Journal of Intelligent Systems and Internet of Things
Vol. 18, No. 01, PP. 169-184, 2026
DOI: https://doi.org/10.54216/JISIoT.180112

Trustworthy and Interpretable AI in IoT-Based Medical Systems: A Review and Framework for CoT-XAI Integration

Faisal Binsar[1,2,*], Sasmoko[1,3]

[1]Professional Engineer Program Department, Faculty of Engineering, Bina Nusantara University, Jakarta 11480, Indonesia

[2]Faculty of Economics and Business, Muhammadiyah University Berau, 77311, Indonesia

[3]Research Interest Group in Education Technology, Bina Nusantara University, Jakarta 11480, Indonesia

Emails: faisal.binsar@binus.ac.id; sasmoko@binus.edu

Abstract

The use of Artificial Intelligence (AI) in medical diagnosis has rapidly evolved with the adoption of large language models and explainability techniques. This study investigates the intersection of Chain-of-Thought (CoT) reasoning and Explainable AI (XAI) in the development of trustworthy diagnostic systems, particularly within Internet of Things (IoT)-enabled healthcare environments. A systematic review of 106 Scopus-indexed publications (2016–2025) was conducted, supported by topic modeling (LDA) and keyword co-occurrence network analysis to identify dominant research themes and gaps. Findings reveal that while CoT and XAI are actively studied, their integration within real-time, distributed, and resource-constrained medical systems remains limited. Most research emphasizes either performance or interpretability in isolation, with minimal efforts to embed step-wise reasoning into deployable clinical AI pipelines. Moreover, few studies address how CoT can function effectively in edge computing or federated learning scenarios common to IoT infrastructures. To address this gap, we propose a multi-layered conceptual framework that integrates CoT reasoning, machine learning predictors, XAI methods, and IoT deployment models. This framework reflects the shift toward user-centric, transparent, and adaptive AI solutions in smart healthcare. It provides a structured path from multimodal data ingestion to clinically interpretable and real-time decision support. This study contributes a novel perspective on reasoning-driven explainability and offers design guidance for future development of interpretable, scalable, and deployable AI systems in medical applications.

Received: March 14, 2025 Revised: June 05, 2025 Accepted: July 14, 2025

Keywords: Explainable AI; CoT Reasoning; Natural Language Processing; Smart Medical Systems; Internet of Medical Things

1. Introduction

Artificial Intelligence (AI) continues to reshape medical diagnosis by offering advanced support through natural language understanding and decision automation [1]. Among the most transformative developments are large language models (LLMs) and transformer-based architectures like BERT, which have enabled AI systems to process complex unstructured medical texts such as clinical notes and patient histories [2], [3]. Despite their impressive capabilities, many AI systems—especially those built on deep learning architectures—remain largely black-box in nature [4]. Their internal decision-making processes are often inaccessible or incomprehensible to end users, clinicians, or stakeholders. This opacity poses significant challenges in medical domains where accountability, trust, and transparency are essential [5], [6]. As such, the lack of explainability in these models

continues to hinder their adoption in real-world healthcare settings, especially where regulatory and ethical requirements demand interpretable and justifiable decisions.

This limitation has catalyzed two parallel research directions: (1) Explainable AI (XAI), which seeks to illuminate how models arrive at decisions through feature attribution methods like SHAP and LIME [7], [8], and (2) Chain-of-Thought (CoT) prompting, a reasoning-based approach that encourages models to generate intermediate steps mimicking human logic [9]. Together, these directions reflect a broader push toward transparent, reasoning-driven AI systems aligned with clinical workflows and ethical accountability [6], [10].

While CoT and XAI have been explored separately, their combined application in intelligent medical systems remains underdeveloped. This is especially important in the context of the Internet of Things (IoT), where medical devices generate real-time, patient-centered data. AI models in such environments must not only infer accurately but explain decisions in ways interpretable to clinicians and patients alike [11]. Moreover, the integration of CoT reasoning into IoT frameworks—such as edge computing or federated learning—offers a promising yet underexplored direction for medical AI [12].

This paper aims to bridge this gap by providing a comprehensive review of the integration between explainability (XAI), reasoning (CoT), and smart medical systems. Unlike previous reviews that treat these topics separately, we synthesize current progress across these three domains and propose a conceptual framework to guide future research in reasoning-driven, explainable AI systems for IoT-enabled healthcare applications.

This paper makes the following contributions:

1. It presents a comprehensive synthesis of research trends at the intersection of CoT, XAI, and medical diagnosis within intelligent and IoT-based systems.
2. It proposes a conceptual framework that links CoT reasoning and XAI explanations to facilitate clinical trust and interpretability.
3. It identifies technical and research gaps, especially concerning real-time inference and integration with IoT architectures.
4. It offers future directions for applying CoT-based reasoning in edge AI, federated learning, and personalized medicine contexts.

The remainder of this paper is organized as follows. Section 2 provides a review of prior studies related to Chain-of-Thought (CoT) reasoning, Explainable Artificial Intelligence (XAI), and their relevance to IoT-based medical systems. Section 3 outlines the research methodology, including the systematic literature review process and the analytical techniques employed, such as topic modeling and keyword network analysis. Section 4 presents the key findings of the study, including bibliometric insights, thematic patterns, and a critical synthesis of seminal works, and introduces a new conceptual framework integrating CoT and XAI within smart healthcare environments. Section 5 offers concluding remarks, followed by Section 6, which discusses the theoretical and practical implications of the study. Finally, Section 7 highlights the limitations and proposes directions for future research.

2. Related Works

Recent advances in machine learning and deep learning have made significant strides in the automation of medical diagnosis, particularly through the processing of unstructured textual data from electronic health records [13], [14]. Pre-trained language models such as BioBERT and ClinicalBERT demonstrate notable performance in tasks like disease classification, symptom detection, and clinical decision support [2], [3]. These capabilities have propelled their use in both research and clinical settings.

However, the lack of interpretability in deep models remains a challenge for clinical acceptance [15]. To address this, XAI techniques have emerged, including LIME and SHAP, which provide insights into the contribution of input features to model predictions [16]. Visual attention maps have also been used to show which parts of clinical text influence decisions [17]. Despite these efforts, explanations generated by XAI often lack the narrative clarity and causal logic preferred by medical professionals [18].

CoT prompting offers an alternative approach, enabling AI to reason through step-by-step inference chains that resemble human logic [9], [12]. Though promising, its application in medical NLP remains limited. Existing studies apply CoT mostly in question-answering and logic-based reasoning, with limited integration into clinical

decision-making pipelines [19]. Moreover, CoT models still face challenges in transparency and user trust, as intermediate steps may not always align with clinical reasoning paths [20].

An important but often overlooked dimension in this context is the deployment of AI in IoT-enabled environments. The proliferation of smart medical devices—ranging from wearable sensors to cloud-connected diagnostic platforms—has increased the volume and velocity of healthcare data [1]. These systems require edge-based AI inference to support real-time diagnostics while maintaining explainability and privacy. However, few studies address how CoT reasoning and XAI can be adapted to these distributed and constrained environments. For example, how can a wearable ECG monitor integrate CoT explanations for irregular heartbeat detection? On the other hand, how might federated learning environments utilize CoT to maintain both interpretability and data locality?

