Vol. 3, No. 1, March 2026

IJAIS International Journal of Artificial Intelligence and Science

Published By Asosiasi Doktor Sistem Informasi Indonesia

AI-Driven Intrusion Detection in Healthcare Systems: A Systematic Review of Techniques, Evaluation Practices and Deployment Readiness, with a Reusable Assessment Framework

Belinda Ndlovu[1]

[1]Informatics and Analytics Department, National University of Science and Technology, Zimbabwe

Email: belinda.ndlovu@nust.ac.zw[1]

Received:
December 20, 2025
Revised:
January 3, 2026
Accepted:
February 3, 2026
Published:
March 20, 2026

Corresponding Author:
Author Name*:
Belinda Ndlovu
Email*:
Belinda.ndlovu@nust.ac.zw

DOI: 10.63158/IJAIS.v3i1.58

Abstract: Cyberattacks threaten healthcare data security and clinical continuity. Although artificial intelligence (AI) promises improved intrusion detection, high accuracy rarely translates into clinical deployment. This systematic review examines evaluation practices and deployment conditions in healthcare intrusion detection research. Following PRISMA 2020, searches across ScienceDirect, Scopus, PubMed, MDPI and SpringerLink covered studies published between 2020 and 2024. Eleven studies met the inclusion criteria. Findings were triangulated against subsequent large-scale reviews to assess consistency across broader evidence. Classical machine learning dominated, with limited adoption of hybrid, federated or transformer-based approaches. Reported accuracy averaged 98.66%, predominantly using benchmark or synthetic datasets under controlled conditions. False positive rates, alert volumes, inference latency and robustness to distribution shift were rarely reported. No included study demonstrated sustained operation in a live clinical network. The review introduces the Healthcare IDS Deployment Readiness Assessment (HIDRA), comprising five dimensions and five ordinal readiness levels, alongside a minimum reporting set and six testable propositions. No study exceeded readiness level 2 of 5. HIDRA offers authors, reviewers and procurement teams a reusable instrument for assessing deployment readiness and identifying gaps between experimental performance and clinical feasibility.

Keywords: Artificial Intelligence; Intrusion Detection Systems; Healthcare Cybersecurity; Deployment Readiness; Systematic Literature Review.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Healthcare information systems are among the most heavily targeted assets in cyberspace, and the consequences of compromise now extend well beyond the loss of confidentiality [1]. The 2017 WannaCry ransomware incident, which disrupted National Health Service facilities across the United Kingdom, established that a cyber incident can degrade clinical service delivery rather than only data governance [2]. The scale of exposure has since grown by an order of magnitude. The February 2024 ransomware attack on Change Healthcare, a United States claims clearing house processing a substantial share of national healthcare transactions, was ultimately reported to the Office for Civil Rights as affecting approximately 192.7 million individuals, making it the largest healthcare data breach recorded to date [32]. The incident is instructive for intrusion detection research in three respects: the initial access vector was a single credential used on a system without multi-factor authentication, the compromise propagated through a business associate rather than a covered entity, and the operational disruption to claims and pharmacy processing was felt across the provider base rather than at one institution.

These characteristics matter because they describe the class of intrusion that detection systems are expected to catch, and they differ from the traffic characteristics of the benchmark datasets on which most published models are trained and evaluated. This mismatch between the threat as it presents operationally and the threat as it is represented in evaluation data is a recurring theme of this review. Attacks on healthcare facilities commonly originate in identity theft, insider misuse of healthcare services, and unauthorised access by malicious actors, including through compromised staff credentials [3]. The resulting harms include denial of service, data loss, breaches of confidentiality, financial loss, data destruction, legal liability and, most consequentially, risk to patient safety [4].

The attack surface has also broadened. The Internet of Medical Things now encompasses infusion pumps, monitoring equipment, imaging systems and wearable sensors, many of which run unpatchable firmware, cannot host an endpoint agent, and communicate over protocols that general purpose network intrusion detection systems parse poorly [20], [29]. Detection in this setting is therefore constrained not only by the difficulty of the

classification problem but by where a detector can physically be placed and what it is permitted to observe.

Dataset development has begun to respond to this. CICIoMT2024 provides multi-protocol capture across a large number of attack categories specifically for medical Internet of Things assessment [23], and represents a substantial improvement on the general-purpose corpora used in much earlier work [24]. It nonetheless remains a laboratory construction rather than operational clinical traffic, and the distinction is central to the argument of this review. Conventional defences have proved insufficient in the healthcare setting [3]. Signature-based intrusion detection performs well against known threats but does not detect zero-day attacks [6], [7]. Anomaly-based detection can in principle identify novel behaviour, but in practice generates high false positive rates and requires large volumes of representative training data to construct reliable profiles of normal activity, which healthcare environments seldom supply [4], [5]. Perimeter defences and general-purpose detection tools interpret clinical communication protocols and medical data sources poorly, and rule-based systems cannot represent the correlations across heterogeneous attributes that characterise intrusion in this setting.

In light of these criticisms, artificial intelligence-based solutions are being recognized as a tool for strengthening security of health care infrastructure. Several systematic literature reviews were identified to explore artificial intelligence-based intrusion detection solutions. Advanced learning-based solutions, deep reinforcement learning, machine learning models for anomaly detection, and data-driven approaches for healthcare security were the topics of focus for these reviews [9], [10], exposing numerous potentials of artificial intelligence as a framework capable of analyzing a refined spectrum of heterogeneous data, attending to subtle patterns in them, and taking proactive measures against potential and emerging threats. In particular, recent advances in this field are not static; data-driven solutions dramatically outperform static solutions, owing to a more reliable and precise detection mechanism for potential threats. Accordingly, this emphasis is laid on the exploration of artificial intelligence for threat and intrusion detection in healthcare infrastructure, taking into consideration the practitioner application of artificial intelligence mechanisms in various stages of intrusion detection, the model type, and their efficacies in evaluating against various

criteria as well as deployment settings and factors that may influence the performance of artificial intelligence-based intrusion detection systems.

Despite the growing number of research studies on AI-based intrusion detection systems, existing work has focused more on algorithmic cataloging and performance evaluation. However, evaluation methodologies and deployment conditions are not thoroughly explored. This study attempts to bridge this knowledge gap by exploring evidence on AI methods, evaluation methodologies, and deployment conditions. By compiling data on algorithmic trends, performance reporting, and deployment conditions, this study aims to contribute to a better understanding of AI-based intrusion detection systems beyond algorithmic performance. Given the increase in the use of artificial intelligence for healthcare intrusion detection and the gaps in evaluation practices, deployment conditions, and constraints, there is a need to synthesize the literature on the subject. To this end, this review is informed by research questions aimed at addressing gaps in the use of artificial intelligence for healthcare intrusion detection. The research questions are important because they provide an analytical framework for the review of the subject matter.

Since this review was conducted, several large-scale syntheses of intrusion detection for medical systems have appeared. Adohinzin and Harrath [20] analyse 53 studies of intrusion detection for the Internet of Medical Things, organising them through a taxonomy spanning classical machine learning, deep learning, transformer-based models, federated frameworks and anomaly detection. Ogunseyi et al. [21] review 129 studies of explainable deep learning intrusion detection for the Internet of Things, with attention to the trade-off between detection performance, computational overhead and explanation quality. Further surveys address datasets, federated learning and blockchain integration for medical intrusion detection [33], [34]. These reviews are larger in corpus and broader in technical scope than the present study, and readers seeking an exhaustive algorithmic taxonomy should consult them.

They fall outside this review's search window and are therefore not part of the synthesised corpus. They are nonetheless discussed here for two reasons: the present contribution should be positioned against them rather than presented in isolation, and

they provide an independent test of whether this review's findings have held as the literature has grown.

The contribution offered here is different in kind rather than in size. Where those reviews organise the literature by technique, this review takes evaluation practice and deployment condition as the unit of analysis, and asks what the published evidence would have to look like before a healthcare provider could justify procurement. That question is not answered by a taxonomy of architectures. It requires an explicit account of what was measured, under what conditions, against what data, and with what reported failure behaviour. The convergence between the present findings and those of the larger reviews is itself informative: the observation that evaluation is accuracy-centric and conducted on benchmark data is reproduced in corpora five and ten times larger, which indicates that the pattern is a property of the field rather than an artefact of a small sample.

Accordingly, the eleven-study corpus, drawn from the 2020 to 2024 window, is used for close synthesis, and the subsequent reviews are used as an external check on whether the patterns identified have persisted as the literature has grown. Where the two disagree, this is reported. The principal output is not the corpus description but the assessment framework derived from it and presented in Section 4. This systematic review addresses the following research questions:

1) RQ1: How is artificial intelligence applied across different stages of intrusion detection pipelines in healthcare systems?
2) RQ2: Which artificial intelligence techniques are most commonly employed for intrusion detection in healthcare environments?
3) RQ3: What benefits do artificial intelligence-based approaches offer for threat and anomaly detection in healthcare cybersecurity?
4) RQ4: How is the performance of artificial intelligence-based intrusion detection models evaluated in healthcare settings, and under what conditions are these evaluations conducted?
5) RQ5: What technical, organisational, and regulatory challenges limit the effectiveness and real-world deployment of AI-driven intrusion detection systems in healthcare

2. METHODS

2.1. Review Design

This study follows a systematic literature review approach for synthesising existing evidence, and is reported in accordance with the PRISMA 2020 statement [13], which updates the original PRISMA reporting guideline [14]. The review was designed to characterise not only which artificial intelligence techniques are applied to healthcare intrusion detection, but under what evaluation conditions their performance is established and what those conditions imply for operational deployment.

