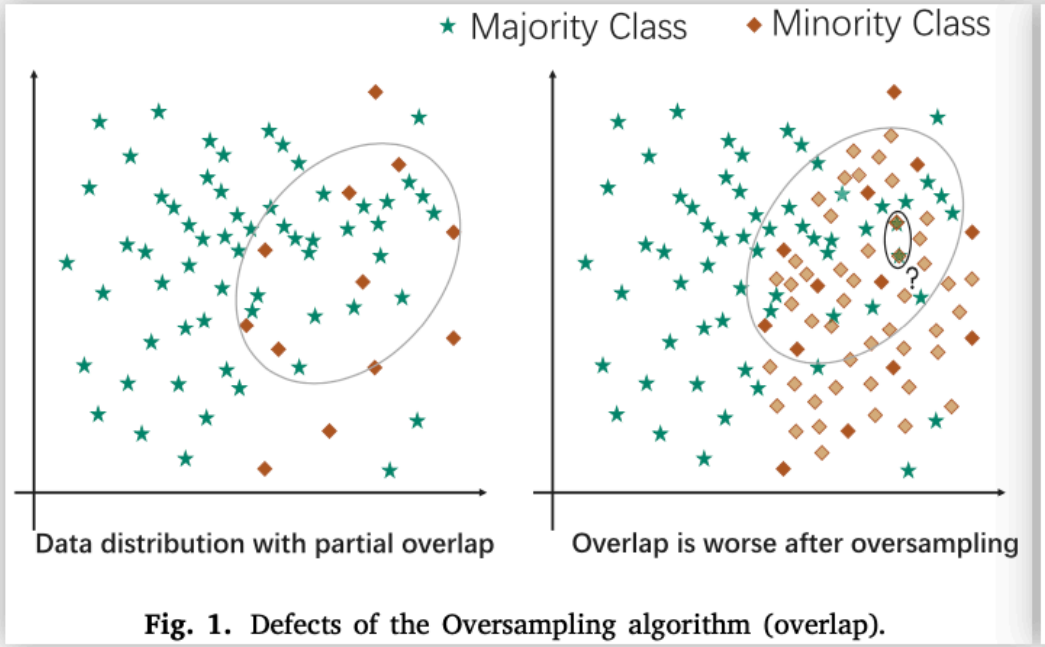


Fig. 1. Defects of the Oversampling algorithm (overlap).

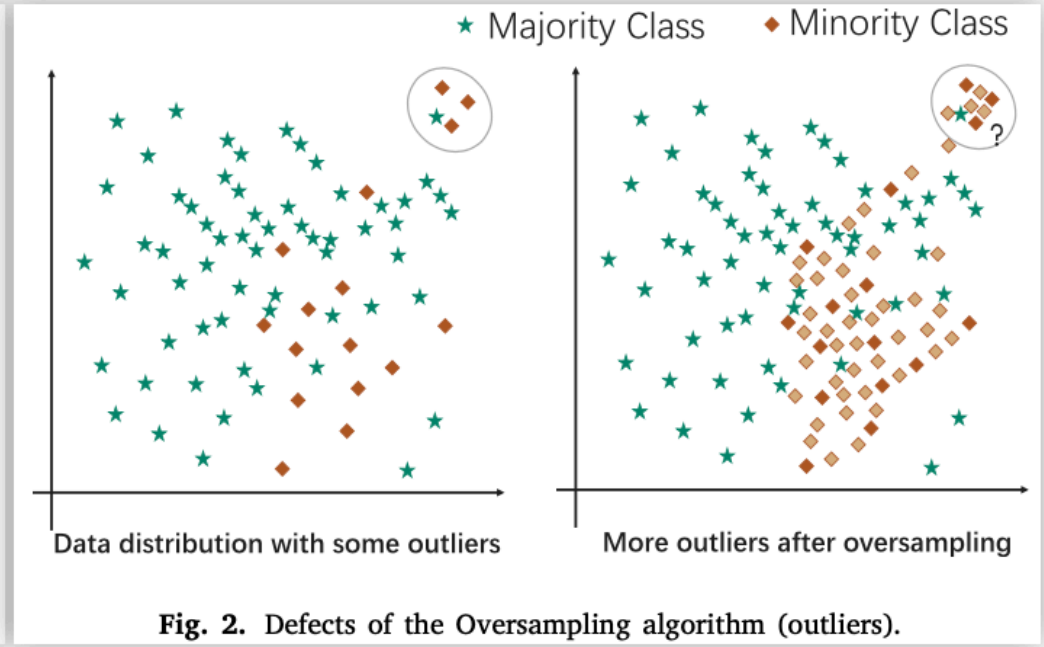


Fig. 2. Defects of the Oversampling algorithm (outliers).