Moreover, even reviews on XAI in healthcare often neglect deployment concerns. Oyebode et al. [11] stress the need for adaptive and personalized systems that offer clear feedback loops between AI decisions and patient data, which are essential in IoT contexts. Yet, few studies explore how XAI explanations can be generated or validated in edge computing scenarios.

In summary, existing literature reveals several blind spots:

1. CoT and XAI are rarely combined in a unified reasoning-explanation pipeline.
2. The role of these techniques in real-world intelligent systems, particularly IoT-enabled healthcare, is underexplored.
3. System-level frameworks to guide their integration are lacking.

This review seeks to bridge these gaps by synthesizing the literature across CoT reasoning, XAI methods, and IoT-based medical applications, and by proposing a framework that aligns reasoning logic, interpretability, and system deployment for trustworthy medical AI.

3. Methodology

This study employs a Systematic Literature Review (SLR) to examine the convergence of Explainable Artificial Intelligence (XAI), Chain-of-Thought (CoT) reasoning, and their integration within smart medical systems, especially in IoT-based healthcare environments. The research process follows five sequential stages as illustrated in Figure 1: Series of Research Methodology Stages, namely: Identification, Screening, Eligibility, Synthesis, and Discussion of Key Findings.

The research process began with formulating focused objectives and guiding questions aimed at understanding how reasoning (CoT) and interpretability (XAI) are addressed in the context of medical diagnosis systems enhanced by IoT technologies. This planning stage also defined the research scope and methodology protocol to ensure transparency and replicability.

The extracted data were analyzed using *Python* to uncover patterns, thematic structures, and collaborative dynamics within the literature [21]. Data preprocessing and manipulation were performed using Pandas, while visualization tools such as Matplotlib and Seaborn supported the generation of analytical graphics. Topic modeling is applied using an algorithm called Latent Dirichlet Allocation (LDA) [22] to understand the existence of frequently mentioned topics and themes. This technique enabled a structured exploration of the conceptual landscape surrounding CoT reasoning and XAI in medical diagnosis.

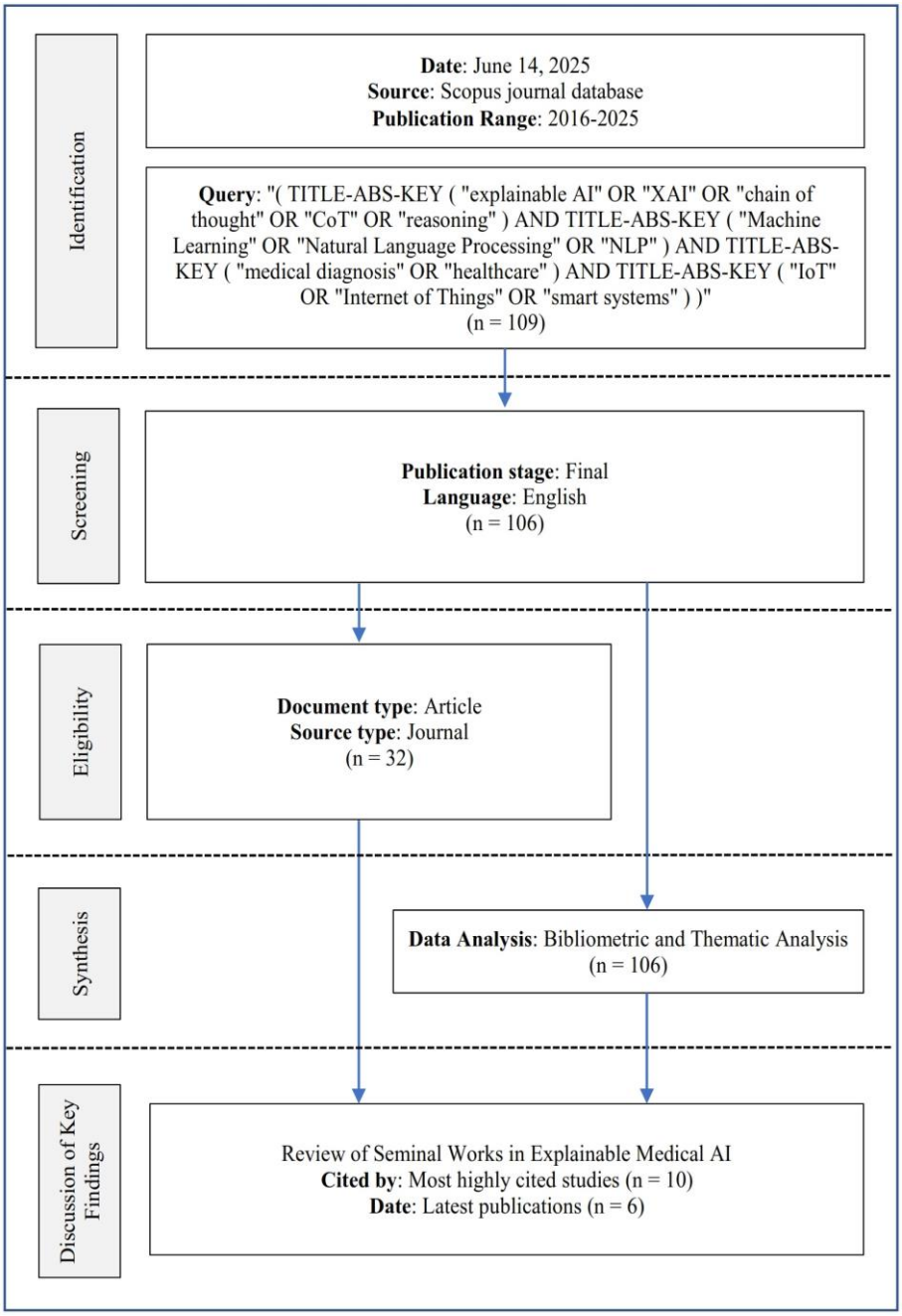

Figure 1. Series of Research Methodology Stages
(Source: Author's work)

A. Identification

The identification phase involved constructing a comprehensive query to extract relevant documents from the Scopus database. The updated search string included terms related to reasoning (*"explainable AI"*, *"XAI"*, *"chain of thought"*, *"reasoning"*), technical paradigms (*"machine learning"*, *"natural language processing"*), medical application domains (*"medical diagnosis"*, *"healthcare"*), and deployment platforms (*"IoT"*, *"smart systems"*, *"Internet of Things"*). The search covered publications from 2016 to 2025.

B. Screening

At the screening stage, initial filtering was conducted by selecting documents with the publication stage set to "*Final*" and language limited to "*English*" to ensure relevancy and quality. The screened documents were then subjected to early data analysis consisting of bibliometric mapping (using *co-occurrence* and *keyword network analysis*) and topic modeling (via *Latent Dirichlet Allocation*) to identify dominant research themes, emerging trends, and conceptual patterns relevant to the intersection of XAI, CoT, and smart medical diagnostics.

C. Eligibility

Following initial analysis, documents were further refined by applying eligibility criteria based on document type, where only journal articles were retained to ensure academic rigor. Additionally, the source type was restricted to peer-reviewed journals, thereby excluding conference proceedings, book chapters, or trade publications. This resulted in a final curated corpus suitable for in-depth synthesis and discussion.

