数 智 健 康 国 际 动 态

北京市卫生健康大数据与政策研究中心 2026.1.26

（一）人工智能在医疗决策中应用

随着人工智能技术的快速发展，医疗领域正经历一场深刻的智能化转型。人工智能不仅具备处理海量、高维医疗数据的能力，还可通过机器学习与逻辑推理，为疾病诊断、预后评估及个体化治疗方案制定等关键临床决策提供有力支持。然而，鉴于医疗决策直接关系患者生命安全与健康结局，人工智能模型在可靠性、安全性与可解释性等方面面临极高要求。当前，人工智能在医疗领域的落地应用仍面临两大核心挑战：其一，如何构建科学、稳健且具临床意义的预测模型性能评估体系；其二，如何在动态、异构、资源受限的真实医疗环境中，实现安全、高效、可信的智能系统部署与持续运行。以下两篇研究分别围绕上述挑战展开，可为相关理论探索与实践应用提供学术参考与方法借鉴。

第一篇文章主要介绍了医疗人工智能预测模型通过估算个体患病风险来辅助临床决策。文章提出了一个包括五大性能领域的预测 AI 模型性能评估体系，并基于案例研究给出了实用性建议。五大性能领域包括：辨别、校准、整体表现、分类及临床效用。辨别关注模型区分事件与非事件的能力（如 AUROC）；校准评估预测概率与实际发生率的一致性；整体表现结合二者；分类基于设定阈值将个体分为高风险/低风险组；临床效用则进一步纳入误分类成本，直接衡量模型辅助决策的实际价值。其中，有效的性能指标应具备两个关键特征：适当性（即正确模型的期望性能最优）和明确焦点（仅反映统计评估或决策分析，不混淆二者）。研究指出，部分常用指标（如 F1 分数、AUPRC）因同时违反这两项原则而不推荐使用。类别不平衡问题常被过度强调，实则应与临床相关的错误分类成本区别对待。最后建议：对于支持医疗决策的预测 AI 模型，评估应聚焦于辨别、校准和临床效用。推荐报告以下核心指标：辨别（AUROC（C 统计量））、校准（平滑校准图（辅以置信区间））、临床效用（净收益及其决策曲线，可直观展示模型在不同决策阈值下相较于常规策略的改善程度）、风险分布图（展示不同结局群体的预测概率分布）。此外，在校准方面应重视外部

验证，因模型在新群体中的校准性能更具临床意义。避免使用不当指标（如分类准确率、F1 等）及焦点模糊的指标（如 AUPRC），以确保评估结果科学、可靠且贴合临床决策需求。总之，医疗 AI 模型的性能评估需遵循方法学严谨性，强调指标适当性与临床相关性，从而推动可靠、可解释的预测工具安全应用于临床实践。

第二篇文章提出一种融合边缘计算、区块链、分块技术与可解释人工智能（XAI）的医疗物联网（IoMT）框架，支持安全、高效、可解释的实时压力检测与临床决策。该框架由多阶段流程构成：IoMT 设备采集多模态生理信号，在边缘端完成预处理与降维；采用内容定义分块（CDC）技术切分数据，生成哈希并上链，保障完整性、防篡改与可追溯；处理后数据加密上传至云端，由 TimesNet 模型（基于 Transformer 的多尺度时序模型）完成压力水平高精度分类；最后通过 SHAP 分析揭示关键影响因素（如时间特征与皮电活性、体温等生理指标），提升临床可信度。研究具有四大创新点：（1）首创将 CDC 分块与区块链结合用于 IoMT 健康数据传输，结合哈希验证与 ECC 签名，重传率低于 5%，实现端到端安全；（2）TimesNet 在护士压力数据集上准确率达 99.6%，显著优于 PatchTST、Autoformer 等模型，在准确率、F1、ROC-AUC 等指标全面领先；（3）边缘预处理降低传输负载，分块策略提升存储与计算效率，区块链仅增加约 0.5%能耗，兼顾安全与效能；（4）SHAP 量化各特征贡献，使压力预测依据透明可验。研究在护士精神压力数据集上进行了实验，实验表明：减小分块尺寸可提升验证效率且不增成本；本框架是首个实现分块、区块链、边缘-云协同与 XAI 端到端整合的方案，在安全性、效率与可解释性上均衡突出。研究提出的框架适用于压力监测、慢病管理等实时临床场景。当前基于仿真与单一数据集，后续需在真实多源 IoMT 环境中验证，并可拓展联邦学习与智能合约，构建更隐私保护、自治化的数字孪生医疗系统。

（徐健编辑）

译文一：

支持医疗决策的人工智能预测模型性能指标评估：概述与指导

Ben Van Calster, Gary S Collins, Andrew J Vickers, Laure Wynants, Kathleen F Kerr, Lasai Barreñada, Gael Varoquaux, Karandeep Singh, Karel GM Moons, Tina Hernandez-Boussard, Dirk Timmerman, David J McLernon, Maarten van Smeden, Ewout W Steyerberg

来源：Digital Health.

时间：2025 年 12 月

链接：https://doi.org/10.1016/j.landig.2025.100916.

1. 介绍

医学文献中充斥着预测性人工智能（AI）模型，这些模型用于估算个体出现（诊断性）或发展（预后性）疾病或健康状态（事件）的概率，也称为临床预测模型。尽管这些模型传统上是通过回归分析等统计方法开发的，但具有更好灵活性的机器学习算法的使用正在增加。例如，传统的逻辑回归模型可能旨在通过人口学、临床和实验室测量，预测左侧阻塞性结肠癌切除患者永久造口的风险。一种更现代的深度学习模型可能旨在基于窦性心律电图预测心房颤动的存在。

无论采用哪种建模方法,面向医疗实践的预测 AI 模型的性能都应得到妥善评估。因此，选择适合医疗预测人工智能的性能指标至关重要，因为表现不佳的模型可能导致错误的临床决策，对患者造成不利影响并增加财务成本。尽管已有许多此类措施被提出，但仍需明确说明。关于医学、统计和机器学习文献中推荐的测量方法，偶尔存在争议。

绩效评估对外部和内部验证研究尤为重要。外部验证研究使用包含目标人群个体参与者数据的数据集来评估模型表现。与训练数据集不同，外部验证数据集包含来自不同地点、时间段或环境的个体数据。相比之下，内部验证通过交叉验证、自助法或（重复）训练-测试拆分等方法，利用与训练数据集同一群体的新个体评估模型表现。因此，内部验证指的不是模型选择，而是对所选模型的独立评估。

在本观点中，我们评估了统计和机器学习文献中的经典和当代模型绩效指标，并为研究人员、终端用户（即医疗人员）及其他利益相关者（如政策制定者）提供建议。我们提出了绩效领域的分类法，描述绩效衡量的关键特征，结合示例案例研究讨论这些指标，并提出建议。

2. 五个性能领域的分类法

我们将绩效指标分为五个领域：辨别、校准、整体表现、分类和临床效用。在这些领域中，前三个基于概率估计来评估性能（附录第 2 页）。

区分关注模型对有事件个体比无事件者更高的概率。辨别力反映了相对表现;也就是说，估计概率的高低并不重要，关键在于它们是否允许区分有事件和无事件的个体。

校准侧重于概率估计与观测事件比例的对应程度。校准通过评估估计值是否过高或过低来反映绝对性能。因此，模型可能有良好的辨别能力但校准较差，反之亦然。

模型的整体表现结合了判别和校准，通过量化概率估计与实际结果 0（无事件）或 1（事件）的接近程度。

第四和第五个性能领域要求对事件的估计风险设定阈值，将个体分为两个互斥的组别：低风险组（估计风险低于阈值）和高风险组（估计风险等于或高于阈值）。这些小组与干预措施（如手术）的决定相关，高风险个体建议手术，低风险者则不建议。因此，该阈值可以称为决策阈值。虽然多重决策阈值可以用来将个体分为三个或更多群体，但我们关注的是常见的单一阈值情况。

第四个表现领域——分类，侧重于个体被正确分类为高风险或低风险的程度。该领域基于列联表或混淆矩阵，即分类（低风险*与*高风险）和结局（事件*与*无事件）的交叉统计。当所有发生事件的个体的概率都高于决策阈值，所有没有事件的个体的概率低于阈值时，分类表现是完美的。分类性能受判别和校准性能的影响。

第五个领域——临床效用，更进一步，在评估个体被划分为低风险和高风险群体时，明确纳入了误分类成本。错误分类成本是一个已确立的术语，广义上指任何类型错误分类所带来的危害，其中错误分类指的是假阳性和假阴性。在生物医学应用中，假阴性的后果（例如，不推荐卵巢恶性女性进行高级手术）几乎总是与假阳性

的后果（将良性肿瘤女性转诊进行高级手术）不同。临床效用基于决策阈值评估决策质量，以及使用模型是否比不使用模型或竞争模型带来更好的决策。因此，决策门槛应具有临床相关性，并与误分类成本相关联。由于注重决策质量，临床效用是最重要的绩效领域。

小组

决策阈值的定义

大多数预测性人工智能模型在医疗实践中的主要目标是支持后续决策。概率估计可能指导医疗专业人员和患者通过避免对低风险患者进行负担沉重且预期效益有限的干预措施，促进高风险患者的干预选择，从而改善健康结局。因此，决策门槛应基于医学而非统计学的标准。

然而，通常通过优化如尤登指数（灵敏度+特异性−1）等统计指标来选择阈值。在最大化尤登指数时，敏感性和特异性同等重要;在医学领域很少出现这种情况。使用统计论证来设定决策阈值与决策理论不符，也脱离了临床医生的实际应用。

相反，一旦模型意图支持的决策被明确定义，应考虑使用该模型支持该决策的四个可能后果：

•真阳性（发生该事件且被归类为高风险者）

•真实阴性（未发生事件且被归类为低风险者）

•假阴性（发生该事件且被归类为低风险者）

•假阳性（无事件且被归类为高风险者）

这些后果的分量可能因干预的性质和效果、医疗系统，或临床医生和患者而异。

本观点的案例研究是患者需要手术切除卵巢肿块的情况。通过评估 adnexa 模型中不同肿瘤的评估，决定是否需要高级手术或保守手术。通常建议恶性概率的决定阈值为 0.1（10%）。建议对根据 ADNEX 预测变量有 10%恶性风险的患者进行高级手术，意味着每 10 名患者中每个真阳性患者都应进行高级手术（即对恶性肿瘤患者进行高级手术）。换句话说，我们接受最多 9 例假阳性（即对最多 9 名良性肿瘤患者进行高级手术）。因此，使用该阈值假设恶性肿瘤的高级手术的医疗益处是良性肿瘤患者不必要手术危害的九倍。在临床效用部分，我们描述了临床效用度量如何将错误分类成本纳入其中。

我们讨论了 32 项表现指标（3 项辨别，6 项校准，9 项总体，11 项分类，3 项临

床效用）指标（表 1），以及相应的视觉评估。

表 1 所讨论的绩效指标总结，评估两个关键特征， 以及 ADNEX 模型在案例研究中重新校准前后的结果

	特征		ADNEX 结果	
	适当性*	重点†	重新校准前	重新校准后
区分项				
AUROC, AUC, or C-statistic	+	+	0.911（0.894 至 0.927）	0.911（0.894 至 0.927）
AUPRC or AP	+	—	0.895（0.862 至 0.921）	0.895（0.862 至 0.921）
pAUROC (sensitivity ≥ 0・8)	+	—	0.141（0.130 至 0.151）	0.141（0.130 至 0.151）
校准				
O:E ratio	+	+	1228（1171 对 1288）	1000（0.955 至 1.046）
Calibration intercept	+	+	0.810（0.619 至 1.006）	0.000（−0.180 至 0.184）
Calibration slope	+	+	0.934（0.833 对 1.051）	1000（0.892 对 1.126）
ECI	+	+	0.105（0.063 至 0.160）	0.002（0.001 至 0.017）
ICI	+	+	0.094（0.074 至 0.118）	0.014（0.009 至 0.038）
ECE	+	+	0.091（0.072 至 0.117）	0.017（0.019 至 0.050）
整体表现				
Loglikelihood	++	+	−370（−407 至−334）	−337（−368 至−307）
Logloss or cross entropy	++	+	370（334 至 407）	377（307 至 368）
Brier score	++	+	0.133（0.118 至 0.147）	0.118（0.106 至 0.131）
Scaled Brier, Brier skill score, or IPA	++‡	+	0.469（0.412 至 0.527）	0.526（0.475 至 0.576）
McFadden R2	++‡	+	0.403（0.343 至 0.461）	0.456（0.405 至 0.504）
Cox – Snell R2	++‡	+	0.427（0.379 对 0.471）	0.469（0.429 至 0.502）
Nagelkerke R2	++‡	+	0.570（0.505 至 0.629）	0.625（0.573 至 0.670）
Coefficient of discrimination or discrimination slope	—	+	0.509（0.478 至 0.540）	0.525（0.495 至 0.556）
Mean absolute prediction error	—	+	0.243（0.226 至 0.260）	0.237（0.222 至 0.252）
分类：总结度量（使用 t=0・1）				