2.2. Database Search Strategy

A systematic literature search was conducted across five electronic databases widely recognised for publishing research in healthcare, cybersecurity, and artificial intelligence. These included ScienceDirect, Scopus, PubMed, MDPI, and SpringerLink. The databases were selected to ensure broad coverage of peer-reviewed literature spanning medical informatics, applied artificial intelligence, and cybersecurity research. The initial search yielded a total of 206 records, comprising 65 articles from ScienceDirect, 56 from Scopus, 37 from PubMed, 30 from MDPI, and 18 from SpringerLink.

Search queries combined keywords relating to artificial intelligence, intrusion detection, anomaly detection and healthcare cybersecurity, joined with Boolean operators to balance retrieval breadth against relevance. Search terms included artificial intelligence, machine learning, intrusion detection, anomaly detection, healthcare cybersecurity and medical systems security. The search window was set at 2020 to 2024. The lower bound was chosen because intrusion detection for medical cyber-physical systems became an identifiable research stream at approximately that point, and the upper bound reflects the date on which the search was executed. In addition, the reference lists of included studies were manually screened to identify relevant publications not captured through the database search. Work published after the window closed is not part of the synthesised corpus, but is used in Section 1.2 to test whether the review's findings have held.

The review protocol was specified in advance of screening and covered the research questions, search strings, inclusion and exclusion criteria, data extraction fields and

synthesis approach. The protocol was not registered with PROSPERO, since that register does not routinely accept reviews outside health-intervention research; this is stated explicitly because unregistered protocols permit post hoc criterion adjustment, and readers are entitled to weigh the review accordingly. Screening was conducted by a single reviewer, which is a further limitation returned to in Section 8. A typical example of a search string used in this study:

(("Artificial Intelligence" OR "AI" OR "Machine Learning" OR "ML" OR "Deep Learning" OR "DL") AND ("Intrusion Detection" OR "Anomaly Detection" OR "IDS" OR "Threat Detection") AND ("Healthcare" OR "Medical" OR "Health"))

2.3. Inclusion and Exclusion Criteria

Studies were included in the review if they met the following criteria:

1) Focused on AI-based intrusion detection or anomaly detection within healthcare or healthcare-related systems.
2) Presented empirical, experimental, or implementation-oriented findings.
3) They were published in peer-reviewed journals or conference proceedings.
4) They were written in English.

Studies were excluded if they:

1) Addressed cybersecurity or intrusion detection in non-healthcare domains without clear applicability to healthcare environments.
2) Were purely conceptual, editorial, or opinion-based in nature.
3) Focused exclusively on cryptographic mechanisms, access control, or policy issues without an intrusion detection component.
4) Represented duplicate publications or extended abstracts.

The English-language restriction was applied to ensure consistency in interpretation and quality assessment. This criterion is acknowledged as a potential limitation and is discussed later in the paper.

2.4. Study Selection and Screening Process

Screening proceeded in three stages. Of the 206 records retrieved, deduplication and removal of clearly irrelevant records left 187 for screening. Title and abstract screening were then conducted against the criteria set out in Section 2.3. This stage eliminated 116

records that did not concern healthcare cybersecurity, did not involve an AI-based intrusion detection component, or fell outside the publication window, leaving 71 records.

A second screening pass over the remaining 71 records assessed the full abstract and, where necessary, the methods section against the empirical-content requirement, since a substantial number of the retained records were conceptual, positional or review articles without primary findings. This pass excluded a further 50 records and produced 21 studies for full-text eligibility assessment. At full-text screening, 10 studies were excluded, giving a final corpus of 11. The complete selection process, with counts at each stage, is shown in the PRISMA flow diagram in Figure 1. The reported counts are 206 records identified, 187 after duplicate removal, 71 after title and abstract screening, 21 sought for retrieval, and 11 included in the synthesis.

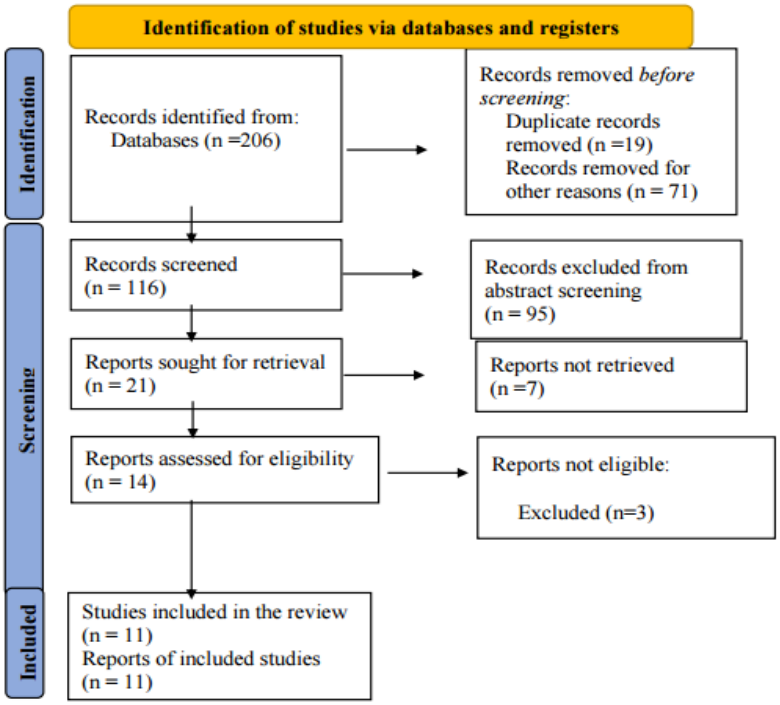

Figure 1. PRISMA Workflow Diagram

2.5. Included Studies

The final review corpus comprised 11 studies that met the inclusion criteria and demonstrated clear relevance to AI-driven intrusion detection in healthcare cybersecurity, of which nine are peer-reviewed as discussed in Section 2.6. Together these studies span a range of artificial intelligence techniques, healthcare system

settings and evaluation strategies, and they form the basis for the synthesis reported in Section 3 and for the readiness assessment in Section 4.

2.6. Quality Assessment

The quality of the eleven research studies was assessed for their methodological rigor using a structured framework for evaluating the quality of research studies on AI-based cybersecurity and healthcare security systems, as discussed in previous systematic reviews on the same topics [8], [10], [18], [19]. The quality evaluation framework considered five key factors for evaluating the quality of research studies on AI-based cybersecurity and healthcare security systems, i.e., research objectives, appropriateness of AI models, transparency in research methodology, evaluation criteria, and the research study's relevance to healthcare cybersecurity scenarios.

Research studies with clear research objectives, appropriate application of AI models, transparent research methodology, and evaluation criteria were considered to be of high quality. The research studies that fall under the high-quality research study category are [4], [6], [8], and [17], as they had discussed their research methodology in detail and had appropriate research study objectives.

Research studies with appropriate application of AI models were considered to be of moderate quality if they had appropriate research methodology, evaluation criteria, and research study objectives, but had shortcomings in other areas such as insufficient discussion on research study objectives, research methodology, and evaluation criteria. The research studies [7], [15], and [16] fall under the moderate quality research study category. A few research studies were considered to be of low quality due to insufficient discussion on research study objectives, research methodology, and evaluation criteria. The research studies on AI-based intrusion detection discussed in [11] and [12] fall into the low-quality research category.

Though the quality evaluation was conducted to understand the research studies, it was not used to filter out research studies for the synthesis phase to avoid ignoring emerging research studies on the same topic. One consequence of retaining lower-quality studies must be stated plainly. Two items in the corpus, [11] and [12], are not published in peer-reviewed journals or indexed conference proceedings: one is a repository preprint and

the other appears in a journal without established indexing. This is inconsistent with the stated inclusion criterion requiring peer review. They are retained because both were assessed as low quality and neither contributes quantitative performance data to the synthesis, so their removal would not alter any reported finding; but their presence in the corpus is a defect in screening execution rather than a design choice, and any reader recomputing the synthesis should exclude them. The nine remaining studies are peer-reviewed.

2.7. Risk of Bias Assessment

The overall risk of bias across all included studies was qualitatively assessed to determine whether there are factors that may affect their reliability and generalizability. Several biases were identified. Dataset bias was identified in a large number of studies that used publicly available or synthetic datasets. Studies that used datasets such as CMU-CERT or synthetic network data may not accurately reflect the complexity and diversity of a healthcare environment. The studies cited may be prone to overfitting and may not adequately capture the complexity and diversity of a healthcare setting. Additionally, dependence on a single set of data may compromise the performance of the model.