D. Synthesis

To manage the extracted information, all data were organized in a structured spreadsheet using Microsoft Excel. This allowed for easy comparison across studies and consistent categorization of themes. The organized dataset served as the foundation for further analysis, helping to identify trends, methodological patterns, and research gaps related to reasoning and interpretability in AI-based medical diagnosis.

The selected articles were synthesized to uncover conceptual overlaps between explainability and reasoning frameworks, especially in the design and deployment of AI in medical IoT environments. This synthesis integrated thematic insights from the previous phase to outline the progression of research and to support the development of a unified framework that incorporates both CoT and XAI principles in smart medical systems.

E. Discussion of Key Findings

This stage presents a detailed thematic discussion of influential works. A Review of Seminal Works in Explainable Medical AI is conducted by selecting two categories of studies:

- Most highly cited studies to identify foundational contributions in the field.
- Latest publications (2025) to capture emerging innovations and trends.

Special attention is given to how these seminal works relate to real-time reasoning, explainability, and deployment in IoT-enabled systems. The discussion reflects on how the reviewed techniques could support practical implementations in intelligent healthcare platforms, sensor-based diagnostics, and federated or edge-computing environments.

4. Results and Discussion

The analysis of selected studies reveals emerging trends, methodological patterns, and research gaps. Emphasis is placed on how CoT enhances reasoning capabilities, while XAI contributes to transparency—together forming the foundation for the proposed conceptual framework.

A. Textual and Thematic Analysis

To understand the conceptual landscape and thematic development surrounding Explainable AI (XAI), Chain-of-Thought (CoT) reasoning, and their applications in smart medical systems, a textual and thematic analysis was conducted on 106 Scopus-indexed documents. Python-based tools and libraries were utilized to perform *Latent Dirichlet Allocation (LDA)* for topic modeling, and *co-occurrence network analysis* for identifying keyword proximities to trace emerging topic trends over time. This dual-method approach offers not only a semantic dissection of the literature but also an illustration of its interconnected topical architecture.

Figure 2 presents the results of Latent Dirichlet Allocation (LDA) topic modeling, which reveals the semantic structure of the corpus analyzed in this study. Among the five detected topics, Topic 1 dominates the discourse, accounting for 25.3% of the total token distribution. The most salient terms in this topic—*such as learning, using, data, system, model, healthcare, network, and detection*—highlight a thematic concentration on machine learning applications for intelligent and data-driven medical systems. The frequent appearance of keywords like *smart, technology, algorithm*, and *framework* indicates a growing focus on the architectural and algorithmic underpinnings of explainable AI (XAI) and reasoning techniques such as Chain-of-Thought (CoT), particularly when applied within healthcare infrastructures. The spatial separation of topics in the intertopic distance map, particularly the distinct positioning of Topic 1, suggests a relatively well-defined cluster, reinforcing its central role in synthesizing research on explainability, medical intelligence, and IoT-driven diagnostics [16], [23].

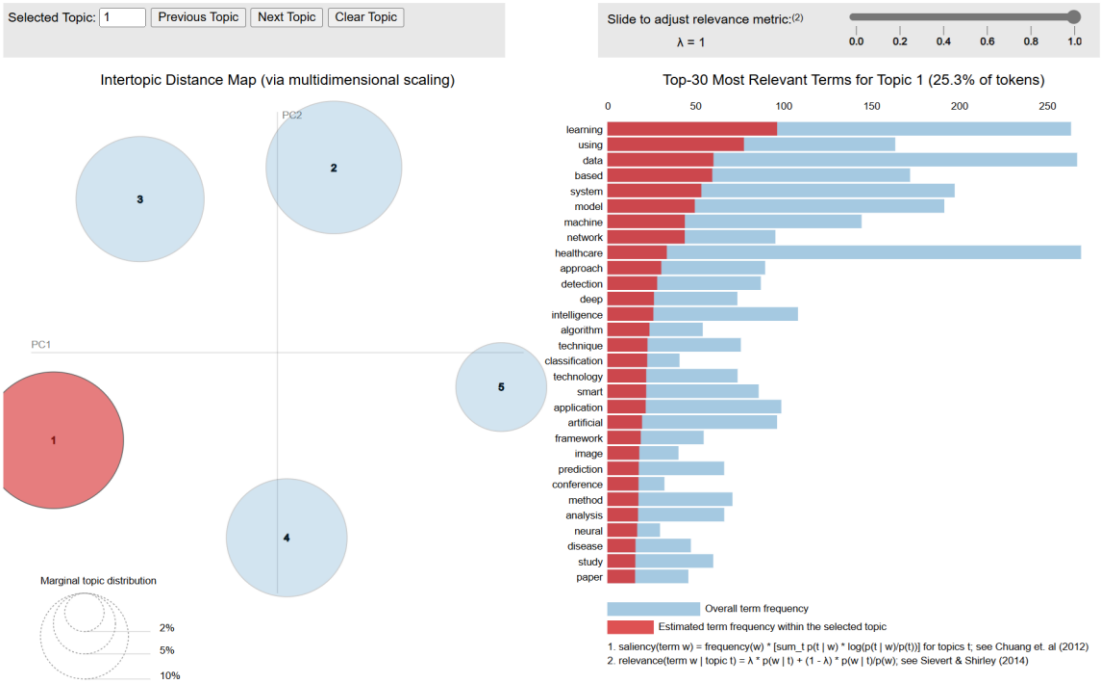

Figure 2. LDA topic modeling

(Source: Author's work)

The spatial distribution of topics in the intertopic distance map reflects a clear clustering pattern, suggesting that each theme maintains conceptual coherence while contributing uniquely to the collective narrative of explainable and intelligent healthcare systems. This suggests that while CoT reasoning and XAI remain theoretically discussed in some clusters; their practical convergence in smart medical systems is increasingly emphasized, albeit still emerging.

In parallel, the keyword co-occurrence network graph in Figure 3 provides a complementary perspective by visualizing semantic proximities and conceptual intersections among key terms in the corpus. Central nodes such as explainable AI, CoT reasoning, machine learning, medical diagnosis, and IoT occupy pivotal positions in the network—demonstrating their role as thematic anchors in this research space [24]. The graph reveals tightly linked sub clusters, such as:

- Interpretability cluster, featuring terms like LIME, SHAP, attention, and trustworthiness, indicating a strong focus on model transparency tools.
- Deployment cluster, featuring federated learning, sensor, data stream, and real-time, highlighting the rising importance of distributed computing and on-device inference in medical contexts.
- Clinical application cluster, where terms like diagnosis, clinical, symptoms, and monitoring converge—pointing to practical use cases of CoT and XAI in patient-centric scenarios.

This network supports the notion that the field is moving from siloed explorations of reasoning or explainability toward a more integrated and contextualized application, particularly in IoT-based diagnostics and edge computing. The network also reinforces about the untapped potential of applying CoT in federated learning environments—a connection implied in the proximity of CoT reasoning and federated learning in the graph, though not yet widely realized in current implementations.

Collectively, both visualizations—LDA topic modeling and keyword network analysis—underscore a growing alignment between CoT reasoning and XAI within intelligent medical infrastructures. However, they also surface gaps in implementation, especially in real-time reasoning, sensor integration, and data stream processing—as noted by both reviewers. These gaps affirm the need for more applied and system-level studies, moving beyond bibliometric trends toward technical prototypes and operational frameworks.

