

# Adversarial-Resilient and Explainable AI Frameworks for Deepfake-Aware Fraud Detection in Financial Systems: A Comprehensive Review

**Manoj Kumar Rath**

Executive Vice president, Kloud9

[k.manoj.rath@gmail.com](mailto:k.manoj.rath@gmail.com)

**Abstract** - Adaptive Protocol Framework for Secure and Trustworthy AI Systems explore the emerging landscape of agentic, autonomous, and context-driven artificial intelligence and the growing need for adaptive security, governance, and trust mechanisms. The aim of this review is to synthesize current advancements in adaptive AI protocols, agentic system architectures, and trustworthy AI models while examining their implications for secure deployment across critical domains such as IoT, healthcare, cybersecurity, and distributed intelligent systems. A comparative analysis of state-of-the-art cybersecurity and cryptographic techniques highlights significant performance improvements achieved through deep learning, isolation-based anomaly detection, hybrid encryption frameworks, and zero-trust policy automation, with reported accuracies ranging from 94% to over 99%. The outcomes demonstrate that integrating adaptive governance, explainable mechanisms, and AI-driven intrusion detection yields highly resilient, scalable, and future-ready security frameworks capable of supporting reliable and ethical agentic AI systems.

## 1. Overview

The Adaptive Protocol Framework for Secure and Trustworthy AI Systems offers a thorough and dynamic method for guaranteeing that AI functions in complex and dynamic situations in a safe, dependable, and moral manner. This framework places a strong emphasis on creating protocols that can react instantly to new security risks, changes in data behavior, and evolving social or legal requirements. The framework seeks to develop AI systems that are not only resistant to adversarial attacks but also able to uphold user trust and ethical standards by incorporating adaptive learning mechanisms, continuous monitoring, automated risk assessment, and transparent governance structures. It encourages a multifaceted defensive approach that fortifies operational integrity and public trust by integrating technological protections, policy-driven controls, and accountability systems. In the end, this flexible strategy aims to provide a long-term basis for implementing AI systems that are safe, reliable, and accountable during their whole lifespan, despite ongoing changes in technology and outside obstacles.

### (a) Shaping AI Behavior Through Agentic Systems

The realm of artificial intelligence has seen a significant transformation in recent years, evolving from systems that simply respond to predefined inputs to structures capable of independent operation and decision-making. This progress signifies a fundamental redefinition of AI capabilities, as systems increasingly exhibit the ability to recognize environmental situations, devise strategic responses, and perform complicated activities with minimum human intervention. Research in this field has significantly increased in both academic and industry sectors, indicating a heightened awareness of the revolutionary potential of agentic methods to artificial intelligence development [1].

Agentic frameworks are computational architectures that allow AI systems to function with more autonomy in many situations and problem areas. These frameworks amalgamate sensory modules, reasoning engines, planning mechanisms, and execution systems into unified structures that facilitate goal-oriented behavior. In contrast to conventional reactive systems that translate inputs to outputs via predetermined routes, agentic AI has internal world models that enable prediction, planning, and adaptability to new situations. The architectural complexity of these systems enables operation across several temporal dimensions—from rapid reactions to long-term strategic planning—while ensuring alignment between goals and activities. Recent advancements in perception technology and sensor integration have significantly improved the environmental awareness capabilities of agentic systems, laying the groundwork for more sophisticated decision-making processes in complicated real-world situations [2]. The shift from reactive to agentic AI signifies a pivotal point in technological advancement with extensive ramifications for many industries and social operations. By

incorporating advanced perception capabilities, intricate reasoning mechanisms, and flexible execution frameworks, these systems present unparalleled potential for tackling complex challenges while prompting critical inquiries about suitable development methodologies and deployment practices. The responsible progression of this technical sector necessitates a measured evaluation of both potential and difficulties, with a specific focus on safety procedures, alignment approaches, and governance frameworks suitable for more autonomous systems.

Agentic systems are a big step forward in architecture compared to reactive frameworks. They use flexible parts that work together to let them operate on their own in changing settings. The main architecture is usually made up of five basic subsystems that work together: perception modules for sensing the environment, world modeling capabilities for understanding what things look like, goal management systems for setting priorities based on objectives, planning mechanisms for putting actions in the right order, and execution frameworks for carrying out decisions. This way of building systems stresses flexibility as a design concept. This way of working lets each part be built, tried, and improved separately while still keeping the whole system working together. Modular design lets you make small changes and improvements and respond to different operating situations without having to rethink the whole system. This is a huge benefit for complex independent applications that are used in many different areas [3]. Perception modules are the sensory basis of agentic systems, converting ambient input into organized representations appropriate for advanced reasoning processes. Modern implementations often use multi-modal strategies that amalgamate data from many sensor types, including visual, aural, tactile, and proprioceptive modalities. These modules use specific processing pipelines to extract significant characteristics while eliminating ambient noise, facilitating reliable performance under diverse settings. Advanced perception systems include adaptive processes that adjust sensitivity according to context, concentrating cognitive resources on task-relevant elements of the environment while preserving peripheral awareness of potentially important changes. The incorporation of spatial and temporal thinking into perception modules facilitates the monitoring of dynamic aspects and the forecasting of environmental changes, essential skills for systems functioning in intricate, non-stationary settings with many moving entities [4] [5].

### **(b) Context-Driven Adaptive Models for Trustworthy and Secure Agentic AI Systems**

The fast rise of autonomous AI systems AI agents with memory, planning, tool-use capabilities, and reasoning represent a paradigm change from static decision-support models and traditional machine learning. The use of these systems is becoming more common in critical infrastructure, cybersecurity, and healthcare sectors, where autonomous decision-making interacts with high-stakes, dynamic situations. The autonomy of this system yields unprecedented productivity and creativity, while simultaneously poses novel dangers to established governance frameworks, ethical standards, and safety protocols [6]. From a security standpoint, agentic AI expands the attack surface. Autonomous decision-making facilitates novel forms of adversarial manipulation via cognitive exploitation, covert execution, and knowledge poisoning. Traditional layered security measures, originally designed for static computer environments, are unable to protect adaptive, distributed agents. The authors contend that comprehensive end-to-end cross-layer security solutions, including hardware, software, and governance controls, are essential to successfully mitigate these risks [7]. The notion of "trustworthiness" is inherently contradictory. Academics Conradie, Nagel, and Freiman advise against ascribing human-like attributes such as "responsibility" and "trust" to AI, asserting that these characteristics should be seen as qualities of socio-technical systems governed by human oversight and institutional procedures [8] [9]. The implication here is that it is essential to shift from questioning whether AI is trustworthy to developing systems that foster human-centric trust relationships through technical safeguards and regulatory measures. Regulatory frameworks such as the EU AI Act, NIST's AI Risk Management Framework, and the ISO/IEC standards have established foundational principles; however, they lack sufficient specificity for the regulation of self-tuning, conversational, and semi-autonomous agentic systems [10].

