Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

A Systematic Review of Agentic AI for Threat Detection and Mitigation in 5G Networks

Kudzaishe Lawal Chizengwe[1], Belinda Ndlovu[2]

[1,2]Informatics and Analytics Department, National University of Science and Technology, Bulawayo, Zimbabwe

Received:
November 4, 2025
Revised:
December 12, 2025
Accepted:
January 1, 2026
Published:
February 14, 2026

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

DOI:
10.63158/journalisi.v8i1.1382

Abstract: Fifth-generation (5G) networks face escalating security challenges driven by decentralised architectures, stringent ultra-low-latency requirements, and rapidly evolving threat landscapes. Agentic Artificial Intelligence (agentic AI) autonomous systems that perceive network conditions, decide on countermeasures, and act in real time offers a promising route toward adaptive defence. This systematic review examines how agentic AI is being applied to detect and mitigate threats within 5G networks. Following PRISMA 2009 guidelines, four databases (IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect) were searched, yielding 22 eligible peer-reviewed studies published between 2020 and 2025, selected for explicit 5G relevance and empirical evaluation. The reviewed evidence clusters into four primary security areas: anomaly detection, DDoS mitigation, network slicing security, and intrusion detection. Across these domains, approaches based on federated learning, deep reinforcement learning, and multi-agent systems generally report stronger detection performance and/or more adaptive response behaviour than conventional, reactive baselines, while supporting privacy-preserving intelligence at the edge. However, key deployment barriers remain: 86% of studies rely on simulation-based validation, scalability beyond 100 nodes is insufficiently characterised, and reported coordination delays (120–180 ms) may conflict with 5G latency constraints in time-critical settings. To consolidate findings, this review proposes a Perception–Decision–Action–Feedback conceptual framework and highlights priorities for real-world validation and deployment-oriented evaluation.

Keywords: Agentic Artificial Intelligence; 5G Networks; Threat Detection; Autonomous Agents; Reinforcement Learning

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The fifth generation (5G) of mobile networks marks a substantial shift in wireless communications, enabling enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC) [1], [2]. At the same time, 5G is not merely a faster radio interface; it is a re-architected, software-intensive ecosystem. The adoption of software-defined networking (SDN), network functions virtualisation (NFV), multi-access edge computing (MEC), and network slicing has made network services more programmable and scalable, but it has also widened the attack surface across both the radio access network (RAN) and the core network [2], [3]. In practice, these capabilities introduce additional trust boundaries, increase inter-component dependencies, and create more operational complexity—conditions that adversaries can exploit through misconfigurations, compromised virtualised functions, slice-level abuse, or attacks that target the edge where decisions must be made rapidly.

In this study, *Agentic Artificial Intelligence* refers to autonomous, goal-directed computational agents capable of perceiving their environment, making independent decisions, executing adaptive actions, and continuously improving through feedback—either individually or cooperatively within multi-agent systems [4]–[6]. This framing differs from traditional reactive AI models that mainly identify patterns and raise alerts after suspicious behaviour is observed. Instead, agentic AI emphasises iterative perception–decision–action loops, enabling systems to adjust behaviours in response to changing conditions and adversarial tactics. Such characteristics are particularly relevant in 5G contexts where security controls increasingly operate at the edge and must respond under strict latency constraints, fluctuating traffic, and mobile user behaviour [7], [8].

However, conventional machine learning (ML)-based security solutions often struggle to keep pace with the dynamic characteristics of 5G networks [9]. Large-scale surveys in [10]–[12] suggest that while ML can improve detection performance, many approaches remain predominantly reactive, with limited capacity to adapt to evolving network states, mobility patterns, or adaptive adversaries. Further studies [13]–[15] reinforce this concern by showing that robustness is frequently compromised under realistic 5G conditions,

particularly at high mobility and at scale, where massive connectivity and heterogeneous devices amplify both uncertainty and attack opportunities.