Thus, this thematic analysis not only validates the conceptual relevance of integrating CoT and XAI into medical IoT systems but also pinpoints actionable directions for future research—such as deploying CoT reasoning in resource-constrained environments (e.g., wearables, mobile health platforms) and evaluating explainability techniques in clinician-facing interfaces for decision support.

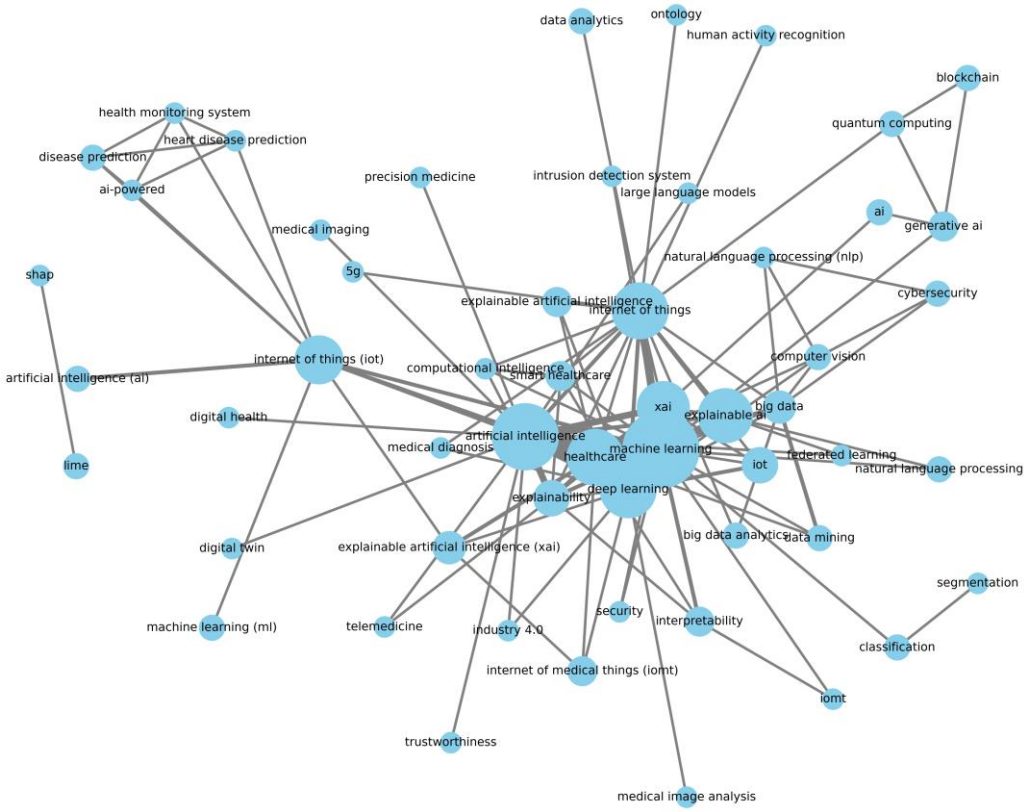

Figure 3. Keyword Co-occurrence Network Graph

(Source: Author's work)

B. Review of Seminal Works in Explainable AI for Medical IoT Applications

The top ten most cited articles identified in this review (see Table 1) represent the foundational literature at the intersection of *Explainable Artificial Intelligence (XAI)*, *machine learning*, and the *Internet of Medical Things (IoMT)*. A key insight emerging from these works is their collective commitment to enhancing both predictive performance and interpretability of intelligent systems for critical medical applications—ranging from chronic illness monitoring to pandemic detection and early diagnosis.

Table 1: Highly Cited Articles in Explainable AI for Medical IoT Applications.

No	Document Title	Authors	Source	Year	Citations
1.	XSRU-IoMT: Explainable simple recurrent units for threat detection in Internet of Medical Things networks	Izhar Ahmed Khan, Nour Moustafa, Imran Razzak, M. Tanveer, Dechang Pi, Yue Pan, Bakht Sher Ali	Future Generation Computer Systems	2022	96
2.	A generative adversarial network (GAN) technique for internet of medical things data	Ivan Vaccari, Vanessa Orani, Alessia Paglialonga, Enrico Cambiaso, Maurizio Mongelli	Sensors	2021	57
3.	Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron	Diogo Gaspar, Paulo Silva, Catarina Silva	IEEE Access	2024	34

4.	Federated Fusion of Magnified Histopathological Images for Breast Tumor Classification in the Internet of Medical Things	Bless Lord Y. Agbley, Jian Ping Li, Amin Ul Haq, Edem Kwedzo Bankas, Cobbinah Bernard Mawuli, Sultan Ahmad	IEEE Journal of Biomedical and Health Informatics	2024	29
5.	Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues	Shahad Sabbar Joudar, A. S. Albahri, Rula A. Hamid, Idrees A. Zahid, M. E. Alqaysi, O. S. Albahri & A. H. Alamoodi	Artificial Intelligence Review	2023	22
6.	Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model	Farhan Ullah, Jihoon Moon, Hamad Naeem, Sohail Jabbar	Journal of Supercomputing	2022	22
7.	Toward explainable AI-empowered cognitive health assessment	Abdul Rehman Javed, Habib Ullah Khan, Mohammad Kamel Bader Alomari, Muhammad Usman Sarwar, Muhammad Asim, Ahmad S. Almadhor, Muhammad Zahid Khan	Frontiers in Public Health	2023	18
8.	Novel IoT framework for event processing in healthcare applications	Naim Shaikh, Kishori Kasat, Rakesh Kumar Godi, V Rama Krishna, Deepak Kumar Chauhan, Jyoti Kharade	Measurement Sensors	2023	17
9.	Reinforcement learning-based dynamic pruning for distributed inference via explainable AI in healthcare IoT systems	Emna Baccour, Aiman Erbad, Amr Mohamed, Mounir Hamdi, Mohsen Guizani	Future Generation Computer Systems	2024	9
10.	Explainable Early Prediction of Gestational Diabetes Biomarkers by Combining Medical Background and Wearable Devices: A Pilot Study With a Cohort Group in South Africa	Şefki Kolozali, Sara L. White, Shane Norris, Maria Fasli, Alastair van Heerden	IEEE Journal of Biomedical and Health Informatics	2024	8

The study by Khan et al. [25] introduces the XSRU-IoMT model, specifically designed to detect cyberattacks in IoMT networks using explainable AI (XAI). Leveraging bidirectional Simple Recurrent Units (SRUs) with skip connections, the model not only accelerates recurrent training but also integrates interpretability to justify system decisions on detected threats. This is crucial, as security models are often perceived as black boxes, particularly within sensitive medical infrastructures where classification errors and privacy violations can have severe consequences. In parallel, Gaspar et al. [26] evaluate the effectiveness of two popular XAI techniques—LIME and SHAP—in the context of an intrusion detection system based on multilayer perceptron models. Their work demonstrates how these interpretive methods assist cybersecurity professionals in interpreting model outputs,

ultimately enhancing the reliability of decision-making in IoT-based health environments. Through perturbation analysis, the study illustrates the comparative strengths of each method, contributing valuable methodological insights.