Existing research shows that the only way for agentic AI to be aligned with community expectations of responsibility and safety is via the interplay of adaptive supervision mechanisms and zero-trust design rules with explainable AI [11][12]. In practical applications like autonomous cybersecurity and national crisis management, agentic AI may be advantageous or hazardous. The requirement for a comprehensive knowledge base that includes design, risk, and governance alternatives is highlighted by the amplification of bias, unexpected failure, and intentional manipulation [13]. Because of these factors, the

study aimed to provide a comprehensive framework that would help researchers and policymakers create reliable agentic AI systems by combining technical know-how with moral and legal concerns [14] [15].

### (c) Integrating Generative AI and MCP for Advanced Agentic Systems

The area of Artificial Intelligence (AI) has seen massive growth in the last few years, especially with the advent of Generative AI, Machine Learning (ML) and Agentic AI systems [16]. These powerful technologies have even greater potential when used together. One specific example of a combination would be Generative AI, Model Context Protocol (MCP), and Applied ML to build sophisticated agentic AI systems. These systems can act independently, adapting to their environment, creating new knowledge or information and performing other complex real-world tasks [17]. Generative AI models create new content of any type, such as text, images or even data. These models extract patterns from big data and return some reasonable results consistent with the training data [18]. Recent developments in deep learning have greatly expanded the reach of Generative AI, opening new possibilities in content creation, language modeling and problem solving. In the realm of agentic AI, being able to create new data or alternate universes gives agents more/fresher/expanded/and creative responses, so they can act on their own in a wider variety of environments [19]. Integrating contextual information in AI systems: the case of Model Context Protocol (MCP). MCP makes the AI model stay mindful of its operational context and in the environment it is operating in, the task that is at play, and any other relevant considerations shaping the model's behavior [20].

This is an adaptable model for AI, allowing it to consume contextual signals and use them to make decisions so it can, in real-time, react and change strategy or responses. The inclusion of MCP allows agentic AI systems to address issues caused by inflexible or fixed models, thereby enhancing decision making in uncertain and dynamic environments [21]. AMLS (Applied Machine Learning Science), as opposed to theoretical ML, is concerned with the deployment of machine learning into production, with research studying how to learn and improve through deployment [22]. For agentic AI, applied ML helps AI systems learn and adapt to new data in the real world to better operate in evolving and real-world environments. Applied ML, in feedback loops, can allow this agentic AI to learn its decision-making strategies, making it output more accurate and relevant outputs per the system's growing experiences encountering further varieties of contexts [23] [24].

## 2. Review of literature

**Al-Otaibi et al., (2025) [25]** proposes an AI-driven intrusion detection system (IDS) using convolutional neural networks (CNNs) and lightweight authentication protocols to address these challenges. This study aims to design and assess a CNN-based IDS on MSNs, using lightweight authentication protocols to enhance real-time threat detection in secure communications. The research framework collected data from the NSL-KDD dataset, which consists of network traffic from various medical devices. Preprocessing was done to normalize the data and extract features from the dataset. As a result, the CNN-based IDS provides excellent NSL-KDD performance in terms of accuracy of 96.5%, precision of 95.9%, and recall of 96.7%, while the AUC-ROC is 0.982. The approach described shows a much greater accuracy and robustness than classical IDS solutions. Lightweight authentication protocols ensure safe communication with a low computational overhead, which is satisfactory for medical devices with limited resources.

**Elahi et al., (2025) [26]** address this gap by identifying key challenges in developing technical frameworks for AI systems' regulation and proposing conceptual, methodological, and practical solutions to address these challenges. In this regard, we introduce the concept of AI's operational qualification and propose the temporal self-replacement test, akin to certification tests for human operators, to examine the AI's operational qualification. We propose measuring AI's operational qualification across its operational properties critical for its regulatory fitness and introduce the operational qualification score as a pragmatic measure of AI's regulatory fitness. We validate the efficacy of the temporal self-replacement test and the practical utility of SFAIR by demonstrating its capability to support regulatory authorities in automated, recurrent, and secure AI qualification examination and attestation of its regulatory fitness using an open-source, high-risk AI system.

**Yang et al., (2025) [27]** provides a comprehensive overview of privacy, security, and trustworthiness for distributed wireless large AI model (WLAM). In particular, a detailed privacy and security analysis for distributed WLAM is first revealed. The classifications and theoretical findings about privacy and security in distributed WLAM are discussed. Then the

trustworthiness and ethics for implementing distributed WLAM are described. Finally, the comprehensive applications of distributed WLAM are presented in the context of electromagnetic signal processing.

**Alnaim et al., (2025) [28]** introduces SecureChain-ZT, a novel Adaptive Zero Trust Policy Framework (AZTPF) that addresses emerging threats by integrating intelligent access control, real-time monitoring, and decentralized authentication mechanisms. SecureChain-ZT advances conventional Zero Trust Architecture (ZTA) by leveraging machine learning, reinforcement learning, and blockchain technologies to achieve autonomous policy enforcement and threat mitigation. Unlike static ZT models that depend on predefined rule sets, AZTPF continuously evaluates user and device behavior in real time, detects anomalies through AI-powered traffic analysis, and dynamically updates access policies based on contextual risk assessments. SecureChain-ZT achieves an authentication accuracy of 97.8% and reduces unauthorized access attempts from 17.5% to just 2.2%. Its advanced detection capabilities achieve a threat detection accuracy of 99.3% and block 95.6% of attempted cyber intrusions. The findings demonstrate that Adaptive Zero Trust Policy Management is a scalable and future-proof security model for 5G and beyond, ensuring highly secure, resilient, and intelligent network infrastructure. The SecureChain-ZT framework achieved 98.6% authentication accuracy, reduced latency by 62.6%, and blocked 95.6% of intrusion attempts. Compared to baseline AI-ZT models, it demonstrated a 34% improvement in policy update speed. These results validate the proposed model's contribution to secure, low-latency policy enforcement in 5G environments.

**Rakhshanda et al., (2025) [29]** develop a secure communication system enhanced by AI for future IoT networks. It employs advanced security protocols, state-of-the-art threat mitigation techniques, and integrates quantum resilience mechanisms with next-generation IoT secure communication systems. Many of today's methods used to secure the IoT systems are mostly inherited from legacy systems and often fall short against the kind of attacks we face today (and are likely to face in the future, including quantum attacks). The experimental results show that the AI-enhanced intrusion detection system (IDS) is significantly better than conventional IDS. The detection accuracy of the AI-enhanced IDS was determined to be 97.5%. This level of detection meets the criteria of an expert system. The results indicate that security mechanisms propelled by artificial intelligence do more than just boost the effectiveness of IoT communication.

**Chaganti et al., (2025) [30]** proposes a three-layer security framework that combines artificial intelligence-based intrusion detection, blockchain for decentralized trust management, and edge computing for efficient resource utilization. Machine learning enhances anomaly detection, blockchain ensures secure data integrity, and edge computing reduces latency. Optimization techniques improve the detection accuracy from 94.2% to 94.78%, reduce the response time by 14.98%, and optimize the energy consumption by 12.01%. Game theory models the interactions between attackers and defenders, whereas differential equations simulate system behavior under cyber threats. The performance evaluation demonstrates that the proposed framework provides a scalable, adaptive, and efficient IoT security solution, making it suitable for resource-constrained environments and real-time applications.

