A Comparative Review of AI, IoT, and Big Data in Healthcare: Towards a Data-Centric Approach for Enhanced Data Quality and Contextual Adaptability

Imane RAFIQ, Zahi JARIR, Hiba ASRI
LISI-Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco

***Abstract*—The convergence of Artificial Intelligence (AI), the Internet of Things (IoT), and Big Data is revolutionizing healthcare by enabling predictive diagnostics, real-time monitoring, and personalized treatment through data-driven analytics and intelligent decision-making. Despite these advancements, the effectiveness of such systems is significantly hindered by poor data quality, including issues such as missing values, noise, bias, and inconsistencies. This study presents a systematic and comparative review of recent research at the intersection of AI, IoT, and Big Data in healthcare, highlighting critical gaps in data quality that undermine model performance and real-world reliability. In response, we introduce the Data-Centric AI (DCAI) paradigm as a promising approach focused on systematic data improvement rather than model complexity. We examine the application of the METRIC framework for assessing data quality dimensions such as completeness, consistency, fairness, and timeliness. Furthermore, we propose future research directions to improve scalability and trustworthiness in AI-driven healthcare, integrating advanced AI techniques such as generative AI and multimodal frameworks with DCAI principles for more ethical AI applications. This work serves as both a comparative synthesis of existing literature and a conceptual foundation for future experimental validation through a case study integrating context-aware data modeling and real-time decision support.**

Keywords*—Data-Centric AI; IoT; Big Data Analytics; healthcare informatics; data quality; bias mitigation; privacy; predictive analytics; machine learning; disease prediction

I. Introduction

The convergence of Artificial Intelligence (AI), Big Data, and the Internet of Things (IoT) is reshaping modern healthcare by enabling real-time monitoring, predictive diagnostics, and personalized treatment pathways. These technologies promise improved patient outcomes, operational efficiency, and data-driven clinical decision-making. However, the performance and reliability of such systems are not solely dependent on advanced algorithms—they are fundamentally tied to the quality and integrity of the underlying data.

Healthcare data, particularly from heterogeneous IoT environments, is often plagued by issues such as missing values, noise, demographic bias, and inconsistencies, which compromise clinical accuracy and fairness [1]. Additionally, challenges around interoperability, privacy, and regulatory compliance continue to impede scalable AI adoption in real-world healthcare settings [2], [3].

Traditional approaches have predominantly focused on model-centric AI (MCAI), prioritizing algorithmic optimization while assuming static, clean datasets. This focus has resulted in unreliable predictions, biased treatment recommendations, and limited generalizability across diverse patient populations [4]. In healthcare—where errors can be life-threatening—this paradigm is no longer sustainable. The increasing complexity of multimodal healthcare data, combined with real-time demands and stringent privacy constraints, requires a shift in focus: from model optimization to systematic data enhancement. As recent findings emphasize that improvements in data quality can often outperform equivalent enhancements in model complexity, especially in critical settings like clinical diagnostics or patient monitoring [5], [6], and that poor data quality, like mislabeled samples, can propagate errors throughout the AI lifecycle, leading to unsafe or inequitable medical decisions[5], [7]. The result is a growing consensus that high-quality, well-curated data must be central to AI development.

Despite widespread research on AI and IoT integration in healthcare, there remains no unified framework that systematically addresses how data quality, fairness, and interoperability should be prioritized to enable robust, ethical AI systems. Data-Centric AI (DCAI) —a paradigm that prioritizes the systematic curation, preprocessing, and contextual enrichment of data to improve AI model performance and fairness— has emerged as a promising approach to filling this void [8], [9]. However, its application within healthcare remains underexplored and unstandardized.

This study offers a conceptual and comparative synthesis of existing AI-IoT-Healthcare literature through the lens of DCAI. While no experimental implementation is conducted, our aim is to theoretically map how DCAI principles can address persistent data quality limitations in healthcare systems. We seek to bridge that gap by comparing twenty-three AI-IoT-Healthcare studies and analyzing DCAI's theoretical and practical potential to enhance data quality. We explore a structured conceptual framework that maps DCAI principles to recurring healthcare challenges—specifically related to data noise, bias, missing values, and contextual inconsistency—, offering pathways toward more scalable, interpretable, and trustworthy AI solutions. Additionally, we identify future directions, including emerging techniques such as generative AI, multimodal integration, and advanced data recovery techniques as promising complements to DCAI in clinical contexts. This study sets the

foundation for future empirical validation through real-world case studies or clinical datasets.

Fig. 1 illustrates the layered integration of IoT, Big Data, and AI within smart healthcare systems. IoT devices serve as the primary data acquisition layer, collecting real-time physiological and behavioral data. This raw data is then aggregated and processed through Big Data infrastructures, which handle high-volume, high-velocity data streams for storage, cleaning, and transformation. The processed data feeds AI algorithms at the decision layer, where predictive models generate insights for diagnostics, risk stratification, and personalized care. This pipeline enables closed-loop feedback mechanisms, supporting adaptive, context-aware healthcare while ensuring scalability, data integrity, and privacy compliance.

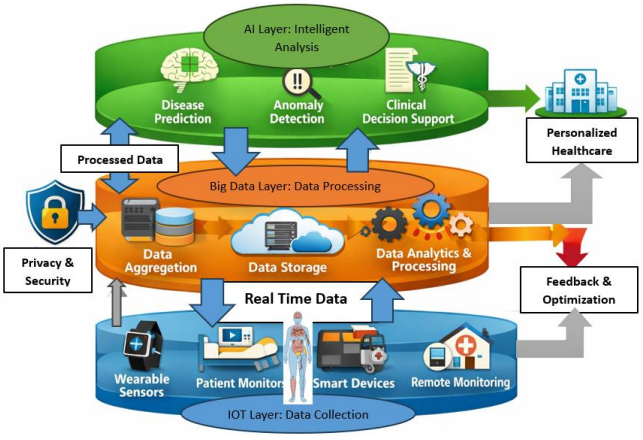

Fig. 1. Integration of AI, IoT, and Big Data in smart healthcare systems.

The remainder of this study is organized as follows: Section II presents related work and the structured literature review methodology. Section III details the technological and operational challenges in AI-driven healthcare. Section IV introduces DCAI concepts and applications. Section V maps DCAI to persistent data quality issues. Section VI outlines relevant data quality metrics. Section VII discusses study limitations. Section VIII explores future research directions, and Section IX concludes with reflections on ethical, scalable AI in healthcare.

II. Background and Related Work

A. Literature Review and Methodology

AI, Big Data, and IoT have driven major advances in healthcare, facilitating predictive diagnostics, real-time monitoring, and data-driven decision-making. Despite the technical sophistication of these systems, their real-world effectiveness is frequently compromised by persistent challenges—including poor data quality, heterogeneity, lack of standardization, and privacy concerns. These issues threaten the reliability, fairness, and scalability of AI applications in healthcare environments.

Data-Centric AI (DCAI) has recently emerged as a promising paradigm to tackle these challenges by emphasizing the quality, structure, and contextual integrity of data over algorithmic complexity. Yet, dedicated review studies that explore DCAI's practical implications in healthcare settings are scarce. This work addresses that gap by conducting a structured review of both foundational DCAI literature and application-oriented studies that highlight real-world limitations in AI, Big Data, and IoT for healthcare.

To ensure scientific rigor, the review process followed a structured literature selection methodology that allows our work to reflect the contributions of prior research on IoT, AI, and Big Data in healthcare, while also incorporating foundational studies on DCAI principles and data quality metrics. Our goal is to highlight the ongoing shift from model-centric to data-centric approaches and to underscore the growing relevance of DCAI-based solutions in healthcare AI.

1) Search strategy and data sources: The literature search was conducted using a multi-database strategy to encompass a broad range of both technical and medical research publications, ensuring comprehensive coverage of both theoretical AI/IoT research and practical healthcare applications. The following academic databases were selected:

- IEEE Xplore
- Scopus
- Web of Science
- PubMed
- ScienceDirect
- Google Scholar

Our search queries combined key terms related to AI, IoT, Big Data, and healthcare, for the selection of the 23 articles such as: "Artificial Intelligence", "Machine Learning", "Deep Learning", "Internet of Things", "Healthcare IoT", "Big Data Analytics", "Context-Aware Systems", and "Smart Healthcare".

Boolean operators were applied to refine results [e.g., ("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning") AND ("Internet of Things" OR "Healthcare IoT") AND ("Big Data Analytics" OR "Smart Healthcare" OR "Real-time Monitoring") AND ("Context-Aware Systems" OR "Privacy" OR "Interoperability" OR "Clinical Decision Support Systems")]. The search was limited to the 2016–2025 publication window. Manual adjustments were made to avoid duplicated retrievals across platforms.

2) Eligibility criteria: The search was conducted on publications between 2016 and 2025, and duplicates were removed. The citation management tool Rayyan was used to facilitate collaborative screening, inclusion/exclusion filtering, and conflict resolution during selection. In Table I, our chosen eligibility criteria that were applied to the various screening steps of our selection process of the twenty-three peer-reviewed application-oriented studies are listed.

3) Selection and analysis process: An initial screening based on titles and abstracts was performed, followed by a full-text review of eligible articles. Twenty-three studies were ultimately selected as representative of prevailing practices and

challenges in AI, IoT, and Big Data-driven healthcare systems for prediction, monitoring, and decision-making.