Despite notable progress in intelligent security for 5G, the literature continues to exhibit several gaps that constrain the development of genuinely autonomous and adaptive defence mechanisms. First, the research landscape remains fragmented: studies in [10]–[12] largely focus on applying traditional ML techniques to 5G security problems, while [13], [16] explore federated learning but often within isolated or narrowly scoped scenarios. Across these bodies of work, there is limited evidence of holistic integration into end-to-end security systems that behave autonomously and adapt continuously across the full 5G stack. Second, there is a lack of agentic synthesis. Only a small subset of studies explicitly addresses how autonomous agents perceive, decide, act, and adapt under adverse 5G conditions—particularly when distributed decision-making, coordination across edge and core, and multi-domain visibility are required. Third, much of the evidence base remains simulation-driven, with limited empirical validation of deployment feasibility, scalability, and adversarial robustness under operational constraints. Finally, theoretical integration is often incomplete: existing work does not consistently unify agentic AI's specific role in autonomous 5G security. For example, [17] examines ML for ICT security, [11] covers AI for 5G security, [14] discusses wireless security, and [16] focuses on reinforcement learning for network security, yet these perspectives are rarely consolidated into a coherent framework centred on agentic autonomy, adaptive action, and distributed decision-making.

To address these gaps, this systematic literature review compiles and synthesises empirical studies that specifically investigate agentic AI approaches for detecting and mitigating threats within the 5G networking framework. The review examines how autonomous agents perceive network conditions, make decisions, execute actions, and incorporate feedback in 5G environments, with the goal of integrating empirical findings with pre-existing theoretical foundations to develop a holistic understanding of the current state of agentic AI for autonomous 5G security.

Accordingly, this research is guided by the following research questions:

1) What are the specific application areas and security threats that agentic AI addresses in 5G networks?

2) Which AI algorithms and architectural models are currently deployed for 5G threat detection and mitigation, and what are their comparative strengths and limitations?
3) What empirical, methodological, and operational gaps persist in the current literature regarding scalability, adversarial robustness, and real-world deployment feasibility?

2. METHODS

The study used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2009) framework. This ensures transparency, reproducibility and methodological rigour [18], [19]. To collectively minimise biases and support evidence-based synthesis, PRISMA divides the review into four sequential phases: Identification, Screening, Eligibility and Inclusion [20].

2.1. Identification

Four major digital libraries were selected for their comprehensive coverage of computer science and telecommunications literature: IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect. A master Boolean expression captured the three conceptual domains of this review: Agentic AI, Threat Detection and Mitigation, and 5G/B5G networks. Table 1 presents the complete search string structure with domain-specific keywords.

Table 1. Search String Components

Domain	Keywords	Rationale
Agentic AI	"agentic AI" OR "autonomous AI" OR "multi-agent system" OR "intelligent agent" OR "cognitive agent"	Captures various terminologies for autonomous systems exhibiting agency, goal-oriented behavior, and adaptive decision-making
Threat Detection & Mitigation	"threat mitigation" OR "attack mitigation" OR "security threat response" OR "threat detection" OR "intrusion detection"	Covers both proactive threat identification and reactive mitigation responses

Domain	Keywords	Rationale
5G/B5G Context	"5G" OR "5th generation" OR "fifth generation" OR "5G network" OR "beyond 5G" OR "B5G"	Ensures focus on fifth-generation and beyond networks, excluding legacy technologies (4G, Wi-Fi)

2.2. Complete Search String:

The following search string was developed to identify studies at the intersection of agentic (autonomous) AI approaches, security response/mitigation, and 5G networking:

(("agentic AI" OR "autonomous AI" OR "multi-agent system" OR "intelligent agent" OR "cognitive agent") AND ("threat mitigation" OR "attack mitigation" OR "security threat response") AND ("5G" OR "5th generation" OR "fifth generation" OR "5G network")).

This query was tailored to meet the functionality and syntax requirements of each database. For IEEE Xplore and the ACM Digital Library, the search was executed across the title, abstract, and keyword fields to improve precision and reduce irrelevant retrieval. For SpringerLink and ScienceDirect, the complete search string was applied, with results constrained to the subject areas of computer science and engineering to maintain topical relevance. Across all databases, the search period was restricted to 2020–2025, reflecting the timeframe associated with the rollout and early deployment of 5G networks. Additional filters were applied to include only English-language, peer-reviewed journal articles and conference papers.

