区块链赋能的可解释人工智能用于可信医疗系统

Md Talha Mohsin
Department of Finance & Operations Management
University of Tulsa
Tulsa, OK 74104, USA

摘要—本文介绍了一个集成区块链的可解释人工智能框架（BXHF），用于应对医疗保健系统中的两个关键挑战：安全的数据交换和易于理解的人工智能驱动的临床决策。我们的架构集成了区块链技术，确保患者记录不可篡改、可审计且防篡改，并结合了生成透明且具有临床相关性的模型预测的可解释人工智能（XAI）方法。通过将安全性保证和可解释性要求整合到统一的优化流程中，BXHF 确保了数据层面的信任（通过验证和加密记录共享实现）以及决策层面的信任（提供可审计并符合临床需求的解释）。其混合边缘-云架构允许在不同机构之间进行联邦计算，从而促进协作分析同时保护患者隐私。我们通过跨境临床研究网络、罕见疾病检测和高风险干预决策支持等用例来展示该框架的应用性。通过确保透明度、可审计性和法规遵从性，BXHF 提升了人工智能在医疗保健中的可信度、接受度和有效性，为更安全可靠的临床决策奠定了基础。

***Index Terms*—区块链，可解释的人工智能（XAI），医疗保健，临床决策支持，数据完整性，模型可解释性，可信人工智能。**

I. 介绍

人工智能（AI）在金融、医疗保健、零售和通信等多个行业中变得无处不在。在医疗保健领域，用户需求至关重要，AI 一直在推动各种应用程序的发展。这些包括疾病诊断和预后、个性化治疗建议、药物开发、提高运营效率等。利用大规模的患者记录、医学影像和基因组数据，AI 展示了识别人类专家通常无法察觉的细微模式的强大能力。然而，在医疗保健中广泛采用 AI 仍然面临两个持久的挑战：缺乏可解释性的人工智能驱动预测以及机构间共享医疗数据的完整性问题。医疗机构严重依赖跨机构的数据交换以提高诊断准确性并加速医学研究。但是，关于数据篡改、未经授权访问和利益相关者之间缺乏信任的担忧阻碍了有效合作。同时，许多尖端的 AI 算法作为“黑箱”操作，在不提供有意义解释的情况下做出正确的预测。这种缺乏透明度经常削弱临床医生的信任和用户友好性，并引发有关负责任的临床决策制定的伦理问题。此外，数据完整性也是当前医疗保健行业最为严重的问题之一 { [1]。通过以分散、安全和透明的方式处理敏感的医疗保健数据，区块链解决了数据安全性、互操作性和患者数据所有权问题 [2]。

区块链技术作为一种潜在的解决方案，用于安全地交换信息和保护数据完整性。该技术于 2008 年在加密货币比特币中首次引入，它是由带有时间戳的区块组成的一条链，并使用了诸如 [3] 等密码学哈希链接在一起。它继承了一些特性，如去中心化、透明度和匿名性 [4]。医疗保健行业以患者为中心的机制使其非常适合采用区块链技术 [5]，这可能有助于实现个性化、可靠且安全的医疗服务，[6]。此外，由于医疗保健组织的基础设施包含连接设备和软件应用程序，并与其它 IT 系统进行通信，它们也受到了区块链和物联网（IoT）使用的影响，[7]。通过在其系统中融入区块链技术和可解释人工智能（XAI），可以大幅改善一些医疗体系所面临的挑战。这些挑战包括药品完整性、临床试验效率、欺诈行为、推进个性化医疗 [8]、访问医疗数据的碎片化和缓慢性、系统兼容性和数据质量 [9]。

为了解决这些问题，我们提出了一种集成区块链的可解释人工智能（XAI）框架用于医疗数据交换。区块链作为一种去中心化、分布式且不可篡改的数字账本，允许机构之间安全、防篡改和可审计地互换病历，支持临床决策支持 (CDS) 系统。另一方面，XAI 使得基于 AI 的建议具有可解释性和临床验证性。通过结合这两个范式，该框架不仅增强了数据完整性，还提高了对 AI 辅助诊断和治疗的信心。