Assessment bias was identified, especially in studies that used accuracy to measure model performance. This may not adequately represent how the model works in a healthcare setting because adversarial cases are not common. Deployment bias was identified, especially in studies that used performance in a controlled environment to measure model performance. This may not adequately represent model performance in a healthcare setting. Algorithmic bias was identified, especially in studies that did not consider adversarial robustness. This may not adequately represent model performance in healthcare settings because adversarial cases may be experienced.

3. RESULTS AND DISCUSSION

The PRISMA workflow diagram below summarizes all the delimitation processes of the research through the final selection of articles. A table has been developed to address the research questions and provide a comprehensive synthesis of the findings from the analyzed and reviewed articles. Table 1 is an important tool for exploring the main aspects of the research in AI-driven intrusion detection in healthcare security, including:

1) The research methodologies employed by researchers.
2) The specific and exact AI techniques and algorithms that are applicable in detecting and addressing security threats.
3) The conclusions drawn regarding the effectiveness, efficiency, and limitations of these AI techniques in protecting healthcare systems from threats and attacks.
4) The main objectives driving the research are protecting patient data, improving system resilience against attacks, and facilitating real-time threat detection and response.

Table 1. Summary of included studies

Author	Country	Research Aim	AI Techniques	Benefits	Challenges	Model Accuracy
[15]	South Africa	Use AI to enhance intrusion detection in Software Defined Wireless Sensor Networks (SDWSNs).	ANN Decision Tree Naïve Bayes	Improved intrusion detection in SDWSNs.	Difficulty in choosing the best anomaly detector due to performance variations across datasets. Limited research on intrusion detection in SDWSNs.	99.94%
[16]	India	Propose HealthGuard, a security framework that uses machine learning to safeguard healthcare systems.	Decision Tree KNN, RF ANN	Enhanced security for healthcare data.	Complexity of implementing robust security measures in healthcare systems	93%
[4]	Taiwan	Develop a hybrid deep learning framework to enhance intrusion	ImmuneNet XGBoost Random Forest, Decision Trees	Improved intrusion detection through deep learning,	Expertise is required to design and train deep learning models	99.63%

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

Author	Country	Research Aim	AI Techniques	Benefits	Challenges	Model Accuracy
		detection accuracy in healthcare systems.			Demanded a lot of computational resources.	
[6]	North America	Compare intrusion detection systems that use medical and network data for improved threat detection in healthcare environments.	KNN SVM ANN Random Forest	Improved threat detection by leveraging both medical and network data, enhancing overall security.	Present technical challenges while integrating diverse data sources for intrusion detection can	99.9%
[17]	Saudi Arabia	Evaluate ensemble learning models for intrusion detection in the Internet of Medical Things (IoMT) to enhance cybersecurity in healthcare.	Stacking Bagging	Enhanced cybersecurity in IoMT by leveraging ensemble learning, improving intrusion detection accuracy and efficiency.	Required careful parameter tuning to design and optimize ensemble learning models for IoMT intrusion detection	98.88%
[7]	Algeria	Propose an intrusion detection a system specifically designed for healthcare applications, leveraging machine learning to	SVM KNN Random Forest MLP	Improved intrusion detection in healthcare applications Enhanced data protection and system integrity through	Tailored intrusion detection solutions are required to address the unique security challenges presented.	99.84%

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

Author	Country	Research Aim	AI Techniques	Benefits	Challenges	Model Accuracy
		enhance security.		machine learning.		
[3]	United Arab Emirates	Utilize machine learning techniques to defend Electronic Health Records (EHRs) against insider threats.	Isolation Forest	Enhanced healthcare data security, improved patient privacy, and data security by identifying anomalies in EHR access patterns.	Data quality and the complexity of defining normal behavior affect the effectiveness of anomaly detection	99.48%
[2]	United Kingdom	Analyze data breaches of patient data in the English National Health Service (NHS) to understand their occurrence and impact.	AI Generalised	Improved understanding of data breaches in healthcare Enables the development of targeted strategies to enhance security and protect patient data.	Underreporting and variations in reporting practices complicate obtaining comprehensive and accurate data on cyberattacks.	N/A
[11]	Switzerland	Explore how AI and machine learning can be used to enhance security in healthcare applications, specifically focusing on mitigating security risks.	AI & ML Generalized	Strengthened security in healthcare applications by leveraging AI and machine learning to detect and respond to threats	Technical and operational challenges whilst integrating AI & ML into existing healthcare infrastructure	N/A

Author	Country	Research Aim	AI Techniques	Benefits	Challenges	Model Accuracy
[12]	Switzerland	Propose using AI-powered threat detection to enhance the security of healthcare data.	AI Generalized	Improved healthcare data security Reduced risk of data breaches and cyberattacks Enhanced patient privacy and trust.	Need for specialized expertise and resources for implementation and maintenance.	N/A
[18]	Norway	Identify, assess, and analyze appropriate AI methods and data sources for modeling and analyzing healthcare security practices.	KNN Bayesian Network Decision Trees (C4.5)	Alleviates the status of healthcare security concerning required security practices.	The accuracy of the ground truth data used to train algorithms can impact their efficiency.	N/A

3.1. Research Origins

Figure 2 shows the geographical distribution of the 11 studies included in this systematic review, providing an overview of the global landscape of AI-based intrusion detection system research in healthcare cybersecurity. As shown in Figure 2, there are four studies each from Europe and Asia, two from Africa, and one from North America. It is also important to note that there were no studies from either South America or Australia, which, rather than indicating a lack of potential research activity in these regions, can be attributed to the databases used in this systematic review.

The results of this systematic review suggest that current research activities in this context are focused in Europe and Asia, with limited representation from other regions of the world. The lack of representation from both South America and Australia can be attributed to the databases used in this systematic review, and this geographical representation can serve as a foundation for analyzing various aspects of AI in this context.

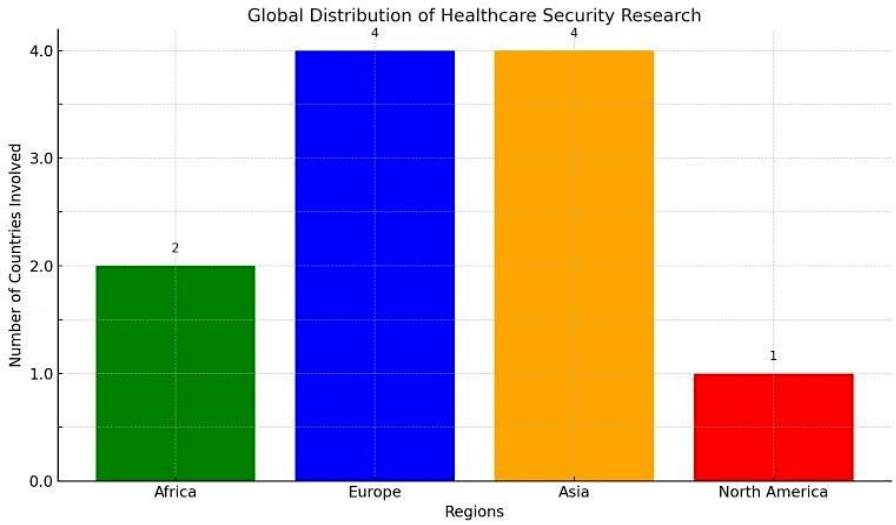

Figure 2. Continental Contribution

3.2. Temporal Distribution of Publications and Research Trends

Figure 3 illustrates the temporal distribution of the 11 studies included in this review across the five years from 2020 to 2024. Research output has not followed a linear progression, but shows periods of limited output followed by more pronounced growth.

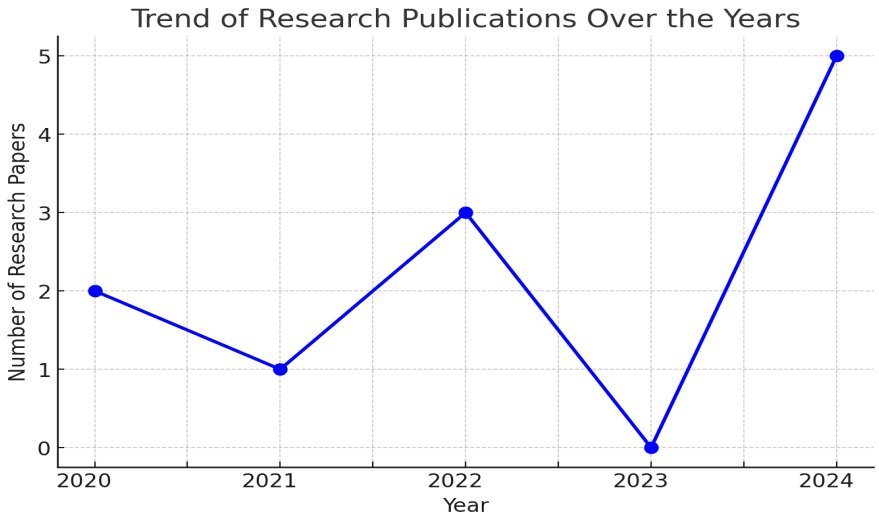

Figure 3. Trend of Research Publications Over the Years

Two studies published in 2020 met the inclusion criteria and were the earliest contributions within the specified time frame. This was then followed by a decline in the number of studies published in 2021, with only one study meeting the inclusion criteria for this review. An increase, however, was recorded in 2022, with three studies on AI-

based intrusion detection systems in healthcare scenarios. No studies were published in 2023 that met the inclusion criteria for this review. On the other hand, the highest number of studies, i.e., 5, was published in 2024, thereby constituting the greatest contribution to the field of AI-based intrusion detection systems within the specified time frame, as depicted in Figure 3. The temporal trend, therefore, shows an increase in the number of studies on AI-based intrusion detection systems in healthcare scenarios, especially in the most recent part of the time frame, without indicating a long-term trend.