To enhance methodological rigour, the search strategy underwent PRESS (Peer Review of Electronic Search Strategies) evaluation [20] and was further reviewed by the supervising lecturer prior to execution. The final searches were conducted between 24 and 31 October 2025. In total, 792 records were retrieved: IEEE Xplore (28), ACM Digital Library (373), SpringerLink (156), and ScienceDirect (235). Following automated and manual deduplication in Mendeley, 781 unique records remained for screening.

2.3. Inclusion and Exclusion Criteria

Table 2 summarises the inclusion and exclusion criteria applied during the screening stage. These criteria were designed to ensure that only studies directly relevant to

agentic (agent-based) AI for threat detection and mitigation in 5G (and beyond-5G) environments were retained, while excluding works outside scope or lacking agentic autonomy.

Table 2. Inclusion and Exclusion Criteria

Inclusion Criteria (Studies Were Included If They …)	Exclusion Criteria (Studies Were Excluded If They …)
1. Addressed Agentic / Agent-Based Ai Within 5G OR B5G Networks	1. Focused On Non-5g Technologies (E.G., 4g, Wi-Fi)
2. Published Between 2020 and 2025	2. Applied Generic ML Models Without Agentic Features
3. Reported Empirical Or Conceptual Work On Threat Detection/Mitigation	3. All Papers That Were Non-Empirical (Position Papers, Editorials)
4. Peer-Reviewed Journals Or Conference Proceedings	4. Grey Literature Or Non-English Sources

The screening rules prioritised studies that explicitly incorporated agentic characteristics—such as autonomous decision-making, environmental sensing, goal-directed behaviour, adaptive action selection, or cooperative multi-agent coordination—within 5G security contexts. This emphasis ensured that included studies went beyond conventional predictive or reactive ML models and instead reflected the capacity for independent, action-oriented security responses. The timeframe of 2020–2025 was selected because it aligns with the period of broad 5G rollout and the emergence of related security research. Finally, restricting the corpus to English-language, peer-reviewed journal and conference publications helped maintain research quality and methodological reliability by focusing on work that has undergone formal scholarly review.

2.4. Screening and Eligibility

The 781 unique records were screened at the title and abstract level against the pre-defined inclusion and exclusion criteria. Studies were excluded at this stage for several recurring reasons, including the absence of agentic or agent-driven AI features (e.g., no autonomous perception–decision–action capability), lack of a 5G or beyond-5G

networking context, publication types outside the scope of primary research (such as commentaries, editorials, or secondary reviews), and sources that did not meet the language and publication requirements (e.g., non-English papers or unpublished/grey literature such as technical reports). Following this initial screening, 22 articles were retained for full-text assessment.

To strengthen screening consistency and minimise selection bias, two independent reviewers assessed each title and abstract. Inter-rater agreement was quantified using Cohen's Kappa coefficient, yielding a value of 0.87, which indicates a high level of reliability between reviewers.

All 22 studies assessed at the full-text stage met the eligibility requirements. Specifically, each study addressed agentic AI within a 5G (or beyond-5G) setting, reported empirical security-related work on threat detection and/or mitigation, and was published in peer-reviewed journal articles or conference proceedings between 2020 and 2025.

2.5. Quality Assessment

Every study was checked with a 5-level scale borrowed from [21] alongside [22]. This assessment evaluated methodological rigor, relevance to the research questions, and the quality of empirical validation rather than serving as an exclusion mechanism. Table 3 presents the quality assessment dimensions and scoring criteria.