Vaccari et al. [27] explore the use of GANs to address the challenge of limited datasets in IoMT, particularly for patients with COPD. They emphasize the importance of validating synthetic data quality using XAI approaches, underlining that the success of IoMT implementation is deeply dependent on the interpretability of both the data and the underlying models. Shaikh et al. [28] propose an event-driven IoT architecture for healthcare, built upon the Event Process Healthcare (EPH) model. Using a Cloud-based Deep Learning (CDN) engine, their work stresses the need for real-time decision-making supported by event-driven analytics. This architecture illustrates the necessary synergy between technical precision and operational efficiency in smart healthcare systems, with explainability playing a critical role in fostering user trust and service sustainability.

Agbley et al. [29] merge federated learning (FL) with explainability in a model for breast cancer classification using histopathological images within IoMT environments. The study shows how federated approaches allow global model training without data sharing, addressing privacy concerns. Visualization tools based on XAI are used to validate the classification processes, increasing physicians' trust in automated diagnostic systems. Similarly, Baccour et al. [30] focus on dynamic pruning and distributed inference in IoMT systems, supported by reinforcement learning-guided XAI strategies. This innovation addresses challenges in local computation and privacy without sacrificing accuracy. By embedding explainability into model pruning processes, the approach paves the way for more efficient and transparent medical systems based on edge computing.

Kolozali et al. [31] propose a predictive IoT-based approach for identifying gestational diabetes mellitus (GDM) risk earlier than conventional diagnoses. The integration of wearable sensor data, medical background, and explainable machine learning models effectively enhances early decision-making. Their study exemplifies how sensor technologies and interpretable AI can improve clinical support systems through real-time, data-driven insights. Ullah et al. [32] utilize a combination of CNNs and ensemble classifiers to detect COVID-19 from CT and X-ray images. Explainability is achieved through Grad-CAM and t-SNE, which visualize and validate the models' classification decisions. This study is a compelling example of how XAI supports clinical confidence in high-resolution image-based diagnoses, especially during global health emergencies.

Finally, Javed et al. [33] present XAI-HAR, an explainable activity recognition approach for detecting cognitive health indicators such as dementia in smart homes. Combining smart sensor data with interpretable methods like LIME, this system offers insights into user activities and supports transparent classification of health conditions. The solution is relevant not only for clinical applications but also for daily monitoring in aging populations. Joudar et al. [34] conduct a systematic review on the use of AI for early diagnosis and triage of Autism Spectrum Disorder (ASD). They stress the need for explainability to foster adoption in a domain historically challenged by delayed or inaccurate diagnoses. Their findings offer both theoretical foundations and practical guidance for developing reliable and transparent systems in neurodevelopmental healthcare.

Collectively, these ten articles form five thematic clusters demonstrating that explainable AI and reasoning-driven models are highly relevant to intelligent medical systems (Table 2). Whether in early detection, privacy-preserving analytics, IoMT security, or patient-centered cognitive support, XAI approaches provide critical transparency and trust. Continued progress in this area will be critical to ensuring the ethical and broad-scale adoption of AI in healthcare.

Table 2: Highly Cited Article Categories in AI for Medical IoT Applications.

No	Category	Article
1.	Explainability for Trust in AI-Powered Medical Systems	[25], [26]
2.	Integration of Explainability in IoT-Based Diagnosis and Monitoring	[27], [28]
3.	Federated Learning and Privacy-Preserving Explainable AI	[29], [30]
4.	Early Disease Detection with Wearables and Data Fusion	[31], [32]
5.	Cognitive and Behavioral Health Support via XAI	[33], [34]

Interestingly, Topic 1 (shown in Figure 2, LDA topic modeling) aligns closely with Category 2 (Integration of Explainability in IoT-Based Diagnosis and Monitoring) and Category 3 (Federated Learning and Privacy-Preserving Explainable AI) from Table 2, which comprise research contributions that bridge explainability with real-world data sources and infrastructure. For instance, papers such as Agbley et al. [29] and Baccour et al. [30] demonstrate precisely the kind of system-level synthesis that Topic 1 represents—federated and distributed systems that fuse interpretability with diagnostic capability.

Based on the six most recent publications from 2025 (Table 3), these works demonstrate an increasing emphasis on integrating advanced machine learning with explainability, edge computing, and IoT infrastructure to support real-time, interpretable medical diagnostics and decision-making.

Table 3: Recent Advances in XAI and CoT for Smart Healthcare (2025)

No	Document Title	Authors	Source	Year
1.	Rise of the Machines - Artificial Intelligence in Healthcare Epidemiology	Lemuel R Non, Alexandre R Marra & Dilek Ince	Current Infectious Disease Reports	2025
2.	Optimized disease prediction in healthcare systems using HDBN and CAEN framework	G. Prabaharan, S.M. Udhaya Sankar, V. Anusuya, K. Jaya Deepthi, Rayappan Lotus, R. Sugumar	Methodsx	2025
3.	Rough Set Theory and Soft Computing Methods for Building Explainable and Interpretable AI/ML Models	Sami Naouali, Oussama El Othmani	Applied Sciences Switzerland	2025
4.	Autonomous intrusion detection for IoT: a decentralized and privacy preserving approach	Vitalina Holubenko, Diogo Gaspar, Rúben Leal, Paulo Silva	International Journal of Information Security	2025
5.	MetaXAI: Metahuman-assisted audio and visual explainability framework for Internet of Medical Things	İbrahim Kök	Biomedical Signal Processing and Control	2025
6.	Explainable AI-Driven Gait Analysis Using Wearable Internet of Things (Wiot) and Human Activity Recognition	Ponugoti Kalpana, Sarangam Kodati, L. Smitha, Dhasaratham, Nara Sreekanth, Aseel Smerat, Muhannad Akram Ahmad	Journal of Intelligent Systems and Internet of Things	2025

Two of the articles—Kalpana et al. [35] and Kök [36]—focus on integrating Explainable AI (XAI) into real-time and sensory-rich environments within the Internet of Medical Things (IoMT). Kalpana et al. [35] present a novel hybrid architecture combining Sparse Gate Recurrent Units (SGRUs) and Devil Feared Feed Forward Networks (DFFFN) for gait analysis using wearable IoT devices. Their model not only achieves superior accuracy in activity recognition across multiple datasets (WHU-Gait, OU-ISIR), but also leverages SHAP models to assess feature contributions—providing interpretability for clinicians. This aligns closely with the demand for transparent, edge-deployable AI in clinical monitoring. Similarly, Kök [36] introduces MetaXAI, a multimodal explainability framework that utilizes SHAP, LIME, and ELI5 methods to produce audio-visual explanations for intrusion detection decisions in IoMT. What sets this work apart is the immersive 3D presentation of AI explanations using virtual reality, which facilitates interpretability among both technical and non-technical users. This shift toward interface-rich, user-centric design reflects growing concerns over trust and usability in AI-driven medical systems.