3.3. AI Techniques for Threat and Anomaly Detection in Healthcare Cybersecurity

An analysis of the eleven studies reveals that a variety of artificial intelligence techniques are being used for threat and anomaly detection in healthcare cybersecurity. In total, twelve different AI techniques were identified in the literature reviewed in this paper. There was a clear emphasis on machine learning techniques. The majority of the reviewed literature, that is, eight of the eleven papers, primarily focused on traditional machine learning techniques. The most commonly utilized machine learning techniques were k-Nearest Neighbours (KNN), Random Forest (RF), and Decision Trees (DT), each of which was utilized in four of the reviewed papers. These techniques are clear favorites in healthcare intrusion detection applications. Artificial Neural Networks (ANNs) were utilized in three of the reviewed papers, while Support Vector Machines (SVMs) were utilized in two of the reviewed papers. A variety of supervised machine learning techniques, namely Multi-Layer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), and Naïve Bayes (NB), were used in one of the reviewed papers.

Apart from traditional machine learning techniques, a small number of papers utilized alternative techniques of artificial intelligence. These techniques include ensemble techniques such as bagging and stacking, probabilistic graphical models such as Bayesian networks, and unsupervised anomaly detection techniques such as Isolation Forest. Each of these techniques was utilized in only one of the reviewed papers. Three of the reviewed papers utilized a broader approach to machine learning-based intrusion detection systems. In these papers, no specific machine learning algorithm was utilized.

3.4. Reported Model Accuracy in Healthcare Threat and Anomaly Detection

Figure 4 presents the reported accuracy values of artificial intelligence models used for threat and anomaly detection in healthcare cybersecurity across the reviewed studies.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

Among studies reporting quantitative performance outcomes, model accuracy was consistently high.

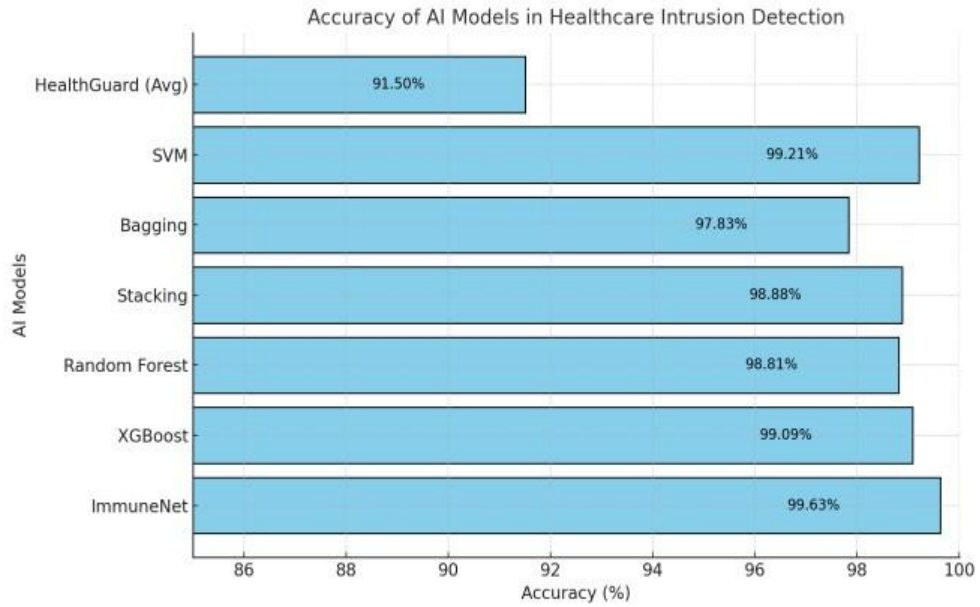

Figure 4. Accuracy of AI Models in Healthcare Threat and Anomaly Detection

A study focused on the application of AI to improve intrusion detection in Software-Defined Wireless Sensor Networks (SDWSNs) achieved an accuracy of 99.94%. Overall, across the seven research studies that reported the average accuracy of the evaluated AI models, the average accuracy was 98.66%. This implies that most AI models used for intrusion detection are effective at classification, as evidenced by the experimental results reported in the research studies. On the other hand, the HealthGuard framework, a machine learning-based security framework, achieved the lowest accuracy of 93%. This framework is used in the security of healthcare systems. In contrast, four research studies did not show any accuracy results. Among these research studies, three focused on artificial intelligence/machine learning approaches in a broad sense, without specifying the performance of the models developed.

3.5. AI Technique Challenges in Healthcare Security Papers

Figure 5 summarises the frequency of challenges associated with the application of artificial intelligence techniques in healthcare cybersecurity, as reported across the 11 included studies. The figure highlights several recurring challenge categories, indicating

that limitations related to data, model behaviour, and operational constraints are widely documented within the reviewed literature.

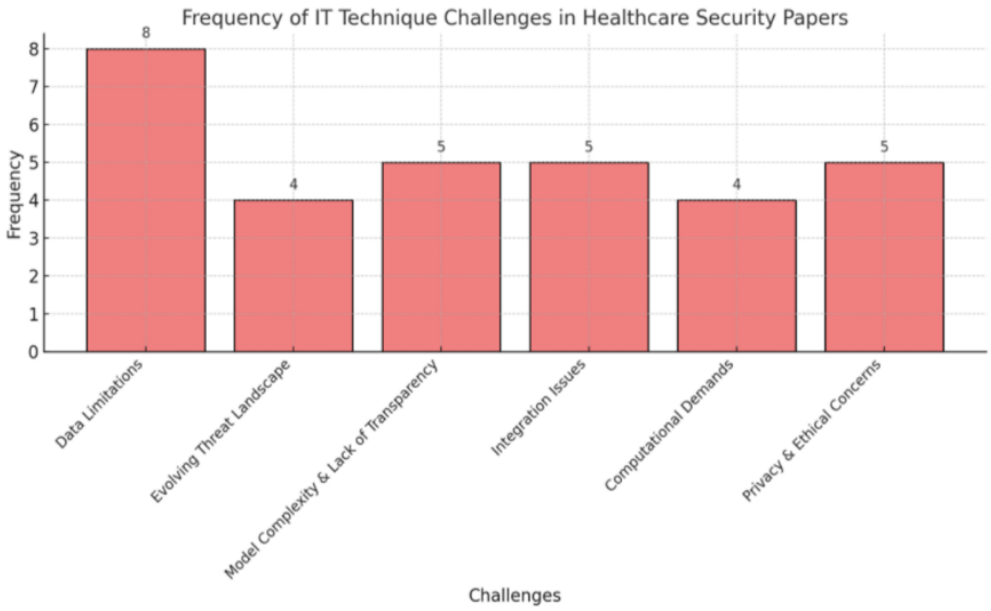

Figure 5. Frequency of AI Technique Challenges in Healthcare Security Papers

This is evident in Figure 5, which shows that data limitations emerged as the most frequently encountered challenge, appearing in 8 studies. This includes challenges related to data availability, data quality, class imbalance, and the lack of labeled data in healthcare security. The second most frequently encountered challenges were model complexity and lack of transparency, reported in five studies. This includes challenges related to understanding, interpreting, and explaining AI model decisions in relation to healthcare security.

Challenges related to integration issues and privacy/ethical concerns emerged as challenges in five studies. This includes challenges related to integration and privacy/ethical concerns when deploying an AI-based IDS in a healthcare environment. Computational challenges related to deploying an AI-based IDS in a healthcare environment emerged in four studies, while evaluation-related challenges emerged in four studies. This shows that challenges related to deploying an AI-based IDS in a healthcare environment are multidimensional, as depicted in Figure 5.

3.6. RQ1: Application of AI Across Intrusion Detection Pipeline Stages in Healthcare Systems

The review indicates that artificial intelligence is applied across several phases of the intrusion detection process in the health care system, and it is not limited to any one phase. Based on the literature reviewed, AI is predominantly used in data monitoring and collection, modeling, and classification within the health care system. This is based on the layered approach to health care cybersecurity infrastructures, as discussed in references [3], [8], [12], and [18]. In the data monitoring phase, AI is predominantly used to analyze data generated by networks of devices connected to the health care system, particularly in Internet of Medical Things (IoMT) environments. In this regard, AI continuously monitors the network and the communication patterns of connected devices, helping detect potential cyberattacks on networked devices in the health care system.