Table 3. Quality Assessment Criteria

Dimension	Scoring Criteria	Weight Rationale
AGENTIC AI RELEVANCE	1.0 = Strong demonstration of agentic characteristics (autonomy, adaptivity, goal-orientation) 0.5 = Partial agentic features (e.g., adaptive but not autonomous)	Ensures alignment with review, focusing on truly agentic systems
5G/B5G CONTEXT	1.0 = Explicitly designed for 5G/B5G networks with 5G-specific features	Verifies relevance to 5G-specific security challenges

Dimension	Scoring Criteria	Weight Rationale
	0.5 = Generic approach applied to 5G context	
METHODOLOGICAL CLARITY	1.0 = Fully transparent methodology enabling reproducibility 0.8 = Mostly transparent with minor ambiguities 0.6 = Partially clear with significant gaps	Assesses reproducibility and methodological rigor
THREAT FOCUS	1.0 = Threat detection/mitigation as primary objective 0.5 = Threat handling as secondary consideration	Ensures centrality of security focus to review scope
EMPIRICAL VALIDATION	1.0 = Real-world deployment or operational testbed 0.8 = Network emulation with realistic parameters 0.6 = Simulation-based validation	Evaluates empirical evidence quality and generalizability

Description: Scoring interpretation is the maximum score is 5.0. The minimum inclusion threshold is 3.5, as scores below that indicate insufficient methodological rigour, unclear agentic characteristics, or minimal empirical validation.

Some studies scored low on validation, indicating a disconnect between simulated settings and real-world use, as seen in RQ3. Still, those results helped shape our interpretation of the overall picture. A few had weaker design descriptions, which made replication more difficult; this was flagged per dimension three ratings. Even so, every paper met the minimum threshold, with scores above 3.5. Scores ran from 3.7 up to 4.8, averaging 4.2 with little spread (SD 0.3). This suggests that most of the work was well established and clearly focused on agentic AI. The quality assessment scores for all 22 studies are presented in Table 4.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 4. Quality assessment scores

Paper	Agentic AI	5G Context	Methodology	Threat Focus	Validation	Total
P1	1	1	1	1	0.8	4.8
P2	1	1	0.8	1	0.6	4.4
P3	1	1	0.8	1	0.6	4.4
P4	1	1	1	1	0.6	4.6
P5	1	1	1	1	0.6	4.6
P6	1	1	0.8	1	0.6	4.4
P7	1	1	0.6	0.5	0.6	3.7
P8	1	1	1	1	0.6	4.6
P9	1	1	0.8	0.5	0.6	3.9
P10	1	1	0.8	0.5	0.6	3.9
P11	1	1	0.8	0.5	0.6	3.9
P12	1	1	1	0.5	0.6	4.1
P13	1	1	1	0.5	0.6	4.1
P14	1	1	0.8	1	0.6	4.4
P15	1	1	1	0.5	0.6	4.1
P16	1	1	0.8	0.5	0.6	3.9
P17	1	1	0.8	1	0.8	4.6
P18	1	1	0.8	1	0.6	4.3
P19	1	1	0.6	0.6	0.6	3.8
P20	1	1	1	1	0.8	4.8
P21	1	1	0.8	1	0.6	4.3
P22	1	1	1	1	0.6	4.6

2.6. Inclusion

Following completion of all PRISMA screening stages, 22 peer-reviewed studies were retained for inclusion in the review and subsequently subjected to qualitative and quantitative synthesis. Figure 1 presents the PRISMA flow diagram, detailing the number of records identified, screened, assessed for eligibility, and included, as well as the reasons for exclusion at each stage.