TABLE I. ELIGIBILITY CRITERIA APPLIED TO THE SCREENING AND FULL-TEXT ASSESSMENT PROCESSES

Criterion	
Inclusion	Peer-reviewed journal articles or chapter book or conference papers
	Publications between 2016 and 2025 published in **English**
	Studies addressing healthcare applications of AI, IoT, and/or Big Data for predictions of diseases or monitoring or real word application
	Articles discussing at least one of the following: challenges related to data quality, context awareness, interoperability, scalability, privacy, Healthcare diagnostics, monitoring, decision support using AI/IoT or Real-time systems.
Exclusion	Non-peer-reviewed sources, editorials, or opinion pieces
	Studies unrelated to healthcare applications
	Works focused exclusively on algorithms without consideration of data characteristics or system-level integration
	Redundant or overlapping publications

To bridge the gap between these practical challenges and the conceptual framework of DCAI, the tool Rayyan was used to facilitate collaborative screening, conflict resolution, and exclusion tagging. This process resulted in the selection of:

- Seven Foundational DCAI Publications outlining data-centric concepts, lifecycle frameworks, data-centric design principles, and metrics for evaluating data quality
- Twenty-Three Application-Oriented Studies demonstrating real-world applications and challenges in AI-IoT-Big Data healthcare systems.

A PRISMA-style flow diagram (Fig. 2) outlines the article selection workflow, culminating in the inclusion of twenty-three peer-reviewed studies focused on AI and IoT applications in healthcare. These articles were analyzed using a unified comparison framework emphasizing data sources, modeling techniques, evaluation outcomes, and reported limitations. This methodology enabled a systematic comparative analysis and the extraction of key limitations, forming a solid foundation for identifying research gaps addressed by the Data-Centric AI (DCAI) paradigm.

To enhance analytical depth and provide objective insights, a quantitative content analysis was also performed. Each selected study was reviewed to extract explicitly stated limitations related to data quality, scalability, interoperability, privacy, and integration. These issues were then categorized by frequency, allowing the construction of a visual summary of common challenges (see Fig. 3). This process not only highlighted recurring barriers but also emphasized the relevance of DCAI as a unifying approach to overcome these limitations.

Table II presents an organized summary of the core literature reviewed in this study, categorized by focus and contribution.

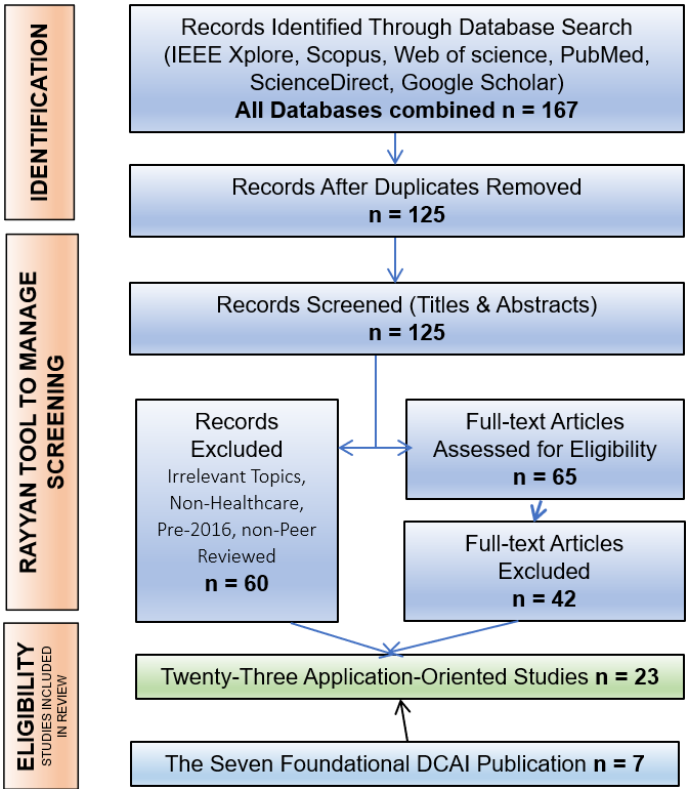

Fig. 2. PRISMA-style flow diagram illustrating the article selection process. A total of 167 records were initially identified from six academic databases. After removal of duplicates and exclusion based on titles and abstracts, 65 full-text articles were reviewed. Finally, 23 articles were selected as application-oriented studies in healthcare AI, IoT, and Big Data and added to 7 foundational DCAI publications. Rayyan was used to manage and streamline the screening process.

TABLE II. SUMMARY OF KEY ARTICLES USED IN THIS STUDY

Category	Article	Focus/Key Contribution
Foundational Articles	[4] Bhatt et al. (2024)	Techniques for data-centric deep learning improvement
	[9] Malerba & Pasquadibisceglie (2024)	DCAI philosophy: prioritizing data refinement over model tuning
	[1] Schwabe et al. (2024)	METRIC framework for healthcare data quality evaluation
	[15] Zha et al. (2023) - A Survey	Taxonomy and automation levels in DCAI
	[5] Jin et al. (2024)	DCAI lifecycle: data operations across AI stages
	[6] Nieberl et al. (2024)	DCAI implementation patterns and research gaps
	[7] Zha et al. (2023) - Perspectives	Strategic challenges in DCAI for training, inference, and deployment
Application-Oriented Articles	[16] Zonayed et al. (2025)	Narrative synthesis of AI models in chronic disease prediction; highlights key trends and challenges in healthcare ML
	[17] Šajnović et al. (2024)	IoT and Big Data analytics in preventive healthcare; synthesizes enablers, gaps, and interoperability issues
	[18] Charfare et al. (2024)	Survey on AI-IoT integration in healthcare; identifies performance trends and system-level limitations
	[19] Alsabah et al. (2025)	Comprehensive review of smart healthcare IoT architectures; technical challenges and future directions
	[10] Zon et al. (2023)	Context-aware data optimization in healthcare systems
	[11] Chung et al. (2020)	Deep learning for context-driven health risk prediction
	[13] Saranya & Fatima (2022)	Context-aware data fusion for IoT patient monitoring.
	[20] Kishor & Chakraborty (2021)	AI-IoT integration for remote health monitoring
	[21] Ghazal et al. (2021)	ML applications in smart city healthcare
	[12] Kim & Chung (2020)	Adaptive health context prediction models
	[22] Kaur (2021)	Heart disease prediction using ML and IoT
	[2] Banerjee et al. (2020)	Big Data and IoT trends in healthcare
	[3] Meraj et al. (2021)	IoT for monitoring infectious diseases
	[23] Vijayalakshmi et al. (2021)	Disease prediction using Big Data tools
	[24] Verma et al. (2019)	Hybrid IoT-cloud architecture for diagnosis
	[25] Aceto et al. (2020)	Personalization through Industry 4.0 healthcare IoT
	[14] Aborokbah et al. (2018)	Context-aware decision systems for chronic diseases.
	[26] Castro et al. (2017)	IoT wearables for activity recognition
	[27] Chui et al. (2019)	Behavior monitoring via Big Data and IoT
	[28] Ngiam & Khor (2019)	ML applications for healthcare analytics
	[29] Shah et al. (2018)	IoT-based systems for monitoring disorders
	[30] Tian et al. (2019)	IoT-cloud platforms for smart healthcare
	[31] Yin et al. (2016)	Overview of IoT use cases in healthcare

A. Comparative Analysis of AI, IoT, and Big Data Healthcare Studies

To assess the state of healthcare applications involving AI, IoT, and Big Data, we conducted a comparative analysis of the 23 peer-reviewed studies published between 2016 and 2025. Based on the defined inclusion and exclusion criteria, these studies were selected for their relevance to disease prediction, patient monitoring, smart healthcare systems, and healthcare infrastructure. Our objective was to extract common practices, recurring challenges, and methodological gaps that inform the need for a shift toward the emerging data-centric paradigm.

1) Categories of reviewed studies: To move beyond a structured descriptive synthesis, the selected 23 articles were classified into four primary domains based on their methodological focus and healthcare application scope. Within each category, we explicitly compare approaches, data dependencies, and reported limitations to identify recurring patterns and systemic gaps.

a) Context-aware systems and adaptive intelligence: Several studies underscore the importance of integrating environmental, behavioral, and physiological context for more accurate and responsive healthcare interventions. In [10], Zon et al underscored the lack of context-aware medical systems, which limits adaptability to patient conditions, addressing the difficulty of integrating context-aware medical systems due to heterogeneous healthcare data, highlighted the lack of a structured understanding of medical contexts, and identified key medical contexts for AI adaptation. Similarly, in [11], Chung et al. proposed an ambient context-based deep neural network that leverages contextual health data for health risk

assessment in patients with chronic diseases, improving prediction accuracy but facing issues with heterogeneous and noisy input IoT data. In [12], Kim & Chung advanced this area by developing a neural network-based adaptive context prediction model for ambient intelligence, which enhanced patient care personalization. Their approach demonstrated an enhanced ability to predict and adapt to various patient states, leading to better patient care, though the systems required further validation due to data integration complexity. In [13], Saranya & Fatima tackled data heterogeneity in IoT-based patient monitoring by developing an improved context-aware data fusion model that enhances security and optimizes data integration from multiple sensors, achieving high precision in real-time health monitoring. In [14], Aborokbah et al. proposed an adaptive computing paradigm that utilizes machine learning techniques, which improved chronic disease management but required further validation for diverse populations.

