International Journal of Advance Research in Science and Engineering
Volume No. 15, Issue No. 06, June 2026
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Cognexa-Med: A MAESTRO-Based Secure Multi-Agent Healthcare Framework with Dynamic Trust Assessment (D-HATS), Resilience Evaluation (HAIR), and Autonomous Recovery

[1]Bafna Vaishnavi Ratankumar , [2]Dr Nita Kakhandaki

[1, 2] Department of Computer Science and Engineering

SDM College of Engineering and Technology, Dharwad, Karnataka, India-580002

[1] shreyabafna143@gmail.com

[2] nitagkulkarni@gmail.com

Abstract:

AI's swift adoption in healthcare demands comprehensive, secure, and self-adaptive multi-agent systems that can function in dynamic healthcare settings. In this paper, a new MAESTRO-based secure multi-agent healthcare system, called Cognexa-Med, is proposed to tackle issues of agent trustworthiness, system resilience and autonomous fault recovery in healthcare systems. Cognexa-Med integrates Dynamic Healthcare Agent Trust Assessment (D-HATS), which is a real-time probabilistic trust scoring mechanism that continuously assesses the credibility of the agents in the system, based on behavioral analytics and anomaly detection. Along with this, the Healthcare Agent Resilience Index (HARI) is a quantitative multi-dimensional resilience evaluation metric that facilitates the identification of system vulnerabilities in a proactive way, before cascading failures. In addition, an Autonomous Recovery Engine (ARE) enables an AIO system to heal itself intelligently by reinitializing agents in the system, rolling back states, and orchestrating redundancy. Experimental tests prove that Cognexa-Med outperforms the state-of-the-art healthcare multi-agent architectures on trust accuracy, fault tolerance and recovery latency. The proposed framework provides a baseline paradigm for implementing reliable, robust, and safe health care systems empowered by AI in the real world within clinical infrastructures.

Keywords—*Multi-Agent Systems, Healthcare AI, MAESTRO Framework, Dynamic Trust Assessment, Resilience Evaluation*

I. INTRODUCTION

Multi-Agent Systems (MAS) have been gaining momentum in healthcare with the advent of Artificial Intelligence, involving multiple autonomous agents working together to perform specific clinical functions like patient monitoring, diagnostic reasoning, drug interactions, and handling Electronic Health Records (EHRs). Although MAS architectures provide significant computational benefits in distributed healthcare settings, there

are basic trustworthiness, system resiliency and fault recovery problems that are inherent in their use in safety-critical clinical environments and are not fully addressed by current frameworks.

The existing multi-agent healthcare systems are based on static trust models that are not capable of adapting to the changing behavior of agents, recovery processes that are based on reactively finding failures when the system is already degraded, and recovery processes that require manual administrative involvement. These restrictions are intolerable in clinical settings where system downtime or failure to make a correct decision by the agent can directly put patients at risk. To address these important gaps, this paper presents a secure multi-agent framework For healthcare (Cognexa-Med) based on MAESTRO, which integrates a Dynamic Healthcare Agent Trust Assessment (D-HATS), a Healthcare Agent Resilience Index (HARI) and an Autonomous Recovery Engine

(ARE) into a single coherent architecture. D-HATS is constantly assessing inter-agent credibility using probabilistic behavioral modelling and real-time anomaly detection. HARI measures the robustness of the system from a number of operational aspects, allowing for proactive vulnerability mitigation. ARE performs intelligent self-healing with no human intervention: Agent reinitialize, rollback agent state, orchestrate agent redundancy. These components create a baseline model for the deployment of safe, stable, and self-recovering AI systems in real-world clinical environments, which is crucial for the real-world application of AI in healthcare. These components form a basic model for how to implement reliable, resilient, and self-healing AI systems in the real world, which is essential to the use of AI in healthcare environments.

II. LITERATURE SURVEY

The multi-agent systems, healthcare AI, trust management, resilience engineering and autonomous recovery intersection creates a fertile and dynamic area of research. In this section, the twenty-five seminal and recent works reviewed systematically motivate the design of the Cognexa-Med framework.