At the behavioral modeling stage, AI is commonly employed to learn the patterns of normal operation within healthcare information systems, especially electronic health record systems. Several studies have demonstrated the application of AI in modeling legitimate user behavior to detect unusual activities, such as unauthorized access, credential misuse, and privilege escalation, within healthcare information systems [12]. The application of AI in healthcare information systems is therefore user-centric. This is because it considers intrusion detection beyond network-centric attacks, incorporating insider and misuse-related attacks that may not be detected by conventional rule-based intrusion detection systems.

At the detection and classification stage, AI is employed to classify malicious activities from legitimate ones according to the patterns learned. The application of AI in intrusion detection is evident in various studies that have used user behavior logs to demonstrate AI-based classifiers' ability to detect internal security threats that conventional security systems may miss. The application of AI in intrusion detection is therefore evident in enhancing the sensitivity of intrusion detection systems within complex healthcare environments [3], [12].

In addition to network-centric and user-centric approaches to intrusion detection within healthcare information systems, the review of various studies revealed the application of AI within software-defined healthcare information systems. One of the applications is

the Software-Defined Wireless Sensor Networks (SDWSNs). Within these systems, AI agents analyze the behavior of the control plane to detect intrusions that may affect data transmission in healthcare sensor networks. The application of AI in intrusion detection within healthcare information systems is evident in all the stages of the intrusion detection pipeline. However, the studies reviewed have been conducted in an experimental environment. The deployment of AI applications within fully operational healthcare intrusion detection systems is not evident.

3.7. RQ2: Commonly Employed Artificial Intelligence Techniques for Healthcare Intrusion Detection

The findings of this review suggest that the existing literature on intrusion detection in health care environments relies primarily on conventional machine learning approaches, whereas the application of advanced artificial intelligence concepts is relatively rare. Conventional supervised learning algorithms, such as machine learning classifiers, have emerged as the most popular approaches to the implementation of intrusion detection in health care environments [3], [6], [8], [15], [18]. Tree and instance-based learning algorithms, such as Random Forest, Decision Tree, k-Nearest Neighbour, and Support Vector Machine, have emerged as the most popular approaches to the implementation of intrusion detection in health care environments, based on the existing literature [3], [6], [7], [15], [17]. In addition, the application of Artificial Neural Networks has also emerged, but to a relatively lesser extent.

Deep learning models are also recognized in the broader intrusion detection literature as viable approaches for modeling complex, high-dimensional data, although their application in healthcare-specific intrusion detection remains limited [19], [21]. In cases where deep learning methods are applied, their application is reported as experimental, and their adoption is recognized as being influenced by computational and implementation factors in healthcare environments [19], [25].

In addition to supervised learning methods, a smaller number of studies also apply anomaly-based and unsupervised methods, such as Isolation Forest, when sufficient training data are not available [3], [6], [10]. Ensemble learning methods are also reported as supplementary methods, although typically not as primary intrusion detection methods [9], [10], [19].

Several studies reported the application of "artificial intelligence" or "machine learning" methods, although the specific methods were not consistently reported. This has previously been recognized as a limitation in healthcare-specific cybersecurity research [8], [10], [18]. In terms of the distribution of methods reported in the literature, classical machine learning methods are predominant, while advanced AI methods are underrepresented in healthcare-specific intrusion detection systems.

3.8. RQ3: Benefits of Artificial Intelligence-Based Approaches for Healthcare Threat and Anomaly Detection

From the reviewed literature, it can be determined that the use of artificial intelligence-based approaches in threat and anomaly detection in healthcare cybersecurity scenarios, which involve heterogeneous data sources, dynamic patterns of threats, and critical operational constraints, has several advantages [3], [5], [8], [18], [25]. Among these advantages, some relate to improving threat and anomaly detection capabilities, while others focus on increasing flexibility, reducing the need for manual interventions, and improving the management of complex behavioral patterns. One advantage of artificial intelligence-based intrusion detection methodologies is their ability to learn a system's normal behavior and identify anomalies. Unlike rule-based intrusion detection methodologies, which often fail to adapt to dynamic patterns of normal system behavior, AI-based methodologies are highly advantageous in healthcare cybersecurity environments that involve highly dynamic normal system behavior [7], [8], [10].

Another significant advantage is the potential of AI-based methods to assist in detecting unknown or insider threats. The use of unsupervised and anomaly-based learning methods is seen as beneficial for the detection of attacks, as these methods do not require the use of known attacks for the detection process, making them more suitable for the healthcare environment where the availability of attack data is limited and the potential for insider attacks is a major concern [3], [12].

In addition, there are potential advantages to using AI-based methods for attack detection, as highlighted in the literature. In particular, the use of deep learning methods for attack detection is seen as beneficial, given their ability to recognize complex relations in data exchanged within the network, thereby identifying long-running intrusions [8], [19], [20]. However, these potential advantages are mostly evident within

experimental scenarios. Finally, the literature reports that ensemble methods combining several AI techniques can improve detection performance in healthcare settings [8], [10]. Overall, the benefits identified relate less to raw detection capability than to adaptability: the capacity to accommodate changing behaviour and heterogeneous data in an environment where both are the norm.

3.9. RQ4: Evaluation Practices and Conditions for AI-Based Intrusion Detection in Healthcare

The results of this review revealed that the performance of artificial intelligence-based intrusion detection models in health care environments is mainly evaluated through accuracy-based performance metrics, which vary significantly. Although high accuracy values are reported by various models, the literature reveals little consistency in how performance is evaluated or under what conditions evaluations are performed [4], [6], [8], [11], [15].

Among empirical research on artificial intelligence-based intrusion detection models, accuracy is the most widely used performance metric. Various models have reported high accuracy above 97%, especially those that use ensemble learning techniques such as stacking and bagging [8], [11]. Other models that use hybrid or anomaly-based approaches, including Isolation Forest and supervised classifiers, report high accuracy values above 99% under controlled experimental conditions [12]. In contrast, models such as HealthGuard report accuracy values of 90%-93% across various experimental conditions [15].

Besides the reported performance values, the review also points out the fact that there is a significant variation in the evaluation conditions. Many studies use simulated evaluation environments or benchmark datasets and synthetic traffic rather than real-world deployment scenarios in healthcare systems [8], [10]. Some studies use well-known benchmark datasets such as the CMU-CERT dataset and allow for comparative evaluations. However, the performance reported in such studies may not accurately reflect the model's effectiveness in the real world, as the complexity and variability of the deployment scenario are not fully represented.

Some studies also point out that the accuracy measure may not be sufficient for the context of the proposed system. In the proposed system, which detects cyber-attacks in the healthcare system, the accuracy measure may not be sufficient because class imbalance and the low prevalence of attacks can lead to poor detection performance. However, the other performance metrics, such as precision, recall, sensitivity, and F1 score are not reported consistently. Based on the evaluation methods reported in the studies included in the review, it can be concluded that although the proposed AI-based intrusion detection model's performance is promising, there is a lack of standardization in the evaluation methods.

3.10. RQ5: Technical, Organisational and Regulatory Challenges Limiting AI-Driven Intrusion Detection in Healthcare

The literature identifies technical, organisational and regulatory constraints that limit the operation of AI-based intrusion detection in healthcare settings [1], [3], [6], [8], [18], [25]. These constraints are the principal reason why systems that perform well in experimental conditions rarely progress to sustained operational use.

The primary challenge in operationalising AI-based intrusion detection in healthcare is data. Studies consistently report a shortage of high-quality security data, arising from the privacy, ethical and legal constraints on releasing clinical network traffic [3], [6], [10], [21]. The field has responded by relying on benchmark and synthetic corpora [1], [2], [18], which substitute availability for realism. A related problem is class imbalance: attack events are rare relative to benign traffic, which depresses both detection reliability and the informativeness of accuracy as a metric [4], [18]. Adversarial manipulation of detection models is a further concern, since it bears directly on how models must be maintained and updated as the threat environment changes [11], [22].

Evaluation practice forms the second constraint. Across the corpus, accuracy is the dominant reported measure, and the false positive rate is seldom reported. This is consequential in healthcare, where a high volume of spurious alerts consumes scarce analyst capacity and erodes confidence in the system [5], [8], [25]. Reviews of the wider literature report the same absence of standardised evaluation approaches for medical settings [9], [10], [19].

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

In the organizational context, the need for integration, along with resource constraints, assumes paramount importance. In the healthcare domain, the fragmented, heterogeneous, and legacy-plagued information technology infrastructure makes the integration of intrusion detection tools that incorporate artificial intelligence a challenging task, thereby increasing implementation costs [12], [17], [19]. The high costs of infrastructure, along with maintenance costs, are other factors that affect the consideration of implementing AI-based intrusion detection systems, especially in the healthcare domain, which is often resource-constrained [12], [18]. All these factors affect the consideration of implementing computationally intensive AI-based intrusion detection systems, despite the positive results obtained during experimentation.

Regulatory and governance factors form the third constraint. Data protection and de-identification requirements limit what a detection system may observe and retain [3], [26]. Adoption is further impeded where systems cannot supply the transparency healthcare institutions require for explanation, audit and accountability [11], [16], [18], [25]. Questions of ethics, bias, privacy and the effect on the security workforce operate as additional constraints on adoption, and are best treated as design requirements rather than as downstream compliance matters [26], [27]. Taken together, these constraints indicate a field with demonstrated technical potential and limited demonstrated operational effect. The obstacles are not principally algorithmic. They are sociotechnical and regulatory, and they are not adequately addressed by research designs that evaluate detection quality in isolation from the environment in which detection must occur.