While all studies acknowledge context as a critical factor, most approaches treat context as an auxiliary feature rather than a first-class data entity. None adopts a systematic data-centric strategy to validate, curate, or standardize contextual information, leading to scalability and robustness limitations.

b) Disease prediction and monitoring: IoT-enabled models for real-time monitoring and diagnosis are prominently featured. In [13], Saranya & Fatima proposed a fusion model for wearable data that achieved 97.9% accuracy in patient monitoring but faced issues of sensor reliability and cybersecurity. The need for real-time disease prediction was explored by Kishor & Chakraborty in [20], who integrated IoT with machine learning classifiers, obtaining high accuracy but encountering real-time processing constraints. Meanwhile, in [22], Kaur investigated the increasing prevalence of heart diseases and the need for predictive models that can provide timely warnings and applied IoT and machine learning for heart disease prediction, demonstrating high accuracy but struggling with data privacy and completeness. Similarly, in [3], Meraj et al reviewed IoT-based infectious disease detection, emphasizing the need for real-time outbreak prediction, though sensor noise and lack of data remained challenges. In [26], Castro et al. focused on wearable-based human activity recognition using IoT and developed a machine learning-based system that classified physical activities, achieving improved movement detection accuracy, but environmental noise remained a limiting factor. In [29], Shah et al. conducted a case study on IoT-based sensing for healthcare applications, focusing on detecting narcolepsy episodes and implemented machine learning classifiers to identify sleep disorder patterns, demonstrating that IoT sensors could effectively improve sleep disorder diagnosis and patient monitoring. Additionally, in [28], Ngiam & Khor reviewed the challenges and opportunities of applying machine learning algorithms in healthcare and emphasized the importance of high-quality data preprocessing and ethical considerations in AI-driven healthcare systems, demonstrating that integrating AI could enhance prediction and diagnostic accuracy while raising concerns about data bias, data labeling, annotation inconsistencies and privacy. The study [16] is a more recent contribution by Zonayed et al. that comprehensively synthesizes the advancements, challenges, and future directions of integrating Machine Learning (ML) and the Internet of Things (IoT) in healthcare. This integration facilitates real-time health monitoring and the analysis of intricate medical datasets, yielding insights that support evidence-based clinical decision-making. The review highlights the high predictive accuracy (85%–95%) achieved by models like CNNs and XGBoost, particularly in diagnostics and chronic disease management. Challenges remain concerning data security, interoperability, and ethical transparency, scalability, and the need for explainable AI to foster clinical trust and ethical transparency. Future direction must prioritize robust security, standardization, and effective human-AI collaboration to fully realize the potential of Healthcare 5.0. Finally, in [30], Tian et al. analyzed the potential of smart healthcare technologies, including IoT, AI, and Big Data, in making medical care more efficient and personalized and identified key areas where smart healthcare solutions could improve hospital management and patient engagement, highlighting the necessity of integrating intelligent systems for optimizing healthcare operations.

Across this category, models are predominantly model-centric, focusing on algorithm selection and accuracy optimization, while data quality issues are treated as secondary concerns. This imbalance limits real-world generalization despite strong experimental results.

c) Big Data analytics and infrastructure challenges: In [23], Vijayalakshmi et al. introduced a Big Data-driven model to enhance disease forecasting using machine learning techniques and leveraged Big Data analytics for chronic disease prediction, improving classification accuracy but struggling with data sparsity and labeling errors. Respectively, a broader examination of IoT and Big Data analytics in biomedical healthcare was conducted by Banerjee et al in the chapter [2], who highlighted standardization challenges in wearable device data processing, where latency remained a key limitation. Both noted substantial improvements in data-driven decision-making, but also highlighted latency issues, sparse data, and infrastructural limitations. In [21], Ghazal et al. applied machine learning for smart healthcare infrastructure, optimizing medical resource allocation, but still faced data quality issues. In [29], Šajnović et al. conducted a large-scale synthetic review that analyzes 2272 publications from the Scopus database on the intersection of the Internet of Things (IoT) and Big Data Analytics in preventive healthcare, observing exponential literature growth since 2012, peaking in 2023. The study identifies eight key themes, including the role of AI in personalized medicine (genetics/genomics) and risk prediction, and the use of big data in public health and epidemiology, alongside critical challenges concerning data security, privacy, interoperability, ethical concerns, and the high cost of the IoMT system. In the comprehensive survey [18], Charfare et al. examine the integration of IoT and AI technologies in healthcare applications. The study reviews 28 AI/ML models across diverse applications, including disease detection (COVID-19, diabetes), patient monitoring, athletic performance tracking, and elderly care. Top-performing models

include LightGBM (99.23% accuracy for activity recognition) and LSTM (99% for fall detection). Key challenges identified include data security, computational constraints, energy efficiency, and interoperability issues. The study emphasizes future research directions in federated learning, adaptive algorithms, and multi-modal data fusion for enhanced healthcare delivery. In [19], Alsabah et al. conducted a comprehensive analysis of IoT-based smart healthcare systems, focusing on the interplay between edge computing, cloud services, and data analytics. They examine AI/ML applications in smart healthcare, analyzing studies covering disease diagnosis (glaucoma, diabetic foot ulcers, skin diseases), medical image encryption, and IoT security. Key technologies include CNNs, deep learning, federated learning, and blockchain-based privacy protection. Models demonstrate enhanced diagnostic accuracy, early disease detection, and secure data transmission. Applications span medical imaging, remote health monitoring, and intrusion detection. Their review highlights technical challenges such as system scalability, latency, and real-time decision-making, particularly when handling massive healthcare datasets. They also underscore the need for high-throughput infrastructure capable of managing heterogeneous medical data while ensuring interoperability and low-latency processing. In [27], Chui et al. focused on patient behavior monitoring using IoT and Big Data analytics and developed a framework to analyze patient movement patterns and predict potential health risks, providing valuable insights for caregivers, improving patient safety, and reducing hospital readmissions.

While Big Data enables broader population-level insights, the absence of standardized data pipelines and quality control mechanisms limits scalability and reproducibility—issues that are not addressed by model improvements alone.

d) Privacy, security, and interoperability: The integration of IoT and AI has exposed healthcare systems to data privacy risks and interoperability hurdles. In [25], Aceto et al. examined the role of IoT and Big Data in Healthcare 4.0, improving data fusion for personalized medical services, while highlighting privacy concerns. Similarly, in [31], Yin et al. provided an overview of the Internet of Things in healthcare, summarizing its applications, challenges, and future directions, and discussed the importance of data security, interoperability, and regulatory considerations in IoT-based healthcare systems, emphasizing the need for robust security frameworks to ensure patient data protection. In [24], Verma et al. developed a hybrid secure cloud IoT framework for disease prediction and diagnosis, which improved data security and accuracy of disease diagnosis but raised concerns regarding scalability and privacy.

These studies clearly demonstrate that privacy and interoperability are not peripheral concerns, but structural constraints. However, they remain loosely coupled to AI model design, reinforcing the need for an integrated data-centric framework.

2) Key observations and emerging gaps: These studies converge on a critical insight: while model-centric innovations have improved local accuracy, they often neglect systemic issues related to data quality and governance. Few solutions prioritize data refinement as a strategic goal, dataset representativeness, noise reduction, or robust data integration—gaps which DCAI directly addresses. As shown in Table III, the reviewed studies highlight recurring limitations (e.g., noisy data, lack of standardization, sensor faults, and unstructured inputs) that underscore the need for a paradigm shift toward DCAI methodologies.

To consolidate the expanded comparative insights derived from the 23 reviewed studies, we performed an updated quantitative synthesis of the key limitations reported in AI-, IoT-, and Big Data-driven healthcare systems. As illustrated in Fig. 3, privacy concerns remain the most dominant challenge, cited in 10 studies (43.5%), reflecting persistent issues related to patient data protection, regulatory compliance, and trust in large-scale digital health infrastructures. Data heterogeneity follows closely, reported in 9 studies (39.1%), highlighting the ongoing difficulty of integrating heterogeneous data sources such as IoT sensors, EHRs, and unstructured clinical data. Noisy data was identified in 6 studies (26.1%), underscoring the impact of sensor inaccuracies, transmission errors, and real-world data variability on model reliability. A second tier of limitations includes security risks and integration challenges, same as noisy data limitation each reported in 6 studies (26.1%), indicating that secure data exchange and seamless system interoperability remain unresolved barriers. Real-time processing constraints, data sparsity, and scalability limitations were each mentioned in five studies (21.7%), reflecting the computational and infrastructural challenges of deploying AI systems in dynamic clinical environments. Missing values were cited in 4 studies (17.4%), pointing to data incompleteness issues arising from sensor failures and fragmented health records. Lower-frequency yet significant concerns include standardization issues and sensor faults (same as missing values limitation every four studies, 17.4%), as well as contextual inconsistencies, bias, ethical concerns, and labeling errors (each reported in two to three studies, ~8.7–13%). Collectively, these findings demonstrate that many of the most critical obstacles are fundamentally data-centric rather than model-centric, reinforcing the necessity of Data-Centric AI (DCAI) approaches that prioritize data quality, governance, interoperability, and scalability as prerequisites for trustworthy and effective healthcare AI deployment.

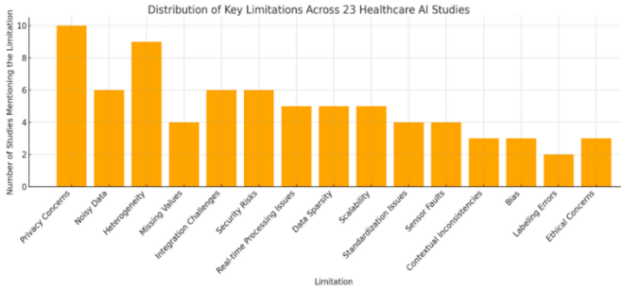

Fig. 3. Distribution of key limitations mentioned across 23 reviewed healthcare AI studies. Privacy, noise, and data heterogeneity emerge as the most cited barriers, underscoring the need for robust data quality and governance frameworks.