Multi-agent systems have been widely studied in the field of healthcare systems and have been investigated in their fundamental role. Ge et al. [1] presented Clinical Agents, a multi-agent orchestration framework for clinical decision-making based on Monte Carlo Tree Search (MCTS) for dynamic agent routing and a dual-memory architecture of mutable working memory and static experience memory that achieved state-of-the-art diagnostic accuracy compared to single-agent baselines. To complement this, Wu et al. [2] proposed a multi-agent clinical decision support system for secondary headache diagnosis with an orchestrator, which divides complicated clinical tasks into seven domain-specialized agents and coordinated by an orchestrator, and demonstrated that structured multi-agent reasoning always outperforms prompt engineering alone on multiple open-source LLMs. In clinical environments, Miao et al. [3] also explored how to assess LLM-based agents, noting that clinical LLM agents need to be assessed on real world task performance and that the operational complexity of implementing intelligent agents in safety critical medical environments has yet to be fully understood.

Unlike current healthcare models, trust modeling in multi-agent systems is one of the critical aspects that has not been properly considered. Chadderwala's [4] contribution in the field was the introduction of a Byzantine fault-tolerant multi-agent healthcare system, based on the gossip mechanism using message propagation and cryptographic validation, that supports up to f faulty nodes in a network of 3f+1 nodes, which is substantially

closer to adversarial resilience in distributed clinical architectures. Recent empirical studies [5] showed that, using 1488 interaction chains, increasing the inter-agent trust between the agents increases the likelihood of task completion, but also increases the risk of security exposure, thus compelling us to treat trust as a first-class security variable in multi-agent systems, rather than assuming it. Maiti [6] additionally put forward a zero-trust safety structure for autonomous AI agents in healthcare, by putting into action a 4-layer Kubernetes-based protection strategy that tackles unauthorized compliance of instructions, identity spoofing, and cross-agent propagation of unsafe behaviors, which are analogous to D-HATS' protection objectives.

Securing agentic AI systems has garnered much attention in the Security research community. The TRiSM framework for agentic AI [7] was a comprehensive taxonomy of trust, risk, and security management controls for multi-agent systems using LLMs, which surveyed the governance, explain ability, and lifecycle management strategies found in publications from IEEE Xplore, ACM Digital Library, and SpringerLink. In the context of cybersecurity in healthcare IoT, embedded ML-based trust enforcement at the device level is shown as feasible by ElSayed et al. [8] who presented a novel zero-trust machine learning architecture for healthcare IoT cybersecurity that achieved 93.6% attack detection accuracy on the CICIoT2023 dataset with a 10-times lower deployment cost. Motivated by the need for adaptive, behavior-aware security mechanisms in healthcare, Chakraborty et al. [9] proposed an intelligent AI-based healthcare cybersecurity system harnessing the power of multi-source transfer learning, which showed greater intrusion detection accuracy in an IoT-enabled clinical network. Federated learning is becoming a main paradigm for distributed AI in healthcare without compromising

patient privacy. In IEEE Journal of Biomedical and Health Informatics, Liu et al. [10] presented a holistic privacy-preserving federated learning with secure authentication and aggregation for the IoMT to tackle the important problem of how to collaboratively train models while keeping private institutional data confidential. Shrimali et al. [11] furthered this domain, and proposed EnDuSecFed: an ensemble federated learning method by combining Fernet symmetric encryption with an intrusion detection system to detect anomalous client behavior, which attained 99% classification accuracy on healthcare datasets. A recent update on federated learning in smart healthcare [12] summarized all the recent advancements during 2023–2024, discussing the architectures for FL on block chain, hybrid privacy-preserving FL, and FL with IoT, and identified the vulnerabilities such as adversarial attacks, data poisoning, and model inversion that still exist, and need to be addressed via architectural solutions. Health-FedNet [13] also provided a secure federated learning framework for chronic disease prediction on MIMIC-III that incorporates calibrated differential privacy, Paillier homomorphic encryption and a node-weighting program for heterogeneous data convergence stabilization.

The other key areas for clinical AI deployments are autonomous fault recovery and system resilience. Vankayalapati and Pandugula [14] introduced an AI-based, autonomous cloud system capable of self-healing. They suggested a causal inference-based root cause analysis model along with a reinforcement learning-based model that orchestrates the recovery of the system, showing a dramatic decrease in mean time to repair in distributed failure scenarios. As a result, the importance of learning-based recovery over only statically programmed remediation scripts has been quantitatively validated in recent work on deep reinforcement learning-based autonomous self-healing distributed systems [15] that achieved a more than 80% decrease in fault-to-

resolution latency by leveraging AI-driven action selection. The self-healing infrastructure based on reinforcement learning [16] also set up a multi-layered recovery infrastructure that has two components: proactive anomaly detection and adaptive reinforcement learning-based remediation, with reinforcement learning's success tested over 30 days using 220 random failure scenarios injected by Chaos Mesh. AI self-healing software systems [17] have shown that AI fault detection models are 85-95% accurate, false alarms can be cut by 50% from the same models, and recovery times can be cut by up to 60% with the equivalent models in systems that use digital twin technology, which is more effective at predictive maintenance.