	特征		ADNEX 结果	
	适当性*	重点†	重新校准前	重新校准后
Loglikelihood	– §	+	0.794（0.768 至 0.819）	0.691（0.661 至 0.723）
Logloss or cross entropy	–¶	+	0.799（0.776 至 0.822）	0.700（0.677 至 0.724）
Brier score	–¶	+	0.597（0.551 至 0.643）	0.399（0.353 至 0.448）
Scaled Brier, Brier skill score, or IPA	—	+	37,400（24,600 至 68,500）	43,300（23,600 至 119,000）
McFadden R2	—	+	0.592（0.544 至 0.639）	0.392（0.346 至 0.442）
Cox – Snell R2	–¶	—	0.818（0.792 至 0.843）	0.756（0.727 至 0.782）
Nagelkerke R2	—	+	0.625（0.581 至 0.667）	0.480（0.438 至 0.522）
分类：部分测量（使用 *t*=0·1）				
Sensitivity or recall	—	+	0.954（0.934 至 0.974）	0.984（0.972 至 0.993）
Specificity	—	+	0.643（0.603 至 0.686）	0.415（0.370 至 0.463）
Positive predictive value or precision	—	+	0.716（0.679 至 0.753）	0.614（0.577 至 0.650）
Negative predictive value	—	+	0.937（0.911 至 0.964）	0.965（0.938 至 0.986）
临床实用性‖（净收益：*t*=0.1;预期成本：成本 9：1）				
Net benefit	+	+	0.443（0.411 至 0.475）	0.444（0.411 至 0.478）
Standardised net benefit	+	+	0.912（0.892 至 0.932）	0.915（0.900 至 0.930）
Expected cost	+	+	0.355（0.274 至 0.376）**	0.355（0.274 至 0.376）**

* 适当性：++，严格纯正;+，半正;——，不合适。

† 重点：+，衡量专注于纯统计或决策分析评估 通过妥善解决错误分类成本；–，度量结合了统计和决策分析 绩效评估因此缺乏明确的焦点。

‡ 这些测量在渐近上是严格正确的。

§ 当 t=0.5 时为半适当，这很少是临床相关阈值。

¶ 当 t 等于真实患病率时，称为半适当，而真实患病率很少是临床相关阈值。

‖ 特别是为了临床应用，使用置信区间和 p 值 临床效用衡量与决策分析的原则相矛盾。

** 原始型号的预期成本在决策门槛 0.06 处被最小化 重新校准模型为 0.15。AP=平均精度。AUC 或 AUROC=面积 受试者工作特征曲线。AUPRC=精确召回面积 曲线。ECE=预期校准误差。ECI=估计校准指数。ICI=综合 校准指数。IPA=预测准确指数。MCC=马修相关系数。 O：E 比率=观测到的比值超过预期。pAUROC=部分 auroc。

3. 信息性绩效衡量的关键特征

我们定义了绩效衡量应满足的两个关键特征：（1）衡量标准应当恰当，（2）应明确聚焦于仅反映统计或决策分析价值，通过妥善考虑差异性错误分类成本。不具备第一个特征的测度不可信，而不具备第二个特征的测度则是模稜两可的。第三个理想的特征是直觉解读。我们不再讨论这一特性，因为可解释性是主观的，受背景知识和熟悉度影响。

3.1 正确性

如果使用正确模型时，其期望值最优，则称该性能指标为适当，该模型基于模型中的预测变量或特征给出正确概率。这里的期望值指的是多次重复验证研究后获得的平均值。在任何给定数据集中，尤其是样本量较小时，由于随机变异，正确模型可能被错误模型超越。适当性的重要性在于，正确的衡量标准无法被欺骗：在预期中，正确的模型不会被错误的模型超越。当一个测度的期望值仅对正确模型最优时，即严格意义上的正当测度。当期望值对正确模型最优且对某些错误模型最优时，测度称为半正值。当一个错误的模型的期望值优于正确模型时，该测度被称为不当，无法信任。这 32 项措施的适当性状态列于表 1，附录（第 3 - 6 页）提供了示意图。

3.2 明确聚焦于统计或决策分析评估

统计评估和决策分析性能评估在医疗实践中具有明确区别。前四个绩效领域（附录第 2 页）关注统计绩效的不同方面，而临床效用领域则关注决策-分析表现。统计绩效指标对模型评估至关重要，但不能用来判断模型是否应在实际应用中：例如，引用良好的辨别和校准，或断定模型可用于辅助卵巢手术决策，均不合适。如果绩效衡量旨在超越统计价值的衡量，应根据决策-分析原则（小组）纳入错误分类成本。

如果错误分类成本以隐性或临时方式影响绩效指标，该指标既不能评估统计绩效，也无法充分评估临床实践决策的质量。

3.3 案例研究：卵巢癌诊断模型的外部验证

作为一个案例研究，我们考虑卵巢肿瘤女性恶性肿瘤的预测。ADNEX 模型由国际卵巢肿瘤分析（IOTA）联盟开发（作者 BVC 和 DT 均为该联盟成员），用于术前估算计划手术的卵巢肿瘤女性恶性肿瘤的概率。该模型可用于决定肿瘤中心接受手术类型（高级*或*保守手术），或其他地方患者是否转诊至肿瘤中心。ADNEX 基于 1999 年至 2012 年间在 10 个国家（意大利、比利时、瑞典、捷克、波兰、法国、英国、中国、西班牙和加拿大）24 个二级和三级护理中心招募的 5909 名个体数据开发。TransIOTA 研究通过外部验证 ADNEX 区分良性与恶性肿瘤的能力，数据来源于 2015 年至 2019 年间在比利时、意大利、捷克和英国四个国家的 1 所二级护理中心和 5 个三级护理中心招募的 894 名女性。有 434 名恶性肿瘤女性（患病率 49%）。IOTA 联盟数据的追溯性使用已获鲁汶大学医院（IOTA 主要伦理委员会）研究伦理委员会批准（S64709）。

出于教学目的，我们利用该数据集计算所有讨论的绩效指标，并采用 95%置信区间（95% CI），并展示所有讨论的可视化内容。置信区间通过百分位自助法对 1000 个自助样本进行了测量。我们通过逻辑调整评估了 ADNEX 的现状及更新后的性能（表 1）。为更新 ADNEX，我们在估计事件概率（线性预测变量）的 logit 上拟合了结果的逻辑回归模型。该方法类似于普拉特标度法，后者是机器学习中一种著名的方法，用于提升预测的校准。逻辑重校准本质上是对线性预测变量进行线性变换。因此，该方法是一种保持排名的方法，基于恶性肿瘤概率估计对患者的排名在更新前后保持不变。

所有 R 和 Python 脚本，以及 894 名参与者的恶性肿瘤风险估算和结果，均可在 GitHub 仓库中获取。

3.4 绩效衡量

本节简要讨论所选措施。关于各项措施的详细描述，包括公式，见附录（第 7 -

23 页）。

区分

区分的定义意味着区分措施应仅依赖于数据集中估计概率的排名。关键指标是一致性概率或 C 统计量。对于二元结果，C-统计量等于接收者工作特征曲线（AUROC）下的面积。多位研究者建议在患病率远低于 0.5（类别失衡）时不应使用 AUROC。当事件罕见时，AUROC 被描述为误导性或过于乐观，因为它忽视了获得可接受阳性预测值（PPV 或精度）和敏感性（或召回率）的难度，或未考虑误分类成本。精度-回忆曲线（PR）及其下面积（AUPRC）常被推荐作为 ROC 曲线和 AUROC 的替代方案。另一种替代 AUROC 的方法是部分 AUROC（pAUROC），该方法关注 ROC 曲线中特异性或敏感性达到特定最低耐受水平的部分。AUROC、AUPRC 和 pAUROC 是半正规的，因为这些基于秩的度量对概率估计的单调变换保持不变。将所有 ADNEX 概率除以 100 并不会改变这些测度的值。

没有理由将 AUROC 贴上误导性或过于乐观的标签。区分措施不应反映差异性的错误分类成本，类别不平衡不应与错误分类成本或医疗相关性混为一谈。与 AUROC 不同，AUPRC 和 pAUROC 没有明确的焦点（第二个关键特征）。AUPRC 和 pAUROC 将统计表现与临床效用相结合，但不遵循决策-分析原则。对于 AUPRC 来说，PR 曲线并不直接考虑真实的负数。虽然真负面在某些非医疗应用中可能无关紧要，但这类误差在医疗应用中通常非常重要。对于 pAUROC，“我们需要至少 90%的敏感度，因为我们希望找到至少 90%的癌症病例”这样的说法，表面上看起来可能合理。然而，根据具体情况和患病率，可能需要不同的决策门槛来将个体归类为高风险。因此，这种方法没有决策-分析基础。

尽管判别能力对预测人工智能至关重要，但仅靠 AUROC 无法用来识别模型的优良或有用性。通过 ROC 或 PR 曲线进行可视化是可接受的，但根据我们的经验，这些图在总结指标（如 AUROC）或相关临床效用指标（如净益处）提供的信息外，并无实用信息。

图 1 展示了 ROC 和 PR 曲线，并展示了该案例研究中关于灵敏度低于 0.8 即不可接受低的论点的 pAUROC。ADNEX 模型的 AUROC 为 0.91（95% CI 0.89 - 0.93），AUPRC 为 0.89（95% CI 0.86 - 0.91）。忽略敏感值低于 0.8，pAUROC 为 0.14（95%置信区间 0.13 - 0.15）。

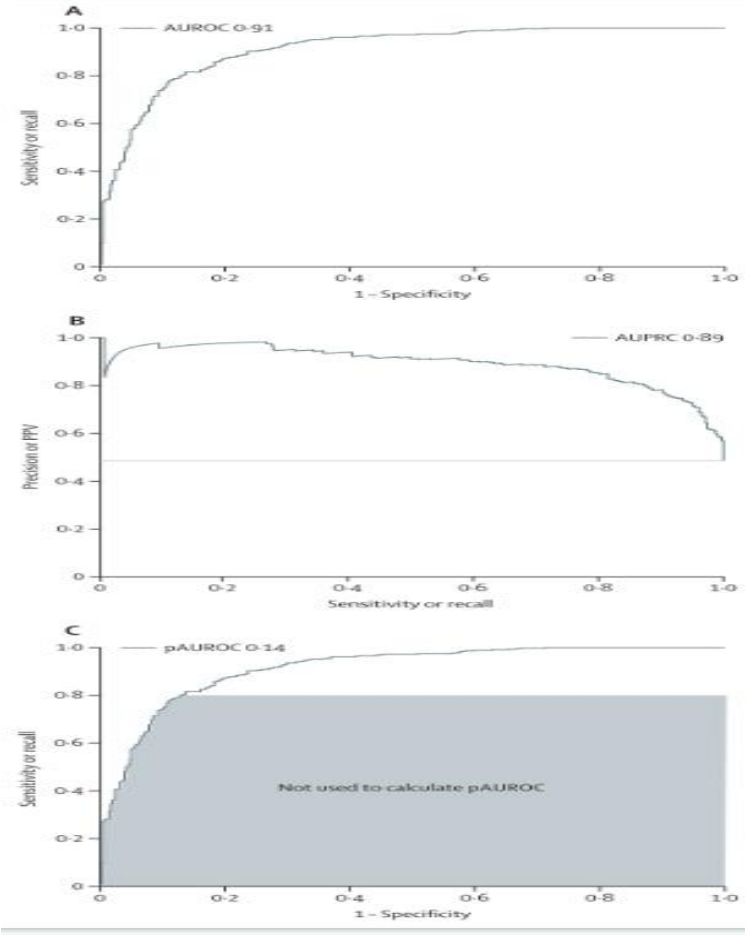

图 1 ADNEX 模型的 ROC 曲线、PR 曲线和 pAUROC 可视化

校准

统计学和机器学习文献中提出了几种方法来应对校准问题。这些方法可分为三个逐渐严格的等级，分别为平均、弱和中度校准。前两个层级主要来自统计文献。关于第四个水平——强校准的量化研究仍在进行中。强校准见附录（第 9 页）讨论。