3) Summary of comparative insights

a) IoT-based healthcare solutions: Many studies emphasize real-time patient monitoring but lack context-aware and adaptive AI models to dynamically adjust to patient conditions [3], [10].

b) AI for predictive healthcare: AI models demonstrate strong disease prediction capabilities, but their accuracy is often limited by poor data quality, missing values, and a lack of contextual understanding [11], [23].

c) Security and privacy concerns: IoT-driven healthcare systems are highly vulnerable to cyber threats, necessitating secure data-sharing and privacy-preserving [13], [24].

d) Interoperability issues: The absence of standardized data formats and frameworks inhibits seamless integration between healthcare systems and IoT devices, creating barriers to efficient data exchange [25], [31].

These studies demonstrate the growing impact of IoT, AI, and Big Data in modern healthcare, particularly in areas such as real-time monitoring, predictive diagnostics, and personalized treatment. However, most adopt a model-centric approach that prioritizes algorithm performance over data integrity. Persistent data issues continue to undermine scalability, fairness, and reliability. In response, this study advocates for the DCAI paradigm, which prioritizes high-quality, standardized, and context-aware datasets as the foundation for trustworthy and adaptive healthcare systems.

TABLE III. COMPARATIVE ANALYSIS OF HEALTHCARE AI-IOT STUDIES

Ref	Model Used	Application	Data Type & Source	Metrics for Evaluation	Results of the Study	Limitations
[10] Zon et al. (2023)	Various Context-Aware Systems (The study primarily reviews various context-aware models used in healthcare applications, without focusing on a specific single model)	Context-aware healthcare applications that utilize user location, demographic information, activity level, time of day, phone usage patterns, lab/vital metrics, and patient history data.	Healthcare IoT & Sensor Data from 25 peer-reviewed articles found in databases	Accuracy, Sensitivity	Identified key medical contexts for AI adaptation, Improved patient-centered healthcare	Context heterogeneity affects reliability and lack of details on methods for determining contexts and on the systems currently used.
[11] Chung et al. (2020)	Deep Neural Networks	Context-basedHealth risk alert system for chronic disease patients	Electronic Medical Record (EMR), Personal Health Record (PHR) & Public Health Data (PHD) & Open API Data (OAD) (Korean Health & Nutrition Survey data)	Root Mean Square Error (RMSE) and comparison of prediction accuracy rates	Improved risk prediction accuracy, Effective ambient context pattern detection	Limited number of data set scope and integration issues, possibly affecting the reliability of predictions.
[13] Saranya & Fatima (2022)	Improved Context-aware Data Fusion (ICDF) and Enhanced Recursive Feature Elimination (ERFE) Model	IoT-Based Patient Monitoring	Biometrics and health data collected from IoT devices and sensors, including RFID sensors	Accuracy, Precision, , recall, F1-score, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) rates	Enhanced real-time monitoring, reduced false alerts (97.9% accuracy)	Challenges in data integration, potential cyber vulnerabilities in IoT systems, reliance on the accuracy of sensor data, need for regular battery replacements, and scalability challenges
[20] Kishor & Chakraborty (2021)	Machine Learning Classifiers that employ seven classification algorithms: Decision Tree, Support Vector Machine, Naïve Bayes, Adaptive Boosting, Random Forest, Artificial Neural Network, and K-	The early and accurate prediction of various diseases including heart disease, diabetes, breast cancer, hepatitis, liver disorder, dermatology, thyroid issues, surgery data, and spect heart conditions.	IoT Data, Electronic Health Record (HER) , UCI Health datasets	Accuracy, sensitivity, specificity, and area under the curve (AUC).	The Random Forest classifier achieved an accuracy of 97.62%, sensitivity of 99.67%, specificity of 97.81%, and AUC of 99.32%.	Limitations related to the quality of data collected through IoT sensors & Real-time processing issues

	Nearest Neighbor.					
[21] Ghazal et al. (2021)	Machine Learning (From the selected existing studies)	Smart Healthcare Systems , Smart Cities & Patient Monitoring	Discussing the use of Big Data and IoT Sensor Data (biodata transmitted in real-time and historical health records from various healthcare settings)	Efficiency, Prediction Accuracy	Enhanced patient monitoring, Improved diagnostic methods through centralized device monitoring + Potential for decentralized care, especially beneficial in underserviced areas	Potential downsides of smart city technologies, such as privacy and security concerns relating to IoT integration in healthcare
[12] Kim & Chung (2020)	Neural Network-Based Adaptive Context Model	Ambient Intelligence (Adaptive Context Prediction for Health Monitoring)	Heterogeneous data such as medical big data, bio big data, lifelog data, and various health-related contexts from personal health devices, IoT devices, environmental data and structured and unstructured medical data.	F1-score, Recall , RMSE, error rate	The neural-network based similarity weight method achieved the highest prediction accuracy when implemented with the specified parameters & Potential in improving healthcare delivery and quality of life through context predictions.	Potential fairness concerns due to contextual inconsistencies and limited scope of data, potential challenges in accurately predicting changes in context and health conditions and the necessity of substantial advanced data integration techniques to handle large volumes of heterogeneous data effectively.
[22] Kaur (2021)	IoT Framework + Machine Learning from various research findings	Heart Disease Prediction (Remote patient monitoring, medical data analysis, and decision support systems)	IoT Wearables, Patient Data (clinical records, medical images, and structured and unstructured sensor data)	Accuracy, Precision & Recall	High prediction accuracy for cardiac risks	Security concerns around patient data privacy and the potential for data leaks can affects model generalization
[2] Banerjee et al. (2020)	Not a specific model but a acknowledging of Big Data Analytics	Book chapter about IoT & AI in Healthcare for Smart Cities & Healthcare Integration	Wearable Devices, Sensor Networks	Data Efficiency, Latency	Improved data standardization for healthcare analytics, disease detection rates and patient outcomes	The chapter does not detail specific limits but identifies challenges concerning the quality of data in developing countries like Wearable device latency issues , Infrastructure-dependent implementation
[3] Meraj et al. (2021)	IoT Sensor Analysis of real-time surveillance systems	Infectious Disease Detection particularly in pandemic contexts	IoT Sensors, Public Health Data	Outbreak Prediction Accuracy	Enabled early outbreak warnings, Improved infectious disease detection rates and healthcare responses, better real-time care management.	IoT infrastructure for continuous large -scale monitoring, potential for data issues
[23] Vijayalakshmi et al. (2021)	Big Data Analytics	Chronic Disease Prediction and Real-time patient monitoring	Data would likely come from healthcare systems such as electronic medical records (EMRs) / EHR, Medical Reports	Classification Accuracy	Improved long-term disease predictions , clinical outcomes and decision-making processes in healthcare through real-time data analysis	Acknowledging challenges related to data quality and the implementation of IoT systems in certain contexts, Data sparsity limits precision, issues with unstructured data processing
[24] Verma et al. (2021)	Hybrid Secure Cloud IoT Model (Hybrid Encryption & Optimization (HEE, GFI-GWALO))	Disease Prediction & Diagnosis in IoT-based Healthcare	IoT Data, Cloud Computing, including structured, semi-structured and unstructured medical records	Security, Accuracy, encryption time, decryption time	Enhanced secure cloud-based diagnosis , Achieved 100% accuracy, encryption time of 80ms, addressing the challenges of managing health data in IoT-based healthcare systems	Privacy concerns in cloud-based storage, data quality issues and integration challenges
[25] Aceto et al. (2020)	Not a specific model but combining	Personalized Healthcare and	Systematic literature review, secondary data	Literature synthesis	Identified challenges & benefits of Industry 4.0 in healthcare , Improved	Privacy concerns limit implementation (Regulatory and transparency concerns)

	research from IoT, cloud computing, Industry 4.0 and big data analytics to assess their impact on healthcare systems	healthcare service delivery	(Smart Devices, Cloud Data with the use of a "snowball" method to analyze the available research)	(Integration Efficiency)	data fusion for patient management and improved clinical and public health operations through better data analysis and decision-making)	+ the challenges of error-free big data analysis and the management of large amounts of medical data
[14] Aborokbah et al. (2018)	Context-Aware Decision Systems (Support Vector Machines (SVM))	Chronic Disease Management (Cardiac Disease Risk Prediction in Smart cities)	Wearable Devices, IoT Data (Physiological sensor data)	Decision Accuracy	Adaptive healthcare decision support , Effective real-time monitoring of heart failure risk	Contextual inconsistencies in patient data, Limited dataset size, Quality of data collected and generalizability of results.
[26] Castro et al. (2017)	IoT-Based Human Activity Recognition (Bayesian & C4.5 Decision Trees)	Health Monitoring , Human Activity Recognition (HAR)	Wearable Sensor Data (HR, acceleration, respiration)	Pattern Recognition Accuracy, classification rate	Improved activity-based patient monitoring (95.83% accuracy for activity recognition)	Small dataset, limited activity categories, potential of environmental noise affects precision
[27] Chui et al. (2019)	ML + Big Data & IoT	Patient Behavior Monitoring and cardiovascular disease detection in healthcare	Healthcare Institution Data (Wearable & implantable sensors)	Anomaly Detection Rate , Sensitivity, specificity	Better patient behavior analytics , Reduced hospital readmissions, improved health outcomes.	Lack of real case studies, Data heterogeneity affects standardization, Privacy concerns in patient monitoring
[28] Ngiam & Khor (2019)	Deep Learning, Neural Networks + Big Data & ML	AI in Healthcare (medical diagnostics, predicting disease risks, analyzing medical images, and supporting clinical decision-making)	EHR, IoT Data, lab results, textual physician notes, Demographic data, imaging data	Prediction Accuracy, precision, recall	Comprehensive survey on AI applications (High predictive accuracy in clinical settings and better personalization of treatments)'	Ethical concerns in AI adoption, potential challenges in data preprocessing and inaccurate labeling
[29] Shah et al. (2018)	IoT Sensing Systems + SVM , KNN and RF	Healthcare Monitoring, Sleep Disorder Detection (Narcolepsy) and other human activities	Wearable Sensors, IoT Networks (Radar-based Band-S signals)	Reliability, Sensitivity , Classification accuracy	Improved remote patient monitoring (90%+ accuracy in detecting sleep disorders), better care for narcolepsy patients through early detection of sleep episodes, reduction of accidents related to sleep attacks.	Issues in reliability for sleep disorder tracking, Small sample size (the study may be limited by the quality of the data collected due to multipath interference and the need for specialized equipment)
[30] Tian et al. (2019)	AI & IoT-based Smart Healthcare from the reviewed studies	Medical Intelligence (Smart Hospitals), AI-assisted Diagnosis, chronic disease management, patient monitoring and treatment personalization.	EHR, IoT Sensors (Clinical trial datasets, literature reviews)	System Efficiency , Qualitative benefits assessment	Enhanced real-time AI-driven healthcare, Smart healthcare systems improve the efficiency of care, enable faster treatment and reduce costs while improving patients' quality of life.	Privacy concerns in real-time applications, Scalability of AI integration in healthcare
[31] Yin et al. (2016)	IoT-Based Healthcare Framework from the reviewed studies (not a specific model)	IoT in Medical Systems to improve remote patient monitoring and personalized rehabilitation services	Big Data, EHR, datasets from various sources such as hospitals, home-care, rehabilitation centers	Adoption Rate, System interoperability, real-time updates	Overview of IoT applications in healthcare, Improved patient tracking & remote monitoring	Integration challenges remain, Security and data privacy challenges, the complexity of managing big data
[16] Zonayed et al. (2025)	Narrative review of studiens that uses (KNN, SVM, Random	A comprehensive narrative review - Detection and prediction of	Healthcare datasets from UCI and hospital-based EHRs	Qualitative synthesis of existing research	Identified current trends, challenges, Key finding: KNN achieved 100% (breast cancer), RF 96.6%	Feature selection, model interpretability, data imbalance, lack of real-time adaptability, Data security,