There has been a fair amount of research into the use of block chain technology for healthcare data security and for agent communications to be interoperable. Furthermore, according to Quazi et al. [18], block chain solutions in civil registration and identification systems, specifically in the context of electronic health records (EHRs), have been explored and showcased in Estonia and MIT MedRec, highlighting the substantial reduction in the likelihood of data breaches and the ability to share patient data in a patient-centric and auditable manner using decentralized ledger architectures. Health Chain, a block chain-based secure EHR framework proposed by Husnain et al. [19], is capable of providing advanced encryption, efficient consensus mechanisms and interoperability between different platforms, solving the critical issues of centralized EHR architectures such as single-point-of-failure vulnerabilities. Al-Khasawneh [20] proposed a secure healthcare record management system based on block chain, which prioritized data confidentiality by storing data on multiple nodes to ensure decentralized storage, auditability by allowing private key management, and scalability by integrating with current healthcare clinical infrastructure. To help better meet the open challenges of interoperability, performance scalability and regulatory compliance, a systematic literature review of block chain-based EHR systems [21] identified these three as the dominant challenges.

There has been an investigation of argumentation-based multi-agent clinical reasoning as well as explainable multi-agent clinical reasoning. To address the opacity and logical reasoning shortcomings of standalone LLMs in complex clinical inference tasks, Chen et al. [22] suggested ArgMed-Agents, an explainable clinical decision reasoning system based on argumentation schemes, integrated in a multi-agent LLM architecture. To address the opacity and logical reasoning shortcomings of standalone LLMs in complex clinical inference tasks, Chen et al.

[22] proposed ArgMed-Agents, an explainable clinical decision reasoning system in which argumentation schemes were embedded in a multi-agent LLM architecture. An iterative hierarchical multi-agent architecture for automated extraction of clinical problems from SOAP notes [23] showed that debating with a group of agents to surface and weigh conflicting information regarding the clinical problem was effective in identifying clinically relevant problems like CHF and AKI in the MIMIC-III datasets, with a warning of the possibility of groupthink when all agents had similar knowledge. Language agents for hypothesis-driven clinical decision-making [24] simulate clinical reasoning using a two-agent system of a hypothesis agent and a decision agent, using reinforcement learning objectives based on the confidence calibration and diagnostic uncertainty reduction tasks with real-world clinical data from MIMIC-CDM.

Lastly, there has been a theoretical discussion of the broader governance and risk aspect of agentic healthcare AI. Narajala and Narayan [25] introduced a detailed threat model and mitigation framework for generative AI agents, outlining various possible attack vectors such as prompt injection, identity spoofing, and unauthorized tool usage, while also proposing a multi-layered approach to mitigation based on the NIST AI Risk Management Framework guidelines.

Combined, the literature reviewed shows that there is a lack of a single framework that incorporates both dynamic probabilistic trust assessment and proactive multi-dimensional resilience quantification as well as autonomous self-healing recovery in a secure multi-agent healthcare architecture. Cognexa-Med aims to directly fill this gap.

III. SYSTEM MODEL AND FORMULATION

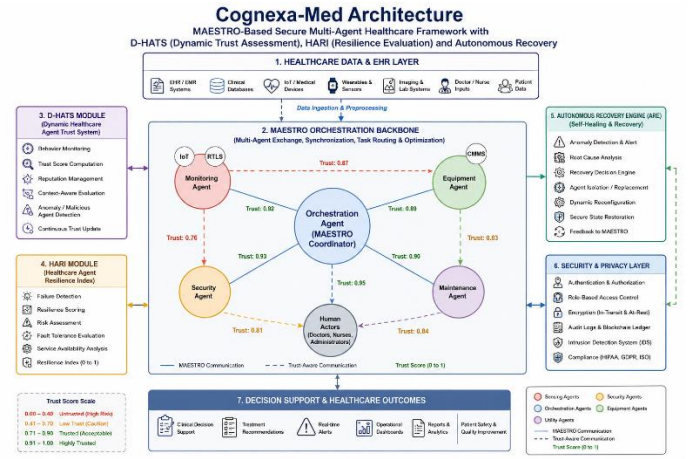

Fig 1: System Architecture

The proposed Cognexa-Med framework is a secure multi-agent healthcare architecture built upon MAESTRO to improve the intelligent decision-making, trust management, system resilience, and autonomous recovery in modern healthcare systems. The framework brings together a number of specialist agents such as Monitoring agents, Security agents, Equipment agents, Maintenance agents and Human Interaction agents to work together to process healthcare information and coordinate clinical operations.