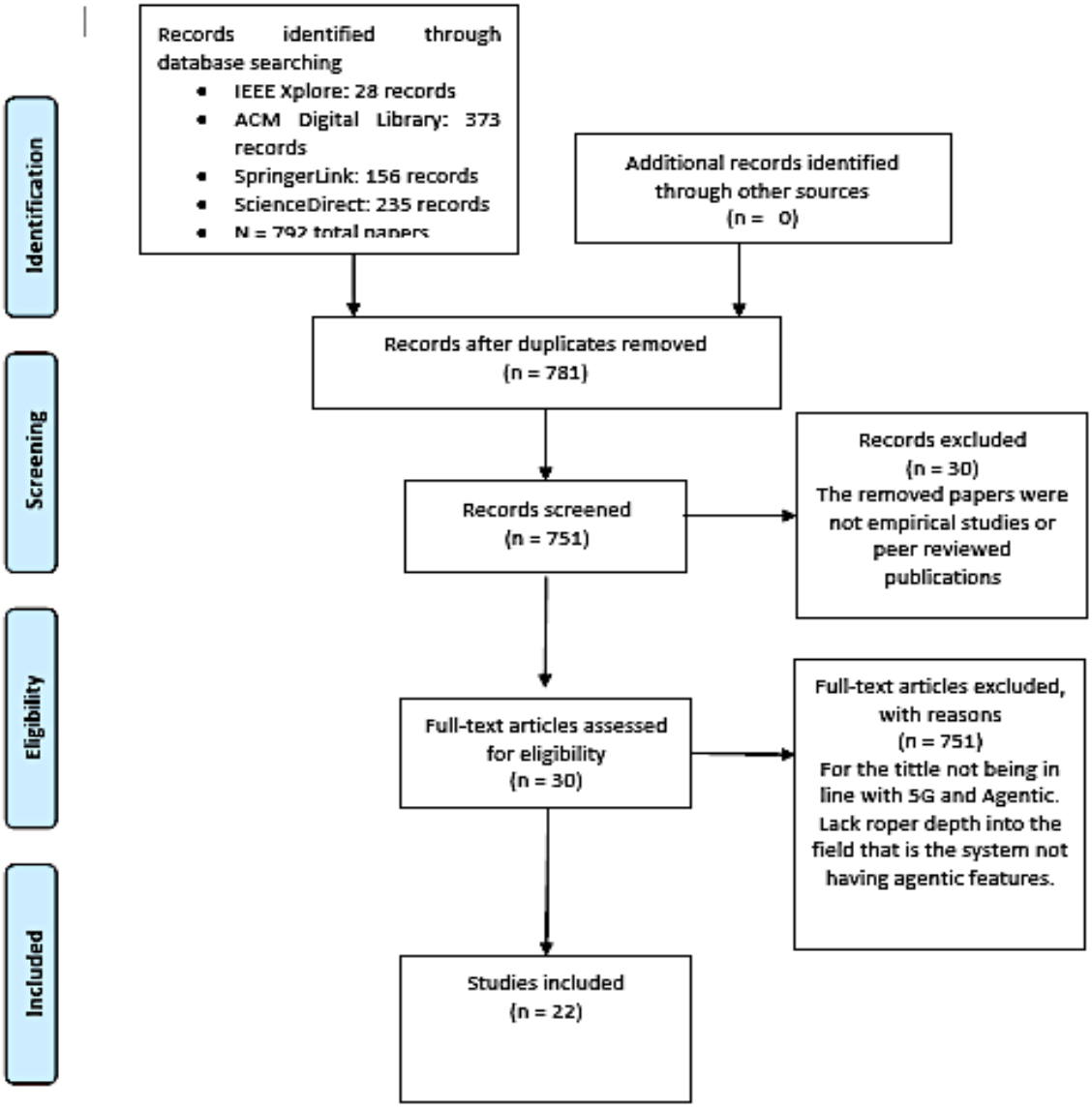

Figure 1. Prisma Flow Diagram

3. RESULTS AND DISCUSSION

The 22 included studies collectively demonstrate a growing—yet still early-stage—research landscape on the use of agentic AI to detect and mitigate security threats in 5G and beyond-5G (B5G) networks. Across the corpus, agentic AI is typically operationalised through autonomous decision-making loops (perception–decision–action), often implemented using reinforcement learning, federated learning, or multi-agent coordination. While the studies vary widely in scope and maturity, they consistently reflect a common motivation: conventional, largely reactive security analytics struggle to

cope with the scale, mobility, heterogeneity, and programmability that characterise modern 5G architectures.

3.1. Summary of Included Studies

The included studies exhibit substantial diversity in publication venue, methodological design, and the specific agentic capabilities emphasised. Some contributions prioritise autonomy and goal-directed behaviour for slice security and orchestration, while others focus on environmental perception and adaptive action in fast-changing threat scenarios such as DDoS, jamming, or encrypted traffic abuse. Table 3 synthesises each paper's core attributes in a structured form, summarising the agentic features reported (e.g., autonomy, adaptivity, perception, and multi-agent behaviour), the primary security focus, the underlying AI model or architecture, the associated 5G/B5G context, and the key limitations that constrain generalisability or real-world applicability.