参考论文: Clustering–based incremental learning for imbalanced data classification (Liu, Wang, et al.)

2.3 我们的对策：智能分群 + 递进式学习

我们不“人工合成”噪声，而是“智能拆解”数据（借鉴 DRIL 框架）

- **步骤一 (DR Step): “两步聚类” – 智能拆分多数类**
  - **做法：** 我们使用“两步聚类法”(Two-Step Cluster)，将海量的“非闭塞”患者（多数类）自动分为几个有代表性的“**亚群**”(Sub-groups)。
  - **比喻：** 不再是‘一整群低风险’，而是‘**亚群1**: T波轻微倒置’、‘**亚群2**: ST段轻微压低’、‘**亚群3**: 正常ECG’...
- **步骤二 (IL Step): “增量学习” – 逐个击破**
  - **做法：** 将“急性闭塞”病例（少数类）**依次**与每个“非闭塞亚群”组合，分阶段训练模型。
  - **临床获益：** 强迫模型在**所有类型的“干扰背景”下**，都能精确识别出“高危闭塞”的共同特征，显著提高模型的鲁棒性和泛化能力。

2.4 方法框架图

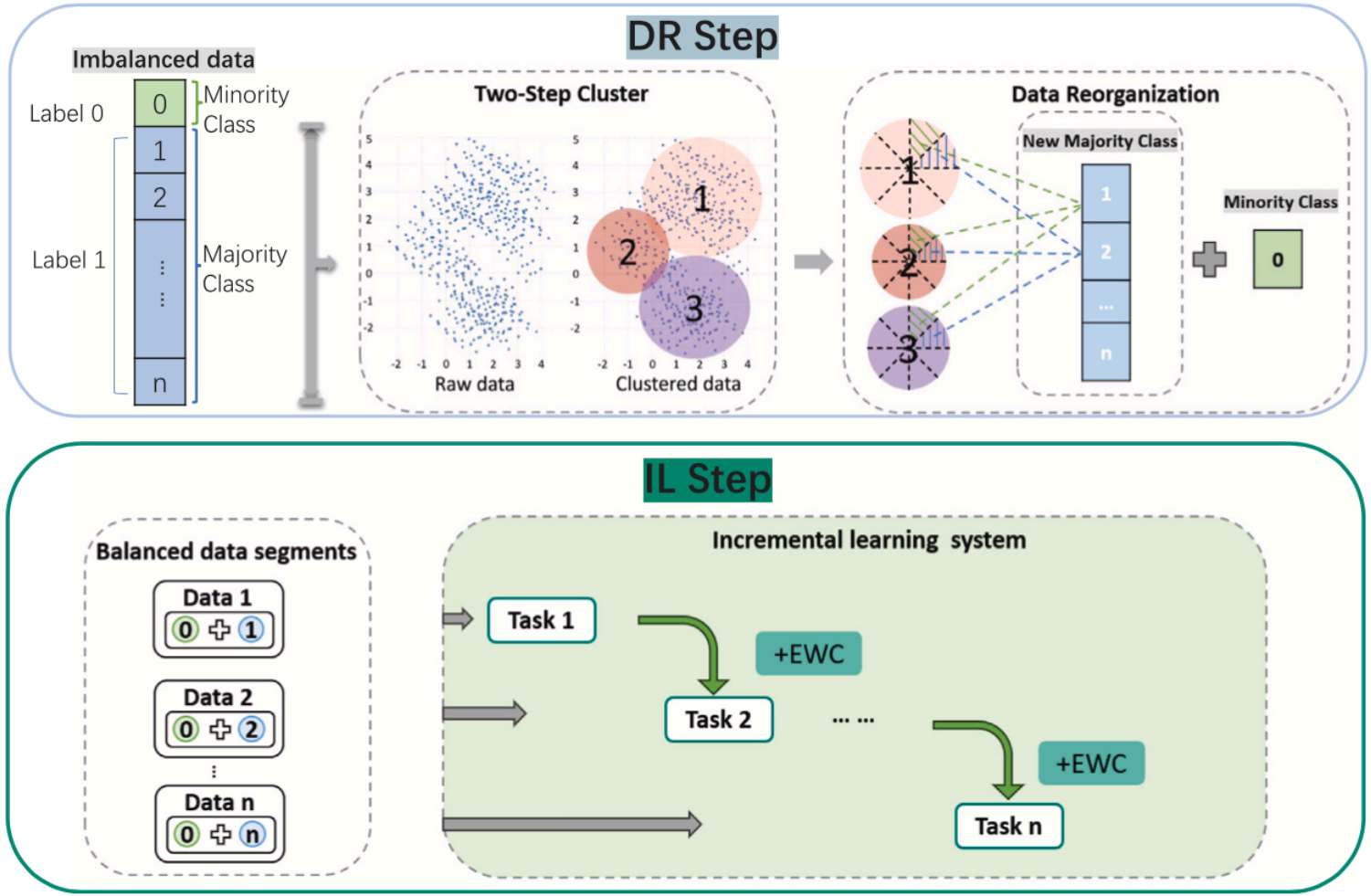


Fig. 4. Framework overview of the DRIL methodology.

3. 数据偏倚与公平性：因果推理

3.1 挑战：算法偏见的“伪装”

问题：AI模型会“投机取巧”，学习临床“捷径”(Shortcuts)

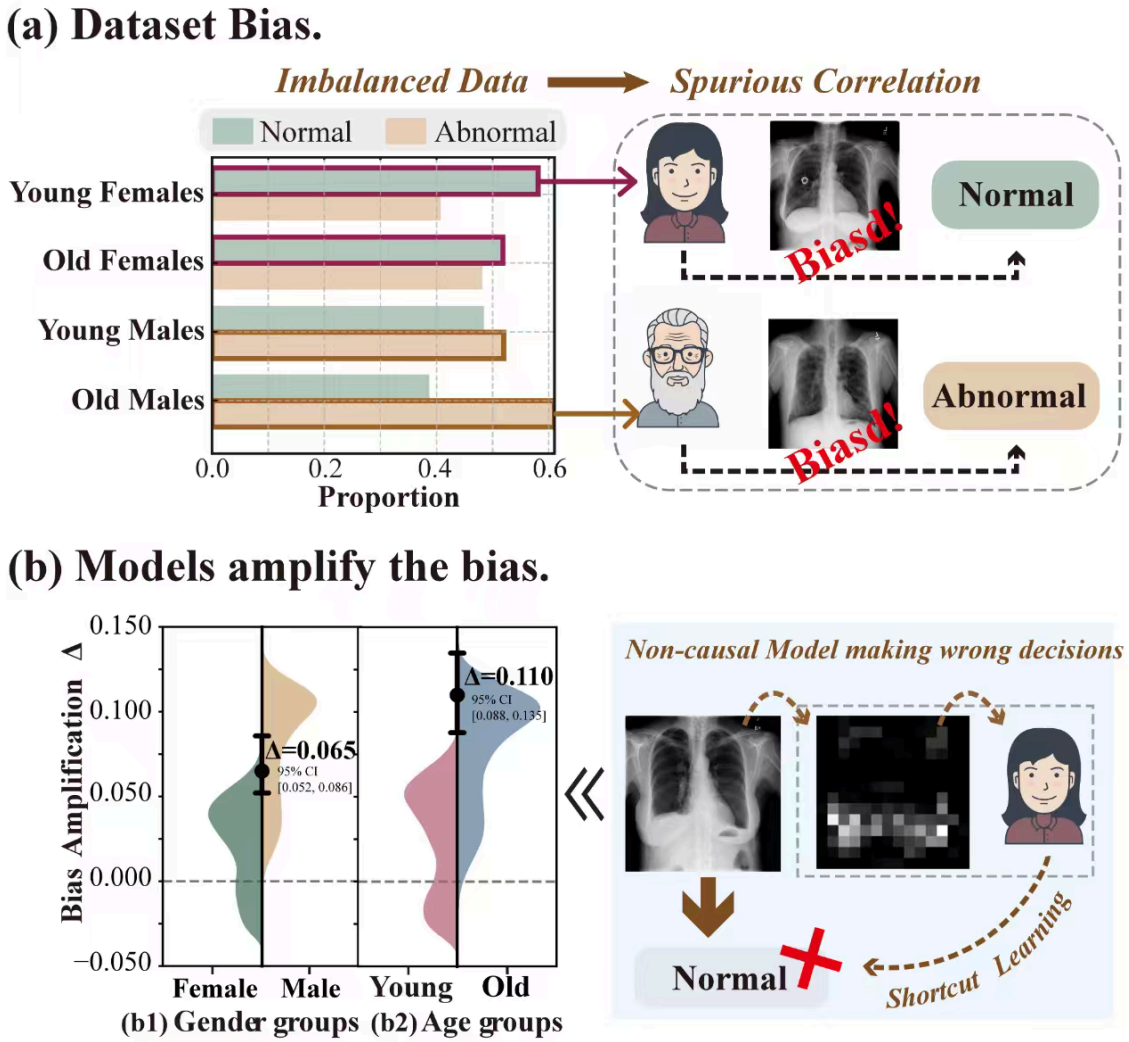
我们的数据中，不可避免地存在**混淆变量 (Confounders)**，如年龄、性别、既往病史等。