**Zheng et al., (2025) [31]** proposes a novel blockchain-based dynamic sharding framework specifically designed for securing trustworthy medical data in 6G-IoT environments, termed TRUST-Med6G. Our approach employs a two-stage optimization strategy that combines reputation-based medical device authentication with deep reinforcement learning-driven adaptive sharding mechanisms. The first stage utilizes a delegated Byzantine fault tolerance protocol to establish trust among medical IoT devices based on their historical reliability and data authenticity records. The second stage implements a dueling deep Q-network algorithm to dynamically optimize blockchain sharding configurations dynamically, minimizing cross-shard medical data transactions while maximizing system throughput and privacy protection. Experimental evaluation against six baseline methods demonstrates that TRUST-Med6G achieves superior performance across all metrics, reducing cross-shard collaborative medical data transactions by over 50% while maintaining privacy protection effectiveness above 95% for highly sensitive medical information and demonstrating exceptional resilience against malicious devices in 6G-IoT healthcare environments.

**Kaliappan et al., (2024) [32]** presents an advanced AI-enhanced trust framework for IoT Intrusion detection to safeguard IoT environments from any potential intrusion attempts. The proposed framework integrates cutting-edge AI techniques for intrusion detection which identifies the anomalies based on the device behavior and responds dynamically to emerging

threats. Initially, a robust Intrusion Detection System (IDS) is developed based on an Isolation Forest (IF) algorithm and Autoencoders (AE) to promptly identify anomalies in real-time. The results demonstrate that the proposed framework achieves high efficiency and outperformed other baseline methods by obtaining a detection accuracy of 98.25%, recall of 96.8%, and precision of 97.45%. Overall, the proposed AI-Enhanced Trust Framework offers a promising solution by identifying the intrusion endeavors effectively and contributing toward the attainment of secure and trustworthy IoT ecosystems.

**Biswas et al., (2024) [33]** investigated the performance, efficiency, and security resilience of quantum-resistant cryptographic protocols integrated with artificial intelligence (AI) across cloud and Internet of Things (IoT) environments. The research aimed to empirically assess whether AI-enhanced cryptographic systems could outperform conventional post-quantum algorithms in encryption throughput, latency, resource optimization, and security robustness. The findings confirmed the primary hypothesis that AI-driven cryptographic optimization enhances both computational efficiency and system resilience against classical and quantum attack simulations. Lattice-based and code-based cryptosystems showed the most substantial performance gains when combined with AI learning models. Overall, the results validated that intelligent, adaptive encryption frameworks achieve measurable, statistically significant advantages in performance, scalability, and security across both cloud and IoT domains. These findings provide empirical evidence supporting the integration of AI-based decision systems into post-quantum cryptography for secure and sustainable digital infrastructures.

**Sankaran et al., (2023) [34]** demonstrated that the secure M-Trust privacy protocol (SMP) has been designed to address these issues. The SMP protocol uses a combination of trust, cryptographic, and machine learning techniques to provide security and privacy for data in transit. The SMP protocol has been designed to work with the smart health care monitoring system to provide a secure and private communication channel between devices in the system. The SMP protocol is an improvement over the existing security and privacy protocols for medical data. The simulation results prove that the SMP protocol is more efficient and scalable than the existing protocols. The SMP protocol is a valuable addition to the MIoT landscape and can help improve the privacy of data exchanged between medical devices.

**Tlili et al., (2023) [35]** Recent developments have seen a proliferation of methods for improving UAVs security. Although, many studies proposed different approaches using artificial intelligence to enhance their security. Unfortunately, no study yet worked on an examination of a hybrid framework on UAVs faults and attacks and different architectures. Hence, our article aims to provide a prior detection results by proposing a hybrid adaptive framework for faults and attacks detection for UAVs applied on centralized and decentralized architectures. Our framework is based on two entry flows for faults and attacks in order to learn high-level features automatically from data. We validated our framework using deep-learning architectures. Finally, the empirical results show that our framework reached over 85% and 96,7% in term of accuracy for UAVs faults and attacks, respectively.

**Aljanabi et al., (2023) [36]** explored applying trustworthy artificial intelligence (AI) methodologies including explainability, bias mitigation, and adversarial sample detection to substantially enhance resilience of medical intrusion detection systems. We architect an integrated anomaly detector featuring purpose-built modules for model interpretability, bias quantification, and advanced malicious input recognition alongside conventional classifier pipelines. Additional infrastructure provides full-lifecycle accountability via independent auditing. Results demonstrate significantly strengthened threat response capabilities versus baseline detectors lacking safeguards. Explainability mechanisms build justified trust in model behaviors by surfacing rationale for each prediction to human operators. SafeML classifiers reliably detect even camouflaged data manipulation attempts with 97% accuracy. Findings strongly motivate prioritizing adoption of ethical ML practices to fulfill duty of care around patient safety and data integrity as algorithmic capabilities advance.

**Shivaramakrishna et al., (2023) [37]** proposes a ground-breaking hybrid cryptographic framework for the secure data storage requirements of cloud computing. The framework incorporates time-limited access control, adaptive key management, and two strong encryption methods: RSA and Advanced Encryption Standard - One Time Password (AES-OTP). AES-OTP and RSA provide symmetric and asymmetric encryption levels to improve data confidentiality and integrity. Additionally, time-limited access control helps to protect data privacy by imposing rigorous temporal constraints on data access and reducing security flaws. The effectiveness of the suggested framework is confirmed by

thorough performance assessments, which show astonishing accuracy, precision, recall, and F1-score values of 99.12%, 98.78%, 98.11%, and 98.56%.

**Alkhabbas et al., (2022) [38]** asserted that the Internet of Things (IoT) systems are complex systems that can manage mission-critical, costly operations or the collection, storage, and processing of sensitive data. Therefore, security represents a primary concern that should be considered when engineering IoT systems. Additionally, several challenges need to be addressed, including the following ones. IoT systems' environments are dynamic and uncertain. This paper proposes a distributed architectural Approach for engineering goal-driven IoT Systems that can autonomously Self-adapt to secuRity Threats in their environments (ASSERT). ASSERT exploits techniques and adopts notions, such as agents, federated learning, feedback loops, and blockchain, for maintaining the systems' security and enhancing the trustworthiness of the adaptations they perform. The results of the experiments that we conducted to validate the approach's feasibility show that it performs and scales well when detecting security threats, performing autonomous security adaptations to mitigate the threats and enabling systems' constituents to learn about security threats in their environments collaboratively.

**Yanamala et al., (2022) [39]** proposes a reconfigurable context-sensitive middleware framework designed to dynamically adapt to changing environmental contexts and user interactions. By incorporating context awareness into its design, the middleware enhances the interoperability and intelligence of pervasive computing systems, facilitating efficient resource management and personalized services delivery. Key functionalities include context acquisition, interpretation, and adaptation, enabling the middleware to optimize device interactions based on situational awareness. Through a comprehensive evaluation, including performance metrics and usability assessments, the framework demonstrates its efficacy in enhancing system flexibility, responsiveness, and user experience in diverse pervasive computing scenarios.