4. DISCUSSIIN AND HEALTHCARE IDS DEPLOYMENT READINESS ASSESSMENT (HIDRA)

The synthesis in Section 3 identifies a consistent pattern: detection performance is reported with precision, while the conditions determining whether a detector could operate in a hospital are reported inconsistently or not at all. This asymmetry cannot be resolved by another literature summary. It requires an instrument that makes the missing information visible, so that a study reporting 99 percent accuracy on a benchmark capture and a study reporting 94 percent on live clinical traffic are not read as equivalent claims. This section specifies such an instrument.

HIDRA assesses a study along five dimensions, each scored on a five-level ordinal scale. The dimensions were derived inductively from the challenge categories reported across

the corpus and cross-checked against the deployment barriers identified in the larger concurrent reviews [20], [21], [33]. They are not proposed as exhaustive, and Section 4.4 states the conditions under which the instrument should be revised.

4.1. Assessment dimensions

D1. Evidential realism. What data was the system evaluated on, and how closely does it resemble the environment in which it would operate? Level 1 denotes evaluation on a general-purpose benchmark unrelated to healthcare. Level 2 denotes a healthcare-adjacent public dataset such as a medical Internet of Things benchmark. Level 3 denotes a purpose-built testbed with clinical protocol traffic. Level 4 denotes retrospective evaluation on capture from a real clinical network. Level 5 denotes prospective evaluation on live clinical traffic.

D2. Evaluation completeness. Which properties were measured? Level 1 denotes accuracy alone. Level 2 adds precision, recall and F1. Level 3 adds the false positive rate expressed as absolute alert volume per unit time, which is the quantity that determines analyst workload. Level 4 adds inference latency and resource footprint measured on the intended deployment hardware. Level 5 adds robustness evidence under distribution shift, adversarial perturbation, or both.

D3. Operational integration. How would the detector connect to what a hospital already runs? Level 1 denotes no consideration. Level 2 denotes a stated architecture without implementation. Level 3 denotes a prototype integration with a security information and event management platform or equivalent. Level 4 denotes demonstrated interoperation with clinical systems, including the electronic health record or device management infrastructure. Level 5 denotes sustained operation within an established security operations workflow.

D4. Interpretability and analyst fit. Can a human act on the output? Level 1 denotes a bare classification label. Level 2 denotes a confidence score. Level 3 denotes post hoc explanation using an established attribution method. Level 4 denotes explanation validated with security analysts rather than assumed to be useful. Level 5 denotes measured effect of the explanation on analyst triage accuracy or time to decision. The distinction between levels 3 and 4 is the one most often collapsed in the literature, and

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

Ogunseyi et al. [21] report the same gap between generating explanations and evaluating them.

D5. Governance and regulatory alignment. Has the legal and ethical operating envelope been addressed? Level 1 denotes no treatment. Level 2 denotes acknowledgement of applicable regimes. Level 3 denotes a stated lawful basis for the processing of personal and clinical data, with data minimisation applied. Level 4 adds auditability sufficient to support incident investigation and regulatory reporting. Level 5 denotes evidence of institutional approval and operation under an approved governance framework. [26] and [27] set out the responsible-AI conditions on which levels 3 to 5 depend, and their argument that privacy and consent must be designed into the system rather than certified afterwards applies directly here.

4.2. Composite readiness level

A study's Deployment Readiness Level is the minimum of its five dimension scores rather than the mean. The minimum is used deliberately. Deployment readiness is limited by the weakest dimension, since a detector that is accurate, fast and interpretable but has no lawful basis for processing the data it requires cannot be deployed at all. Averaging would allow strength in the dimensions researchers find tractable to conceal absence in the dimensions that determine feasibility. The composite is therefore conservative by construction, and is intended to be.

4.3. Retrospective application to the corpus

Applying HIDRA to the eleven studies places every one of them at readiness level 1 or 2, and none above level 3 on any single dimension. The distribution is concentrated rather than dispersed: studies score at level 1 or 2 on evidential realism because evaluation used benchmark or synthetic data; at level 1 or 2 on evaluation completeness because accuracy dominates and absolute alert volume is almost never reported; at level 1 on operational integration because integration is discussed as a challenge rather than demonstrated; at level 1 or 2 on interpretability; and at level 1 or 2 on governance, where regulatory requirements are cited as obstacles rather than addressed as design constraints. Because the composite is the minimum, no study in the corpus exceeds level 2.

This result should not be read as a criticism of the individual studies, most of which do competently what they set out to do. It is a statement about what the field collectively measures. A research programme that optimises exclusively for dimension D2 at level 1 will produce increasingly accurate models and no deployable systems, which is an accurate description of the present state of healthcare intrusion detection research.

4.4. Conditions for revising the instrument

HIDRA is a proposal, not a standard, and it should be revised under stated conditions rather than defended. Three in particular. First, if a dimension is found to be consistently scored at the same level across a large and diverse corpus, it is not discriminating and should be removed or resolved into finer levels. Second, if the minimum rule produces composites that experienced practitioners judge inconsistent with their own assessment of deployability, the aggregation rule requires revision, and we would regard practitioner disagreement as evidence about the instrument rather than about the practitioners. Third, the dimensions were derived from a corpus in which no study reached sustained live operation; once such studies exist, dimensions that matter only in operation, such as model maintenance under concept drift and the cost of retraining, will need to be added. An inter-rater reliability study across independent assessors is a necessary next step and has not been conducted here.

4.5. The Regulatory Environment As A Design Constraint

The corpus treats regulation almost uniformly as an obstacle to deployment. That framing is now out of step with the regulatory position, which has moved from general data protection duties toward specific and auditable technical obligations that bear directly on how a detection system must be built. Three developments are consequential. In the United States, the Department of Health and Human Services issued a proposed rule in January 2025 to strengthen the HIPAA Security Rule, moving toward mandatory rather than addressable implementation specifications in areas including multi-factor authentication, encryption, asset inventory and network segmentation [35]. In the European Union, the NIS2 Directive brings healthcare providers within a common cybersecurity risk-management and incident-reporting regime, while the Artificial Intelligence Act establishes obligations on high-risk AI systems covering risk management, data governance, logging, transparency and human oversight [36], [37].

Medical device regulation adds cybersecurity requirements across the device lifecycle for connected equipment.

For intrusion detection research the implication is specific rather than general. Logging and traceability obligations mean that a detector must be able to reconstruct why an alert was raised, which converts interpretability from a desirable property into a compliance artefact. Human oversight obligations mean that fully automated response is constrained, which makes analyst-facing output quality a functional requirement rather than a usability nicety. Data governance obligations bear on training data provenance, which is the point at which the field's reliance on synthetic and benchmark corpora becomes a regulatory as well as a scientific problem. These map onto HIDRA dimensions D4 and D5 and are the reason those dimensions are included.

Architectural guidance has moved in parallel. Zero trust models, which assume no implicit trust based on network location and require continuous verification, are increasingly the assumed context for detection in segmented clinical networks [31]. A detector designed for a flat perimeter-defended network makes assumptions about lateral traffic visibility that do not hold under segmentation, and this is a further respect in which benchmark evaluation may mislead.

For low- and middle-income settings the constraint binds differently. Where regulatory capacity is developing and security budgets are constrained, the practical requirement is for detection that is inexpensive to run, operable by small teams, and defensible without a dedicated compliance function. This is an argument for lightweight and interpretable models over maximum-accuracy architectures, and it is under-served by a literature that reports accuracy without reporting resource footprint. Two of the eleven corpus studies originate in Africa, and the wider regional research base is developing [28], [30], but the deployment economics of these settings are not yet reflected in the evaluation practices of the field.

4.6. Research Gaps

Four gaps follow from the synthesis and from the HIDRA assessment, and each is stated so that its closure would be recognisable. The evidential realism gap. Detection models are trained and evaluated on benchmark or synthetic corpora whose traffic

characteristics are not established to resemble clinical networks. The gap closes when studies report performance on capture from operational clinical environments, or, where that is not obtainable, when they characterise the distance between their evaluation data and clinical traffic rather than leaving it unstated.

The evaluation completeness gap. Accuracy is reported almost universally; absolute alert volume, inference latency on target hardware, and behaviour under distribution shift are reported rarely. Because attacks are rare relative to benign traffic, a model with excellent accuracy can still generate an unworkable alert load, and no study in the corpus reports the figure that would allow this to be checked. The gap closes when the minimum reporting set in Section 9.2 is routinely met.

The interpretability evaluation gap. Explanation is increasingly generated but seldom evaluated. The relevant question is not whether an attribution method produces a plausible output but whether it improves the decisions of the analyst who receives it, and that question requires a study with human participants. This gap is identified independently in the larger review by Ogunseyi et al. [21], which strengthens confidence that it is real. The deployment evidence gap. No study in the corpus reports sustained operation in a live clinical network, and none reports what happened to detection quality over time. Concept drift, model maintenance cost and the organisational work of running a detector are consequently absent from the evidence base. This is the gap that HIDRA is designed to make visible, since it is invisible in any synthesis organised by algorithm.