- **常规AI的陷阱：学习“伪相关性”**
  - 模型会发现一个“捷径”：“70岁以上男性” (Z) 这个特征，与 “发生心梗” (Y) 高度相关。
  - 它会**放弃学习**复杂的 ECG波形 (X)，转而依赖 年龄/性别 (Z) 来进行预测。

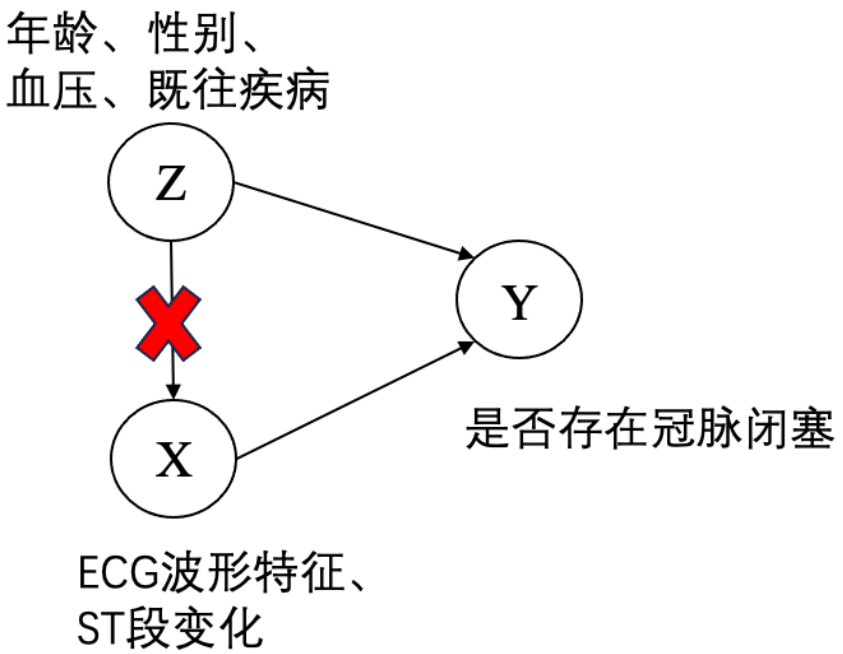
3.2 临床的灾难性后果

1. **漏诊 (False Negative):** 面对一个**40岁的女性**（不符合“捷径”特征），即使其ECG有典型闭塞波形，模型也可能将其误判为“低风险”。
2. **误诊 (False Positive):** 面对一个**70岁的男性**（符合“捷径”特征），即使其ECG正常，模型也可能过度敏感，将其判为“高风险”。

- **我们的目标：** 斩断“捷径”，强迫模型从ECG信号本身进行因果推断。



3.3 我们的对手：“后门路径” (Back-Door Path)



AI模型有两条路径可以“抄作业”：

1. ✅ **真实诊断路径 (Causal Path)**
   - $X \rightarrow Y$
   - ECG波形 **导致/反映** 冠脉闭塞。
   - **这是我们希望AI学习的唯一路径。**
2. ❌ **虚假偏倚路径 (Back-Door Path)**
   - $X \leftarrow Z \rightarrow Y$
   - 年龄(Z) 既影响 ECG波形(X)，也影响 冠脉闭塞(Y)。
   - 这在X和Y之间建立了一条“虚假”的统计关联。**（图中红色X标记的路径就是这条“后门”的关键干扰）**

3.4 我们的对策：“后门调整” (Back-Door Adjustment)