The work by Prabaharan et al. [37] contributes a robust, high-performance hybrid architecture—HDBN and CAEN—for adaptive feature extraction and disease prediction in healthcare and edge computing environments. With metrics such as 93% accuracy and 95% specificity, this study benchmarks a new level of performance in real-world scenarios. While it doesn't explicitly apply to a specific medical device, the framework's scalability, combined with its future integration plan for XAI, positions it as a strong candidate for deployment in smart medical platforms. The framework's modularity further supports future explainability, even if currently

theoretical. In a related vein, Naouali and El Othmani [38] propose a rough set theory-based feature selection framework that dramatically improves classifier accuracy and interpretability, especially within cardiovascular diagnostics. Their MLSpecialReduct technique, when coupled with traditional classifiers, achieves accuracy levels up to 0.99. This paper emphasizes not only accuracy but also the transparency of the model's inner workings—underscoring the growing interest in interpretable architectures for health-related AI models integrated with IoT.

The work of Non et al. [39] expands the domain of XAI beyond individual-level diagnostics to a population-level application—healthcare epidemiology. The paper surveys AI applications in infection surveillance, antimicrobial stewardship, and resistance prediction, emphasizing the importance of explainable models in ensuring trust among healthcare professionals. While this study is more exploratory than technical, its emphasis on data transparency and ethical governance echoes calls for XAI as a core requirement in future AI-based public health infrastructure.

Lastly, Holubenko et al. [40] tackle a critical but often overlooked dimension of smart medical systems—security. They present a privacy-preserving, federated learning-based Host Intrusion Detection System (HIDS) tailored for IoT healthcare environments. The system analyzes system call traces and incorporates XAI to elucidate its high accuracy predictions (~98%). This contribution is vital given the sensitivity of medical data and the need for trustworthy, decentralized AI solutions.

Collectively, these six articles reflect a convergence of efforts toward explainable, scalable, and context-aware AI in IoT-integrated healthcare systems. Whether through immersive visualization [36], feature optimization [38], real-time analysis [35], or federated security [40], the trajectory of research is unmistakably moving toward practical, deployable, and human-centric AI systems. These insights provide critical validation and direction for the framework proposed in this paper—positioning Chain-of-Thought reasoning and explainability not as theoretical luxuries but as operational necessities in modern smart medical environments.

C. Conceptual Framework

Based on the thematic analysis (including LDA topic modeling and keyword co-occurrence network) and the review of seminal and recent publications in the fields of Explainable Artificial Intelligence (XAI), Chain-of-Thought (CoT), and the Internet of Medical Things (IoMT), it is evident that the adoption of reasoning-driven explainability in intelligent medical systems still faces several critical gaps in terms of design, implementation, and practical deployment.

One of the primary gaps is the lack of approaches that unify explicit reasoning (via CoT) with model interpretability (via XAI) in real-time, IoT-based healthcare environments. Many studies still treat predictive modeling and interpretability as separate efforts, without offering an integrated pipeline. Furthermore, how CoT reasoning can be efficiently executed in edge computing, federated learning, or wearable sensor environments remains underexplored. Yet, the success of AI systems in modern healthcare services increasingly depends on their ability to provide interpretable logic in resource-constrained settings and with a strong alignment to end-user needs—namely clinicians and patients.

To address these gaps, the new conceptual framework illustrated in Figure 4: Conceptual Framework is proposed. It presents a systematized architecture that integrates multiple functional layers required to support explainable, trustworthy, and deployable AI systems in connected smart medical ecosystems.

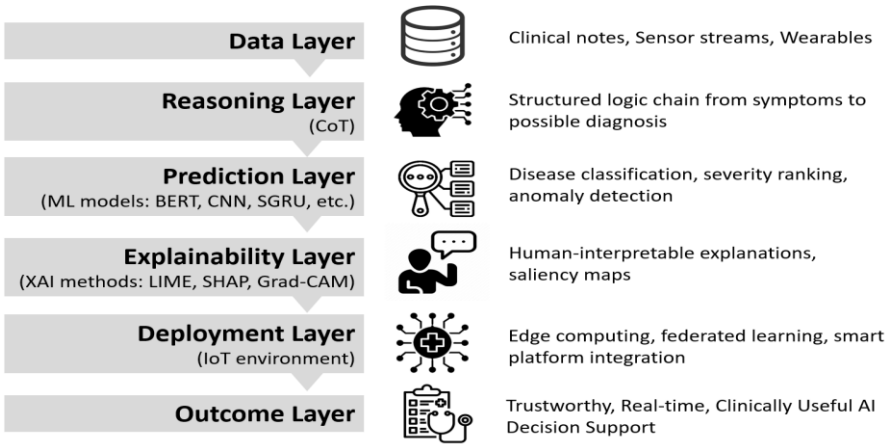

Figure 4. Conceptual framework

(Source: Author's work)

The framework consists of six core functional dimensions: Data Layer, Reasoning Layer, Prediction Layer, Explainability Layer, Deployment Layer, and Outcome Layer. Each layer is designed to be modular, yet interdependent, forming a cohesive flow from raw input data to transparent and real-time medical decision support.

- The Data Layer serves as the foundation for ingesting multimodal clinical and sensor data.
- The Reasoning Layer applies Chain-of-Thought logic to create structured inferences between symptoms and diagnosis.
- The Prediction Layer uses advanced machine learning models to perform classification, severity assessment, or anomaly detection.
- The Explainability Layer transforms model outputs into human-readable explanations.
- The Deployment Layer ensures that the entire pipeline operates within IoT-based, distributed, and privacy-aware environments.
- The Outcome Layer delivers trustworthy, real-time, and clinically actionable AI-supported decisions.

D. Data Context and Information Sources

The system is designed to accommodate multimodal medical data, including unstructured text such as clinical notes, patient summaries, and observational reports; signal-based data like ECG and heart rate from wearable sensors; and contextual information from patient environments. In this context, interoperability and real-time accessibility are critical. For example, wearable IoT devices generate continuous data streams that require instant analysis, while textual medical records demand natural language processing (NLP) to extract clinical meaning.

This data also varies significantly in terms of quality, frequency, and format, requiring the system to perform data fusion and automated pre-processing. Therefore, the initial pipeline must standardize input data to ensure consistency and readiness for downstream processes.

E. Core Technological Components

The framework's technological foundation consists of three major components:

- A CoT-based Reasoning Engine, which produces structured logic paths (e.g., "If symptom A is present, then disease B is likely due to factor C"). This component serves as a bridge between black-box modeling and white-box interpretability.
- Machine Learning Predictors, including models such as BERT (for textual data), CNN (for image or signal analysis), and SGRU (for sequential inputs). These models execute tasks such as disease classification, severity ranking, and anomaly detection based on CoT outputs.
- XAI Modules, including SHAP, LIME, and Grad-CAM, which justify model predictions through feature attribution, attention maps, or explanatory narratives. These tools make the model’s decision process traceable and clinically verifiable.

The integration of CoT in this pipeline provides an explanation-first approach that aligns with the trust, transparency, and accountability required in healthcare decision-making.

F. Smart and Distributed Infrastructure

The framework is optimized for deployment within IoT environments using edge computing and federated learning, both of which are highly relevant to modern demands for privacy, low latency, and system scalability.

- In edge computing scenarios, reasoning and interpretability are executed directly on devices such as wearables or mobile edge nodes. This allows for localized analysis that is fast, efficient, and reduces bandwidth demands.
- In federated learning environments, models are trained and updated across decentralized nodes without transferring patient data. This design supports compliance with privacy regulations (e.g., HIPAA, GDPR) and addresses the challenges of geographically distributed healthcare systems.