平均校准（或称大规模校准）评估模型的平均估计概率是否等于数据集中观察到的流行率。校准大面积的两个指标是观测到的超预期（O：E）比值和校准截距。ADNEX 的 O：E 比为 1.23（95%置信区间 1.17 – 1.29），表明观察到的事件比模型预期多出 23%（表 1）。ADNEX 模型的校准截距为 0.81（95% CI 0.62 – 1.01），表明平均概率被低估（截距>0）。O：E 比比校准截距的解释更直观。

如果大规模校准良好且估计概率平均分布不大也不偏小（由校准斜率量化），则该模型校准较弱。扩散过大的估计概率平均过于接近 0 和 1（斜率<1），而估计的扩散过小的概率平均过于接近流行率（斜率>1）。在内部验证过程中，校准斜率小于 1 可能表明可能存在过拟合。在我们的案例研究中，ADNEX 模型的校准斜率为 0.93（95%

置信区间 0.83 – 1.05），表明概率分布充足。

中度校准意味着在估计概率为 x 的人中，观察到的事件比例也等于 x。评估中度校准最常见的方法是使用校准图，也称为信度图。校准图可基于个体分组或平滑生成。图 2 展示了案例研究中使用的数据的分组（十组大小相同）和平滑（使用局部估计的散点图平滑或黄土）图。这些图大多位于对角线以上，表明概率在整个范围内都被低估了。一个可能原因是验证研究中六个参与中心中有五个是三级护理中心，导致恶性肿瘤患病率较高（49%）。分组图无法全面处理中度校准，因为估计概率差异很大的个体仍可能归入同一组。

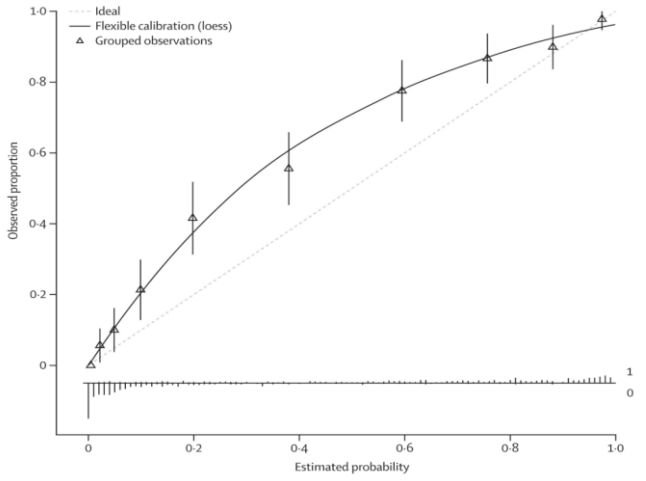

图 2 ADNEX 模型校准图，使用十组相同样本量，并使用黄土平滑器对估计概率进行分析

对于校准图，提出了若干总结性度量，如分组图的预期校准误差（ECE），以及平滑图的估计校准指数（ECI）和积分校准指数（ICI）。类似于 Hosmer – Lemeshow 检验等统计检验，所提出的总结测度无法准确判断校准失准的方向。此外，ECE、ECI 和 ICI 依赖于所用的分组或平滑方法，且存在统计一致性问题。改进的汇总指标正在研究中。因此，校准图（包括置信区间）是评估校准的关键工具，通过可视化基于估计风险的校准表现。

所有讨论的校准措施均为半合格的，重点关注统计表现（第二关键特征）。

整体表现

整体性能的基本衡量包括基于似然的度量、对数损失（也称为交叉熵或负对数似然）参数和 Brier 评分。表示相对于零模型表现的指标包括尺度 Brier（也称为 Brier 技能评分或预测准确指数）和解释变异比例的 R 平方测量，如 McFadden、Cox – Snell

和 Nagelkerke 的 R 平方。较少使用的总体指标包括辨别斜率（也称为判别系数或概率 AUROC）和平均绝对预测误差。

对数似然、对数和 Brier 评分严格正确，缩放后的 Brier 和 R 平方测度是渐近严格适当（即当样本量较大，例如超过 100 时严格正确），而判别斜率和平均绝对预测误差则不正确。所有讨论的总体指标都侧重于统计表现。

我们案例研究中使用的整体绩效指标见表 1。整体性能指标的图显示了事件和非事件估计概率的分布。图 3 显示了 ADNEX 模型的 Violin 图，表明良性肿瘤患者大多估计恶性肿瘤的概率非常低。恶性肿瘤患者大多估计概率为中高，峰值分布较小。

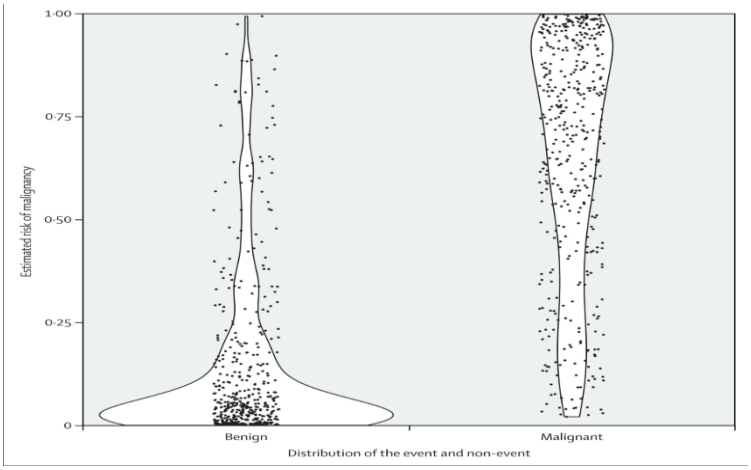

图 3 基于 ADNEX 模型的恶性肿瘤概率估计图

分类措施

在我们案例研究中，通常推荐的 10%阈值， ADNEX 将 578 名患者归类为高风险，其中 414 人为恶性肿瘤（真阳性），164 人为良性肿瘤（假阳性）。其余 316 例被归类为低风险，其中 296 例为良性肿瘤（真阴性），20 例为恶性肿瘤（假阴性）。

分类指标分为摘要测量和描述性部分测量。常见的部分测量包括敏感度（或召回）、特异性、PPV（或准确率）和阴性预测值（NPV）。敏感性和特异性根据观察到的结果进行分类评估，而预测时结局尚未知。PPV 和 NPV 在临床上更具相关性，因为它们评估结果时会根据风险分类来评估。

作为总结指标，我们讨论分类准确率、平衡准确率、尤登指数、kappa、诊断比值比、F1 和马修相关系数（MCC）。F1 和 MCC 的引入旨在解决类别不平衡带来的挑战，即在事件罕见时将所有患者归类为低风险，从而提高分类准确性。F1 与 AUPRC 相似，并有几个缺点：（1）F1 忽略真阴性，（2）F1 没有直观的解释，（3）F1 的

绝对值仅通过简单交换结果标签（即 1 变为 0,0 变为 1）改变。这些问题适用于更一般的 F_{text} 其中 F1 是其中的特例。与 F1 类似，MCC 没有直观的解读。

在给定的相关决策阈*值* t 下，所有分类指标均为不当（附录第 3 - 6 页）。某些分类指标（平衡准确性、尤登和 F1）在 t=0 • 5（分类准确性）或 t 等于真实流行率时是半适当的;然而，这些 T 值很少是临床上最相关的阈值。F1 是唯一一个没有明确关注统计表现的汇总指标，因为它混淆了分类与临床效用。

与分类性能相关的图包括 ROC 和 PR 曲线，显示了所有可能决策阈值上的部分分类度量。这些图的一个局限是阈值不易被看到（见图 1）。另一种图是分类图，其概率阈值在横轴，一个或多个分类指标在纵轴（附录第 24 页）。

在阈值为 10%时，ADNEX 模型的分类准确率为 0.79（95% CI 0.77 - 0.82），F1 评分为 0.82（0.79 - 0.84），MCC 为 0.63（0.58 - 0.67）（表 1）。

临床应用

与经典决策-分析理论一致，临床效用侧重于基于对应临床相关阈值的模型分类决策质量。为了评估效用，会明确考虑错误分类成本。在医疗预测研究中，临床效用最常用的指标是净收益。净收益的最大值等于患病率。标准化净收益等于净收益除以患病率，其最大值为 1。净收益基于错误分类成本设定决策门槛，遵循两者之间的经典联系。设定错误分类成本并不简单，且可能对成本应定额存在分歧（附录第 13 页）。因此，净收益或标准化净收益会在一系列合理的决策阈值的决策曲线中绘制。净收益和标准化净收益是半正当的;低于阈值的概率估计可以是任何值，只要低于阈值，概率估计值高于阈值则同理。

一个相关的指标是预期成本。与净收益相对，预期成本会寻找在错误分类成本下最小化成本的决策阈值。模型的误校准可能反映在期望成本最小的决策阈值上，而净收益固定了阈值，使误校准降低净收益值。期望成本是半正的，因为它对概率的秩保持变换不敏感。如果我们将成本归一化为 1，就可以绘制出合理归一化成本范围内的期望成本。

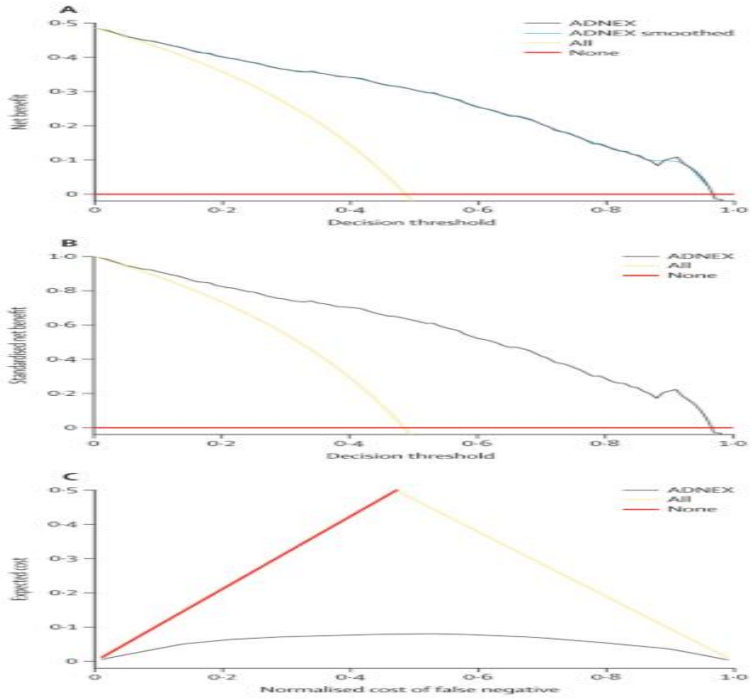

图 4 ADNEX 模型的决策曲线及其净收益、标准化净收益及预期成本

根据决策理论，关键关注点是检查模型是否优于参考策略（要么治疗所有人，要么不治疗）以及（如果相关）竞争模型，具有更好的效用。在我们的 ADNEX 案例研究中，如果我们同意对每个真阳性最多十名患者进行干预，我们认为真阳性（或假阴性的危害）的益处是假阳性的九倍。这些错误分类成本的相关决策阈值为 0.1。对于期望成本，假阴性的归一化成本为 0.9（*而*假阳性为 0.1）。我们进一步假设合理阈值范围为 0.05 至 0.40（误阴性归一化成本介于 0.60 至 0.95 之间）。在所有合理决策阈值下，ADNEX 的（标准化）净收益优于参考策略（见图 4A－B）。预期成本曲线给出同样的印象（见图 4C）。该模型在 t=0.1 时的净收益为 0.44。当假阴性归一化成本为 0.9 时，期望成本在 t=0.06 时最小化为 0.35。