	Forest, XGBoost, AdaBoost, CNN, En-RfRsK, HFSDLF)	chronic diseases: breast cancer, diabetes, hepatitis, Alzheimer's, lung cancer			(heart), LR 93.18% (hepatitis), AdaBoost highest in Alzheimer's, CNN+XGBoost 97.43% (lung)	patient privacy ,Scalability , data heterogeneity interoperability issues, context-aware access control and ethical
[29] Šajnoví et al. (2024)	Not applicable (review paper)	Preventive healthcare via IoT and Big Data Analytics	Synthetic review of 2272 papers from Scopus between 1985 and June 10, 2024., including real-world IoT health data and big data systems	Not empirical; thematic synthesis used	Identified major research trends, technologies, and enablers (IoT devices, data lakes, wearable sensors); proposed future directions for intelligent preventive systems	Lack of standardization in data collection, ,heterogeneous data, need for interoperability among IoMT devices.privacy concerns, data integration challenges, limited real-world validation
[18] Charfare et al. (2024)	Multiple AI techniques (e.g., deep learning, ML algorithms) integrated with IoT platforms	Broad spectrum of healthcare applications including remote monitoring, predictive diagnostics, elderly care, and emergency response	Literature-based review of current deployments and innovations across real-world IoT-healthcare systems	Qualitative synthesis (no experimental metrics provided)	Presents a synthesized landscape of IoT-AI convergence, highlighting current innovations, benefits, and key healthcare domains highest performance (LightGBM for HAR (99.23%) and LSTM for fall detection (99%))	Lack of experimental data; challenges discussed include latency, security vulnerabilities, data heterogeneity, insufficient interoperability and need for real-time adaptability
[19] Alsabah et al. (2025)	Not a single model; comprehensive review of IoT architectures integrated with AI, Big Data analytics, edge/fog computing, and cloud platforms	Smart healthcare systems including remote patient monitoring, disease diagnosis, hospital management, emergency response, and personalized healthcare services	Secondary data from peer-reviewed IoT-based healthcare studies, including sensor data, EHR/EMR data, wearable data, and cloud-based medical datasets	Qualitative evaluation based on system scalability, latency, interoperability, data processing efficiency, security, and reliability	Identified key enabling technologies and architectural patterns for smart healthcare; highlighted the role of AI and Big Data in enhancing decision-making, real-time monitoring, and service personalization; provided a consolidated view of technical challenges and future research directions	Challenges remain in scalability, heterogeneity, interoperability, privacy preservation, noisy data, real-world deployment costs and the need of explainable ethical Ai models

III. Challenges in Integrating AI, Big Data, and IoT in Healthcare

Despite the transformative potential of AI, Big Data, and IoT in healthcare, their integration faces substantial barriers that impact reliability, scalability, and ethical deployment. A recurring limitation is data quality, particularly in real-time environments driven by sensor-generated streams.

Challenges stem from noise, missing data, inconsistencies, interoperability issues, privacy concerns, and bias, all of which hinder AI reliability and scalability in healthcare. Sensor noise, such as movement artifacts or calibration errors in ECG devices, distorts predictions and triggers false alarms [13]. Missing data further weakens AI models, leading to biased and unreliable decision-making, as gaps in patient records degrade deep learning performance [1]. Inconsistencies across IoT devices, caused by variations in data formats and protocols, complicate seamless integration, requiring standardized frameworks [21]. Without robust standardization, data fusion from disparate sources remains a technical bottleneck. Bias in IoT-generated datasets, often due to demographic imbalances or inconsistent labeling, results in inequitable AI predictions that disproportionately impact underrepresented populations [12]. Additionally, privacy risks, such as data anonymity and pseudonymization, affect dataset usability and AI predictive accuracy, making privacy governance essential in AI applications [1]. Scalability and real-time adaptability remain obstacles, with many AI models struggling to dynamically adjust to changing patient conditions [12]. Furthermore, interoperability issues between IoT devices, AI models, and hospital systems create siloed environments that restrict data sharing and hinder holistic patient care [30], [31].

Poor data quality in AI-driven healthcare systems can lead to severe outcomes, including misdiagnosis, inequitable care delivery, and loss of clinical trust. According to Jin et al. in [5], imperfections such as incorrect labels, missing values, and anomalies not only degrade model performance but risk overfitting and undermine generalizability—particularly critical in healthcare where model predictions influence life-altering decisions. In [6], Nieberl et al. emphasize that data cascades, which occur when initial quality issues propagate through the pipeline, are a hidden yet profound threat to AI reliability.

Building on our literature-driven hypothesis, we propose that Data-Centric AI (DCAI) offers a viable pathway to overcoming these challenges by emphasizing data quality, fairness, security, and standardization, ensuring more reliable AI-driven healthcare outcomes [4], [9]. Fig. 4 illustrates the major data-related challenges encountered in the integration of AI, Big Data, and IoT for Healthcare.

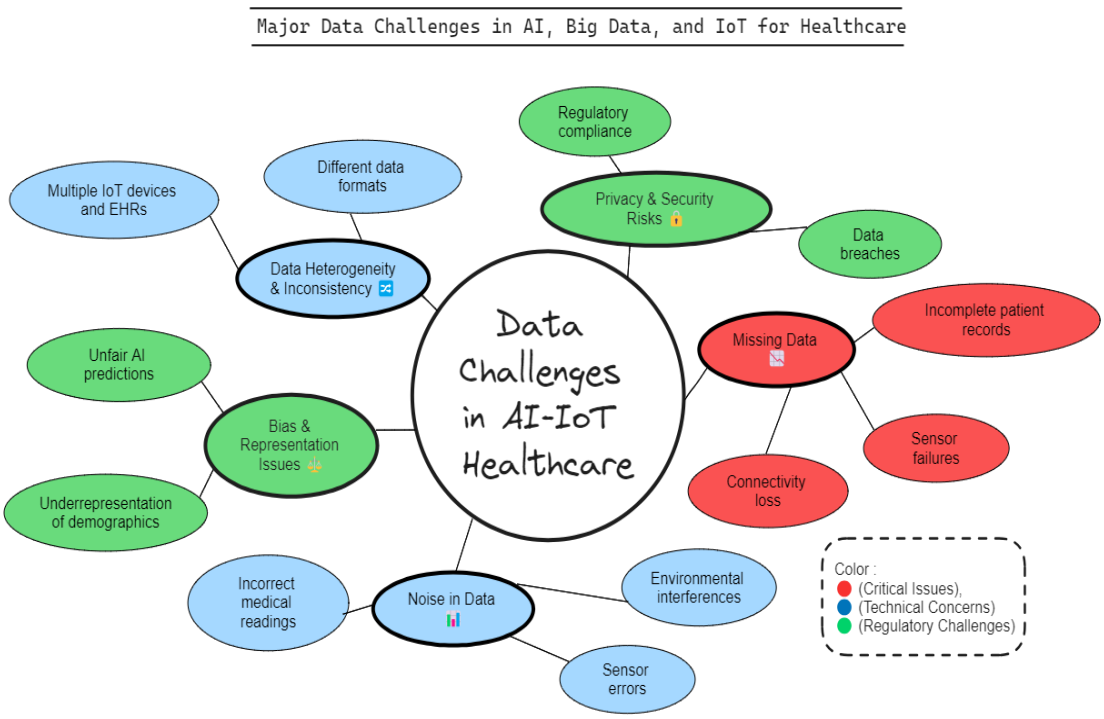

Fig. 4. Major data challenges in AI, Big Data, and IoT for healthcare.