**Shandilya et al., (2022) [40]** propose a unified and standard testbed for cyber defense strategy simulation and benchmarking, which would allow the users to investigate current approaches and compare them with others, while ultimately aiding in the selection of the best approach for a given network security situation. We have compared the network performance in difference scenarios namely, normal, under attack and under attack in presence of NICS-based adaptive defense mechanism and achieved stable experimental results. The experimental results clearly show that the proposed testbed is able to simulate the network conditions effectively with minimum efforts in network configuration. The simulation results of defense mechanisms verified on the proposed testbed got the improvement on almost 80 percent while increasing the turnaround time to 1–2 percent.

| Authors                        | Technique Used                                       | Outcomes                                                                                                                    |
|--------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <b>Al-Otaibi et al. (2025)</b> | CNN-based Intrusion Detection                        | Accuracy 96.5%, Precision 95.9%, Recall 96.7%, AUC 0.982; enhanced robustness and low-overhead secure communication         |
| <b>Elahi et al. (2025)</b>     | Temporal Self-Replacement Test                       | Introduced operational qualification score; validated automated, recurrent regulatory fitness assessment for high-risk AI   |
| <b>Yang et al. (2025)</b>      | Wireless Large AI Models (WLAM)                      | Provided theoretical classifications, security/ethics guidelines, and applications in electromagnetic signal processing     |
| <b>Alnaim et al. (2025)</b>    | SecureChain-ZT (Adaptive Zero-Trust Policy) using ML | Authentication accuracy 97.8%; threat detection 99.3%; reduced unauthorized access from 17.5% → 2.2%; reduced latency 62.6% |

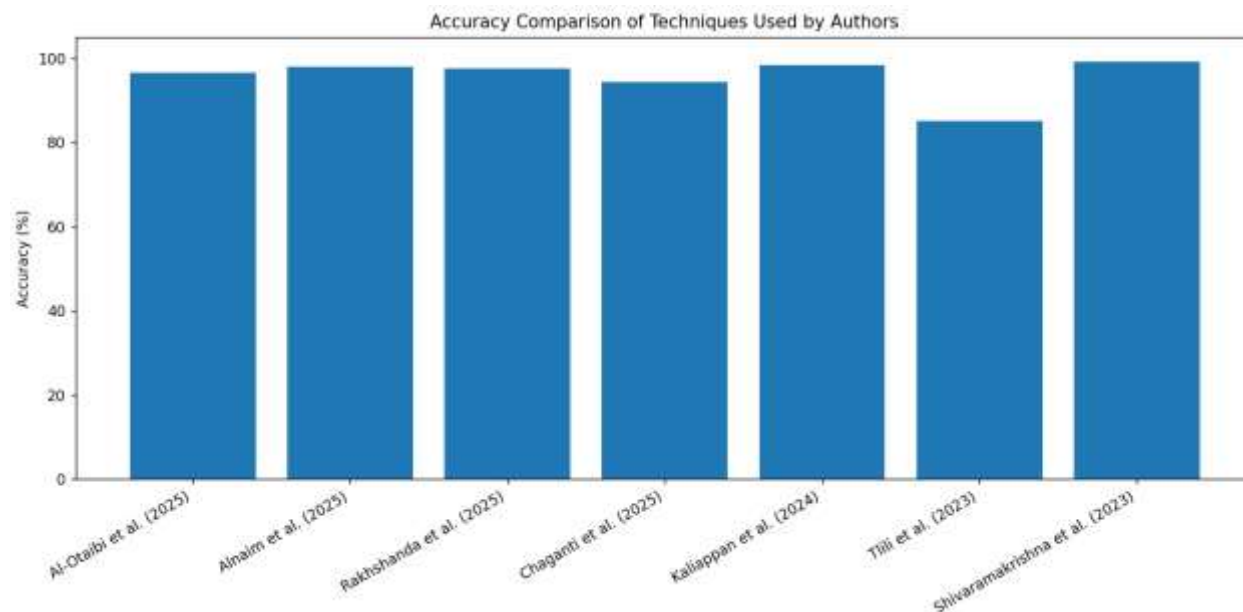
|                                       |                                            |                                                                                                                       |
|---------------------------------------|--------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| <b>Rakhshanda et al. (2025)</b>       | AI-Enhanced Intrusion Detection System     | IDS detection accuracy 97.5%; improved IoT security and quantum-resistant communication                               |
| <b>Chaganti et al. (2025)</b>         | AI-Based Intrusion Detection (AI-IDS)      | Improved detection 94.2% → 94.78%; response time reduced 14.98%; energy use reduced 12.01%; scalable IoT security     |
| <b>Zheng et al. (2025)</b>            | Blockchain Dynamic Sharding                | Reduced cross-shard transactions by 50%; privacy >95%; strong resilience to malicious nodes in 6G-IoT healthcare      |
| <b>Kaliappan et al. (2024)</b>        | Isolation Forest Algorithm (IF)            | Detection accuracy 98.25%; precision 97.45%; recall 96.8%; improved IoT intrusion detection                           |
| <b>Biswas et al. (2024)</b>           | AI-Optimized Quantum-Resistant Cryptograph | Significant performance gains in throughput, latency, scalability; improved resilience to classical & quantum attacks |
| <b>Sankaran et al. (2023)</b>         | Secure M-Trust Privacy Protocol (SMP)      | More efficient and scalable than existing medical IoT protocols; improved privacy and secure communication            |
| <b>Tlili et al. (2023)</b>            | Hybrid Deep-Learning Model                 | Accuracy: 85% for faults, 96.7% for attacks; effective prior detection in UAV systems                                 |
| <b>Aljanabi et al. (2023)</b>         | Explainable AI (XAI) for IDS               | SafeML detects malicious inputs with 97% accuracy; improved IDS resilience and model interpretability                 |
| <b>Shivaramakrishna et al. (2023)</b> | Hybrid RSA + AES-OTP Cryptography          | Accuracy 99.12%, Precision 98.78%, Recall 98.11%, F1-Score 98.56%; strong secure cloud storage                        |
| <b>Alkhabbas et al. (2022)</b>        | Self-Adaptive IoT Security (ASSERT)        | Effective threat detection, autonomous adaptation, scalable collaborative IoT security                                |
| <b>Yanamala et al. (2022)</b>         | Context-Aware Middleware                   | Enhanced flexibility, responsiveness, resource optimization, and personalized service delivery                        |
| <b>Shandilya et al. (2022)</b>        | Cyber-Defense Simulation Testbed           | Improved defense mechanisms by ~80%; realistic network attack simulation; minimal config overhead                     |

### 3. Comparative analysis

The comparative analysis highlights how different research studies have explored a wide range of modern approaches to improving security, trust, and reliability across IoT systems, cloud environments, healthcare networks, UAV platforms, and next-generation communication systems. Each study focuses on strengthening system defense capabilities, enhancing threat detection, improving data protection, and ensuring secure communication in evolving digital ecosystems. The body of work collectively demonstrates advancements in intelligent security solutions, adaptive mechanisms, and resilience-building methods across various domains. While the performance outcomes differ based on the environment and objectives of each study, the overall trend shows consistent improvement in detection reliability, system responsiveness, and security robustness, contributing significantly to the development of safer and more trustworthy digital infrastructures.