3.5 重新校准后的结果

重新校准模型的性能图形显示见附录（第 25－29 页）。校准图在验证数据集中重新校准后更接近对角线。表 1 提供了 ADNEX 模型在重新校准前后的所有性能指标。所有严格正确的测量方法在重新校准后均有所改善。半适当的措施要么有所改善，要么保持不变。例如，由于等级保持更新方法，重新校准无法改善如 AUROC 等基于等级的判别指标。分类的不当总结指标（诊断比值比除外）显著恶化。部分分类指标有所改善（敏感性和净值），而其他指标则变差（特异性和 PPV）。整体绩效的不

当衡量有所改善。大多数不当绩效指标在重新校准后恶化，说明了“正确性”概念的重要性。

4. 讨论

我们评估了 5 个性能领域（辨别、校准、整体表现、分类和临床效用）中 32 项经典和当代性能指标，用于用于医疗实践的预测人工智能模型。在验证预测模型的性能时，我们警告不要使用不当的指标（13 个指标）或没有明确关注统计或决策分析表现的指标（三个指标;表 2）。值得注意的是，F1 是唯一一个同时违反这两个特征的指标。不当的测量可能会误导研究人员，而非澄清模型的表现。将统计与决策分析绩效混为一谈、未妥善考虑错误分类成本的指标存在歧义，应被专门的临床效用衡量标准所取代。

表 2 在验证预测模型以支持临床决策的背景下，针对不同测量和图表的建议和说明

	推荐	**备注**
区分		
AUROC	推荐阅读	该指标量化了辨别力，而辨别力是统计模型表现的关键组成部分。
AUPRC and pAUROC	不建议	这些指标试图超越统计评估，但违反决策-分析原则。
ROC curve and PR curve	既不建议也非必需	这些图块提供的关于 AUROC 的信息有限。
校准		
O:E ratio	既不建议也非必需	该指标可解释，但仅部分评估校准;内部验证时，O：E 比率通常为 1 或接近 1。
Calibration intercept and calibration slope	既不建议也非必需	这些指标难以解释，只能部分评估校准;在内部验证过程中，校准斜率可用于评估过拟合情况。[66]
ECI, ICI, and ECE	不是必须的	这些指标总结了校准图，掩盖了校准错误的性质和方向，并且在统计一致性上存在困难。
Calibration plot or reliability diagram	推荐阅读	该指标是评估校准最有洞察力的方法，尤其是在使用平滑而非分组时;内部验证时，优先使用图，但仅报告校准斜率也可接受;对于外部验证，强烈建议使用校准图，并标注不确定性（例如 95%置信区间）。
整体表现		

	推荐	备注
Loglikelihood, Brier, R2 measures	既不建议也非必需	我们建议分别评估辨别和校准。这些指标对于模型选择任务具有高度相关性，而这些任务超出了本观点的范围。
Discrimination slope and MAPE	不建议	这些措施不当;也就是说，对于错误模型，数值可能比对正确模型更好。
Risk distribution plots	推荐阅读	展示每个结果类别的风险估计分布，可以为模型行为提供宝贵的见解。
分类		
Classification accuracy, balanced accuracy, Youden index, DOR, kappa, F1, and MCC	不建议	这些措施在临床相关决策阈值下是不合适的;此外，有些度量难以解释。
Sensitivity (recall) and specificity	不是必须的;如果一起报告，可以具有描述性	虽然单独看不合适，但如果一起报道，可以描述性地呈现。然而，这些指标大多是理论性的，因为它们取决于预测结果。
PPV (precision) and NPV	不是必须的;如果一起报告，可以具有描述性	虽然单独看不合适，但如果一起报道，可以描述性地呈现。PPV 和 NPV 是非常实用的衡量标准，因为它们取决于排名。
Classification plot	既不建议也非必需	分类图可以描述性地展示，显示敏感性和特异性，或按阈值分别显示 PPV 和净现值。
临床应用		
NB or standardised NB (with a decision curve) and EC (with a cost curve)	推荐阅读	衡量做出多大更好决策的重要指标。NB 决策曲线使人们能够在各种临床相关决策阈值上展示潜在的临床效用，相较于默认决策（及竞争模型）。

AUPRC=精度-回忆曲线下的面积。AUROC=受试者工作特征曲线下的面积。DOR=诊断比值比。EC=预期成本。ECE=预期校准误差。ECI=估计校准指数。ICI=集成校准指数。MAPE=平均绝对预测误差。MCC=马修相关系数。NB=净收益。NPV=阴预测值。O：E 比率=观测到的比值超过预期。pAUROC=部分 auroc。PPV=阳性预测值。PR=精确 - 回忆。ROC=接收机工作特性。

我们认为，针对医疗实践的预测性人工智能模型的性能评估应聚焦于辨别、校准和临床效用。辨别和校准有助于建模师和临床医生理解模型如何改进。辨别力差表明可以选择其他有助于区分有无事件个体的预测因子。校准错误可能导致系统性过度治疗或不足治疗，从而影响预测性人工智能的应用。校准错误往往不仅是模型本身的问题，更是我们需要加深对模型验证和应用的各种情境理解的信号。遗憾的是，校准措施仍然被低估。整体性能指标结合了辨别力和校准性能，使其信息量不如区

分力和校准性能的单独评估。临床效用侧重于决策者和患者，评估模型是否平均带来更好的临床决策。

我们建议报告以下核心测量和图表：AUROC、平滑校准图、临床效用指标如带决策曲线的净收益，以及每个结局类别概率分布图（表 2）。在内部验证预测型 AI 模型时，校准可能不那么重要，因为模型开发和内部验证基于完全相同的人群。校准对于外部验证更为重要，因为模型在不同情境和群体中被评估。虽然校准图在内部验证中很有用，但仅仅利用校准斜率和可能的 O：E 比进行有限评估即可;但我们预期成熟模型的 O：E 比接近 1。除了推荐的核心集外，PPV 与 NPV 结合，或敏感性与特异性结合，也可以描述性报告。单独使用这些措施是不合适的。报告的测量和图表应尽可能附带置信区间，临床效用指标除外，因为不确定性的量化是近期争论和研究的话题。

阶级不平衡在模型开发和绩效评估中备受关注。我们认为，阶级不平等并不像常被说的那么严重。阶级不平衡的程度在数学上与错误分类成本的不平衡程度成正比。阶级失衡与目标人群相关的流行病学特征，而错误分类成本则是与决策背景相关的临床概念。误分类费用取决于医疗干预的性质和效果（例如是否进行手术的决定）。因此，我们建议不要使用 F1、AUPRC 或 pAUROC，而采用专门的临床效用指标。值得注意的是，当真正的阴性特征尚未明确定义时，我们不对其他医疗情况作出索赔，比如病变检测。

与绩效评估相关的三个主题值得重点关注：样本量、绩效异质性和报告透明度。首先，足够的样本量对于评估性能具有足够精确度非常重要。此前的建议是，最小结局类别中至少包含 100 到 200 名个体。针对回归模型，现在有更具体的样本量计算方法。在比较不同模型的校准时，通常需要更多数据。其次，模型性能的异质性应基于不同地点、环境或时间段间的种群和测量方法差异。荟萃分析和荟萃回归方法可用于量化和理解外部验证研究中性能的异质性。对使用不同外部数据集验证的模型进行简单比较，反映了不同背景下的不同人群，可能导致错误结论。第三，全面报告预测性 AI 建模研究至关重要，这可以通过遵循 TRIPOD+AI 及相关报告指南来实现。为避免性能黑客，应更加重视提前发布协议，并在合理范围内共享分析代码和数据。

该观点的一个局限是我们仅关注二元结果的绩效指标。然而，这些原则同样适用

于其他类型的结果，如名义结果、序数结果、事件发生时间或竞争风险结果。第二个限制是我们可以深入讨论其他几个话题。我们没有涉及反事实预测（假设干预下的预测），这一概念最近理所当然地受到了重视。此外，讨论所有指标是不可能的，绩效指标的研究仍在进行中。例如，校准是一个活跃的研究领域，重点关注强校准、误校准程度的量化以及不确定性等方面。此外，我们并未直接讨论模型比较，尽管在同一外部验证数据集上对竞争模型进行一对一比较尤为重要。与模型比较相关的一个具体话题是评估向现有模型添加新预测变量的增量价值。虽然竞争模型可以使用相同的核心测量和可视化来评估，但对于模型选择和比较等任务，适当的整体测量方法更为有趣。专门的衡量方法，如广泛使用但不当的净重分类改进，可用于评估竞争模型。

总之，我们认为绩效衡量应当适当，明确聚焦于纯统计或决策分析评估。为了评估医疗实践中的预测 AI 模型，推荐适用于大多数情况的核心绩效指标集包括 AUROC、校准图、临床效用度量如带决策曲线分析的净收益，以及显示风险估计分布的图。

译文二：

利用基于可解释、可信赖、用转换器模型和区块链集成块建立的人工智能 IoMT 的高效临床决策支持框架

Kübra Arslanoğlu, Mehmet Karaköse

来源：Diagnostics (Basel).

时间：2025 年 12 月

链接：https://doi.org/10.3390/diagnostics16010007.

1. 引言

物联网（IoT）的概念极大地扩展了人们获取信息的能力，而云计算则为物联网技术带来了强大的信息处理能力，并提升了物联网系统的效率。医疗物联网（IoMT）是一种基于物联网的技术，通常用于开发物联网支持的医疗系统，用于监测心电图、脉搏、血压和脑信号等多种生命体征。如今，IoMT 已成为一项有望彻底改变医疗行业发展的技术。IoMT 通过互联网连接医疗设备及其应用，显著提升了医疗服务的质量、可及性和个性化。这些智能设备能够持续监测患者的健康状况并收集其健康数据，构成了实时健康监测和干预的基础。然而，要让这项技术发挥全部潜力，收集的数据需要被有效处理和分析。这是一项重大挑战，尤其考虑到庞大的数据量和复杂的数据结构。

云计算和边缘计算技术为处理 IoMT 系统中收集的大量数据提供了解决方案。云计算实现了健康数据的集中处理，拥有强大的数据处理能力和大容量存储空间，而边缘计算则支持设备本身的数据处理，降低延迟并减轻数据传输负担。这两项技术的整合有望提升智能医疗系统中 IoMT 的有效性，并实现更快、更准确的临床决策。通过利用人工智能（AI），可以实现实时健康监测、疾病预测和预防，同时提升用户体验和系统可靠性。区块链因其去中心化、不可篡改性、可追溯性、匿名性和透明度等优越特性，是大幅提升云生态系统功能及安全性/隐私的理想技术。区块链与云

计算的集成有潜力提升身份验证与访问控制、数据安全与隐私、交易公平性、去中心化数据共享以及供应链管理应用的效率。区块链是一种创新技术，根据认证和安全规则向节点提供分布式数据账本。区块链在 IoMT 系统中的主要目的是通过最大限度减少数据被篡改或未经授权访问的风险，确保数据的完整性和安全。

传统的卸载方案可能不足以满足不断变化的网络环境和低延迟的需求。人工智能技术，尤其是深度学习和强化学习，在该领域取得了重大突破，研究人员基于深度强化学习开发了云端-边缘系统的资源分配方案。考虑安全的服务部署方法应使决策者理解为何某一部署最佳，因此 XAI 技术因其建议易于理解且可验证，受到安全社区越来越关注，吸引了患者、医生和监管机构等不同利益相关者。