目标：强迫AI忽略“虚假路径”，只学习“真实路径”

- **理念：** 我们不能直接计算 $P(Y|X)$，因为这个概率被年龄(Z)污染了。我们必须计算 $P(Y|do(X))$ —— 即在“干预”了ECG后，闭塞发生的真实概率。

- **“后门调整”的临床转译：**
  我们通过对混淆变量 (Z) 进行统计分层，强迫模型回答一个更精确、更公平的临床问题：

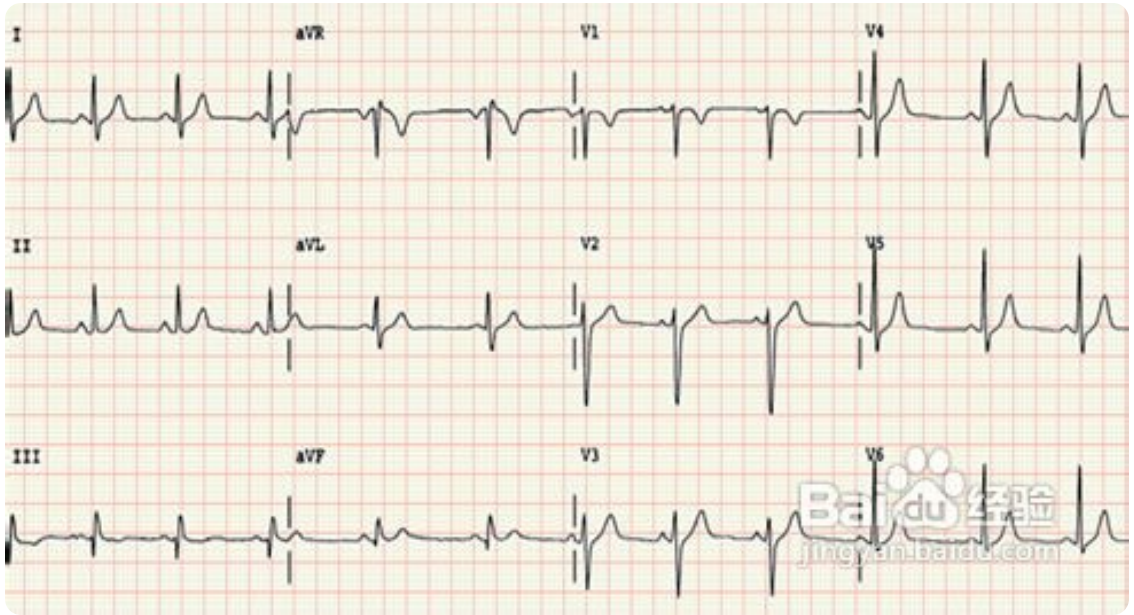
  “假设在年龄、性别、血压(Z)完全相同的两个患者中，
  ECG波形(X)的‘何种差异’，
  真正导致了其中一人发生闭塞(Y)？ ”

4. 时序建模：基于Prompt的时间序列大模型

ECG的双重身份 (视角A: 时序信号)

ECG的本质：12导联的高维、非平稳时序信号

- **数据特征：** 12通道（导联）高度耦合，10秒 @ 500Hz 采样 = **60,000个数据点**。
- **病理特征：**
  - **局部 (Local)：** 单个导联的P–QRS–T波形态、ST段偏移。
  - **全局 (Global)：** 12个导联间的协同变化、长时依赖 (Long–Range Dependency)。