II. 背景及相关工作

A. 区块链在医疗保健中的应用

区块链是一种经过验证的去中心化和公开的数字账本，它使用加密技术在多个网络上记录交易，因此任何涉及的记录都不能在事后被篡改而不改变后续的区块 [10] [11]。区块链的一个关键属性是去中心化；由于没有中央权威控制添加到区块链的内容，传递到区块链的条目是在对等网络中达成共识的 [12]。其去中心化的账本有助于防止篡改记录保持，而不可变性、可追溯性和智能合约则提高了透明度并对访问权限进行了改进。由于需要以患者为中心的方法来处理医疗保健系统，并连接不同的系统以及增加电子健康记录（EHR）的准确性，区块链在医疗领域具有巨大的潜力 [3]，并且区块链在医疗领域的应用一直是众多研究的主题。区块链的去中心化特性、开放性和无需许可的特点为医疗提供了独特的解决方案 [10]，因为数据共享和访问是与市民健康记录相关的固有问题 [15]。区块链的一些关键属性，如不可变性、去中心化和透明度，可以潜在地解决医疗保健中的紧迫问题，包括患者对自己健康信息的访问以及在护理点不完整的记录等问题 [13]。区块链也可以用来规避传统医疗保健架构中的问题，以实现电子健康记录 [14] 的安全存储、共享和检索。

由于区块链是一种具有去中心化和防篡改特性的分布式架构，可以为保护个人健康记录共享系统提供一种新的、更好的方式 [16]。

鉴于在共享电子健康记录时隐私和安全的保存是至关重要的 [7]，将区块链整合到医疗技术中赋予患者对其个人信息更大的控制权；从而提高保密性和隐私性 [11]。

此外，区块链不仅将提供最大程度的隐私保护，还确保适当用户可以轻松地添加并访问永久信息记录 [17]。

B. 医疗领域的可解释人工智能

除了可靠的数据交换外，AI 模型的可解释性已成为临床应用的关键。虽然深度学习模型在医学成像、基因组学和风险分层中实现了出色的预测准确性，但它们的“黑箱”特性限制了医生的信任并引发了监管担忧。作为一个 AI 系统应该有能力在整个医疗路径 [18] 中承担责任，可解释的人工智能（XAI）通过特征归属、可视化、规则提取和反事实解释等方式来提供模型推理的见解从而解决这一问题。由于 XAI 的卓越效果，自然地，有一个日益增长的努力来增强临床医学中基于 AI 技术的信任并提高其接受度 [19]。将科学或因果解释的意义融入日常临床实践中最重要的原因之一是通过构建更稳健的世界模型来改进未来护理的巨大潜力 [20]。

尽管在区块链和 XAI 方面各自都有所发展，但将这两种范式整合到一个统一框架中的努力却很少。现有研究要么集中在安全、分散的数据共享上，要么集中在人工智能驱动预测的可解释性上，但没有两者兼顾。这种分割阻碍了真正值得信赖的医疗保健系统的发展，在这样的系统中，临床医生和机构可以同时依赖共享数据的完整性和 AI 辅助决策的透明度。

III. 提出的框架

我们提出了一种统一的**区块链 – 可解释人工智能医疗框架（BXHF）**，它同时确保了安全的数据来源和可解释的临床预测，以支持医疗系统中的临床决策支持（CDS）。与以往将基于区块链的医疗安全和可解释的人工智能视为孤立研究方向的做法不同，BXHF 建立了一个数学基础的管道，在该管道中，安全性保证和可解释性约束被共同编码到预测过程中。

A. 系统架构

BXHF 系统由五个相互依赖的层次组成，如下图 1 所示。

1) 数据层（加密患者记录）： 患者记录定义为：

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

其中 x_i 表示多模态医疗特征（EHR、影像、实验室结果），y_i 表示临床结果。每个 x_i 使用同态加密进行加密。由于框架禁止直接访问原始数据，计算是在隐私保护查询协议下进行的，从而确保保密性并符合 HIPAA 和 GDPR 等法规。