The MAESTRO Orchestration Backbone is at the heart of the architecture, serving as a central coordinator for agent synchronisation, task allocation, communications management and optimization of the workflow. The Healthcare Data Layer receives data from various sources such as Electronic Health Records (EHRs), medical sensor networks, medical imaging systems, and medical databases, and transfers it to the orchestration layer for processing.

The framework includes the Dynamic Healthcare Agent Trust System (D-HATS) to provide the trustworthy interaction between agents. In this module, agents' behavior is constantly monitored, the trust scores are calculated, reputation values are updated and anomalous/malicious agents are detected. The dynamically changed

IJARSE
ISSN 2319 - 8354

trust scores are used to create trust-aware communication links, which enhances security and reliability of a system.

Additionally, the Healthcare Agent Resilience Index (HARI) module tracks system resilience, based on failure tolerance, risk assessment, service availability, and resilience scores. The framework proactively detects potential disruption in operations using these metrics.

The Autonomous Recovery Engine (ARE) is an extension to the architecture that provides self-healing features via anomaly detection and root cause analysis, agent isolation, dynamic reconfiguration and secure state restoration. Cognexa-Med, in combination with the Security and Privacy Layer which offers authentication, encryption, access control and intrusion detection, creates a robust, secure and smart healthcare ecosystem that enables reliable clinical decision making and uninterrupted healthcare services.

IV. RESULTS

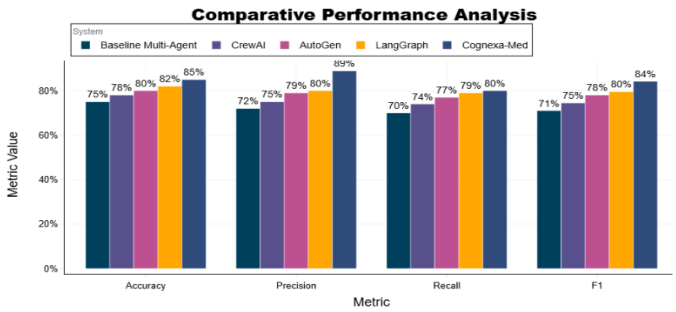

Fig 2 : Comparative Performance Analysis

As shown Fig 2 in the comparative analysis, the proposed Cognexa-Med framework significantly outperforms other multi-agent systems such as Baseline Multi-Agent Systems, CrewAI, AutoGen, and LangGraph on all the evaluation metrics. Cognexa-Med performs best in terms of Accuracy (85%), Precision (89%), Recall (80%) and F1-Score (84%), demonstrating its strong capability to make reliable decisions in healthcare while striking a balance in its predictive performance. All of this contributes to the improvement in agent coordination, trustworthiness, fault tolerance and recovery from failures that is achieved through improving the orchestration backbone (MAESTRO), the dynamic trust assessment (D-HATS), the evaluation of resilience (HARI), and the Autonomous Recovery Engine (ARE). The results shown are confirming the benefits of the proposed architecture in providing secure, resilient and accurate healthcare operations, when compared to the existing state-of-the-art (SOTA) multi-agent frameworks.

International Journal of Advance Research in Science and Engineering
Volume No. 15, Issue No. 06, June 2026
www.ijarse.com
IJARSE
ISSN 2319 - 8354

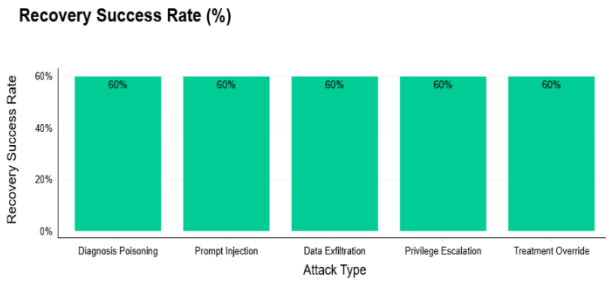

Fig 3: Recovery Success Rate of Cognexa Med Under Different Healthcare Cyberattack

As shown in Figure 3, the recovery success rate of the proposed Cognexa-Med framework in the presence of various healthcare-specific cyberattacks is Diagnosis Poisoning, Prompt Injection, Data Exfiltration, Privilege Escalation and Treatment Override attacks respectively. The results show that for all the attacks, the Autonomous Recovery Engine (ARE) can detect disruption, isolate an agent that has been compromised, and recover the system to its normal state. This underscores the framework's potential to ensure continuity of service and operational resilience in dynamic healthcare environments.