IV. Data-Centric AI: A Paradigm for Addressing Data Challenges in Healthcare

As highlighted in the previous section, the core contribution of our study lies in positioning Data-Centric Artificial Intelligence (DCAI) as a foundational paradigm for addressing longstanding challenges in healthcare AI systems—specifically, issues related to data quality, context-awareness, standardization, and ethical scalability. Recent advances highlight that high-quality, well-annotated, and context-rich data can outperform more complex models trained on flawed datasets, reshaping our understanding of what drives AI success. Unlike traditional AI methodologies, which focus on hyperparameter tuning and deep learning architectures, Data-Centric AI (DCAI) enhances model reliability through systematic data curation [4] (see Table IV).

Despite this shift, few studies have comprehensively mapped how DCAI principles can be applied to healthcare-specific challenges, particularly in the context of IoT-generated data and Big Data analytics. As Jin et al. in [5] argue, a unified, structured synthesis of DCAI methodologies tailored to healthcare is lacking, especially one that considers clinical interoperability, fairness, and regulatory constraints. In healthcare, where the stakes of predictive errors are high and datasets are often noisy, incomplete, heterogenous or biased, DCAI-driven methodologies offer a critical corrective and provide robust solutions to mitigate these challenges [9]. By ensuring high-quality, well-labeled, and diverse datasets, DCAI enables AI systems to generalize effectively across different populations and healthcare environments. As Bhatt et al. in [4] demonstrate that smaller models trained on high-quality datasets can outperform larger models trained on flawed data. Similarly, Malerba et al. in [9] argue that DCAI methodologies—ranging from noise filtering and imputation to data augmentation and bias mitigation—are key to building scalable and ethical healthcare AI.

TABLE IV. Comparison of Data-Centric AI (DCAI) and Model-Centric AI in Healthcare AI

Aspect	Model-Centric AI	Data-Centric AI (DCAI)
Primary Focus	Optimizing AI models and algorithms [4]	Improving data quality, completeness, and fairness [1]
Approach to AI Performance	Improves performance by refining model architecture and hyperparameters [4]	Enhances performance by refining datasets rather than just improving models [1]
Data Handling	Assumes data is fixed and optimizes models to extract insights [9]	Systematic preprocessing, data augmentation, and bias mitigation [4]
Model Complexity	Requires complex models to compensate for suboptimal data [1]	Simpler models trained on high-quality data perform as well as or better than complex models [4]
Scalability	Limited scalability due to dependency on large models and retraining [9]	Highly scalable as data improvements generalize across different models [1]
Bias Mitigation	Bias correction is applied post hoc through model adjustments [1]	Bias is addressed at the dataset level, ensuring fairer and more representative AI outputs [4]
Challenges	Sensitive to data drift, requires frequent model updates, struggles with unreliable data [1]	Requires investment in systematic data collection, labeling, cleaning, and augmentation [4]

DCAI facilitates the standardization and structuring of IoT-generated health data, improving interoperability between AI models, IoT devices, and hospital infrastructures to reduce healthcare system fragmentation. High-quality data is essential for accurate diagnostics, equitable treatment recommendations, and compliance with ethical and regulatory standards. To improve data integrity and to address persistent issues in healthcare, DCAI integrates techniques such as bias detection, noise filtering, and data augmentation [1]. By treating data as a dynamic, evolving infrastructure, DCAI ensures AI systems remain fair, trustworthy, and effective. As highlighted by Zha, Bhat, Lai, Yang, Jiang, Zhong & Hu in [15], DCAI emphasizes continuous data improvement throughout the AI lifecycle, transforming data from a static input into a dynamic, strategic asset. Additionally, real-time adaptive AI models leveraging DCAI principles can dynamically adjust to context-aware decision-making, enhancing patient monitoring and disease prediction by utilizing high-quality labeled datasets that reflect dynamic changes in patient states.

Our study seeks to map recurring limitations identified in AI-IoT healthcare studies we previously discussed, such as inconsistency, bias, and lack of contextual adaptation, to specific DCAI techniques and solutions. We emphasize that issues like heterogeneity, fairness, and reliability are not solely technical but deeply tied to data design. By applying frameworks such as the METRIC framework [1], developers can evaluate healthcare datasets across dimensions like completeness, consistency, timeliness, and contextual fit. This paradigm shift is particularly critical in healthcare, where reliable AI-driven decisions directly impact patient safety and system credibility. Furthermore, robust decision-making relies on high-quality health data to generate consistent and evidence-based insights. By shifting the emphasis from algorithmic optimization to data integrity, interoperability, and real-time adaptability, this study explores the theoretical potential of Data-Centric AI (DCAI) to transform AI-driven healthcare systems into more scalable, ethical, and efficient solutions. Although our work does not empirically validate DCAI interventions, we propose that its principles are particularly well-suited for dynamic, context-aware, real-time healthcare environments, such as remote monitoring or emergency diagnostics. Crucially, DCAI supports the development of adaptive, context-aware AI systems that can respond dynamically to evolving patient conditions using continuously curated and context-rich datasets. This is particularly vital in applications such as remote monitoring and emergency diagnostics, where decision-making must be both rapid and reliable. By reframing the development pipeline around data quality, our study positions DCAI as a catalyst for building more ethical, transparent, scalable, and resilient AI systems in healthcare.

V. Addressing Data Quality Challenges Through Data-Centric AI (DCAI)

DCAI provides a structured framework for overcoming the previously outlined persistent data quality issues in healthcare AI systems. By prioritizing the curation, augmentation, and governance of data over merely tuning models, DCAI supports the development of more accurate, fair, and context-aware AI tools. For instance, bias mitigation is achieved through dataset audits, augmentation strategies, and synthetic data generation, as seen in [4], where underrepresented populations were better included in disease prediction models. Likewise, missing data—often resulting from sensor failures or transmission gaps in IoT-based health monitoring—is addressed through statistical and machine learning-based imputation techniques. In [9], Malerba et al. showcased robust hybrid approaches for managing data gaps, especially in critical use cases like glucose level monitoring.

Furthermore, to tackle noise and inconsistency in heterogeneous healthcare data, DCAI recommends pipelines for standardization and noise filtering, as highlighted by Bhatt et al in [4]. These processes enhance model robustness across diverse data sources and clinical settings. While privacy and security are not traditionally part of data quality metrics, they are integral to data usability and regulatory compliance. In [1], Schwabe et al. underscore that privacy should be treated as a governance concern, though technical solutions like federated learning and differential privacy can still support secure AI deployment. Taken together, these DCAI methods form a foundational toolkit for building scalable, ethical, and effective AI systems in healthcare—aligning data practices with real-world demands for transparency, interoperability, and patient-centered decision-making.

VI. Key Metrics for Evaluating Data Quality in Healthcare AI

The foundational literature on Data-Centric AI (DCAI) offers critical insights into data quality metrics and governance dimensions essential for trustworthy AI systems in healthcare. These include completeness, consistency, accuracy, bias mitigation, and timeliness, alongside data management principles such as privacy, documentation, and alignment with Findable, Accessible, Interoperable, and Reusable (FAIR) principles (see Fig. 5). These metrics are central to developing reliable, ethical, and high-performing healthcare AI systems. In [4], Bhatt et al. highlight the importance of data preprocessing and augmentation in reducing bias and improving dataset robustness. In [1], Schwabe et al. introduce the METRIC framework, which formalizes the assessment of data quality in medical AI applications, particularly in medical imaging and electronic health records (EHRs). In [9], Malerba et al. emphasize standardization, integration, and privacy-preserving AI techniques as foundational to secure and reliable healthcare AI models.

A. Core Data Quality Metrics

- Completeness ensures datasets are comprehensive, preventing unreliable predictions due to missing vitals or sensor gaps, with Schwabe et al. in [1] advocating for imputation techniques to address these issues.
- Consistency focuses on standardizing data formats across heterogeneous sources, as Malerba et al. in [9] emphasize normalization techniques to improve interoperability.
- Accuracy is critical in avoiding errors from faulty sensor readings, with Bhatt et al. in [4] recommending noise filtering methods like Kalman filters to enhance data reliability.

- Bias detection mitigates demographic imbalances in datasets, where techniques such as GAN-based synthetic data generation help improve fairness [4].
- Timeliness ensures real-time decision-making in critical care scenarios, with Schwabe et al. in [1] promoting edge computing to minimize delays.

B. Data Governance and Security Considerations

Though not typically framed as "quality" metrics, data privacy, security, and documentation are crucial for regulatory compliance and patient protection for ethical deployment of healthcare AI. In [9], Malerba et al. emphasize the need for:

- Federated learning to enable secure model training without centralizing sensitive data.
- Differential privacy to ensure anonymity and protect patient confidentiality during analysis and model deployment.

Together, these factors shape trustworthy, high-performance AI systems, ensuring data-driven healthcare solutions are reliable, fair, and efficient. These governance mechanisms are essential for legal compliance (e.g., GDPR, HIPAA), fostering patient trust and ensuring ethical alignment of AI solutions.

C. Standardized Frameworks: The Role of METRIC

The METRIC framework, introduced by Schwabe et al. in [1], provides a structured approach to evaluate medical datasets across key dimensions—completeness, consistency, timeliness, relevance, interoperability, and contextual fit. This framework is particularly useful for multi-source, privacy-sensitive environments like EHR systems and remote IoT monitoring platforms.

Our study underscores the need for standardized, interpretable, and scalable evaluation frameworks such as METRIC, particularly in multi-source, privacy-sensitive healthcare environments. They form the backbone of resilient healthcare AI systems capable of delivering equitable, transparent, and patient-centered outcomes. However, deploying a DCAI Framework in clinical settings is not without challenges—cost, data governance gaps, and system interoperability all pose real-world implementation barriers (see Section VIII (B).