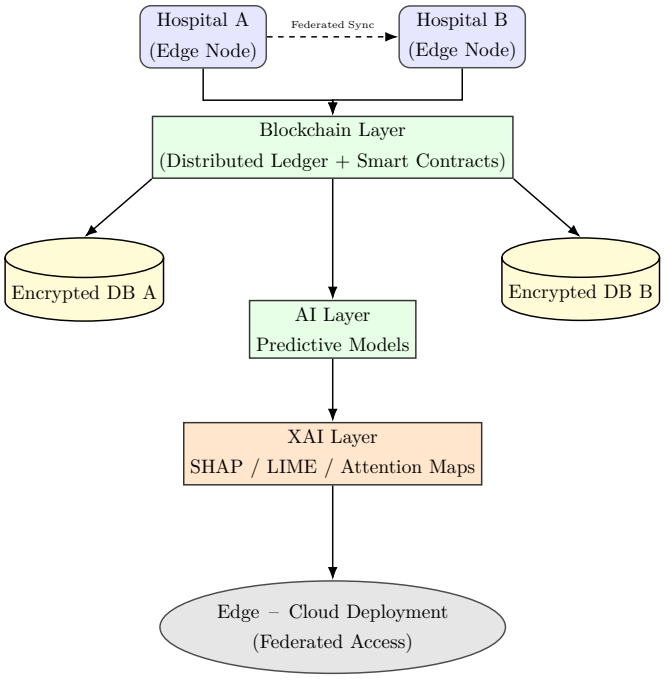

图 1. 区块链-XAI 医疗框架（BXHF）的工作流架构。

2) 区块链层（不可变审计和访问控制）： 分布式账本 L 维护加密数据标识符与访问事务之间的基于哈希的映射。智能合约 ϕ 被部署以强制执行访问策略：

$$\phi(u,d) = \begin{cases} 1, & \text{if user } u \text{ is authorized to access data } d \\ 0, & \text{otherwise.} \end{cases}$$

每次读/写操作都是永久可审计的，这在利益相关者之间建立了问责制和数据来源。

3) AI 层（预测建模）： 设 $f : X \to Y$ 是在安全数据检索下训练的预测模型。对于给定患者输入 x，预测结果为：

$$\hat{y} = f(x), \quad f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right]$$

其中 ℓ 是任务特定损失函数（例如，分类中的交叉熵或预后中的均方误差）。模型训练利用了联合节点，确保原始数据不离开机构边界。

4) 可解释人工智能层（形式化可解释性）： 解释函数 g 提供可理解的见解：

$$g : (x, f(x)) \mapsto E$$

其中，E 表示人类可理解的表现形式（例如特征归属分数 α_j、显著性图或基于规则的集合）。特征重要性必须满足：

$$\sum_j \alpha_j = f(x).$$

与事后方法不同，BXHF 引入了约束增强训练：

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}\big[\ell(f(x), y)\big] + \lambda \cdot \Omega(g(f(x)))$$

其中 Ω 惩罚缺乏临床合理性的解释。这确保了解释性在优化过程中被强制执行而不是之后附加的。

5) 部署层（混合边缘—云）： 敏感计算（加密、初步推理）在医院拥有的边缘设备上执行。大规模训练由联合云节点处理。区块链保证了节点之间的信任和一致性，消除了对中央权威的需求。

6) 集成信任保证（安全+可解释性）： BXHF 框架将区块链安全和可解释人工智能（XAI）整合到一个以数学为基础的目标中。该框架确保预测准确性、可解释性和数据安全性得到同步优化，而不是分别处理这些方面。

$$\mathcal{J}(f, D) = \min_{f \in \mathcal{F}} \underbrace{\mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]}_{\text{Loss}} + \underbrace{\lambda_1 \Omega(g(f(x)))}_{\text{Interp}} - \underbrace{\lambda_2 \mathcal{S}(D)}_{\text{Sec}},$$