随着云服务的普及，数据安全与管理已成为最重要的问题之一。为解决这一问题，广泛使用加密技术来保护敏感数据。安全云数据销毁和文件加密系统主要专注于在云计算基础设施中开发文件加密和销毁系统。在医疗领域，越来越依赖第三方云服务提供商存储数据，这使得保障信息的安全性、完整性和可访问性变得至关重要。由于传统的数据保护技术已不够，基于分块的方法被提出作为有效应对网络威胁的解决方案。此外，先进的基于分块的数据碎片框架通过结合指纹识别、碎片和加密技术，可以减少传输和存储开销。此外，人工智能领域的发展也支持这一结构。变换器模型的出现展示了其处理时间数据的强大能力，并在多变量时间序列领域不断发展应用。变换器模型并行处理序列数据，并利用自注意机制计算不同序列点之间的依赖关系和相关性，提取序列内部信息。

本研究提出一个框架，实现 IoMT 数据的安全、可解释性、成本效益高且节能的处理。该方法结合了边缘设备的数据预处理、云环境下的模型训练、区块链的安全、分块化的数据负载降低以及转换器的时间序列分析。实验结果显示，在加载时间、验证时间、通信量和成本方面均有显著提升。此外，基于转换器的模型高精度以及 SHAP 的可解释性表明该框架在技术和临床准确性方面具有强大的潜力。由于医疗数据日益异质化，以及对安全、经济且实时决策支持的紧迫需求，医疗系统正面临严峻挑战。尽管最近开发的 IoMT 框架改善了医疗水平，但现有研究大多聚焦于非集成单结构技术，并主要强调准确性性能。这些研究常常忽视通信效率、能耗、可解释性和端到端安全等关键方面。为解决这些不足，我们提出了基于区块链和分块技术的新型边缘云框架。该研究的重要性通过以下贡献得到了强调：

- 这项研究首次通过将内容定义分块和区块链技术整合在一起，实现了健康数据的安全、可靠和高效的传输，采用边缘云 AI 架构。
- 分段方法结合区块链完整性验证，防止了错误或不完整的数据块被接受，从而降低重试率和相关传输成本。
- 文献中分块仅用于去重目的，但本研究将分块方法作为传输模块实现，以优化 IoMT 应力数据的可靠传输。结合区块链日志，确保了可验证的可靠性和可追溯的完整性。
- 本文旨在通过在拟建架构中使用基于转换器的模型进行多变量时间序列应力检测，从而比传统机器学习和深度学习方法实现更高的准确性和训练效率。
- 基于 SHAP 的解释方法通过使临床估算透明且易于理解，提高了医疗提供者的信心。
- 该架构在实时压力检测和临床决策支持方面做出了重要贡献，这对于基于 IoMT 的可扩展、安全且可解释的决策支持系统至关重要。

在该架构中，健康数据在边缘设备上预处理，拆分成优化块，传输到云端，通过区块链验证，并由基于转换器的模型处理。这种设计不仅提高了精度，还降低了延迟、能耗和运营成本。此外，它还保证安全和透明。该工作的主要贡献是将分块技术、区块链、边缘云计算和可解释人工智能（XAI）整合到一个框架中，既具备技术效率，也兼具临床可靠性，并具实时验证的价值。因此，这与以往分别处理这些领域维度的研究有明显不同。

2. 方法

本节详细介绍了所提方法的主要组件和架构。在我们的研究中，我们提出了一个基于边缘云、支持区块链和分块、基于转换器的压力检测框架，能够实现物联网健康数据的安全、高效且可解释的处理。所提出的方法论如图 1 所示。

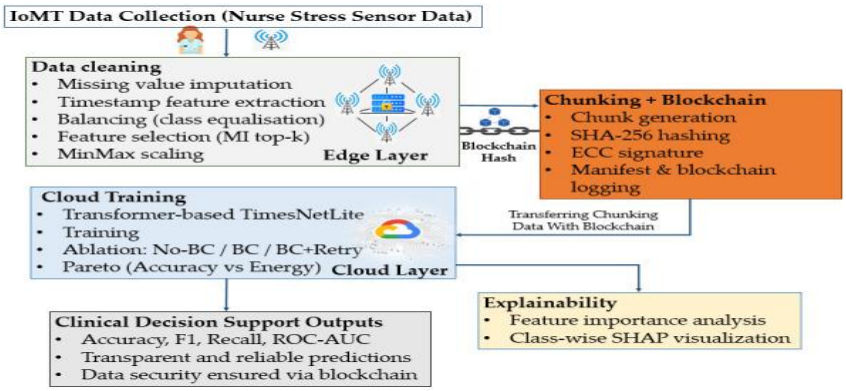

图 1 总体结构

根据研究的总体结构（如图 1 所示），其流程包括以下阶段：

1. 物联网设备的数据采集：护士精神压力数据集被用于模拟医疗领域的 IoMT 生态系统。该数据集包含反映应力水平的多变量传感器数据。数据首先被导入到边缘设备中。

2. 边缘设备上的预处理和特征选择：数据在边缘设备上处理，去除不必要的列，并应用降维（PCA）和数据均衡（SMOTE）方法。因此，通信负载降低，数据不平衡消除了降低模型性能的问题。

3. 区块链与分块整合：数据通过内容定义分块方法被分割成小块。每个区块的哈希值都被生成并存储在区块链上，从而防止数据重复，确保数据完整性、不可篡改性和可追溯性。通过实验记录了区块大小和哈希计算，随后利用这些参数计算通信和存储开销。

4. 安全云传输：边缘设备处理并通过区块链验证的数据被安全地传输到云端环境。通过检查哈希值的兼容性，可以保证传输数据的完整性。

5. 基于转换器的模型训练：基于转换器的深度学习模型，具有高能力分析多变量时间序列，在云环境中进行了训练。

6. XAI 的集成：SHAP 的集成以确保模型输出的透明度。因此，临床医生和终端用户可以解读模型决策所依据的特征。

7. 成本和能效计算：为评估系统的实际适用性，进行了上传时间、验证时间、通信量、区块大小、区块链验证时间和能耗（焦耳）等测量。

本研究开发的所有系统实验实现均在使用 Python 3.10 语言的 Google Colab Pro+ 环境中进行。实验中使用了 NVIDIA T4 GPU（16GB 显存）、12GB 内存和 Intel Xeon 2.20 GHz CPU 资源。建模过程使用 PyTorch（v2.2）、Scikit-learn（v1.4）、

SHAP（v0.45）、NumPy（v2.0.2）、Pandas（v2.2.2）、Matplotlib（v3.10.0）、Seaborn（v0.13.2）完成。边缘云架构是在基于仿真的环境中实现的，而非真实硬件;数据处理阶段在边缘层进行，模型训练在云端层面进行。区块链集成采用了基于SHA-256 的哈希函数、带有 ECC（椭圆曲线密码学）的数字签名验证流程以及 Python 哈希利布模块。

2.1. 数据集与 IoMT 上下文

IoMT 通过传感器、可穿戴设备和远程监测技术，促进了患者的随访以及医疗领域的诊断和治疗流程加速。这些系统提升了医疗服务的可及性，尤其是在欠发达地区，并由低能耗技术支持。如今，5G/6G 连接基础设施实现了低延迟、实时的数据传输，这使得 IoMT 系统能够与云端结构集成。然而，将该结构整合进系统是否成功，取决于有效解决数据安全、隐私、大数据管理和系统可扩展性等因素。IoMT 设备被归类为可穿戴设备、家庭医疗技术和院内智能设备。

文献中已有许多研究分析 IoMT 数据。例如，Wu 等人提出了基于人工智能的云端与物联网集成架构。Wang 等人开发了一个基于云端的平台，用于管理网络-物理-社会系统的复杂性。另一项研究探讨 5G 和物联网在健康诊断流程中的作用，强调大数据分析得益于联邦学习。此外，丁等人提出了基于深度学习的认知服务，涵盖云层和边缘层，采用 CloudCNN－EdgeCNN 结构。在一项关于居家健康监测的研究中，利用联邦学习开发了保护隐私的个性化模型。在 PRE-ACT 项目中，结合了联合学习和移动应用来预测放疗相关副作用。

本研究使用了基于多模态生理信号的数据集，该数据集最初由 Hosseini 等人在 Dryad 库发布并加入 Kaggle 平台，用于显示护士的压力水平。该数据集来自 15 名年龄在 30 至 55 岁之间的女性护士，均在常规医院轮班期间采集。该数据集包含约 1150 万条带时间戳记录，包含 9 个属性：X、Y、Z、EDA、HR、TEMP、ID、DATETIME 和 label。Empatica E4 可穿戴传感器在数据收集过程中持续测量电皮活性（EDA）、心率（HR）和皮肤温度（TEMP）;同时，记录了三轴加速度计（X、Y、Z）的数据，用于追踪身体运动和方向。每条记录标注三种压力等级之一——低、正常或高，并对应一个独特的匿名参与者 ID。该研究获得大学机构伦理委员会（FA19－50 INFOR）批准，并在数据收集前获得所有参与者的知情同意。该数据集为连续应力检测、工

作负载分析及基于人工智能的可解释建模等领域的研究提供了全面的基础。该多维且基于时间序列的数据集代表了现实世界的 IoMT 场景。其特性适合用于边缘设备进行预处理、功能选择和平衡，反映了在数据传输到云端前处理高维度、异构健康数据的需求。

表 1 文献中应用于护士压力数据集的方法概述

参考文献	方法	计算范式	数据处理与列车-测试分段	模型	绩效评估
[32]	基于 TinyML，结合近距离 1 级平衡和极大归一化	Edge（树莓派 RP2040）	近失控欠采样，随机状态：42，最小最大缩放器，train/val/test：60/20/20	KNN	Val_Accuracy：98%Val_Precision：98%Val_Recall：98%Val_F1：98%Test_Accuracy：98%Test_Precision：98%Test_Recall：98%Test_F1：98%
[33]	特征排名与 ROC 验证	集中式	1%子抽样用于预测，Random_state：42，MinMaxScaler,Z 分数归一化，缺失值处理，训练/测试：80/20	KNN	Val_Accuracy：90%Val_Precision：91%Val_Recall：90%Val_F1：91%Test_Accuracy：90%Test_Precision：91%Test_Recall：90%Test_F1：91%
[34]	联合学习	云端-边缘+联邦学习	重复删除、皮尔逊相关性、时间戳、Z 分数归一化、客户分成：80/20，全局评估/测试：50/50	神经网络	全球准确率：90% 精度：85% 召回率：85% F1：85% CL 准确率：97% FL 准确率：93%

表 1 提供了使用同一数据集进行的研究详情。在 Abu-Samah 等人的研究中，数据在基于 TinyML 的框架中处理以适应嵌入式系统，数据不平衡通过 NearMiss-1 方法纠正，XGBoost 模型嵌入在树莓派 RP2040 上，实现了约 86%的准确率。在 Chauhan 和 Singh 的研究中，同一数据集在应力分析范围内进行了评估，EDA 和 HR 信号被确定为最强的预测变量，使用 KNN 和随机森林算法报告的准确率为 83 - 90%。Liu、Xue 和 Hou 专注于数据隐私，提出了基于联邦学习（FL）的方法，结合了使用 FedAvg 算法的局部神经网络模型，并实现了超过 90% 的准确率。这些研究表明，使用同一数据集开发的模型在不同计算范式下具有高准确性和普遍性。尽管这些研究使用相