This deployment layer is not only technical but also socio-technical, bridging AI functionalities with real-world clinical workflows in hospitals, community health centers, and even home-based monitoring systems.

G. Expected Use Cases and Outcomes

The framework is designed to support practical and translational outcomes, including:

- Explainable AI-driven diagnosis, where predictions and their underlying reasoning can be clearly understood by both clinicians and patients.
- Real-time clinical decision-making in applications such as chronic disease monitoring or emergency response via wearable IoT systems.

- Personalized and adaptive clinical interventions, as the reasoning engine can tailor its logic to the unique medical context of each patient.
- Increased trust in AI systems, as transparent explanations reduce fear of black-box decision-making and improve clinical adoption.
- Healthcare efficiency, through faster diagnosis, reduced workload for healthcare professionals, and lower error rates.

Referring to recent studies such as Kalpana et al. [35], Holubenko et al. [40], and Kök [36], this framework aligns with the trajectory of AI innovation toward user-centered, secure, and interpretable healthcare systems.

In summary, this framework brings together the dimensions of reasoning, interpretability, and system connectivity into a unified functional pipeline. It addresses not only technical challenges but also social and ethical considerations in the development of future medical AI systems. With its flexible structure and adaptability to emerging technologies, this framework serves as a robust foundation for future research and real-world implementation in healthcare environments.

5. Conclusion

This study has conducted a comprehensive and systematic literature review to explore the convergence between Chain-of-Thought (CoT) reasoning, Explainable Artificial Intelligence (XAI), and smart medical systems, particularly within the context of IoT-enabled healthcare environments. By synthesizing 106 Scopus-indexed publications through thematic analysis, topic modelling (LDA), and keyword co-occurrence network analysis, the research reveals several critical insights and structural patterns that define the current state and future trajectory of explainable reasoning in medical artificial intelligence.

The findings indicate that while CoT and XAI have independently garnered substantial scholarly attention, their integrated application in medical diagnostic systems—especially those deployed in real-time and resource-constrained IoT settings—remains largely underdeveloped. Many studies have emphasized either interpretability or predictive performance in isolation, yet little address the operational need for reasoning-aware AI systems that offer transparent, justifiable, and deployable outputs within clinical workflows. The review also surfaces a growing shift toward user-centric AI, as evidenced by the increasing presence of explainability-focused methods such as SHAP, LIME, and Grad-CAM in recent literature.

To address these challenges and research gaps, this paper introduces a new conceptual framework that organizes AI-based clinical decision-making into six interrelated layers: Data, Reasoning, Prediction, Explainability, Deployment, and Outcome. The framework is designed to support trustworthy and interpretable medical diagnostics by integrating reasoning mechanisms with model interpretability techniques, and by situating the entire pipeline within distributed, privacy-preserving infrastructures such as edge computing and federated learning. This layered architecture not only synthesizes theoretical findings but also provides a foundation for practical system development, particularly in the design of IoT-based diagnostic platforms.

In conclusion, this study makes an original contribution by positioning explainable reasoning as a central component of intelligent medical systems and by highlighting the practical requirements for deploying such systems in real world, connected healthcare environments. The framework proposed here offers actionable insights for AI researchers, system architects, clinical practitioners seeking to design human-aligned, ethically robust, and scalable diagnostic technologies.

6. Implications

The conceptual and analytical contributions of this study offer significant implications for both research and clinical practice in the domain of medical artificial intelligence. First, from a research standpoint, the integration of CoT reasoning with XAI methods fills an important conceptual void in the development of interpretable machine learning systems. As AI systems become more embedded in clinical workflows, the ability to trace not only the "what" but also the "why" behind predictions is essential. This study’s framework presents a novel model for incorporating sequential logic and human-like reasoning into diagnostic pipelines, thereby pushing the boundaries of current XAI research.

Second, the implications for clinical practice and health system implementation are substantial. The proposed framework enables the creation of AI tools that are not only high performing but also clinician-aligned—providing explanations that mirror clinical reasoning pathways. In environments where trust, accountability, and explainability are regulatory and ethical imperatives, this design approach ensures greater transparency and facilitates smoother adoption among healthcare providers. For instance, CoT-based logic chains allow doctors to scrutinize systematic AI inferences in a manner consistent with differential diagnosis processes.

Third, the deployment-oriented design of the framework—incorporating IoT, edge computing, and federated learning—underscores the growing importance of scalable, real-time, and privacy-preserving AI systems in modern healthcare. As wearables, sensors, and remote diagnostics become the norm, this framework guides developers in embedding explainability into the system's architecture, not as a post hoc feature but as a core function.

Finally, the study also has policy and regulatory implications, especially for agencies seeking to standardize AI use in clinical settings. The framework can serve as a reference model for evaluating the trustworthiness and interpretability of AI-based healthcare technologies, aligning technical development with ethical and legal standards.

7. Limitations and Further Research

This study provides a theoretical framework for integrating Chain-of-Thought (CoT) reasoning and Explainable AI (XAI) in IoT-based medical systems. However, several limitations must be acknowledged. First, the framework has not yet been empirically validated in real-world clinical environments. Its operational effectiveness, especially in edge or federated settings, remains to be tested.

Second, the literature review, though comprehensive, may exclude emerging insights from non-indexed or grey literature sources, which could further enrich the framework's applicability. Third, the framework focuses on technical integration and does not extensively address human factors, such as clinician interaction, cognitive load, or user trust.

Future research should aim to implement and evaluate the framework in various healthcare contexts. Empirical studies are needed to assess its impact on decision-making, interpretability, and user acceptance. Collaboration between AI developers and healthcare professionals will be essential to adapt the framework for domain-specific use cases and ensure practical relevance.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

Data Availability: "This study uses data obtained from the Scopus website in the form of CSV files. Other authors can use this data for further research purposes and not for business interests by mentioning the original source in their research documents. Data can be obtained by downloading it from the Zenodo website at: https://doi.org/10.5281/zenodo.15666988, Creative Commons Attribution 4.0 International License (CC BY 4.0) is the licensing body."

Author Contributorship: "FB (Corresponding Author): Conceptualization, research design, literature review, methodology development, data analysis, interpretation of findings, and drafting of the initial manuscript. SAS: Supervision, critical review and refinement of the manuscript structure, especially in the theoretical framework and discussion sections, methodological guidance, handling journal submission and correspondence, and preparation of publication-related documentation including the cover letter."

Acknowledgment: "The author expresses sincere gratitude to the Professional Engineer Program Department, Faculty of Engineering, Bina Nusantara University, for their continuous academic support. Appreciation is also extended to the Research Interest Group in Education Technology at Bina Nusantara University and Muhammadiyah University of Berau for their valuable insights, contributions, and collaborative support throughout the research process."

References

[1] A. Smith and B. Johnson, "A Comprehensive Study on the Impact of AI in Financial Services," J. Financial Technol., vol. 12, no. 3, pp. 45–60, 2022.

[2] E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[3] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[4] C. Brown and D. White, "Exploring the Role of Machine Learning in Environmental Sustainability," Int. J. Sustainable Dev., vol. 9, no. 4, pp. 201–215, 2023.

[5] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit. Heal., vol. 3, no. 11, pp. e745–e750, Nov. 2021.