其中：

- $f \in \mathcal{F}$ 是来自允许模型集合 $\mathcal{F}$ 的预测模型
- x 表示患者输入特征（电子健康记录、影像、实验室结果），
- y 代表相应的临床结果，
- $\ell(f(x), y)$ 是任务特定的预测损失函数，例如分类中的交叉熵或回归中的均方误差。
- $g(f(x))$ 是提供预测可解释性见解的解释函数，
- $\Omega(g(f(x)))$ 惩罚不一致或缺乏临床可信度的解释，
- $\mathcal{S}(D)$ 是一种安全措施，用于量化存储在区块链上的数据 D 的完整性、来源和可审计性。
- $\lambda_1 > 0$ 和 $\lambda_2 > 0$ 是权衡超参数，控制可解释性、安全性和预测性能之间的平衡。

此公式创建了一种双层信任机制：

i. **数据级信任：** 区块链确保所有患者数据的不可变性、可追溯性和可审计性。每次访问和更新都会被记录为基于哈希的交易，提供篡改证据来源。

ii. **决策级信任：** XAI 确保预测具有可解释性。这些解释本身与区块链加密绑定，保证它们生成后无法被篡改。

通过联合优化 $\mathcal{J}(f, D)$，BXHF 确保只有在同时满足可解释性和数据安全约束的条件下才能实现高预测性能。这从数学上强制执行了一个原则，即值得信赖的人工智能不仅需要准确的预测，还需要透明的推理和可验证的数据来源。

B. 工作流示例

为了证明所提出的集成区块链的可解释医疗框架（BXHF）的操作可行性，我们描述了一个工作流程，突出了其在数据交换、模型调用和解释完整性方面的多层方法。与线性管道不同，BXHF 创建了一个可审核且可验证的过程，将预测和解释链接到一个不可篡改的账本中。这为医疗决策创造了不可变的数据来源。

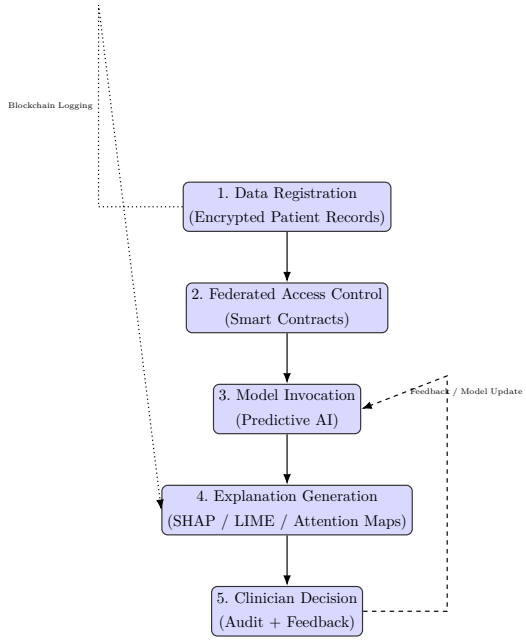

图 2. 用例流程 của BXHF

图 2展示了安全数据注册、联合访问、预测建模、可解释输出和临床医生反馈。区块链日志确保在多个阶段具有可审计性。**第一步：数据注册**
通过独特的加密哈希，所有患者记录，包括电子健康记录（EHRs）、医学影像和测试结果，都已登记在区块链上。
步骤 2：联合访问控制
当一个机构或医生希望访问患者信息时，智能合约强制执行基于同意且符合法规的标准（例如 HIPAA、GDPR）。联合访问通过限制分享给验证实体来保护敏感信息。
步骤 3：带有出处的模型调用
数据检索后，使用预测模型进行诊断或治疗建议。同时计算解释向量，如 SHAP 值、特征重要性分数或符号规则跟踪。预测和解释被加密绑定，并作为新的决策块记录在区块链上。
步骤 4：多机构验证
其他机构可以独立验证所使用的模型是否使用了正确的数据输入并且解释准确地与预测结果相关联。此阶段将解释从临时产物转换为可供联盟范围内审计的实体。
步骤 5：临床界面
临床医生通过特定界面获取框架的预测和可理解的理由。由于每个输出都包括区块链支持的完整性证书，因此确保了解释在事后没有被编辑。此外，框架的来源链保证了数据和 AI 驱动决策背后推理的信任。