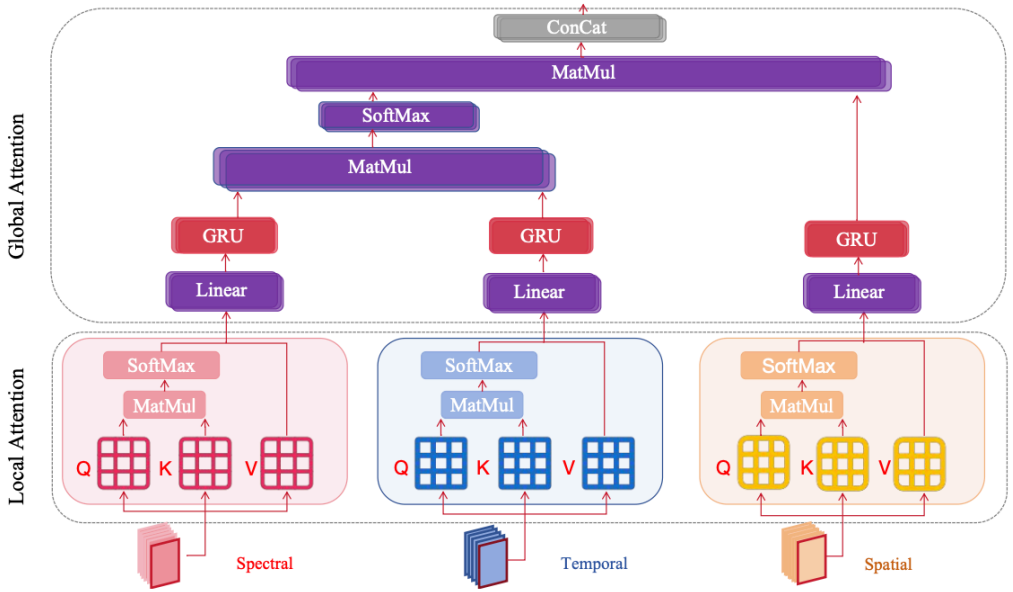


建模的“双重困境”

1. **轻量级模型 (1D–CNN/LSTM)：**
   - **“管中窥豹”：** 擅长捕捉**局部**形态，但易丢失**全局**信息。
   - **临床局限：** 无法有效建模12个导联间的局部协同变化。
2. **重量级模型 (Transformer)：**
   - **“全局视野”：** 自注意力机制 (Self–Attention) 擅长捕捉全局依赖。
   - **挑战：** 60,000个点的序列过长，计算量巨大，且对局部噪声敏感。

“全局-局部”融合框架

- **局部特征 (Local Attention)**
  - **轻量级 (CNN)：** 快速扫描12导联，捕捉**局部**的形态特征（如QRS增宽、ST段偏移、T波倒置）。
  - **临床意义：** 识别独立的“病理波形”。
- **全局特征 (Global Attention)：**
  - **重量级 (Transformer)：** 整合所有“局部”信息，分析它们在**全局**（跨12导联）的**协同关系**。
  - **临床意义：** 识别“心梗模式”（如下壁、前间壁）。



解决“6万个数据点”的算力与噪声挑战

问题： 10秒的ECG（60,000+数据点）对AI来说是“天文数字”，直接处理会“算不动”且“被噪声淹没”。

我们的对策（借鉴 APaLLM 论文）：先“提纯”信号，再“打包”读取。

1. 信号提纯 (Decomposition)

我们首先将原始时序信号分解为三部分：

- **趋势项 (T)：** 整体的大趋势。
- **季节项 (S)：** 周期性的变化信号。
- **残差项 (R)：** T+S与真实信号的残差项。

获益： 我们只让AI分析“R项”。这就像在嘈杂的房间里只听心跳声，极大排除了干扰。

2. 打包读取 (Patching)

AI不会“一个点一个点”地读6万次，而是“一段一段”地读。

- **做法：** 我们将“纯净信号”打包成 Patches（小片段）。
- **类比：** AI不再是“逐字阅读”，而是“逐词阅读”。

获益： 序列长度从60,000急剧缩短到几百，AI可以**高效地看到全局**，理解QRS波、T波等“波形模式”。

5. 图像建模：光学压缩式ECG视觉表示

ECG的双重身份 (视角B: 图像信号)

一个反直觉的思路：为什么不直接“看图”?

- **传统观念：** 将ECG图（12导联波形）视为图像，是“信息丢失”的。
- **大模型（VLM）的新观念：** 视觉是一种**高效的“信息压缩”**。
- **临床类比：**
  - 60,000个点的**时序数据**，就像一本6万字的“流水账”。
  - 1张12导联**ECG图像**，就像一张“总结图表”。
- **AI的进化：**
  - 以前的AI（轻量级模型）看不懂图，只能处理“流水账”。
  - 现在的视觉大模型（VLM）**更擅长“看图”**，能一眼识别出图表中的“关键模式”。

“光学压缩”的ECG视觉表示

借鉴 DeepSeek-OCR 论文：用“少量视觉Token”压缩“海量文本”

- **DeepSeek-OCR 的核心洞察：**
  - VLM（视觉大模型）可以用**极少数**的“视觉Token”（如100个），来完美解码（Optical Character Recognition）一张包含**数千个**“文本Token”的复杂图像。
  - 这证明了“光学2D映射”是一种**极高效率的信息压缩**。
- **我们的临床转译与应用：**
  - **“海量文本”（1000+ Tokens）→“ECG原始信号”（60,000+ Points）**
  - **“压缩图像”（100 Tokens）→“12导联ECG图”（用少量视觉Token表示）**