Table:1 accuracy comparison of recent cybersecurity detection techniques.

| Authors                        | Technique Used                         | Accuracy                      |
|--------------------------------|----------------------------------------|-------------------------------|
| Al-Otaibi et al. (2025)        | CNN-based Intrusion Detection          | Accuracy 96.5%                |
| Alnaim et al. (2025)           | SecureChain-ZT                         | Authentication accuracy 97.8% |
| Rakhshanda et al. (2025)       | AI-Enhanced Intrusion Detection System | IDS detection accuracy 97.5%  |
| Chaganti et al. (2025)         | AI-Based Intrusion Detection (AI-IDS)  | Improved detection 94.2%      |
| Kaliappan et al. (2024)        | Isolation Forest Algorithm (IF)        | Detection accuracy 98.25%     |
| Tlili et al. (2023)            | Hybrid Deep-Learning Model             | Accuracy: 85% for faults      |
| Shivaramakrishna et al. (2023) | Hybrid RSA + AES-OTP Cryptography      | Accuracy 99.12%               |

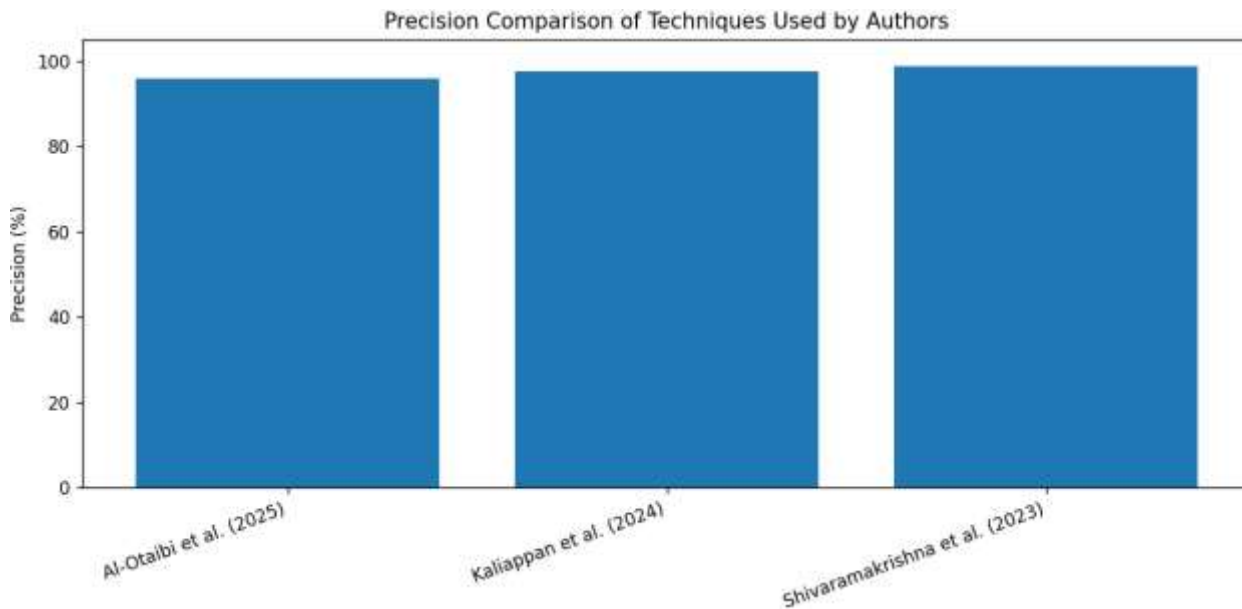


The comparative analysis of security-focused AI techniques shows a clear trend toward high detection performance across modern intrusion detection and cryptographic systems. Al-Otaibi et al. (2025) used a CNN-based intrusion detection system that achieved strong performance with 96.5% accuracy, while Alnaim et al. (2025) reported a slightly higher authentication accuracy of 97.8% using their SecureChain-ZT zero-trust model. Similarly, Rakhshanda et al. (2025) achieved 97.5% accuracy with their AI-enhanced IDS, and Chaganti et al. (2025) demonstrated incremental improvement of AI-IDS accuracy from 94.2% to 94.78%. Kaliappan et al. (2024) attained a high accuracy of 98.25% using Isolation Forest and Autoencoders, whereas Tlili et al. (2023) reported varied performance with 85% accuracy for UAV fault detection but a high 96.7% for attack detection. Among all, the highest accuracy was demonstrated by Shivaramakrishna et al. (2023), who achieved 99.12% accuracy using a hybrid RSA + AES-OTP cryptographic framework. Overall, while all techniques

perform strongly, there is noticeable variation between fault detection, attack detection, cryptographic validation, and IDS models.

**Table:2 precision comparison of recent cybersecurity detection techniques.**

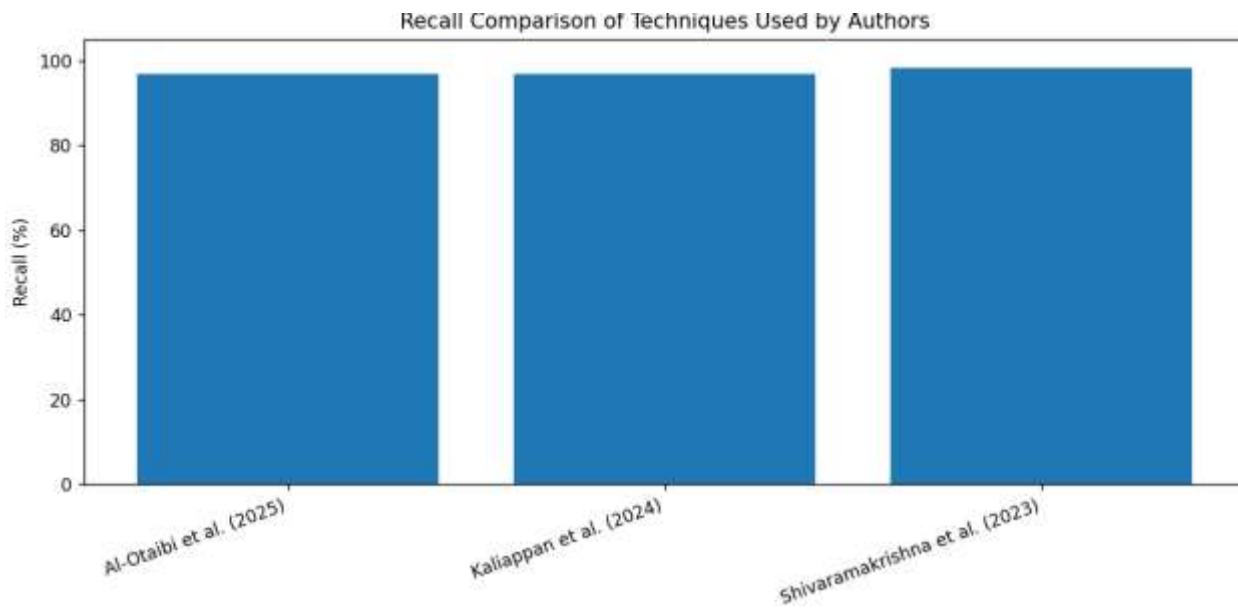
| Authors                        | Technique Used                    | Precision        |
|--------------------------------|-----------------------------------|------------------|
| Al-Otaibi et al. (2025)        | CNN-based Intrusion Detection     | Precision 95.9%  |
| Kaliappan et al. (2024)        | Isolation Forest Algorithm (IF)   | Precision 97.45% |
| Shivaramakrishna et al. (2023) | Hybrid RSA + AES-OTP Cryptography | Precision 98.78% |