例如，假设一名患者因疑似心脏问题被送入医院。实验室结果以及加密的心电图被上传到系统中。临床医生随后利用 BXHF；区块链记录查询日志、检查访问权限并提供加密的数据引用。模型然后预测心力衰竭，并提供可解释的说明：识别出较高的肌钙蛋白水平和异常的心电图信号是重要因素。账本随后记录了这些预测和解释，确保透明度和可审计性。

C. BXHF 的新颖性

BXHF 将区块链整合到医疗数据共享中，并采用可解释的人工智能（XAI）来提高模型透明度，从而整体推进这两个领域的发展。BXHF 的独特之处在于它在数据和决策两个层面提供了双重信任。

- **链上解释完整性**：我们架构的内在质量确保记录中包含模型生成解释的加密哈希以及预测。用户可以检查这些解释，以确保它们有效、未被篡改，并且与模型的决策过程一致。
- **两层信任机制**：BXHF 提出了一种两层信任范式：
  - **数据级信任：**框架中的区块链方面通过不可变性、来源追踪和受监管的敏感医疗数据共享为用户提供数据级别的信任。
  - **决策层面的信任：**模块的人工智能可解释性方面使得模型预测结果可以被解读，而 BXHF 链上存储的解释证明确保了所有利益相关者都能验证决策理由。

- **联合安全-可解释性形式主义**：BXHF 具有联合安全可解释性形式化方法，将安全性和可解释性集成到单一优化管道中，而不是作为单独的问题处理。
- **通过设计实现监管一致性**：随着可审计性被编织进每个模型输出和解释的决策过程中，它生成了一条区块链轨迹，确保了包括 HIPAA、GDPR 和 FDA 指南在内的透明监管合规性。
- **不可变说明来源**：由于所有解释都保存在区块链上并可审核，这确保了禁止任何对事后解释输出的操纵。

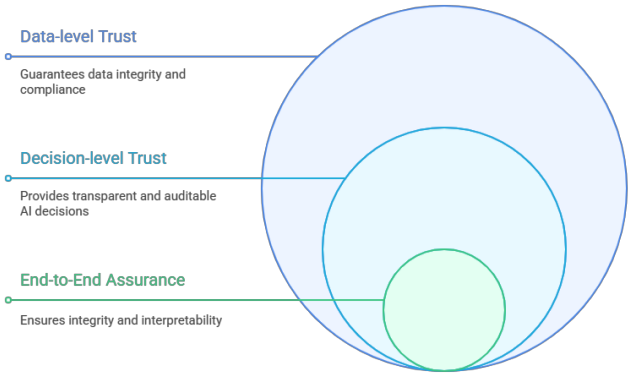

图 3. 两层信任

IV. 关键特点和优势

我们提出的集成区块链的可解释人工智能框架（BXHF）具有多种优势，可以缓解医疗数据共享和 AI 应用中的一些关键挑战：

- **安全且防篡改的数据共享**
  区块链确保跨机构交换的所有患者数据不可篡改、可追溯和可审计。每次交易都会被记录，这可以防止未经授权的修改和数据泄露。
- **透明和可解释的 AI 预测**
  系统提供的不是不透明的结果，而是诸如哪些临床变量影响了决策等解释，以便医生可以判断建议是否合理。
- **双层信任**
  该框架支持对基础记录的准确性以及模型自身的推理都充满信心。大多数现有的系统只解决其中一个问题，但 BXHF 将两者结合起来（图 3）。
- **支持合规性**
  由于所有的预测及其理由都被记录下来，该框架自然生成了一个可审计的历史记录，这有助于遵守 HIPAA 或 GDPR 等法规。
- **部署的可扩展性**
  混合边缘-云架构允许敏感计算在本地进行，从而减少延迟并保护隐私，同时利用云端节点进行大规模模型训练和跨机构协作。此设计平衡了效率、安全性和灵活性。
- **互操作性**
  通过实现数据和模型输出的安全交换，该框架使得医院或研究中心更容易共同开展更大规模的研究或共享临床试验。
- **临床可靠性和患者安全**
  将可验证的记录与可解释的建议相结合，可以降低由于数据损坏或无法解释的模型输出而导致错误的风险。