同的护士压力数据集，但它们的预处理步骤、训练-测试分段比、归一化策略和验证协议存在显著差异。

2.2. 云端与区块链分块压力数据框架

本研究利用 Kaggle 提供的护士压力数据集，基于护士的生理信号分析其压力水平。该多维且基于时间序列的数据集代表了现实世界的 IoMT 场景。其特性适合用于边缘设备进行预处理、功能选择和平衡，反映了在数据传输到云端前处理高维度、异构健康数据的需求。重复去重过程包括四个基本阶段：分块、指纹识别、索引和写入。分块作有两种类型;文件级分块处理将整个文件视为单一部分，具有较低的重复去重率，而块级分块则可实现固定大小（FSC）或内容定义（CDC）格式，并为用户提供更高效率。最新研究表明，基于分块的数据处理方法不仅在存储方面发挥关键作用，还在安全数据传输、计算成本降低和区块链性能优化方面发挥关键作用。在研究中，SmartChunk 开发了一种基于哈希的混合内容定义分块（CDC）方法，以实现高压缩比和更低重复数据率。在另一项使用类似加密过程的研究中，通过在云环境中独立加密每个区块来实现隐私和访问控制。同样，一项研究证明，区块大小直接影响区块链链的大小、验证时间和交易成本。在医疗领域，Subramani 和 Jothi 提出了基于分块的 RAID 加密模型，以优化区块链网络中的数据隐私和可扩展性。此外，区块链优化研究通过自适应压缩和高级数据结构减少了账本规模，提升了验证性能。另一项涉及块-云、SLO 和输入数据的研究中，规模感知的动态块配置可将成本降低多达 61%。

另一方面，医疗应用中对实时数据处理的需求也提升了云端架构的重要性。边缘层施加的工艺处理数据靠近源端，缩短响应时间并减轻网络带宽压力。更大、复杂且耗时的任务则在云端执行，以利用可扩展资源。云架构为集中存储、分析和与不同系统的数据共享提供了灵活的结构。云架构提供三种基本服务模型的多样化运营：SaaS、PaaS 和 IaaS。得益于这种结构，可以计算端到端的总处理时间、边缘、云和通信延迟，并在决策过程中同时提供速度和效率。

深度学习领域的最新研究也表明，分块不仅对存储至关重要，还对内存和计算效率至关重要。其中一些研究在长序列推断中将激活记忆减少了 80%以上，并将最大序列长度增加到三倍，且速度仅损失 10%。掩码区块处理技术将 GPU 内存占用减少了

三倍以上，能够处理长达 16 小时的音频输入，并将字错误率提升了高达 7.7%。这些结果清楚地表明，我们在研究中支持的基于块的策略在存储和计算效率方面优于其他方法。提出的云-边缘和区块链分块压力数据模型，与文献中的研究不同，将分块与区块链集成结合，重点关注 XAI 和健康数据的性能-安全平衡。因此，它在准确性和可靠数据传输方面，为基于 IoMT 的临床决策支持系统做出了独特贡献。

在与区块链的整合层面，即提供多级访问控制的物联网数据安全共享系统中，区块链不仅通过数据加密提升安全性，还通过不可篡改性、可追溯性和透明度。此外，区块链系统中使用的基本密码学构建模块还包括公钥密码学、零知识证明技术以及基于 SHA-256 的哈希函数。这些结构通过创建默克尔树来保证链的完整性。图 2 展示了该系统的架构，该系统集成了基于块的数据拆分、区块链日志以及性能-成本分析模块，以确保压力数据的处理安全、高效且可追溯。

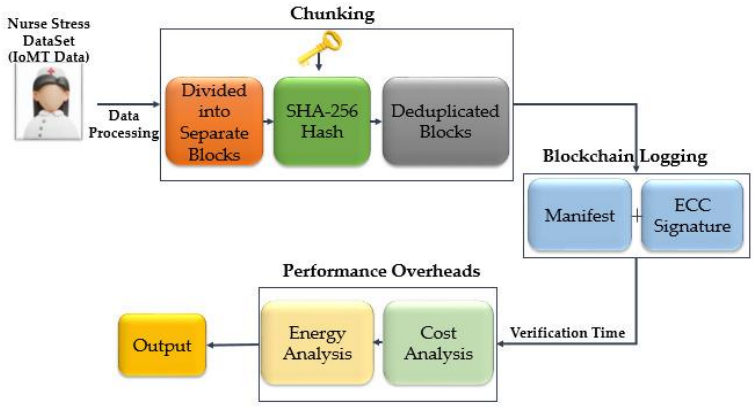

图 2 框架展示了数据分块、区块链日志以及应力数据的性能开销分析

在拟议工作中，块区块链结构与云端架构集成，实现压力数据的安全高效处理。边缘层执行低延迟的预处理任务，而云层则执行更复杂的模型训练和 XAI 应用。区块链集成确保数据受到不可篡改性、完整性和可追溯性保护。虽然偏好基于分块的方法的主要原因是提高存储和内存效率，但当与区块链结合使用时，它提供了一个可靠、可扩展且成本效益高的框架。分块+区块链机制的数学表述是研究中最重要的部分，见方程（1）－（5）。

$$D = \bigcup_{i=1}^{n} C_i \tag{1}$$

D：整个数据集。

n：总块数。

Ci：数据的第 i 块。

数据流被分割成小块。这一过程能够检测重复的数据块。

$$H_i = SHA-256(C_i) \tag{2}$$

$$E_i = AES_k(C_i) \tag{3}$$

H_i：第 i 个分块的哈希值

Ei：加密的数据块

k：分块加密密钥

数据通过固定大小（FSC）或内容定义（CDC）算法被划分为小块。通过对每个部分进行哈希，可以去除重复部分，降低存储和传输成本。

$$H_{\text{manifest}} = SHA256(H_1 \| H_2 \| \ldots \| H_n) \tag{4}$$

$$T = (H_{\text{manifest}},\ \text{meta}),\ B \leftarrow B \cup \{T\} \tag{5}$$

H_{manifest}：由所有分块哈希组合而成的哈希值

T：写入区块链的交易

meta：额外信息，如时间戳、用户 ID。

B：区块链账本

所有哈希值被组合起来生成一个显式哈希。它被记录在区块链上，并以 ECC（椭圆曲线密码学）签名。所有区块都通过写入链中生成的清单，从而在单一哈希下获得安全。

方程（6）定义了构成边缘云架构中总延迟的基本组成部分。因此，系统中的瓶颈可以被分析。

$$L_{\text{total}} = L_{\text{split}} + L_{\text{proc}} + L_{\text{link}} + L_{\text{verif}} \tag{6}$$

Ltotal：系统总延迟

Lsplit：分块延迟

Lproc：处理延迟

Llink：通信（传输）延迟

Lverif: 区块链验证延迟

在方程（7）中，给出了在块丢失情况下系统的重载速率和总成本计算。

$$U_{ploadCost} = DatasetSize\ (GB) \times CloudStoragePrice \quad (7)$$

$$ExtraCost = RetryRatio \times U_{ploadCost} \quad (8)$$

$UploadCost$: 将数据上传到云端的成本

$CloudStoragePrice$: 每 1GB 云存储价格

$ExtraCost$：传输过程中块丢失或重新加载时的额外费用

$RetryRatio$：装填率

方程（9）和（10）展示了模型训练和区块链验证过程中总能耗的计算。

$$Energy\,(\mathrm{J}) = Power\ (\mathrm{W}) \times Time\ (\mathrm{s}) \quad (9)$$

$$E_{\mathrm{total}} = E_{\mathrm{train}} + E_{\mathrm{verify}} \quad (10)$$

P$ower$ (W): GPU 的平均功耗（瓦特）

$Time$ (s): 训练时间（秒）

Etrain: 训练阶段的能量消耗

Everify: 区块链验证过程中额外消耗的能量

这些方程构成了该系统在数据安全和计算效率方面的基础。基于碎片和基于哈希的验证方程通过防止数据重复，减少了可能产生的存储负载，而与能源、成本和验证时间相关的指标则使得评估系统的性能、可靠性和可持续性维度成为可能。该结构将研究的理论基础设施表达为区块链支持的安全数据传输与边缘云 AI 架构之间整合关系的数学表达。

2.3. 传感器数据的变压器模型

变换器架构超越了经典的顺序结构，能够模拟长期依赖关系，这一机制最早由 Vaswani 等人提出提出。该方法极大地提升了自然语言处理和时间序列预测，并行计算优势且更有效地捕捉全局上下文。尤其是在基于物联网和物联网 MT 的应用中，基于变压器的方法在处理大量传感器数据时，在延迟、能量和数据丢失等重大约束下，已获得创新。

TimesNet 是由 Wu 等人开发的一种时间序列变换器导数。与变压器模型不同，TimesNet 以“时频”空间表示数据，而非“时空”。该结构以二维变异格式处理时间序列，基于多个周期的卷积层。这一基本组件称为“TimesBlock”，能够同时捕捉短期的突发变化和长期的生理趋势。在模型设置中，TimesNet 模型在财务系列、能耗预测和健康传感器数据上进行了测试，结果发现其准确率相比经典的变压器、LSTM 和 Informer 模型提升了 4–8%。在最近的研究中，变压器架构不仅被设计用于提升预测性能，还满足了安全性、能效和可解释性等关键需求。

例如，采用混合加密的变换器模型结合了 AES + ECC 混合和 Swin 变换器架构，用于基于 IoMT 系统的安全数据传输和异常检测。Greylag Goose 优化算法执行超参数优化，实现 97.3%的准确率和低能耗。此外，该模型通过解决去重问题，减少了健康数据中的不必要重复;在这方面，它属于我们研究中提出的基于区块链分块的数据管理策略的参考模型。同样，FogMedX-Transform 开发了一个基于转换器的雾化物联网物联网框架，利用特殊的注意力机制实现任务互作性，实现云计算与边缘设备之间的任务共享。该研究实现了高能效，异常检测准确率达 97.5%，任务互作性达 98.7%。

Kalakoti 及其同事提出了一种可解释的基于转换器的 IoMT 网络攻击检测架构。基于 SHAP 和 LIME 的可解释模块被集成到系统中，使模型决策具有可解释性。转换器模型在研究中使用的 CICIoMT2024 数据集中获得了 96.2%的 F1 分数。此外，报告称该模型相比经典 CNN 和 LSTM 方法实现了 7%的整体准确率提升。本研究为我们提出的系统提供了基础参考，特别是在阐述基于 SHAP 的可解释层的理论基础方面。本研究在相同实验条件下系统评估了 TimesNet、Autoformer、PatchTST、非固定变压器（NST）和经典变压器架构。所有模型均采用相同的数据处理步骤和评估指标进行测试。此外，还进行了综合比较，考虑了 GPU 内存使用和训练时间等计算效率指标。TimesNet 的多尺度卷积滤波器比其他基于转换器的方法提供了更高的准确性，因为它们有效模拟了短期应力波动和长期趋势。

表 2 展示了基于 TimesNet 的应力检测模型中使用的超参数。所有超参数均与其他基于转换器的模型保持一致，确保比较可重复且公平。

采用统一的实验方案，确保研究中使用的所有模型均可重复评估。数据集首先通过删除重复数据并用平均值填充缺失值来处理。在研究中，数据集采用分层抽样方法分为训练、验证和测试子集，70%用于训练，20%用于验证，10%用于测试。对训练

子集进行了共 25 次交叉验证，包括 5 次重复和 5 次重复。在每个折叠中应用基于互信息的特征选择，以确定信息增益最高的九个特征。缩放过程仅拟合在相关折叠的训练部分，并在验证部分应用这些参数，以完全防止数据泄露。所有模型都使用相同的超参数集训练，以确保在相同条件下评估。在计算出交叉验证获得的平均性能值后，最终模型对整个训练数据进行了重新训练，并使用了预留用于性能评估的验证和测试子集。

表 2　建模超参数

参数	价值
嵌入维数	128
层次	3
时期	(3, 5, 7, 9, 11)（仅限 TimesNet）
退出	0.1
批次大小	32
学习率（LR）	0.001
优化器	亚当
时代	20
交叉验证	5 × 5 重复分层 k 重形（25 重）
分裂	70%训练，20%验证，10%测试
功能选择	9