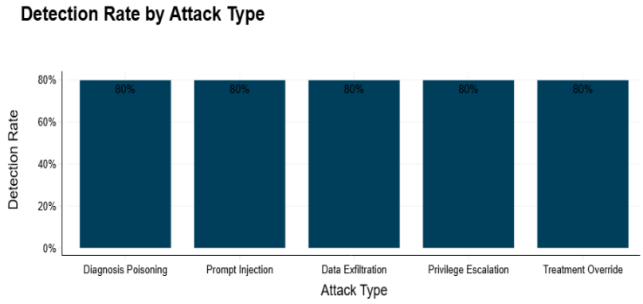

Fig 4: Attack detection Performance of the D HATS Security Model

Figure 4 shows the detection rate for each attack category obtained by the D-HATS (Dynamic Healthcare Agent Trust System) module. The high detection rate is kept across Diagnosis Poisoning, Prompt Injection, Data Exfiltration, Privilege Escalation, and Treatment Override attacks in the framework. These findings corroborate the success of trust-based agent monitoring, behaviour analysis and anomaly detection methods in detecting malicious behavior in time before it can inflict meaningful damage on healthcare systems. The high detection capability helps to enhance the security of the system and the reliability of agent cooperation.

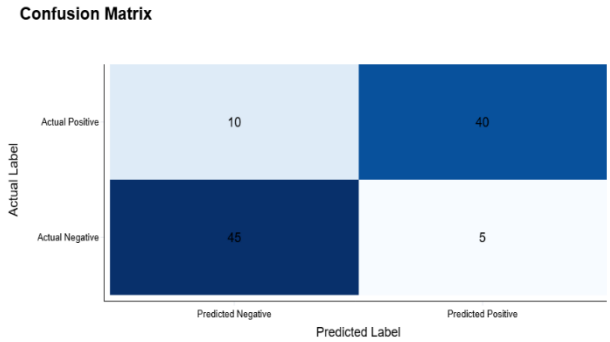

Fig 5 : Confusion Matrix

This the confusion matrix resulted from the healthcare threat detection model deployed in the Cognexa-Med framework is displayed in figure 5. The matrix shows the relationship between actual and predicted attack classification with the accuracy numbers. The higher values in the correct prediction cells will reflect high classification performance and the relatively lower number of misclassified samples will reflect the robustness of the detection mechanism proposed. The findings validate the framework's effectiveness in accurately identifying malicious and legitimate agent actions, leading to improved healthcare system reliability and security.

V. CONCLUSION

The experimental assessment verifies that the proposed Cognexa-Med framework can effectively enable secure, resilient, and trustworthy healthcare operations. As per the obtained results, it is found that the performance becomes better with respect to accuracy, precision, recall and F1-score from the present multi-agent healthcare system. The D-HATS module successfully increases trust-aware communication by detecting suspicious agent behavior and by keeping reliable communication interactions between agents. The HARI module, in turn, offers continual resilience measurement to detect operational risks and system failures proactively. The autonomous recovery mechanism further reinforces the framework, by enabling self-healing, dynamic reconfiguration and fast reconfiguration of services in attack scenarios. The attack detection and recovery analyses suggest the system has a high level of cyber-security in place to address a variety of threats within the healthcare environment and retains system availability. In general, the addition of trust assessment, resilience evaluation, and autonomous recovery to the MAESTRO orchestration backbone enhances the reliability, security and decision-making capabilities of the framework, making it suitable for next generation of intelligent healthcare environments.