V. 用例

集成区块链的可解释人工智能框架（BXHF）旨在适应并适用于需要数据完整性和可解释性的多种医疗场景。以下示例展示了其潜在影响：

i. **多家医院合作诊断罕见疾病**
   来自多家医院的患者数据经常被合并，以获得对罕见疾病具有统计显著性的见解。BXHF 促进了机构之间敏感患者记录的安全共享，同时确保由 AI 生成的诊断预测是可解释的。临床医生可以验证输入数据并了解模型为什么推荐特定诊断的原因，这增强了跨机构决策的信心。

ii. **全球临床研究合作**．我们的框架将允许跨越国界的科研合作。由于临床数据共享经常受到不同法规的制约，BXHF 的混合边缘云架构搭配区块链技术将确保数据来源和访问合规；XAI 将确保人工智能模型输出可审计且标准化，这将使研究人员能够评估结果、重复分析并增强对研究的信任。

iii. **联邦学习用于预测分析**
   BXHF 将使医院能够在不将其患者数据移出自身系统的情况下共同训练模型。每家医院对其记录保持完全控制，而区块链层跟踪更新来源和谁有权限使用这些数据。XAI 层然后帮助医生和研究人员评估这些模型的结果及其背后的推理过程，从而使这些更新在临床上有用且值得信赖。

iv. **为高风险干预提供临床决策支持**在器官移植、肿瘤治疗和重症监护等高风险治疗中，数据完整性和模型可解释性至关重要，BXHF 框架可以安全地整合患者病史、影像和实验室检查结果，同时还能解释基于 AI 的风险估计并减少可能危及患者安全的错误可能性。

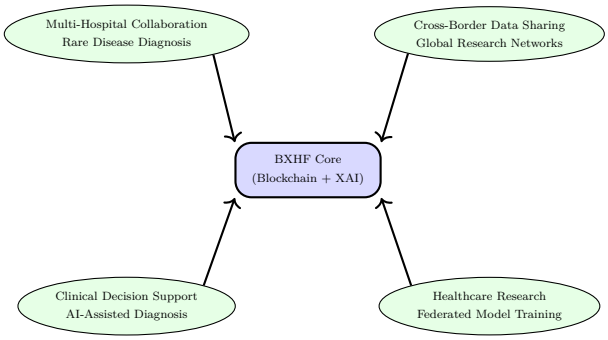

图 4. BXHF 框架的应用案例展示了在多个医疗场景中安全且可解释的人工智能应用。

VI. 结论

区块链集成的可解释人工智能框架（BXHF）解决了医疗保健中两个重要的挑战：安全的数据共享和可解释的人工智能驱动的临床判断。BXHF 在促进便捷的医疗合作和决策方面具有巨大的潜力。由于互操作性问题、用户界面复杂性和数据安全性顾虑等问题阻碍了现有的电子健康记录系统 [21]，我们的框架促进了多医院数据交换、远程医疗、国际研究伙伴关系以及关键临床干预，同时确保遵守法规并保障患者安全。该框架的灵活设计使其能够在需要对数据完整性和人工智能可解释性双重信任的其他领域实施。

总结来说，BXHF 表明在医疗领域可以实现安全、可解释和可审计的 AI 系统，为可靠的人工智能实施、改善患者结果和加强机构间合作奠定了基础。框架的多层架构促进了数据和决策层面的强大信任。区块链的颠覆性特征也将对医疗市场中不同参与者之间的权力平衡产生积极影响 [5]。后续的实施和试点研究将进一步证实该框架在实际临床环境中实用性和可扩展性的优势。

参考文献

[1] M. Zarour, M. Alenezi, M. T. J. Ansari, A. K. Pandey, M. Ahmad, A. Agrawal, R. Kumar, and R. A. Khan, "Ensuring data integrity of healthcare information in the era of digital health," *Healthcare Technology Letters*, vol. 8, no. 3, pp. 66–77, 2021.