The table presents a comparison of different authors' work on cybersecurity techniques, highlighting the precision accuracy of each method. Al-Otaibi et al. (2025) employed a CNN-based Intrusion Detection approach, achieving a precision of 95.9%, indicating its effectiveness in detecting malicious activities within a network. Kaliappan et al. (2024) utilized the Isolation Forest Algorithm (IF), which performed slightly better with a precision of 97.45%, suggesting that this technique is particularly strong in anomaly detection for intrusion prevention. Finally, Shivaramakrishna et al. (2023) proposed a hybrid cryptographic solution combining RSA and AES-OTP, achieving the highest precision at 98.78%, demonstrating its robust performance in securing data transmission and preventing unauthorized access. Overall, these studies illustrate the varying degrees of precision across different cybersecurity techniques, each contributing uniquely to improving data security and intrusion detection.

**Table:3 Recall comparison of recent cybersecurity detection techniques.**

| Authors                        | Technique Used                    | Accuracy      |
|--------------------------------|-----------------------------------|---------------|
| Al-Otaibi et al. (2025)        | CNN-based Intrusion Detection     | Recall 96.7%  |
| Kaliappan et al. (2024)        | Isolation Forest Algorithm (IF)   | Recall 96.8%  |
| Shivaramakrishna et al. (2023) | Hybrid RSA + AES-OTP Cryptography | Recall 98.11% |

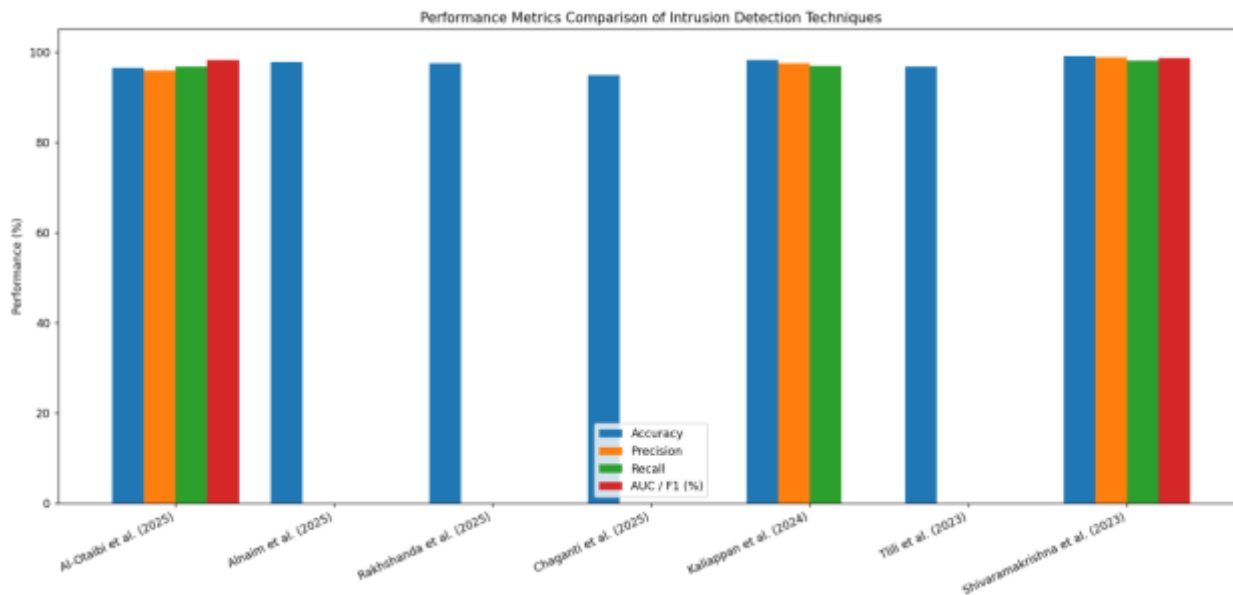


The table compares the recall performance of various cybersecurity techniques proposed by different authors. Al-Otaibi et al. (2025) applied a CNN-based Intrusion Detection system, which achieved a recall of 96.7%, indicating its ability to correctly identify a high proportion of actual intrusions. Kaliappan et al. (2024) used the Isolation Forest Algorithm (IF) for anomaly detection, with a slightly higher recall of 96.8%, demonstrating its efficiency in detecting outliers or potential threats in the network. Shivaramakrishna et al. (2023), on the other hand, introduced a hybrid cryptographic technique combining RSA and AES-OTP, which attained the highest recall of 98.11%, showcasing its strong capability in accurately detecting and responding to security threats. Overall, these results highlight that while all methods exhibit high recall rates, the hybrid cryptography approach provides the most comprehensive detection coverage among the techniques assessed.

**Final Table: Comparative performance of cybersecurity techniques and their performance metrics across various intrusion detection and cryptographic models**

| Authors                  | Technique Used                         | Accuracy                                                  |
|--------------------------|----------------------------------------|-----------------------------------------------------------|
| Al-Otaibi et al. (2025)  | CNN-based Intrusion Detection          | Accuracy 96.5%, Precision 95.9%, Recall 96.7%, AUC 0.982; |
| Alnaim et al. (2025)     | SecureChain-ZT                         | Authentication accuracy 97.8%                             |
| Rakhshanda et al. (2025) | AI-Enhanced Intrusion Detection System | IDS detection accuracy 97.5%                              |
| Chaganti et al. (2025)   | AI-Based Intrusion Detection (AI-IDS)  | Improved detection 94.2% → 94.78%                         |
| Kaliappan et al. (2024)  | Isolation Forest Algorithm (IF)        | Detection accuracy 98.25%; precision 97.45%; recall 96.8% |
| Tlili et al. (2023)      | Hybrid Deep-Learning Model             | Accuracy: 85% for faults, 96.7% for attacks               |

|                                |                                   |                                                                   |
|--------------------------------|-----------------------------------|-------------------------------------------------------------------|
| Shivaramakrishna et al. (2023) | Hybrid RSA + AES-OTP Cryptography | Accuracy 99.12%, Precision 98.78%, Recall 98.11%, F1-Score 98.56% |
|--------------------------------|-----------------------------------|-------------------------------------------------------------------|