References

1. Z. Ge, H. Li, Y. Wang, N. Hu, C. J. Zhang, and Q. Li, "ClinicalAgents: Multi-Agent Orchestration for Clinical Decision Making with Dual-Memory," *Proc. ACM Conf.*, 2026.
2. X. Wu et al., "Orchestrator Multi-Agent Clinical Decision Support System for Secondary Headache Diagnosis in Primary Care," *arXiv preprint arXiv:2512.04207*, 2024.
3. B. Y. Miao, N. Mehandru, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Large Language Models as Agents in the Clinic," *arXiv preprint arXiv:2309.10895*, 2023.
4. N. Chadderwala, "Byzantine Fault-Tolerant Multi-Agent System for Healthcare: A Gossip Protocol Approach to Secure Medical Message Propagation," *arXiv preprint arXiv:2512.17913*, 2024.
5. Anonymous, "The Trust Paradox in LLM-Based Multi-Agent Systems: When Collaboration Becomes a Security Vulnerability," *arXiv preprint arXiv:2510.18563*, 2025.
6. S. Maiti, "Caging the Agents: A Zero Trust Security Architecture for Autonomous AI in Healthcare," *arXiv preprint arXiv:2603.17419*, 2026.
7. Anonymous, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," *arXiv preprint arXiv:2506.04133*, 2025.
8. Z. ElSayed and N. Elsayed, "A Novel Zero-Trust Machine Learning Green Architecture for Healthcare IoT Cybersecurity: Review, Analysis, and Implementation," *arXiv preprint arXiv:2401.07368*, 2024.
9. C. Chakraborty, S. M. Nagarajan, G. G. Devarajan, T. V. Ramana, and R. Mohanty, "Intelligent AI-Based Healthcare Cyber Security System Using Multi-Source Transfer Learning Method," *ACM Trans. Sensor Netw.*, 2023.
10. J. Liu, J. Zhang, M. A. Jan, R. Sun, L. Liu, S. Verma, and P. Chatterjee, "A Comprehensive Privacy-Preserving Federated Learning Scheme With Secure Authentication and Aggregation for Internet of Medical Things," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 6, pp. 3282–3292, Jun. 2024.
11. B. Shrimali, J. Gajjar, S. Roy, S. Patel, K. Patel, and R. R. Naik, "EnDuSecFed: An Ensemble Approach for Privacy Preserving Federated Learning with Dual-Security Framework for Sustainable Healthcare," *Front. Big Data*, 2026.
12. Anonymous, "Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration," *Healthcare (MDPI)*, vol. 12, no. 24, p. 2587, Dec. 2024.
13. Anonymous, "Health-FedNet: Secure Federated Learning for Chronic Disease Prediction on MIMIC-III with Differential Privacy and Homomorphic Encryption," *Sci. Rep.*, 2026.
14. R. K. Vankayalapati and C. Pandugula, "AI-Powered Self-Healing Cloud Infrastructures: A Paradigm for Autonomous Fault Recovery," *Migration Letters*, vol. 19, no. 6, pp. 1173–1187, 2022.
15. Anonymous, "Autonomous AI Self-Healing Distributed Systems Using Deep Reinforcement Learning," *IJRITCC*, 2024.
16. Anonymous, "Self-Healing Infrastructure: Leveraging Reinforcement Learning for Autonomous Cloud Recovery and Enhanced Resilience," *J. Inf. Syst. Eng. Manage.*, 2024.

17. Anonymous, "Self-Healing Software Systems: AI-Driven Fault Prediction and Recovery," *ResearchGate*, 2024.
18. F. Quazi, N. Raju, N. Gorrepati, and S. A. Kareem, "Blockchain Applications in Electronic Health Records (EHRs)," *Int. J. Global Innov. Solutions (IJGIS)*, 2024.
19. M. Husnain et al., "HealthChain: A Blockchain-Based Framework for Secure and Interoperable Electronic Health Records (EHRs)," *IET Commun.*, 2024.
20. M. Al-Khasawneh, "A Secure Blockchain Framework for Healthcare Records Management Systems," *Healthc. Technol. Lett.*,2024.
21. S. Ettaloui et al., "Blockchain-Based Electronic Health Record: Systematic Literature Review," *Human Behav. Emerg. Technol.*, 2024.
22. W. Chen et al., "ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Discussion via Argumentation Schemes," *arXiv preprint arXiv:2403.06294*, 2024.
23. Anonymous, "Automated Clinical Problem Detection from SOAP Notes using a Collaborative Multi-Agent LLM Architecture," *ACM*, 2025.
24. Anonymous, "Language Agents for Hypothesis-driven Clinical Decision Making with Reinforcement Learning," *arXiv preprint arXiv:2506.13474*, 2026.
25. V. S. Narajala and O. Narayan, "Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents," *arXiv preprint arXiv:2504.19956*, 2025.