[2] T. Mazhar, S. Khan, T. Shahzad, M. A. Khan, M. M. Saeed, J. B. Awotunde, and H. Hamam, "Generative AI, IoT, and blockchain in healthcare: Application, issues, and solutions," *Discover Internet of Things*, vol. 5, no. 1, p. 5, 2025.

[3] M. Hölbl, M. Kompara, A. Kamišalić, and L. N. Zlatolas, "A systematic review of the use of blockchain in healthcare," *MDPI*, vol. 10, 2018.

[4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[5] M. Mettler, "Blockchain technology in healthcare: The revolution starts here," in *Proc. IEEE 18th Int. Conf. e-Health Networking, Applications and Services (Healthcom)*, 2016, pp. 1–3.

[6] A. A. Siyal, A. Z. Junejo, M. Zawish, K. Ahmed, A. Khalil, and G. Soursou, "Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives," *MDPI*, vol. 3, 2019.

[7] A. Farouk, A. Alahmadi, S. Ghose, and A. Mashatan, "Blockchain platform for industrial healthcare: Vision and future opportunities," vol. 154, Elsevier, 2020.

[8] H. Omidian, "Synergizing blockchain and artificial intelligence to enhance healthcare," vol. 29, Elsevier, 2024.

[9] "MedRec: Using blockchain for medical data access and permission management," IEEE Xplore, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/7573685

[10] M. Prokofieva and S. J. Miah, "Blockchain in healthcare," vol. 23, Australasian Association for Information Systems, 2019.

[11] A. Haleem, M. Javaid, R. P. Singh, R. Suman, and S. Rab, "Blockchain technology applications in healthcare: An overview," vol. 2, Elsevier, 2021.

[12] A. Hasselgren, K. Kralevska, D. Gligoroski, S. A. Pedersen, and A. Faxvaag, "Blockchain in healthcare and health sciences—a scoping review," vol. 134, Elsevier, 2020.

[13] P. Zhang, D. C. Schmidt, J. White, and G. Lenz, "Blockchain technology use cases in healthcare," in *Advances in Computers*, vol. 111, Elsevier, pp. 1–41, 2018.

[14] J. Jayabalan and N. Jeyanthi, "Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy," vol. 164, Elsevier, 2022.

[15] A. Roehrs, C. A. da Costa, and R. da Rosa Righi, "OmniPHR: A distributed architecture model to integrate personal health records," *Journal of Biomedical Informatics*, vol. 71, pp. 70–81, 2017.

[16] S. Wang, D. Zhang, and Y. Zhang, "Blockchain-based personal health records sharing scheme with data integrity verifiable," vol. 7, IEEE, 2019.

[17] M. A. Engelhardt, "Hitching healthcare to the chain: An introduction to blockchain technology in the healthcare sector," vol. 7, 2017.

[18] R. Procter, P. Tolmie, and M. Rouncefield, "Holding AI to account: Challenges for the delivery of trustworthy AI in healthcare," *ACM Trans. Computer-Human Interaction*, vol. 30, no. 2, pp. 1–34, 2023.

[19] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, and D. Goldgof, "Explainable AI for medical data: Current methods, limitations, and future directions," ACM Computing Surveys, vol. 57, no. 6, pp. 1–46, 2025.

[20] R. L. Pierce, W. Van Biesen, D. Van Cauwenberge, J. Decruyenaere, and S. Sterckx, "Explainability in medicine in an era of AI-based clinical decision support systems," vol. 13, Frontiers Media SA, 2022.

[21] N. Rathore, A. Kumari, M. Patel, A. Chudasama, D. Bhalani, S. Tanwar, and A. Alabdulatif, "Synergy of AI and blockchain to secure electronic healthcare records," *Security and Privacy*, vol. 8, no. 1, e463, 2025.