[6] B. Heinrichs and S. B. Eickhoff, “Your evidence? Machine learning algorithms for medical diagnosis and prediction,” Hum. Brain Mapp., vol. 41, no. 6, pp. 1435–1444, 2020.

[7] S. U. Hassan, S. J. Abdulkadir, M. S. M. Zahid, and S. M. Al-Selwi, “Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review,” Comput. Biol. Med., vol. 185, p. 109569, 2025.

[8] M. R. Santos, A. Guedes, and I. Sanchez-Gendriz, “SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis,” Machine Learning and Knowledge Extraction, vol. 6, no. 1, pp. 316–341, 2024.

[9] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022.

[10] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, “Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making,” IEEE Comput. Intell. Mag., vol. 17, no. 1, pp. 72–85, 2022.

[11] O. Oyebode, J. Fowles, D. Steeves, and R. Orji, “Machine Learning Techniques in Adaptive and Personalized Systems for Health and Wellness,” Int. J. Hum. Comput. Interact, vol. 39, no. 9, pp. 1938–1962, 2023.

[12] Z. Li, Z. Cao, P. Li, Y. Zhong, and S. Li, “Multi-Hop Question Generation with Knowledge Graph-Enhanced Language Model,” Applied Sciences, vol. 13, no. 9, 2023.

[13] W. Jiageng et al., “Clinical Text Datasets for Medical Artificial Intelligence and Large Language Models — A Systematic Review,” NEJM AI, vol. 1, no. 6, May 2024.

[14] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” Healthc. (Basel, Switzerland), vol. 10, no. 3, Mar. 2022.

[15] F. J. Boge, P. Grünke, and R. Hillerbrand, “Minds and Machines Special Issue: Machine Learning: Prediction without Explanation?” Minds Mach., vol. 32, no. 1, pp. 1–9, 2022.

[16] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, “Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models,” in 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 7–12.

[17] X. Chen, “The Advance of Deep Learning and Attention Mechanism,” in 2022 International Conference on Electronics and Devices, Computational Science (ICEDCS), 2022, pp. 318–321.

[18] U. Peters, “Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque,” AI Ethics, vol. 3, no. 3, pp. 963–974, 2023.

[19] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review,” J. Am. Med. Inform. Assoc., vol. 26, no. 4, pp. 364–379, Apr. 2019.

[20] S. Kruschel, N. Hambauer, S. Weinzierl, S. Zilker, M. Kraus, and P. Zschech, “Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models,” Bus. Inf. Syst. Eng., 2025.

[21] C. Albon, Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning, First Edit. Sebastopol, CA: O’Reilly, 2018.

[22] T. Dillan and D. H. Fudholi, “LDAViewer: An Automatic Language-Agnostic System for Discovering State-of-the-Art Topics in Research Using Topic Modeling, Bidirectional Encoder Representations From Transformers, and Entity Linking,” IEEE Access, vol. 11, no. April, pp. 59142–59163, 2023.

[23] F. Binsar, T. N. Mursitama, M. Hamsal, and R. K. Rahim, “Determinants of Digital Adoption Capability for Service Performance in Indonesian Hospitals: A Conceptual Model,” J. Syst. Manag. Sci., vol. 14, no. 2, pp. 188–213, 2024.

[24] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, “Human-centered XAI: Developing design patterns for explanations of clinical decision support systems,” Int. J. Hum. Comput. Stud., vol. 154, 2021.

[25] I. A. Khan et al., “XSRU-IoMT: Explainable simple recurrent units for threat detection in Internet of Medical Things networks,” Futur. Gener. Comput. Syst., vol. 127, pp. 181–193, 2022.

[26] D. Gaspar, P. Silva, and C. Silva, “Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron,” IEEE Access, vol. 12, pp. 30164–30175, 2024.

[27] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, and M. Mongelli, “A Generative Adversarial Network (GAN) Technique for Internet of Medical Things Data,” Sensors, vol. 21, no. 11, 2021.

[28] N. Shaikh, K. Kasat, R. K. Godi, V. R. Krishna, D. K. Chauhan, and J. Kharade, “Novel IoT framework for event processing in healthcare applications,” Meas. Sensors, vol. 27, p. 100733, 2023.

[29] B. L. Y. Agbley et al., “Federated Fusion of Magnified Histopathological Images for Breast Tumor Classification in the Internet of Medical Things,” IEEE J. Biomed. Heal. Informatics, vol. 28, no. 6, pp. 3389–3400, 2024.

[30] E. Baccour, A. Erbad, A. Mohamed, M. Hamdi, and M. Guizani, “Reinforcement learning-based dynamic pruning for distributed inference via explainable AI in healthcare IoT systems,” Futur. Gener. Comput. Syst., vol. 155, pp. 1–17, 2024.

[31] Ş. Kolozali, S. L. White, S. Norris, M. Fasli, and A. van Heerden, “Explainable Early Prediction of Gestational Diabetes Biomarkers by Combining Medical Background and Wearable Devices: A Pilot Study With a Cohort Group in South Africa,” IEEE J. Biomed. Heal. Informatics, vol. 28, no. 4, pp. 1860–1871, 2024.

[32] F. Ullah, J. Moon, H. Naeem, and S. Jabbar, “Explainable artificial intelligence approach in combating real-time surveillance of COVID-19 pandemic from CT scan and X-ray images using ensemble model,” J. Supercomput., vol. 78, no. 17, pp. 19246–19271, 2022.

[33] A. A. Nassani, A. Javed, J. Rosak-Szyrocka, L. Pilar, Z. Yousaf, and M. Haffar, “Major Determinants of Innovation Performance in the Context of Healthcare Sector,” Int. J. Environ. Res. Public Health, vol. 20, no. 6, 2023.

[34] S. S. Joudar et al., “Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues,” Artif. Intell. Rev., vol. 56, no. 1, pp. 53–117, 2023.

[35] P. Kalpana, R. Kumar, and S. Gupta, “Edge-Deployable Sparse Gated Recurrent Networks for IoT-Based Clinical Gait Analysis,” Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom), Kyoto, Japan, 2024, pp. 1-10.

[36] İ. Kök, “MetaXAI: Metahuman-assisted audio and visual explainability framework for Internet of Medical Things,” Biomed. Signal Process. Control, vol. 100, p. 107034, 2025.

[37] G. Prabaharan, S. M. Udhaya Sankar, V. Anusuya, K. Jaya Deepthi, R. Lotus, and R. Sugumar, “Optimized disease prediction in healthcare systems using HDBN and CAEN framework,” MethodsX, vol. 14, p. 103338, 2025.

[38] S. Naouali and O. El Othmani, “Rough Set Theory and Soft Computing Methods for Building Explainable and Interpretable AI/ML Models,” Applied Sciences, vol. 15, no. 9, 2025.

[39] L. R. Non, A. R. Marra, and D. Ince, “Rise of the Machines - Artificial Intelligence in Healthcare Epidemiology,” Curr. Infect. Dis. Rep., vol. 27, no. 1, p. 4, 2025.

[40] V. Holubenko, D. Gaspar, R. Leal, and P. Silva, “Autonomous intrusion detection for IoT: a decentralized and privacy preserving approach,” Int. J. Inf. Secur., vol. 24, no. 1, p. 7, 2025.