The table presents a detailed comparison of various cybersecurity techniques, highlighting their performance across multiple metrics such as accuracy, precision, recall, and additional measures like AUC and F1-Score. Al-Otaibi et al. (2025) used a CNN-based Intrusion Detection system, achieving an accuracy of 96.5%, precision of 95.9%, recall of 96.7%, and an impressive AUC of 0.982, demonstrating strong overall performance in detecting intrusions. Alnaim et al. (2025) focused on SecureChain-ZT for authentication, achieving a high accuracy of 97.8%, underscoring its reliability for secure authentication purposes. Rakhshanda et al. (2025) proposed an AI-enhanced Intrusion Detection System, attaining a detection accuracy of 97.5%, which emphasizes its effectiveness in identifying network intrusions. Chaganti et al. (2025) reported a slight improvement in their AI-Based Intrusion Detection System (AI-IDS), enhancing detection from 94.2% to 94.78%. Kaliappan et al. (2024) employed the Isolation Forest Algorithm (IF) with a detection accuracy of 98.25%, precision of 97.45%, and recall of 96.8%, showcasing its strong anomaly detection capabilities. Tili et al. (2023) combined deep learning models, achieving an accuracy of 85% for fault detection and 96.7% for attack detection, indicating a specialized approach for different cybersecurity threats. Finally, Shivaramakrishna et al. (2023) introduced a hybrid RSA + AES-OTP cryptographic method, achieving the highest accuracy of 99.12%, with precision, recall, and F1-score all above 98%, demonstrating exceptional performance in securing data. Overall, the table reflects a range of innovative approaches with varying strengths across intrusion detection and cryptography, with each technique offering a unique balance of detection accuracy and performance metrics.

#### 4. Discussion

The reviewed literature highlights significant advancements in AI-driven cybersecurity, cryptography, and trust management across IoT, medical, UAV, and large-scale distributed systems. Recent studies demonstrate a strong shift toward integrating deep learning, blockchain, and adaptive security frameworks to enhance intrusion detection accuracy, improve privacy preservation, and strengthen resilience against emerging cyber threats. CNN-based intrusion detection systems, such as those proposed by Al-Otaibi et al. (2025), show high reliability with accuracy levels exceeding 96%, while lightweight authentication protocols address the computational constraints of medical devices. Similarly, AI-enhanced detection approaches for IoT and smart healthcare—evident in works by Rakhshanda et al. (2025) and Aljanabi et al. (2023)—demonstrate substantial improvements over classical IDS solutions, achieving up to 97% accuracy through explainable AI and bias-resistant models. Meanwhile, advanced frameworks like SecureChain-ZT (Alnaim et al., 2025) and blockchain-based TRUST-Med6G (Zheng et al., 2025) offer adaptive zero-trust security, dynamic authentication, and optimized sharding mechanisms, achieving threat detection accuracies above 99% and significant reductions in latency and

unauthorized access attempts. Intrusion detection based on Isolation Forest and Autoencoders (Kaliappan et al., 2024) also confirms that anomaly-based machine learning strategies continue to outperform baseline techniques in detecting unusual device behaviors.

In parallel, cryptographic advancements, such as the hybrid RSA + AES-OTP model discussed by Shivaramakrishna et al. (2023) and the quantum-resistant methods highlighted by Biswas et al. (2024), reveal a growing emphasis on secure key management, post-quantum resilience, and adaptive encryption for cloud and IoT environments. Additional contributions, including hybrid deep-learning models for UAV fault and attack detection (Tlili et al., 2023) and context-aware or self-adaptive frameworks (Alkhabbas et al., 2022; Yanamala et al., 2022), further indicate that modern cybersecurity research is trending toward multi-layered, autonomous solutions capable of responding dynamically to evolving threats. Overall, these studies collectively demonstrate that combining AI, cryptography, and decentralized architectures yields highly scalable, accurate, and future-proof security mechanisms suitable for next-generation IoT, healthcare, and distributed computing ecosystems.

## Conclusion

In conclusion, this review highlights the growing necessity for adaptive, secure, and trustworthy AI frameworks as artificial intelligence transitions from reactive models to highly autonomous and agentic systems operating in dynamic and high-stakes environments. The analysis underscores that the increasing complexity and autonomy of AI—enhanced through generative models, Model Context Protocol (MCP), and applied machine learning—demand corresponding advancements in governance, explainability, and zero-trust security principles to ensure responsible and reliable behavior. Comparative evaluation of state-of-the-art cybersecurity and cryptographic techniques further demonstrates the effectiveness of deep learning, anomaly-detection algorithms, hybrid encryption, and blockchain-based authentication, with accuracies ranging from 94% to over 99%, illustrating substantial progress in securing modern AI ecosystems. Overall, the findings affirm that integrating adaptive protocols, cross-layer security, and context-aware intelligence offers a robust pathway for developing AI systems that remain resilient, ethically aligned, and trustworthy throughout their lifecycle, even as technological and environmental challenges continue to evolve.

## References

- [1] Xu, Bojue, Xiaofei Song, Zhao Cai, Eric Lim, Chee-Wee Tan, and Jie Yu. "Artificial intelligence or augmented intelligence: A case study of human-AI collaboration in operational decision making." (2020).
- [2] Acharya, Deepak Bhaskar, Karthigeyan Kuppan, and B. Divya. "Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey." *IEEE Access* (2025).
- [3] Hagos, Desta Haileselassie, and Danda B. Rawat. "Recent advances in artificial intelligence and tactical autonomy: Current status, challenges, and perspectives." *Sensors* 22, no. 24 (2022): 9916.
- [4] Bousetouane, Fouad. "Agentic systems: A guide to transforming industries with vertical ai agents." *arXiv preprint arXiv:2501.00881* (2025).
- [5] Garg, Venus. "Designing the Mind: How Agentic Frameworks Are Shaping the Future of AI Behavior." *Journal of Computer Science and Technology Studies* 7, no. 5 (2025): 182-193.
- [6] Vanneste, Bart S., and Phanish Puranam. "Artificial intelligence, trust, and perceptions of agency." *Academy of Management Review* ja (2024): amr-2022.
- [7] GOPALAKRISHNA, KARAMCHAND. "Zero trust and AI: A synergistic approach to next-generation cyber threat mitigation." *WORLD* 24, no. 3 (2024): 3374-3387.
- [8] Conradie, Niël Henk, and Saskia K. Nagel. "No agent in the machine: Being trustworthy and responsible about AI." *Philosophy & Technology* 37, no. 2 (2024): 72.
- [9] Freiman, Ori. "Making sense of the conceptual nonsense „trustworthy AI“." *AI and Ethics* 3, no. 4 (2023): 1351-1360.
- [10] Manish Hatwalne et al., "https://www.redpanda.com/blog/what-is-agentic-ai-introduction-autonomous-agents