我们不直接处理60,000个点，而是处理压缩后的“视觉ECG图”。

“全局-局部”视觉编码器

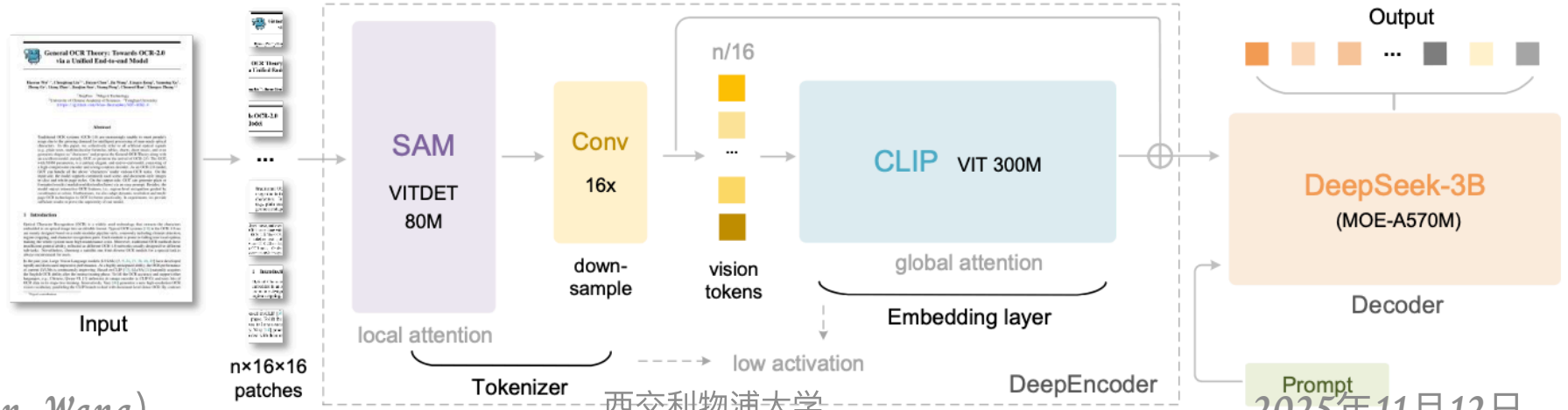
为了让AI“看懂”这张高分辨率的ECG图，我们采用了类似的“全局–局部”视觉框架：

1. 局部感知 (Local Attention)

- **模型：** 窗口注意力（Window Attention）。
- **任务：** 像放大镜一样，精细“感知”ECG图上的**局部细节**（如单个ST段的形态、T波是否倒置）。

2. 压缩与全局认知 (Global Attention)

- **模型：** 16x压缩器 + 全局注意力。
- **任务：** 将“局部细节”压缩后，进行“全局认知”，理解12个导联在图像上的**空间布局和协同关系**（如“胸前导联的集体抬高”）。



6. 模型可解释性：反事实强化学习解释器

AI的“黑盒”——如何信任它?

最大的临床障碍：AI如何取得医生的信任?

- **问题：** 无论模型（轻量级或大模型）预测得多准，如果它只是一个“黑盒”，临床医生就无法信任它，更不敢基于它的建议（如“立即手术”）采取行动。
- **临床医生的核心疑问：**
  i. **“你凭什么这么说？”**(归因, Attribution)
    - AI是根据哪个导联、哪个波形的微小变化，做出的“NSTEMI–急性闭塞”判断?
  ii. **“如果……会怎样？”**(反事实, Counterfactual)
    - “如果V2导联的T波**没有**这个微小抬高，你还会认为是闭塞吗？”
    - “我需要将ECG改变**多少**，才能让你的模型改变诊断？”

我们的目标： AI不仅要提供答案，更要提供“**可操作的证据**”。

反事实解释器 (Counterfactual)

我们引入一个AI“侦探”(RELAX)，让它来审查“黑盒模型”的工作。

这个“侦探”的任务是玩一个“**What-if**”游戏：
“我需要对这份‘高危心梗’ECG做‘多小’的改动，才能让‘黑盒模型’将其改判为‘低风险’？”