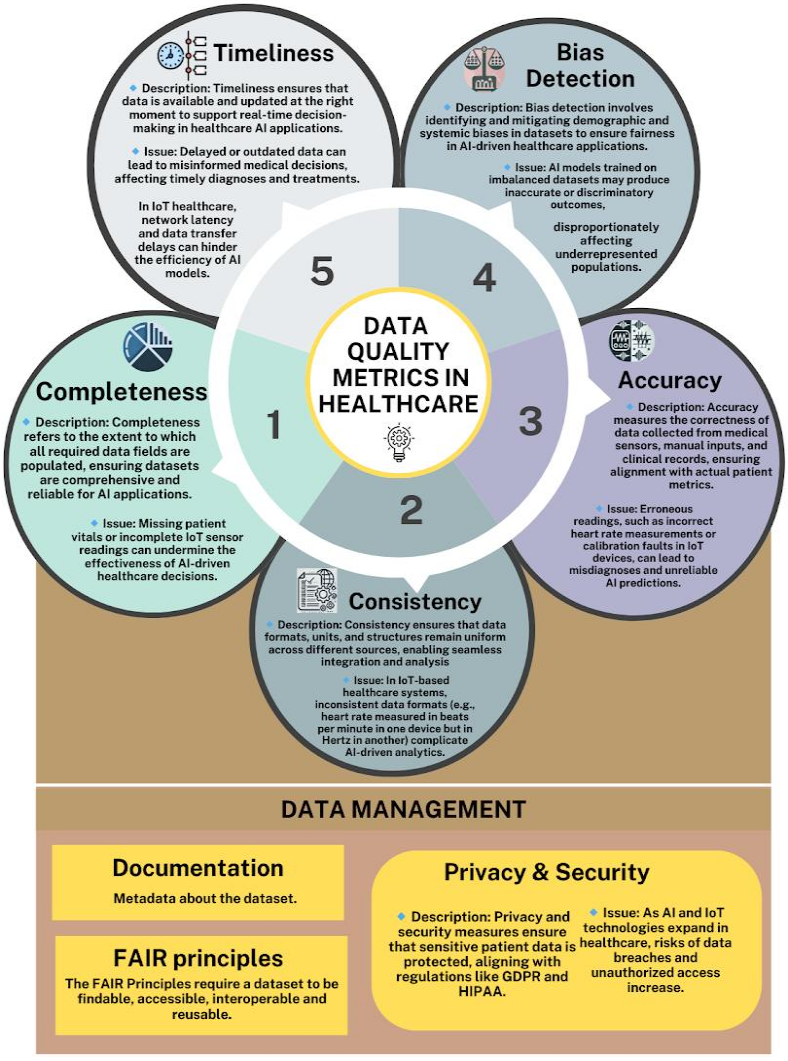

Fig. 5. Data quality metrics in healthcare AI (from the METRIC framework).

D. Summary: Core Principles of DCAI in Healthcare

- Enhancing Data Completeness and Consistency
- Ensure no missing values in critical health datasets and maintain consistency across diverse healthcare sources.
- Mitigating Bias and Ensuring Fairness.
- Promote diverse and well-balanced datasets and eliminate demographic bias in AI-driven medical decision-making.
- Standardizing Healthcare Data and Enabling Interoperability.
- Harmonize heterogeneous data formats and support seamless integration of multiple IoT sources and clinical data to facilitate smooth data exchange.
- Strengthening Data Security and Privacy.
- Apply robust privacy-preserving mechanisms to protect sensitive patient information to ensure confidentiality and prevent unauthorized access or exposure of sensitive medical information.

VII. Challenges and Limitations

Despite the potential of Data-Centric AI (DCAI) to improve healthcare AI systems, several challenges and limitations must be addressed to ensure scalability, fairness, and trustworthiness. These issues stem from both technical and systemic barriers within healthcare data ecosystems.

First, quantifying fairness in AI systems remains inherently complex due to demographic disparities across global healthcare environments. Biased training data can lead to inequitable outcomes, particularly for underrepresented populations, and current bias metrics often fail to capture these nuances accurately.

Second, data variability and quality continue to pose significant hurdles. Real-time data streams from IoT devices require constant monitoring to ensure consistency, contextual relevance, and timeliness. However, as Schwabe et al in [1] highlight, pervasive data issues—including noise, missing values, and inconsistency across heterogeneous sources—complicate integration and reliability of AI models.

Third, privacy and regulatory constraints (e.g., HIPAA, GDPR) restrict access to comprehensive patient data, hindering bias detection and correction. While privacy-preserving methods such as federated learning and differential privacy provide alternatives [9], their implementation is still limited by technical and legal challenges.

Fourth, computational overhead poses another challenge, as evaluating data quality at scale demands significant resources, especially in large IoT and Big Data systems [4].

Fifth, explainability and trust in AI-driven decisions remain critical for clinical adoption, as healthcare professionals require interpretable models and datasets with documentation, annotation, and data provenance for decision support.

Lastly, real-world deployment presents practical barriers. Integrating AI systems with existing hospital IT infrastructure, obtaining regulatory approvals, and ensuring seamless interoperability across data sources all remain major impediments to widespread adoption.

Importantly, while our study synthesizes findings from foundational DCAI literature and healthcare-focused research, it does not incorporate empirical experimentation or advanced AI techniques (e.g., generative AI, self-supervised learning, explainable AI). These promising areas are earmarked for future exploration.

- Key Findings and Gaps
  - Persistent Data Quality Issues: Noise, missing values, and bias remain critical obstacles to reliable healthcare AI.
  - Limited Implementation of DCAI: Although promising, DCAI is rarely applied systematically in healthcare settings.
  - Lack of Standardized Evaluation Metrics: Inconsistent metrics across studies hinder cross-model comparisons and benchmarking.
  - Interoperability Challenges: Diverse data sources (e.g., EHRs, IoT, imaging) often lack harmonization, limiting seamless AI integration.
  - Difficulty Quantifying and Mitigating Bias: Demographic disparities in data create systemic bias that is difficult to measure and correct.
  - Limited Multimodal Integration: Many current models fail to combine diverse healthcare data types (e.g., clinical text, images, sensor streams), reducing contextual awareness and diagnostic performance.

VIII. Future Work

As healthcare AI systems evolve, Data-Centric AI (DCAI) practices—such as data augmentation, bias mitigation, and context-aware preprocessing—are becoming essential for developing ethical, interpretable, and adaptive solutions [6], [15]. While this review provides a conceptual synthesis of DCAI principles, bridging foundational literature and application studies through a structured methodology, significant research opportunities remain. These involve both enhancing core DCAI practices and integrating emerging technologies to improve scalability, fairness, real-time adaptability, and clinical trustworthiness.

A. Key Future Research Directions

Below, we outline key future research directions to extend the impact and applicability of DCAI in AI-driven healthcare systems:

1) Generative AI for data augmentation: Explore the use of generative models such as GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) to create realistic synthetic healthcare data. This approach could reduce data scarcity, correct demographic imbalances, and improve model fairness and generalizability.

2) Multimodal AI frameworks: Investigate the integration of diverse data modalities—including EHRs, IoT sensor streams, medical imaging, and genomics—to build comprehensive and context-rich patient profiles. Evaluate their impact on diagnostic accuracy, clinical decision support, and patient stratification.

3) Advanced bias mitigation techniques: Develop and validate fairness-aware training strategies and bias auditing tools that operate across the entire AI pipeline. This includes addressing hidden biases in data collection, labeling, and model inference stages.

4) Transformer-based and hybrid AI models: Assess the role of cutting-edge architectures such as transformers and graph neural networks (GNNs) in modeling relational and temporal dependencies in complex healthcare environments, particularly where patient data is sparse, noisy, or interconnected.

5) Context-aware and self-adaptive AI systems: Design AI systems capable of dynamically adjusting to patient-specific contexts and environmental changes in real-time. Such models could enhance remote patient monitoring, emergency diagnostics, and chronic disease management by incorporating adaptive decision logic.

6) Explainable and trustworthy AI (XAI): Integrate explainability mechanisms into DCAI workflows to improve clinical interpretability and transparency. This involves applying interpretable models or post-hoc explanation tools (e.g., SHAP, LIME) to enhance clinician trust and facilitate regulatory approval.

7) Privacy-preserving AI methods: Implement and empirically validate federated learning, homomorphic encryption, and differential privacy to ensure regulatory compliance (e.g., HIPAA, GDPR) while maintaining model utility and protecting sensitive patient data.

8) Standardized IoT interoperability: Advance the development of universal communication protocols and interoperability standards for integrating heterogeneous IoT devices and healthcare platforms. This is essential for real-time data sharing, multi-system alignment, and scalable deployment.

B. Practical Considerations and Conclusion of the Section

While Data-Centric AI (DCAI) offers a promising path to improving fairness, accuracy, and scalability in healthcare AI, its implementation is hindered by practical barriers such as high resource requirements, lack of data governance structures, and limited interoperability in clinical settings. Many hospitals operate with fragmented IT systems, unstructured data silos, and unclear data ownership, which complicate systematic data curation and annotation workflows required for DCAI adoption [1], [5]. Implementing privacy-preserving techniques like federated learning or synthetic data generation demands not only technical expertise but also legal and institutional coordination, which may be lacking in resource-constrained environments [4], [9]. Moreover, most DCAI frameworks are designed for static datasets and are not easily adaptable to dynamic, real-time healthcare scenarios where context, data quality, and ethical oversight must constantly evolve. Addressing these challenges requires interdisciplinary collaboration, scalable governance models, and cost-effective toolkits tailored for healthcare infrastructure, ensuring that DCAI moves from conceptual promise to real-world impact.