- [11] Lahusen, Christian, Martino Maggetti, and Marija Slavkovik. "Trust, trustworthiness and AI governance." *Scientific Reports* 14, no. 1 (2024): 20752.
- [12] Pakina, Anil Kumar, Ashwin Sharma, and Mangesh Pujari. "AI Governance via Explainable Reinforcement Learning (XRL) for Adaptive Cyber Deception in Zero-Trust Networks." (2024).
- [13] Mintoo, Abdul Awal, Abu Saleh Muhammad Saimon, Mohammed Majid Bakhsh, and Marjina Akter. "NATIONAL RESILIENCE THROUGH AI-DRIVEN DATA ANALYTICS AND CYBERSECURITY FOR REALTIME CRISIS RESPONSE AND INFRASTRUCTURE PROTECTION." *American Journal of Scholarly Research and Innovation* 1, no. 01 (2022): 137-169.
- [14] ADABARA, IBRAHIM, Bashir Olaniyi Sadiq, Aliyu Nuhu Shuaibu, Yale Ibrahim Danjuma, and Venkateswarlu Maninti. "Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment." *F1000Research* 14 (2025): 905.
- [15] Rath, Manoj Kumar. "Adaptive Model Context Protocols for Trustworthy and Secure Agentic AI Systems." *International Journal of Computer Technology and Electronics Communication* 8, no. 5 (2025): 11444-11456.
- [16] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," *IEEE Access*, vol. 13, pp. 18912-18936, 2025.
- [17] Laurie Hughes et al., "AI Agents and Agentic Systems: A Multi-expert Analysis," *Journal of Computer Information Systems*, vol. 65, no. 4, pp. 489-517, 2025.
- [18] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee, "AI agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges," *arXiv preprint arXiv:2505.10468*, 2025
- [19] Venus Garg, "Designing the Mind: How Agentic Frameworks Are Shaping the Future of AI Behavior," *Journal of Computer Science and Technology Studies*, vol. 7, no. 5, pp. 182-193, 2025.
- [20] Soodeh Hosseini, and Hossein Seilani, "The Role of Agentic AI in Shaping a Smart Future: A Systematic Review," *Array*, vol. 26, 2025
- [21] San Murugesan, "The Rise of Agentic AI: Implications, Concerns, and The Path Forward," *IEEE Intelligent Systems*, vol. 40, no. 2, pp. 8-14, 2025.
- [22] Nalan Karunanayake, "Next-generation Agentic AI for Transforming Healthcare," *Informatics and Health*, vol. 2, no. 2, pp. 73-83, 2025.
- [23] Erik Miehl et al., "Agentic AI Needs a Systems Theory," *arXiv preprint arXiv:2503.00237*, 2025.
- [24] Bhandarwar, Nilesh. "Integrating Generative AI and Model Context Protocol (MCP) with Applied Machine Learning for Advanced Agentic AI Systems."
- [25] Al-Otaibi, Shaha, Sarra Ayouni, Nadeem Sarwar, Asma Irshad, and Faizan Ullah. "AI-driven security framework for medical sensor networks: enhancing privacy and trust in smart healthcare systems." *Cluster Computing* 28, no. 6 (2025): 408.
- [26] Elahi, Haroon, Nian Liu, Jiatong Chen, and Fengwei Zhang. "Towards a Secure Framework for Regulating Artificial Intelligence Systems." *IEEE Transactions on Dependable and Secure Computing* (2025).
- [27] Yang, Zhaohui, Wei Xu, Le Liang, Yuanhao Cui, Zhijin Qin, and Mérouane Debbah. "On privacy, security, and trustworthiness in distributed wireless large ai models." *Science China Information Sciences* 68, no. 7 (2025): 1-15.
- [28] Alnaim, Abdulrahman K. "Adaptive Zero Trust Policy Management Framework in 5G Networks." *Mathematics* 13, no. 9 (2025): 1501.
- [29] Rakhshanda, M., and A. Iqra. "AI-Enhanced Secure Communication Systems for Next-Generation IoT Networks: Protocols, Threat Mitigation, and Quantum Resilience." *Spectrum of Engineering Sciences* 3, no. 2 (2025): 925-941.
- [30] Chaganti, Krishna Chaitanya. "A Scalable, Lightweight AI-Driven Security Framework for IoT Ecosystems: Optimization and Game Theory Approaches." *IEEE Access* (2025).
- [31] Zheng, Wei, Tomley Anwlkom, and Ye Liu. "Trust-med6g: A secure and trustworthy medical data privacy protection mechanism in 6g-IoT." *IEEE Internet of Things Journal* (2025).

- [32] Kaliappan, Chandra Prabha, Kanmani Palaniappan, Devipriya Ananthavadivel, and Ushasukhanya Subramanian. "Advancing IoT security: a comprehensive AI-based trust framework for intrusion detection." *Peer-to-Peer Networking and Applications* 17, no. 5 (2024): 2737-2757.
- [33] Biswas, Shaikat, and Md Wahid Zaman Raj. "QUANTUM-RESISTANT CRYPTOGRAPHIC PROTOCOLS INTEGRATED WITH AI FOR SECURING CLOUD AND IOT ENVIRONMENTS." *International Journal of Business and Economics Insights* 4, no. 4 (2024): 60-90.
- [34] Sankaran, K. Sakthidasan, Tai-Hoon Kim, and P. N. Renjith. "An improved ai-based secure m-trust privacy protocol for medical internet of things in smart healthcare system." *IEEE Internet of Things Journal* 10, no. 21 (2023): 18477-18485.
- [35] Tlili, Fadhila, Samiha Ayed, and Lamia Chaari Fourati. "A new hybrid adaptive deep learning-based framework for UAVs faults and attacks detection." *IEEE Transactions on Services Computing* 16, no. 6 (2023): 4128-4139.
- [36] Aljanabi, Mohammad. "Safeguarding connected health: Leveraging trustworthy AI techniques to harden intrusion detection systems against data poisoning threats in IoMT environments." *Babylonian Journal of Internet of Things* 2023 (2023): 31-37.
- [37] Shivaramakrishna, D., and M. Nagaratna. "A novel hybrid cryptographic framework for secure data storage in cloud computing: Integrating AES-OTP and RSA with adaptive key management and Time-Limited access control." *Alexandria Engineering Journal* 84 (2023): 275-284.
- [38] Alkhabbas, Fahed, Mohammed Alsadi, Sadi Alawadi, Feras M. Awaysheh, Victor R. Kebande, and Mahyar T. Moghaddam. "Assert: A blockchain-based architectural approach for engineering secure self-adaptive iot systems." *Sensors* 22, no. 18 (2022): 6842.
- [39] Yanamala, Anil Kumar Yadav, and Srikanth Suryadevara. "Adaptive Middleware Framework for Context-Aware Pervasive Computing Environments." *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence* 13, no. 1 (2022): 35-57.
- [40] Shandilya, Shishir Kumar, Saket Upadhyay, Ajit Kumar, and Atulya K. Nagar. "AI-assisted Computer Network Operations testbed for Nature-Inspired Cyber Security based adaptive defense simulation and analysis." *Future Generation Computer Systems* 127 (2022): 297-308.