- **临床解读：**
  i. **输入 (Original Example):** 一份被“黑盒模型”诊断为 **“高危闭塞”** 的ECG。
  ii. **“侦探” (Counterfactual Generator):** RELAX 解释器开始工作，它尝试 **“最小程度”** 地修改ECG波形（如“降低V3的T波”、“拉平aVL的ST段”...）。
  iii. **输出 (Optimal Counterfactual):** “侦探”找到了一个 **“最小修改方案”**（比如：*仅将V3的T波幅度降低0.2mV*），这份新ECG就被“黑盒模型”改判为 **“低风险”** 了。
- **最终结论：** 这份“反事实”报告精确地告诉了医生——**V3导联的T波**，就是导致模型做出“高危”诊断的 **“罪魁祸首”**。

反事实解释器框架图

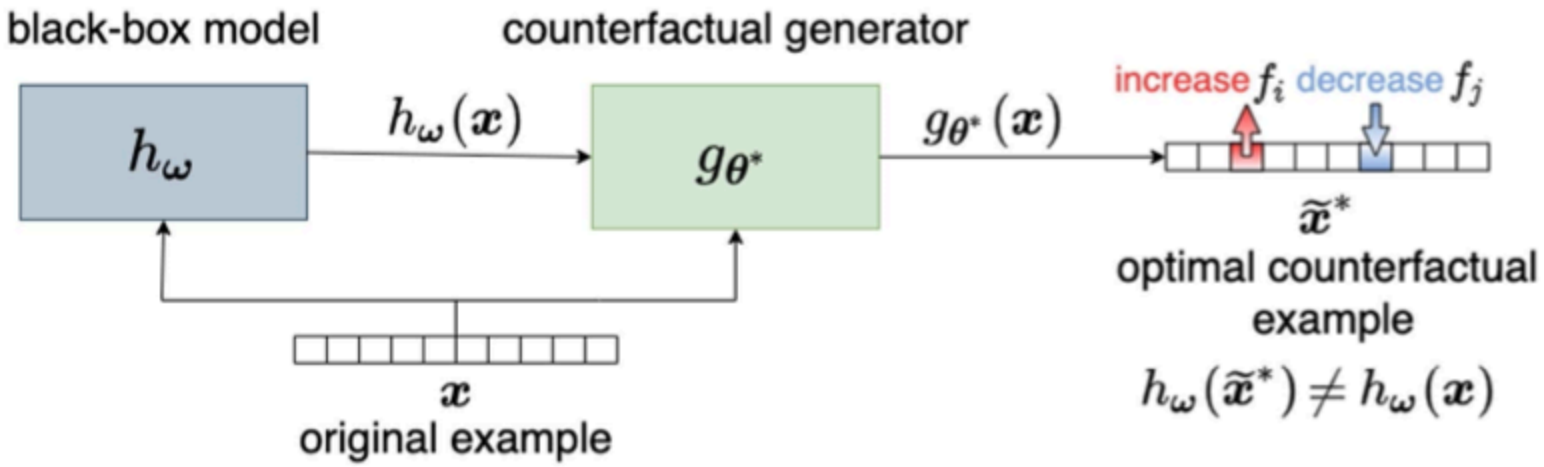


Fig. 1. Overview of a generic counterfactual explainer.

Chen, Z., et al. Explain the Explainer: Interpreting Model–Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent. IEEE Transactions on Artificial Intelligence.

“解释”我们的“侦探”

问题： RELAX “侦探”本身也是一个复杂AI（深度神经网络），它不也是“黑盒”吗?

答案： 是的，所以我们更进一步——“**解释这个解释器**”。
我们使用“**知识蒸馏**”技术，将这个“侦探”的复杂“大脑”

……“蒸馏”成一个医生能看懂的“简单决策流程图”（Distilled Decision Tree Policy）。

- **临床价值：**
  - 我们最终交付给医生的，**不是另一个黑盒**。
  - 而是这样一个**透明的、可被验证的“IF–THEN”逻辑**：

> - “**IF** V2导联T波 > 0.1mV (且未被修改过)...”
> - “**THEN** ‘侦探’的首选动作是：修改 **V2导联**。”

- **最终：** 整个诊断（黑盒模型）和解释（侦探模型）的流程，都变得**完全透明、可信赖**。

欢迎交流 ~

- 邮箱：jia.wang02@xjtlu.edu.cn
- 欢迎老师学术交流
- 欢迎学生RA访问