基于 TimesNet 的分类架构的总体结构，是转换器模型中最合适的时间序列模型，并能带来最高性能，见图 3。该模型采用变压器方法，通过多尺度卷积块增强，以捕捉短期和长期应力模式。模型的输入层直接接收来自护士压力数据集的多变量传感器信号。这些信号首先通过线性投影层，并转换为更高维的表示。然后，构成特征提取过程核心的 MultiPeriodConv 块被激活。在该块内，连续的 TimesBlock 层对不同周期长度为 3、5、7、9 和 11 的数据进行多尺度过滤，成功学习突变和长期趋势。获得的中间表示会经过 LayerNorm 层以提高统计稳定性，随后通过 Dropout 层减少过拟合。学到的特征随后被转移到稠密层的分类空间，Softmax 函数计算每个应力水平的概率。在最后阶段，输出层生成最终应力类别预测。

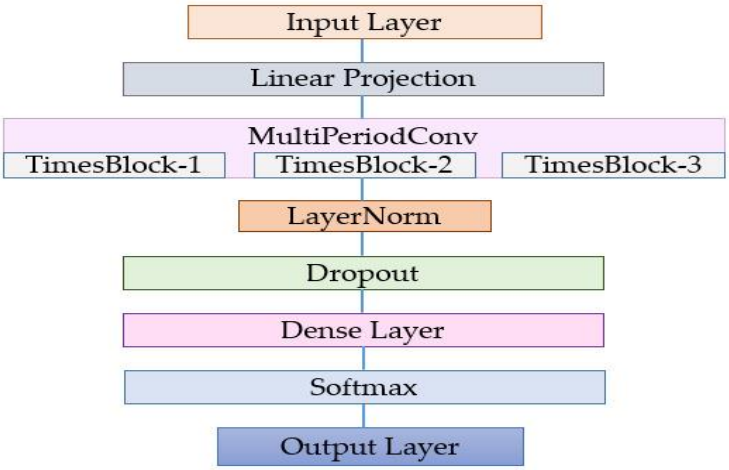

图 3 基于 TimesNet 的拟议模型架构

可解释性分析采用 SHAP Kernel Explainer 方法进行，以明确研究范围内的数据和模型可解释性。可靠性和成本分析通过基于区块链的分块机制完成。数据块范围从 64 到 1024 行，采用 SHA-256 哈希、清单验证、ECC 签名占位符以及 0.05 的重试概率进行控制。能量测量尽可能通过 NVIDIA-SMI 实时通过 GPU 功耗值获得。该实验方案提供了一个公平、可重复且计算高效的比较环境。该研究旨在将 TimesNet 模型确立为参考模型，因其在物联网数据上的高精度，尤其是其时间感知能力，相较其他架构显示出更高的准确性和稳定性。

3. 实验结果

本节介绍了本研究中使用的不同基于转换器模型的性能结果、可解释性分析以及基于区块链分块的可靠性性能值。

3.1. 模型比较与选择

研究中使用的所有基于转换器的架构均在同一训练协议下评估，采用平衡数据集、基于 MI 的特征选择和共同超参数。为更可靠地衡量模型的泛化性能，采用了 5 × 5 重复分层交叉验证方法，计算每个模型的准确性、精度、敏感度、F1 分数和 ROC-AUC 指标的平均和标准差值。所得结果见表 3。

表 3 基于变压器的模型性能

模型	准确率（%）（平均性病）±	精准度（%）（平均±STD）	召回率（%）（平均±性病）	F1 评分（%）（平均性病±）	ROC-AUC（%）（平均性±病）
时代网	**99.6 ± 0.08**	**99.6 ± 0.08**	**99.6 ± 0.08**	**99.6 ± 0.08**	**99.8 ± 0.13**
PatchTST	99.5 ± 0.06	99.5 ± 0.06	99.5 ± 0.06	99.5 ± 0.06	99.8 ± 0.04
转换器 Encoder	99.5 ± 0.04	99.5 ± 0.04	99.5 ± 0.04	99.5 ± 0.04	9.8 ± 0.01
自耦成型	99.3 ± 0.08	99.3 ± 0.08	99.3 ± 0.08	99.3 ± 0.08	99.7 ± 0.06
国家标准	99.1 ± 0.07	99.1 ± 0.07	99.1 ± 0.07	99.1 ± 0.07	99.7 ± 0.03

通过查看表 3，观察到所有模型的准确率均超过 99%，基于转换器的方法在应力分类任务中展现了较高的泛化成功率。在模型中，TimesNet 是研究中最成功的架构，在所有指标上表现出最高的性能。特别是，它提供了极为平衡的分类性能，准确率、精度、灵敏度和 F1 评分均达 99.6%。ROC-AUC 值为 99.8，表明模型在类别间具有极高的区分能力。尽管 PatchTST 和转换器 Encoder 模型在 TimesNet 之后表现良好，TimesNet 的多尺度卷积结构使其能够更有效地模拟短期和长期模式。尽管 Autoformer 和 NST 型号也达到了高精度水平，但其性能相较其他型号相对较低。对模型在 20%验证和 10%测试集上的表现进行了评估，结果基于指标的表 4 显示。

表 4 基于转换器模型的验证/测试集性能

模型	**准确率% Val/Test**	**精度% Val/Test**	**回忆率% Val/Test**	**F1 分数% Val/Test**	**ROC-AUC % Val/Test**
时代网	99.6/99.6	99.6/99.6	99.6/99.6	99.6/99.6	99.9/99.9
PatchTST	99.6/99.6	99.6/99.6	99.6/99.6	99.6/99.6	99.9/99.9
转换器 Encoder	99.6/99.6	99.6/99.6	99.6/99.6	99.6/99.6	99.9/99.9
自耦成型	99.4/99.4	99.4/99.4	99.4/99.4	99.4/99.4	99.8/99.8
国家标准	99.2/99.2	99.2/99.2	99.2/99.2	99.2/99.2	99.7/99.7

通过查看表 4，可以观察到所有基于转换器的模型在 20%验证和 10%测试集中都表现出极为接近且高的性能。所有模型均达到准确性、精度、灵敏度，且 F1 分数超过 99%，表明所用数据处理流水线和训练协议运行可靠。在这一紧凑的性能分布中，TimesNet 在所有指标上略高于其他模型，并凭借其多尺度卷积结构，在分类成功率上保持稳定优势，更有效地建模短期和长期模式。总体而言，验证和测试结果相互支持，表明模型具有较高的泛化能力，支持选择 TimesNet 作为参考模型。图 4 显示

了 TimesNet 模型训练过程的损失曲线和准确率曲线，以及验证和测试集的性能。图表显示，模型在 20 个时期内训练和验证损失均呈稳步下降，准确率值迅速提升。

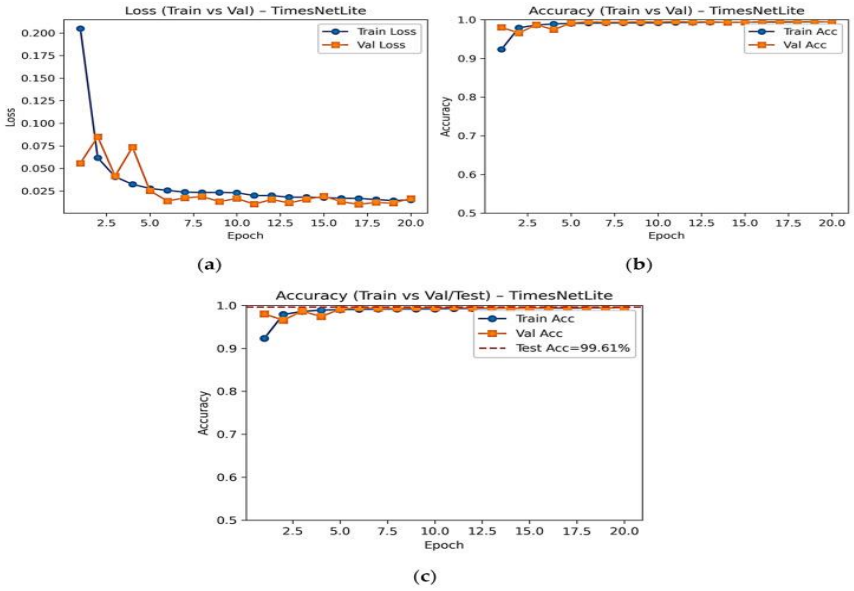

图 4 TimesNet 模型的训练与验证损失（a）、性能（b）以及训练－验证－测试准确率（c）曲线

图 5 详细说明了 TimesNet 模型在验证和测试数据集上的分类成功情况。验证集上的混淆矩阵展示了模型在学习过程中的泛化能力，而测试集矩阵则展示了实际性能在独立数据上的表现。两个矩阵的高准确率证明模型提供了稳定可靠的性能。

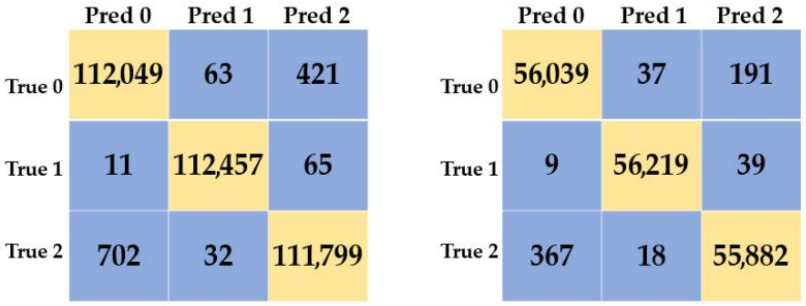

图 5 TimesNet 模型的验证集混淆矩阵（a），测试集混淆矩阵（b）

3.2. 可解释性分析

为提升模型的可解释性，采用了 SHAP（SHapley 加法解释）分析，以确定每个属性对模型预测的贡献，尤其是针对临床医生。图 6 展示了基于 TimesNet 模型 SHAP 值的属性重要性排序。

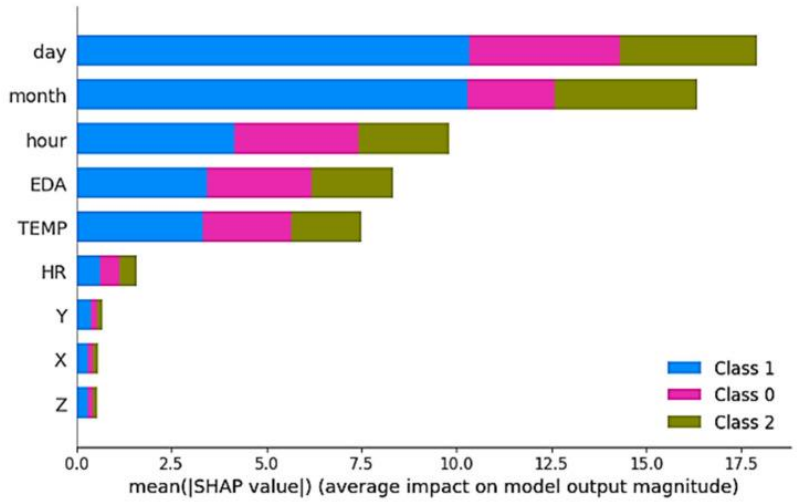

图 6 TimesNet 模型 SHAP 属性重要性排序

根据图 6，天、月和小时变量对时间维度影响最大，而生理参数，尤其是皮肤电导度（EDA）和温度（TEMP），在模型决策中起决定性作用。心率（HR）和加速度计轴（X、Y、Z）贡献较小。颜色显示属性对不同类别的影响（0 级-低压力，1 级-中等压力，2 级-高压力）。这些发现表明，模型的决策机制可以通过考虑环境特征和生理信号在应激预测中进行生物学解释。

图 7 显示了 TimesNet 模型在 0 类（低应力）SHAP 值下的属性重要性分布。虽然时间维度属性如日、月和小时很重要，但人们认为，尤其是生理信号的 EDA 和 TEMP 值对低应激分类有显著贡献。该类别的决策机制中，EDA 值较低的表现非常明显。