IX. Conclusion

This study positions Data-Centric Artificial Intelligence (DCAI) as a transformative paradigm for advancing trustworthy, scalable, and ethical AI-driven healthcare systems. Unlike traditional model-centric approaches that emphasize algorithmic optimization, DCAI redirects focus toward systematic improvements in data quality—enhancing completeness, consistency, standardization, fairness, and interoperability. Through our literature-based analysis of twenty-three AI-IoT healthcare studies and foundational DCAI works, we identified persistent challenges related to noise, missing values, demographic bias, contextual inconsistency, and privacy concerns. DCAI addresses these limitations by integrating practical techniques such as data augmentation, noise filtering, bias mitigation, and standardized preprocessing, while frameworks like METRIC offer structured tools for evaluating data quality across key dimensions. Although our study is theoretical in nature, it establishes a foundational synthesis linking DCAI principles to healthcare-specific data challenges. Future directions should focus on integrating DCAI principles with cutting-edge technologies such as explainable AI (XAI), generative AI, and multimodal learning, while also advancing real-time, context-aware AI–IoT convergence in smart healthcare environments. Such integration will enable adaptive systems capable of continuous learning, secure data handling, and interpretable decision support—crucial for delivering equitable and high-quality care at scale.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Special appreciation is extended to the researchers cited throughout the study. Gratitude is also extended to the academic mentors and reviewers whose comments helped to refine the manuscript.

References

[1] D. Schwabe et K. Becker, « The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review | npj Digital Medicine », NPJ Digital Medicine, vol. 7, no 203, 2024, Consulté le: 28 octobre 2024. [En ligne]. Disponible sur: https://www.nature.com/articles/s41746-024-01196-4

[2] A. Banerjee, C. Chakraborty, A. Kumar, et D. Biswas, « Chapter 5 - Emerging trends in IoT and big data analytics for biomedical and health care technologies », in Handbook of Data Science Approaches for Biomedical Engineering, vol. 5, V. E. Balas, V. K. Solanki, R. Kumar, et M. Khari, Éd., Academic Press, 2020, p. 121-152. doi: 10.1016/B978-0-12-818318-2.00005-2.

[3] M. Meraj, S. A. M. Alvi, M. T. Quasim, et S. W. Haidar, « A Critical Review of Detection and Prediction of Infectious Disease using IOT Sensors », in 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), août 2021, p. 679-684. doi: 10.1109/ICESC51422.2021.9532992.

[4] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, et W. El-Shafai, « A Data-Centric Approach to improve performance of deep learning models », Sci Rep, vol. 14, no 1, p. 22329, sept. 2024, doi: 10.1038/s41598-024-73643-x.

[5] W. Jin et al., « DCAI: Data-centric Artificial Intelligence », in Companion Proceedings of the ACM Web Conference 2024, Singapore Singapore: ACM, mai 2024, p. 1482-1485. doi: 10.1145/3589335.3641297.

[6] M. Nieberl, A. Zeiser, et H. Timinger, « A Review of Data-Centric Artificial Intelligence (DCAI) and its Impact on manufacturing Industry: Challenges, Limitations, and Future Directions », in 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore: IEEE, juin 2024, p. 44-51. doi: 10.1109/CAI59869.2024.00018.

[7] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, et X. Hu, « Data-centric AI: Perspectives and Challenges », 4 avril 2023, arXiv: arXiv:2301.04819. doi: 10.48550/arXiv.2301.04819.

[8] M. H. Jarrahi, A. Memariani, et S. Guha, « The Principles of Data-Centric AI », Commun. ACM, vol. 66, no 8, p. 84-92, août 2023, doi: 10.1145/3571724.

[9] D. Malerba et V. Pasquadibisceglie, « Data-Centric AI | Journal of Intelligent Information Systems », Journal of Intelligent Information Systems, vol. 39, no 4, 2024, Consulté le: 28 octobre 2024. [En ligne]. Disponible sur: https://link.springer.com/article/10.1007/s10844-024-00901-9

[10] M. Zon, G. Ganesh, M. J. Deen, et Q. Fang, « Context-Aware Medical Systems within Healthcare Environments: A Systematic Scoping Review to Identify Subdomains and Significant Medical Contexts », International Journal of Environmental Research and Public Health, vol. 20, no 14, Art. no 14, janv. 2023, doi: 10.3390/ijerph20146399.

[11] K. Chung, H. Yoo, et D.-E. Choe, « Ambient context-based modeling for health risk assessment using deep neural network », J Ambient Intell Human Comput, vol. 11, no 4, p. 1387-1395, avr. 2020, doi: 10.1007/s12652-018-1033-7.

[12] J.-C. Kim et K. Chung, « Neural-network based adaptive context prediction model for ambient intelligence », J Ambient Intell Human Comput, vol. 11, no 4, p. 1451-1458, avr. 2020, doi: 10.1007/s12652-018-0972-3.

[13] S. S. Saranya et N. S. Fatima, « IoT-Based Patient Health Data Using Improved Context-Aware Data Fusion and Enhanced Recursive Feature Elimination Model», IEEE Access, vol. 10, p. 128318-128335, 2022, doi: 10.1109/ACCESS.2022.3226583.

[14] M. M. Aborokbah, S. Al-Mutairi, A. K. Sangaiah, et O. W. Samuel, « Adaptive context aware decision computing paradigm for intensive health care delivery in smart cities—A case analysis », Sustainable Cities and Society, vol. 41, p. 919-924, août 2018, doi: 10.1016/j.scs.2017.09.004.

[15] D. Zha et al., « Data-centric Artificial Intelligence: A Survey », 13 juin 2023, arXiv: arXiv:2303.10158. doi: 10.48550/arXiv.2303.10158.

[16] M. Zonayed et al., « Machine learning and IoT in healthcare: Recent advancements, challenges & future direction », Advances in Biomarker Sciences and Technology, vol. 7, p. 335-364, 2025, doi: 10.1016/j.abst.2025.08.006.

[17] U. Šajnović, H. B. Vošner, J. Završnik, B. Žlahtič, et P. Kokol, « Internet of Things and Big Data Analytics in Preventive Healthcare: A Synthetic Review », Electronics, vol. 13, no 18, p. 3642, sept. 2024, doi: 10.3390/electronics13183642.

[18] R. H. Charfare, A. U. Desai, N. N. Keni, A. S. Nambiar, et M. M. Cherian, « IoT-AI in Healthcare: A Comprehensive Survey of Current Applications and Innovations », IJRCS, vol. 4, no 3, p. 1446-1472, août 2024, doi: 10.31763/ijrcs.v4i3.1526.

[19] M. Alsabah et al., « A comprehensive review on key technologies toward smart healthcare systems based IoT: technical aspects, challenges and future directions », Artif Intell Rev, vol. 58, no 11, p. 343, août 2025, doi: 10.1007/s10462-025-11342-3.

[20] A. Kishor et C. Chakraborty, « Artificial Intelligence and Internet of Things Based Healthcare 4.0 Monitoring System », Wireless Personal Communications, vol. 127, p. 1615-1631, 2021, doi: 10.1007/s11277-021-08708-5.

[21] T. M. Ghazal et al., « IoT for Smart Cities: Machine Learning Approaches in Smart Healthcare—A Review », Future Internet, vol. 13, no 8, Art. no 8, août 2021, doi: 10.3390/fi13080218.

[22] B. Kaur, « IOT Framework for Heart Diseases Prediction Using Machine Learning », IJATCSE, vol. 10, no 3, p. 2036-2041, juin 2021, doi: 10.30534/ijatcse/2021/781032021.

[23] S. Vijayalakshmi, A. Saini, A. Srinivasan, et N. K. Singh, « Disease prediction over big data from healthcare institutions », in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), mars 2021, p. 914-919. doi: 10.1109/ICACITE51222.2021.9404567.

[24] A. Verma, G. Agarwal, A. K. Gupta, et M. Sain, « Novel Hybrid Intelligent Secure Cloud Internet of Things Based Disease Prediction and Diagnosis », Electronics, vol. 10, no 23, Art. no 23, janv. 2021, doi: 10.3390/electronics10233013.

[25] G. Aceto, V. Persico, et A. Pescapé, « Industry 4.0 and Health: Internet of Things, Big Data, and Cloud Computing for Healthcare 4.0 », Journal of Industrial Information Integration, vol. 18, p. 100129, juin 2020, doi: 10.1016/j.jii.2020.100129.

[26] D. Castro, W. Coral, C. Rodriguez, J. Cabra, et J. Colorado, « Wearable-Based Human Activity Recognition Using an IoT Approach », Journal of Sensor and Actuator Networks, vol. 6, no 4, Art. no 4, déc. 2017, doi: 10.3390/jsan6040028.

[27] K. T. Chui, R. W. Liu, M. D. Lytras, et M. Zhao, « Big data and IoT solution for patient behaviour monitoring », Behaviour & Information Technology, vol. 38, no 9, p. 940-949, sept. 2019, doi: 10.1080/0144929X.2019.1584245.

[28] K. Y. Ngiam et I. W. Khor, « Big data and machine learning algorithms for health-care delivery », The Lancet Oncology, vol. 20, no 5, p. e262-e273, mai 2019, doi: 10.1016/S1470-2045(19)30149-4.

[29] S. A. Shah et al., « Internet of Things for Sensing: A Case Study in the Healthcare System », Applied Sciences, vol. 8, no 4, Art. no 4, avr. 2018, doi: 10.3390/app8040508.

[30] S. Tian, W. Yang, J. M. L. Grange, P. Wang, W. Huang, et Z. Ye, « Smart healthcare: making medical care more intelligent », Global Health Journal, vol. 3, no 3, p. 62-65, sept. 2019, doi: 10.1016/j.glohj.2019.07.001.

[31] Y. Yin, Y. Zeng, X. Chen, et Y. Fan, « The internet of things in healthcare: An overview », Journal of Industrial Information Integration, vol. 1, p. 3-13, mars 2016, doi: 10.1016/j.jii.2016.03.004.