4.7. Implications

For researchers, the implication is that marginal accuracy improvements on established benchmarks have low expected value for the field, because accuracy is not the binding constraint on deployment. Studies that report a lower headline accuracy under more realistic conditions, with alert volume and latency stated, contribute more to the evidence base than studies reporting higher accuracy under controlled conditions. Reviewers and editors are in a position to change this incentive, and HIDRA is intended in part as a reviewing aid.

For practitioners and procurement teams, the implication is that published accuracy figures do not transfer to institutional environments and should not be treated as

procurement evidence. A vendor claim of 99 percent detection accuracy is compatible with an unusable alert load. The questions that discriminate between systems are what data the evaluation used, what the false positive rate was in absolute alerts per day at the deployment scale, what the inference latency is on the hardware the institution owns, and what the system does when it encounters traffic unlike its training data. HIDRA can be used directly as a procurement questionnaire. For policymakers, the implication is that the emerging regulatory obligations on logging, transparency and human oversight are more achievable if evaluation practice changes in advance of enforcement. Requiring reporting of deployment-relevant metrics in publicly funded research is a low-cost intervention with a direct effect on the quality of the evidence base available to health systems. For low- and middle-income health systems, the implication is that the current literature does not answer the question they need answered, which concerns detection that is affordable to run and operable by small teams. Building that evidence requires evaluation on the hardware and staffing such systems actually have.

4.8. Limitations

This review has limitations that bear on how its conclusions should be used. The corpus is small. Eleven studies met the inclusion criteria, which is a narrow base for claims about a field. This is why the findings are triangulated in Section 1.2 against reviews analysing 53 and 129 studies; the convergence observed there is the principal reason for confidence in the patterns reported, and readers should weight that convergence more heavily than the corpus size. Screening was conducted by a single reviewer without independent duplicate assessment, so selection error cannot be excluded and no inter-rater agreement statistic can be reported. The protocol was not prospectively registered. Both are departures from best practice in systematic review and are stated rather than mitigated.

Two included items do not meet the stated peer-review criterion, as set out in Section 2.6. Neither contributes quantitative findings, but their inclusion is an execution defect. The search covered five databases and excluded IEEE Xplore and the ACM Digital Library, which are substantial venues for intrusion detection research. This will have depressed retrieval of engineering-oriented work and is the most likely explanation for the absence of transformer-based and federated approaches from the corpus, given that reviews searching those databases do retrieve them [20].

The English-language restriction excludes work published in other languages, which is a particular concern for evidence from low- and middle-income settings. The search window closes at 2024, so the corpus does not reflect work published subsequently. This is mitigated but not removed by the triangulation in Section 1.2, which indicates that the patterns identified persist in later and larger syntheses. Readers should nonetheless treat the corpus description as characterising the 2020 to 2024 literature, and the framework in Section 4 rather than the corpus as the portable contribution. HIDRA itself is unvalidated. It was derived from this corpus and applied by its author, so the retrospective scoring in Section 4.3 is an illustration of the instrument rather than an independent test of it. Inter-rater reliability, construct validity and predictive validity against actual deployment outcomes all remain to be established, and until they are, HIDRA should be treated as a structured prompt rather than a measurement.

4.9. Future Research Agenda

4.9.1. Testable Propositions

The following propositions are stated so that they can be tested and, if wrong, shown to be wrong. Each is offered as a null that the field should attempt to reject. P1. Detection performance does not degrade materially when models trained on public medical intrusion benchmarks are evaluated on capture from operational clinical networks. This is the proposition most likely to be rejected, and its rejection would be the single most useful result the field could produce, because it would quantify the benchmark transfer penalty that is currently assumed but unmeasured. P2. At clinically realistic base rates of attack, models reporting accuracy above 99 percent on benchmark data generate an alert volume that a typical hospital security team can process within its shift capacity. P3. Post hoc explanations supplied with alerts do not improve analyst triage accuracy or reduce time to decision relative to alerts supplied without them. Rejecting this proposition requires a study with security analysts as participants, of which the corpus contains none. P4. Federated approaches, which allow institutions to train collaboratively without exchanging patient data, achieve detection performance equivalent to centralised training on pooled data in non-independent and non-identically distributed clinical settings.

Recent federated work provides the starting point for testing P4: reviews of federated intrusion detection set out the state of the approach [38], while knowledge distillation

under non-independent and non-identically distributed conditions [40], deep learning detectors purpose-built for medical settings [39], and federated explainable methods [41] indicate the specific variants worth evaluating in clinical deployment. P5. Lightweight models deployable on constrained hardware do not differ materially in operational detection quality from larger architectures once alert volume and latency are accounted for. This proposition matters disproportionately for resource-constrained health systems. P6. Detection quality is stable over time in the absence of retraining. Rejecting this would establish the maintenance cost that must be included in any procurement decision, and no study in the corpus reports the longitudinal data required to test it. Testing P6 requires the concept drift methods established in the wider machine learning literature [42], which have not yet been applied systematically to healthcare intrusion detection.

4.9.2. A Minimum Reporting Set

To make the propositions above testable across studies, the following should be reported as a matter of routine. Provenance and characteristics of evaluation data, including the base rate of attack traffic. Precision, recall and F1 alongside accuracy, with the confusion matrix. False positive rate expressed as absolute alerts per unit time at a stated deployment scale, not as a proportion. Inference latency and memory footprint measured on the intended target hardware. Performance under at least one distribution shift or adversarial perturbation. Where explanations are generated, the method used and any evaluation of their effect on human decisions. The lawful basis and governance arrangements for the data used. Compute cost of training and retraining. None of these is burdensome to report, and their routine absence is what prevents the field's results from accumulating.

4.9.3. Research Directions

Beyond the propositions, four directions follow from the gaps. Cross-institutional data collaborations are required to produce clinical intrusion corpora with realistic base rates, and federated and privacy-preserving learning offer a route to model development where data cannot be pooled [33], [34]. Human-factors research on analyst interaction with detector output is largely absent and is the precondition for progress on HIDRA dimension D4. Adversarial robustness in the medical setting, where an evasion may have physical consequences for a patient, warrants treatment distinct from the general adversarial machine learning literature [22]. Finally, longitudinal deployment studies,

however small, would contribute more than further benchmark comparisons, because a single well-documented year of operation in one hospital would supply evidence the field currently lacks entirely.

5. CONCLUSION AND RECOMMENDATIONS

This review synthesised evidence on artificial intelligence for intrusion detection in healthcare, with evaluation practice and deployment condition rather than algorithmic technique as the object of analysis. The corpus is dominated by classical machine learning, reports a mean accuracy of 98.66 percent, and obtains that figure almost entirely on benchmark or synthetic data under controlled conditions. False positive rate expressed as alert volume, inference latency on target hardware, and robustness under distribution shift are rarely reported, and no study reports sustained operation in a live clinical network. The same patterns appear in concurrent reviews with corpora five and ten times larger, which indicates that they characterise the field rather than this sample. The gap between reported performance and demonstrated deployability is therefore not an accident of incomplete reporting. It reflects what the field has chosen to measure. A research programme optimising for accuracy on fixed benchmarks will continue to produce accurate models and few deployable systems, and the incentive to change this sits with reviewers and funders as much as with authors. The contribution offered in response is the Healthcare IDS Deployment Readiness Assessment, a five-dimension rubric with five ordinal levels and a deliberately conservative aggregation rule, accompanied by a minimum reporting set and six testable propositions. Applied retrospectively, it places every study in the corpus at readiness level 1 or 2 of 5. HIDRA is itself unvalidated, and its own reliability and validity are among the first things that should be tested. It is offered on that basis: as an instrument others can apply, criticise and improve, rather than as a finding. The recommendation to the field is narrow and follows directly. Report what deployment requires, evaluate explanations against the humans who must use them, and publish at least some studies from operational environments even where the resulting numbers are less impressive than those obtainable on a benchmark. The less impressive number obtained under realistic conditions is the more valuable scientific result.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

ACKNOWLEDGMENT

We would like to thank the reviewers for taking the time to review this manuscript

REFERENCES

[1] H. and F. J. Anthony Minnaar, "Cyberattacks and the Cybercrime Threat of Ransomware to Hospitals and Healthcare Services During the COVID-19 Pandemic," *Acta Criminol. Afr. J. Criminol. Vict.*, vol. 34, no. 3, p. 3, 2021, doi: 10.10520/ejc-crim.

[2] S. Ghafur, S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin, "A retrospective impact analysis of the WannaCry cyberattack on the NHS," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–7, 2019, doi: 10.1038/s41746-019-0161-6.

[3] M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique, "Anomaly-based threat detection in smart health using machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02760-4.

[4] M. Akshay Kumaar, D. Samiayya, P. M. D. R. Vincent, K. Srinivasan, C. Y. Chang, and H. Ganesh, "A Hybrid Framework for Intrusion Detection in Healthcare Systems Using Deep Learning," *Front. Public Health*, vol. 9, no. January, pp. 1–18, 2022, doi: 10.3389/fpubh.2021.824898.