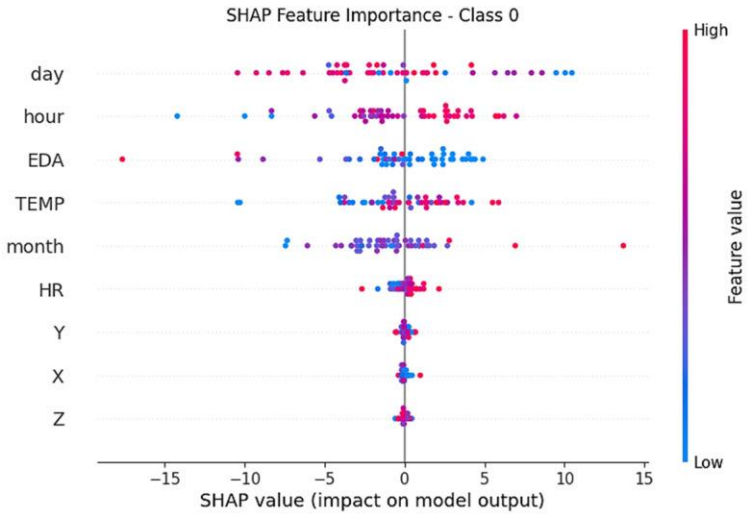

图 7 TimesNet 模型 0 类（低压力）的 SHAP 值

图 8 显示了 TimesNet 模型在 1 类（中等应力）SHAP 值下的属性重要性分布。在这门课中，时间属性（月份、日、小时）和生理属性（EDA、TEMP）共同起着决定性

的作用。尤其是高 EDA 和 TEMP 值对中度应激的分类产生积极影响。研究还观察到心率（HR）的部分贡献增加。

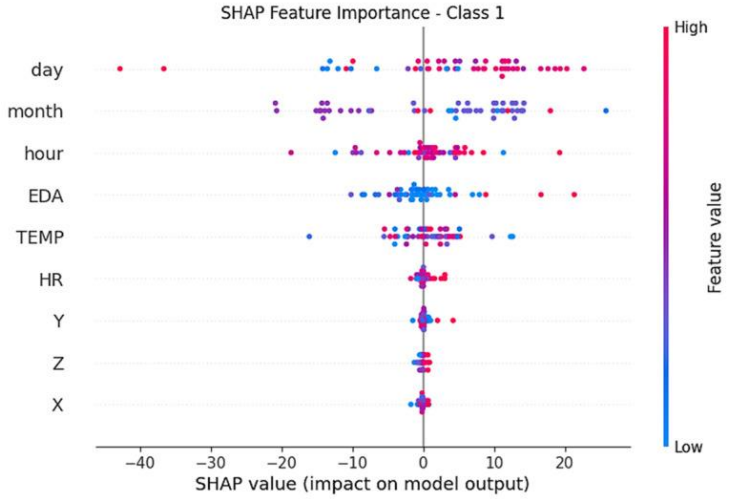

图 8 TimesNet 模型 1 类（中等应力）的 SHAP 值

图 9 显示了 TimesNet 模型在 2 级（高应力）SHAP 值下的属性重要性分布。在高压力类别中，EDA 和 TEMP 属性尤为突出。虽然时间因素（月、日、小时）也有效，但尤其是高 EDA 值对模型高应力类别预测有强烈影响。这种情况支持了皮肤导电率随着压力增加而增加的现象，这取决于生理基础。

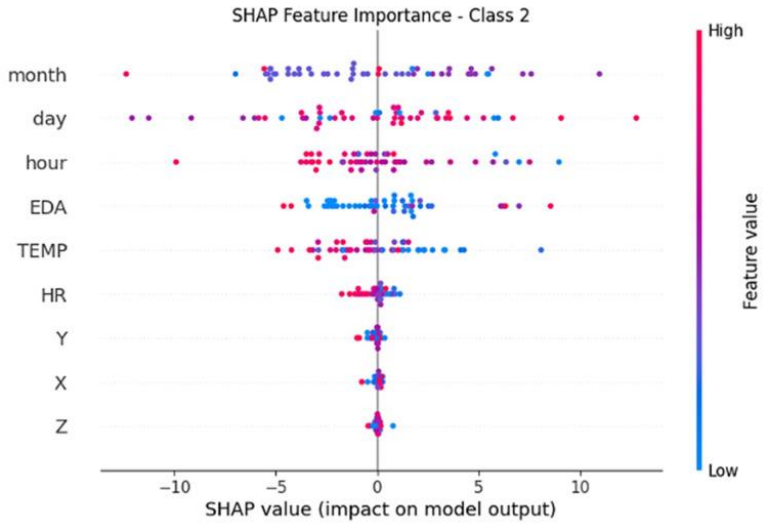

图 9 TimesNet 模型 2 类（高压力）的 SHAP 值

3.3. 区块链 + 分块 + 能源/成本分析

在我们的研究中，采用了基于区块链的日志机制来确保数据安全和完整性。区块链日志确保模型输出和数据块以不可篡改的方式记录，从而提高系统的可靠性并实

现追溯验证。表 5 展示了在拟议系统区块链集成过程中产生的 manifest hash、ECC 签名和时间戳信息。这些值确认每个区块的记录都可靠。

表 5 区块链集成过程中生成的清单哈希、ECC 签名和时间戳

参数	价值
manifest_hash	0824bab5176126c7e3fe9be96eb20163d7018a751610c70a30d6c7b102bfeelf
ecc_signature	42248d1FD75BC733D152bb28087fbb56EA374829945EA7C05CE4E1B378CB6BE
时间戳	1,759,346,521.7381518

在表 6 中，分析了区块大小对系统性能的影响，以保护数据完整性并确保成本效益。由于区块大小直接影响验证时间和重试率，需要评估不同的情景。根据表 6，观察到随着区块大小减小，验证时间增加，但成本（根据 Google Cloud 账户成本计算）保持不变。

表 6 不同区块大小的系统性能

区块大小	区块数量	重赛	重试比率	验证时间	上传成本（美元）
64	26,374	1318	0.049973	0.024687	0.002984
128	13,187	659	0.049973	0.013822	0.002984
256	6593	329	0.049971	0.006675	0.002984
512	3296	164	0.049757	0.003138	0.002984
1024	1648	82	0.049757	0.001657	0.002984

在表 7 中，进行了一项消融研究，以评估基于区块链的数据安全和重试机制对模型性能、训练时间、能耗和云计算成本的影响。研究以 TimesNet 为基础模型，评估了三种不同的情景。第一种情景中，模型仅由 TimesNet 组成。在第二种情景中，TimesNet 与区块链结构同时运行。在第三种情景中，除了区块链结构外，还使用了重试机制以减少数据传输的丢失。通过这种方式，系统分析了基于区块链的日志和重试方法在防止数据丢失时的开销和益处。

结果显示，区块链和重试机制对准确率的影响微乎其微，TimesNet 模型在所有场景下均达到 99%以上。尽管在区块链和重试场景中训练时间和能耗略有增加，但差距不到 1%。此外，基于云端的上传成本保持不变，而重试机制的额外成本非常低。因此，增加区块链和重试步骤以提升系统可靠性和数据完整性，在性能损失或成本负担方面几乎没有开销，是可行的。

表 7 在不同场景下测试研究中使用的 BC + 区块架构

剧情	准确率（%）	列车时间	平均威力	GPU 能量（J）	验证时间	重试比率	上传成本（美元）
时代网	99.5	4804.817	30.178	141,831.937	0	0	0.0029
时代网+BC	99.3	4822.324	30.131	142,009.955	0.016	0	0.0029
TimesNet+BC+重试	99.3	4852.261	30.062	142,642.564	0.017	0.049	0.0029

表 8 将比较扩展到架构维度，并评估不同研究在数据完整性、可解释性和系统安全性方面的方法。根据表格，文献中的大多数研究仅聚焦于存储效率或联邦学习安全。Chunk - RAID 方法在医疗领域提供基于区块链的完整性检查，但大多停留在数据存储层面，未涵盖数据传输和模型训练过程。我们提出的工作是首次弥合这一空白，实现分块区块链在边缘云架构中的整合。

表 8 拟议建筑与文献研究的比较

参考文献	数据集	方法	安全性/数据完整性	模型	赛
拟议工作	护士压力（IoMT 传感器）	Chunk - SHA256 - ECC，区块链日志，TimesNet	区块级完整性、ECC 签名、区块链清单	变形金刚（TimesNet）	SHAP
[37]	Linux/TREC	混合 CDC + 哈希去重	没有数据完整性，偏向压缩	-	-
[39]	链上数据集	分块哈希 + 显现组合	SHA-256 哈希 + manifest 验证	-	-
[40]	健康（电子健康记录）	分块 - RAID + AES 加密 + 区块链	AES-256 + RAID 完整性	基础机器学习	-
[41]	云存储	自适应压缩 + 高级分块	SHA-256 + ECC 验证	-	-
[60]	物联网传感器数据	区块链+联邦学习	FL 参数在区块链上注册	基于变压器的 FL	赛

该系统提供可靠的数据传输，具备块级数据完整性、基于 ECC 的数字签名和区块链日志，同时通过基于 TimesNet 的变压器模型实现高低能耗。此外，决策过程通过基于 SHAP 的可解释性分析实现透明，系统通过上传时间、重试成本和 GPU 能耗等参数量化系统的能量和成本平衡。总之，本研究提出了一个全面且创新的框架，将分区、区块链、边缘-云协作和可解释的人工智能整合为单一架构，实现安全数据传输，支持透明、可靠的临床决策。

4. 讨论与结论

本文提出了基于云-边缘-区块链分块的集成 AI 架构，用于基于 IoMT 的压力检测。结果显示，该系统不仅因其高精度，还因其安全、可追溯、节能且可解释的结构而与现有文献中存在差异。特别是，分块方法与区块链的整合使该方法（文献中通常仅用于数据复制或压缩）定位为一种安全的数据传输和验证机制。通过这种方式，数据传输过程中错误或缺失块的重试率降至低于 5%，并通过 ECC 签名和清单验证确保系统的完整性。在时间序列建模方面，基于转换器的 TimesNet 架构在捕捉短期应力波动和长期趋势方面优于其他转换器衍生工具（Informer、PatchTST、NST）。基于 SHAP 的可解释性分析使模型的决策机制具有临床可解性。这不仅在技术精度上具有创新性，也体现在基于生物学基础的 XAI 输出方面。

区块链与分块的整合并未对系统安全基础设施带来显著的性能负担。获得的能源和成本指标显示，区块链日志和重试机制仅使总能耗增加 0.5%，同时完全保证数据可靠性。这一结果表明，区块链分块组合可以在容错性低的系统（如健康数据传输）中成功实现高安全性/低成本的权衡。此外，该系统在云端架构上的适用性在降低延迟和维持能耗平衡方面具有显著优势。当研究文献中类似方法时，大多数研究关注存储效率或加密性能，却未涵盖数据传输、能源优化、可解释性和边缘云集成等整体因素。在这方面，本研究是首批将分块+区块链+边缘云+XAI 组件整合在单一框架中的实例之一，实现端到端安全、可解释且成本效益高的 IoMT 数据处理。

本研究存在某些方法论上的局限性，研究结果应在此框架内进行评估。所用的护士压力数据集来自单一机构，不同 IoMT 环境中传感器配置和用户行为的不同可能影响模型性能。能源和成本分析是通过模拟计算的，而非实际系统测量，因此可能无法完全反映实际应用中出现的变异性。还模拟了区块链和内容定义分区流程，以展示该架构的整体功能。此外，基于固定数据分离的模型验证无法完全反映传感器错误、数据异常和分布变化等真实 IoMT 流中可能发生的情况。未来研究计划利用来自多个 IoMT 来源的实时数据进行更全面的验证过程。在此框架下，当前研究结果支持所提系统的可靠性及其综合结构的有效性。总之，本研究为医疗领域开发可靠、可解释且节能的 IoMT 架构奠定了坚实基础。该系统有望在技术准确性和患者安全方面，开创下一代可靠的基于人工智能的医疗系统，应用于临床决策支持应用。基于

这些方向，旨在使所提系统不仅能在学术层面应用，也能应用于应用健康信息系统。尤其是结合联邦学习和基于智能合同的验证集成，该研究旨在发展为一个完全自主、隐私保护且可解释的数字孪生基础设施。