[5] A. Si-Ahmed, M. A. Al-Garadi, and N. Boustia, "Survey of Machine Learning based intrusion detection methods for Internet of Medical Things," *Appl. Soft Comput.*, vol. 140, pp. 0–2, 2023, doi: 10.1016/j.asoc.2023.110227.

[6] A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, "Intrusion Detection System for Healthcare Systems Using Medical and Network Data: A Comparison Study," *IEEE Access*, vol. 8, pp. 106576–106584, 2020, doi: 10.1109/ACCESS.2020.3000421.

[7] M. Khaldi, N. Mahammed, M. A. Lahmar, and F. D. Daouadji, "An Intrusion Detection System for Healthcare Applications using Machine Learning," *CEUR Workshop Proc.*, vol. 3694, pp. 94–101, 2024.

[8] S. B. Weber, S. Stein, M. Pilgermann, and T. Schrader, "Attack Detection for Medical Cyber-Physical Systems—A Systematic Literature Review," *IEEE Access*, vol. 11, no. April, pp. 41796–41815, 2023, doi: 10.1109/ACCESS.2023.3270225.

[9] K. Arshad et al., "Deep Reinforcement Learning for Anomaly Detection: A Systematic Review," *IEEE Access*, vol. 10, no. October, pp. 124017–124035, 2022, doi: 10.1109/ACCESS.2022.3224023.

[10] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," *IEEE Access*, vol. 9, pp. 78658–78700, 2021, doi: 10.1109/ACCESS.2021.3083060.

[11] S. Badi, "Mitigating Security Risks in Healthcare Applications through AI and Machine Mitigating Security Risks in Healthcare Applications through AI and Machine Learning," no. August, 2024, doi: 10.13140/RG.2.2.32500.36485.

[12] S. Arefin, "Strengthening Healthcare Data Security with Ai-Powered Threat Detection," no. October, 2024, doi: 10.18535/ijsrm/v12i10.ec02.

[13] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021, doi: 10.1136/bmj.n71.

[14] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *J. Clin. Epidemiol.*, vol. 62, no. 10, pp. 1006–1012, 2009, doi: 10.1016/j.jclinepi.2009.06.005.

[15] S. M. Wa Umba, A. M. Abu-Mahfouz, and D. Ramotsoela, "Artificial Intelligence-Driven Intrusion Detection in Software-Defined Wireless Sensor Networks: Towards Secure IoT-Enabled Healthcare Systems," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, 2022, doi: 10.3390/ijerph19095367.

[16] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E. M. Tag-ElDin, and A. U. Rehman, "HealthGuard: An Intelligent Healthcare System Security Framework Based on Machine Learning," *Sustainability*, vol. 14, no. 19, 2022, doi: 10.3390/su141911934.

[17] T. Alsolami, B. Alsharif, and M. Ilyas, "Enhancing Cybersecurity in Healthcare: Evaluating Ensemble Learning Models for Intrusion Detection in the Internet of Medical Things," *Sensors*, vol. 24, no. 18, 2024, doi: 10.3390/s24185937.

[18] P. K. Yeng, L. O. Nweke, A. Z. Woldaregay, B. Yang, and E. A. Snekkenes, "Data-Driven and Artificial Intelligence (AI) Approach for Modelling and Analyzing Healthcare Security Practice: A Systematic Review," *Adv. Intell. Syst. Comput.*, vol. 1250 AISC, no. March, pp. 1–18, 2021, doi: 10.1007/978-3-030-55180-3_1.

[19] J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," *IEEE Access*, vol. 9, pp. 101574–101599, 2021, doi: 10.1109/ACCESS.2021.3097247.

[20] O. Adohinzin and Y. Harrath, "A Systematic Review of Intrusion Detection Systems for Internet of Medical Things: Performance, Efficiency, Explainability, and Generalization," *Digit. Threats Res. Pract.*, 2026, doi: 10.1145/3816026.

[21] T. B. Ogunseyi, G. Thiyagarajan, H. He, V. Bist, and Z. Du, "Performance Analysis of Explainable Deep Learning-Based Intrusion Detection Systems for IoT Networks: A Systematic Review," *Sensors*, vol. 26, no. 2, p. 363, 2026, doi: 10.3390/s26020363.

[22] R. Kalakoti, S. Nomm, and H. Bahsi, "Explainable Transformer-Based Intrusion Detection in Internet of Medical Things (IoMT) Networks," in *2024 Int. Conf. Mach. Learn. Appl. (ICMLA)*, IEEE, 2024, pp. 1164–1169.

[23] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT," *Internet Things*, vol. 28, p. 101351, 2024, doi: 10.1016/j.iot.2024.101351.

[24] G. Krishnamoorthy and S. M. K. Sistla, "Exploring Machine Learning Intrusion Detection: Addressing Security and Privacy Challenges in IoT—A Comprehensive Review," *J. Knowl. Learn. Sci. Technol.*, vol. 2, no. 2, pp. 114–125, 2023, doi: 10.60087/jklst.vol2.n2.p125.

[25] I. Bala, I. A. Pindoo, M. M. Mijwil, M. Abotaleb, and W. Yundong, "Ensuring Security and Privacy in Healthcare Systems: A Review Exploring Challenges, Solutions, Future Trends, and the Practical Applications of Artificial Intelligence," *Jordan Med. J.*, vol. 58, no. 2, pp. 250–270, 2024, doi: 10.35516/jmj.v58i2.2527.

[26] B. Ndlovu and K. Maguraushe, "Balancing Ethics and Privacy in the Use of Artificial Intelligence in Institutions of Higher Learning: A Framework for Responsive AI Systems," *Indones. J. Inform. Educ.*, vol. 9, no. 1, p. 39, 2025, doi: 10.20961/ijie.v9i1.100723.

[27] T. Kapuya, W. Kubiku, M. Dube, and T. Mukudu, "Ethical AI Frameworks: Balancing Privacy, Consent and Responsible Use," 2024, doi: 10.46254/EU07.20240168.

[28] M. Moyo and B. Ndlovu, "AI-Driven Intrusion Detection Systems: Algorithms, Key Applications, Efficacy," in *Lect. Notes Netw. Syst.*, Springer Nature Switzerland, 2026, pp. 40–53.

[29] M. A. Alsoufi et al., "Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review," *Appl. Sci.*, vol. 11, no. 18, 2021, doi: 10.3390/app11188383.

[30] A. Maramba and B. Ndlovu, "AI-Driven Threat and Incident Detection in Healthcare Cybersecurity: A Stacked Ensemble Model," in *2025 4th Zimbabwe Conf. Inf. Commun. Technol.*, 2025, pp. 1–9, doi: 10.1109/ZCICT67553.2025.11602039.

[31] A. Dube, B. Mpande, F. Dzehonye, and T. Chazuza, "Zero Trust Architecture: Redefining Security," in *Lect. Notes Netw. Syst.*, Springer Nature Switzerland, 2026, pp. 26–39, doi: 10.1007/978-3-032-20533-9.

[32] U.S. Department of Health and Human Services, Office for Civil Rights, "Change Healthcare Cybersecurity Incident Frequently Asked Questions," HHS.gov, 2025.

[33] A. Salehpour, M. A. Balafar, et al., "Intrusion detection system for IoMT in smart hospitals: a systematic literature review," *Internet Things Cyber-Phys. Syst.*, 2026, doi: 10.1016/j.iotcps.2026.01.006.

[34] Y. Wang et al., "A comprehensive survey on intrusion detection in internet of medical things: Datasets, federated learning, blockchain, and future research directions," *Meas. Sensors*, 2025, doi: 10.1016/j.measen.2025.101896.

[35] U.S. Department of Health and Human Services, "HIPAA Security Rule To Strengthen the Cybersecurity of Electronic Protected Health Information: Notice of Proposed Rulemaking," *Fed. Regist.*, vol. 90, no. 4, pp. 898–1002, Jan. 2025.

[36] European Parliament and Council, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)," *Off. J. Eur. Union*, L 333, pp. 80–152, 2022.

[37] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, 2024.

[38] B. Buyuktanir, S. Altinkaya, G. Karatas Baydogmus, et al., "Federated learning in intrusion detection: advancements, applications, and future directions," *Cluster Comput.*, vol. 28, p. 473, 2025, doi: 10.1007/s10586-025-05325-w.

[39] A. Berguiga, A. Harchay, and A. Massaoudi, "HIDS-IoMT: a deep learning-based intelligent intrusion detection system for the internet of medical things," *IEEE Access*, 2025.

[40] H. Peng, C. Wu, and Y. Xiao, "FD-IDS: Federated learning with knowledge distillation for intrusion detection in non-IID IoT environments," *Sensors*, vol. 25, no. 14, p. 4309, 2025, doi: 10.3390/s25144309.

[41] R. Kalakoti, S. Nomm, and H. Bahsi, "Federated learning of explainable AI (FedXAI) for deep learning-based intrusion detection in IoT networks," *Comput. Netw.*, p. 111479, 2025.

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

[42] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, 2018, doi: 10.1109/TKDE.2018.2876857.