Table 3. Included studies on Agentic AI for threat detection in 5G Networks

ID	Ref	Agentic features	Security focus	AI model / architecture	5G context	Key contribution (short)	Main limitation
P1	[23]	Au, Go, Ac	Anomaly detection	Federated Learning + policy engine	B5G; distributed FL	Scalable orchestration framework for FL-based anomaly detection	Scalability evaluation limited
P2	[24]	Au, Go	Slice isolation; anomaly detection	FL + DNN	5G network slicing	Secure slicing architecture using FL-enabled detection	Simulation-only
P3	[25]	Au, Pe	DDoS mitigation	Deep Q-Network (DQN)	SDN in 5G	RL agent enables adaptive DDoS response actions	Limited attack coverage
P4	[26]	Au, Go, Ac	Intrusion detection	Dueling DQN	5G/B5G wireless	DRL-based IDS tailored for wireless conditions	Limited scalability; single-agent focus
P5	[27]	Au, Pe, Ac	Cyberattack detection (O-RAN)	FL with distributed agents	B5G; O-RAN	FL framework supporting B5G/O-RAN security analytics	Simulation-based; no deployment validation

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

ID	Ref	Agentic features	Security focus	AI model / architecture	5G context	Key contribution (short)	Main limitation
P6	[28]	Au, Go, Ac	IIoT intrusion	FL + GRU-based DRL	IIoT over 5G	Combines DRL + FL for federated IIoT intrusion detection	IIoT-specific; limited generalisability/attack breadth
P7	[29]	Au, Pe, Ac	Jamming mitigation	Federated DRL (FDRL)	5G HetNets	Federated DRL mitigates dynamic jamming in HetNets	Federated coordination complexity underexplored
P8	[30]	Au, Go	Slice security; isolation	Asynchronous FL	5G slicing	Async FL improves responsiveness & privacy of slice policies	Simulation-only
P9	[31]	Au, Ac	Slicing security	CNN-based DL framework	5G/B5G slicing	DL framework to protect slicing mechanisms	No agent coordination; limited online learning
P10	[32]	Au, Go, Pe, Ac, MA	QoS-driven slice optimisation (resource protection)	Federated MARL (DDPG + FedAvg)	LoRa/5G slicing	Multi-agent Federated RL optimises QoS/isolation	LoRa setting limits direct mapping; security is secondary
P11	[33]	Au, Go, Ac	Slice optimisation; load balancing	Federated LSTM	5G slicing + RAN	Fed-LSTM improves slice load prediction for management	Limited attack modelling; optimisation-centric
P12	[34]	Au, Pe, Go	Slice isolation; anomaly detection	FL framework	5G slicing security	FL-based anomaly detection to strengthen slice security	High-level; limited deployment details
P13	[35]	Au, Go, Ac, MA	Resource allocation; QoS assurance	MARL	5G V2X; vehicular slicing	MARL for resource allocation (security treated implicitly)	Security treatment minimal

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

ID	Ref	Agentic features	Security focus	AI model / architecture	5G context	Key contribution (short)	Main limitation
P14	[36]	Au, Go	DDoS / flash events	MARL (MADDPG)	5G-enabled SDN-IoT	MADDPG applied for attack detection/response in SDN-IoT	Scalability unclear
P15	[37]	Au, Go, Ac	Cyberattack detection	P2P Federated Learning	B5G protection (distributed)	Peer-to-peer FL for distributed protection	Simulation-only
P16	[38]	Au, Go, Ac	Slice security	Two-layer FL + mean-field game	5G slicing	FL + game theory for adversarial slice protection	Controlled simulations; limited large-scale validation
P17	[39]	Au, Ac, MA	General cyber defence	MARL (DDPG/TD3 variants)	Cyber-range (CAGE), not 5G-specific	Shows MARL potential vs coordinated attacks	Highly simulation-centric; not 5G-grounded
P18	[40]	Au, Go, Ac	Honeypot deployment; deception	RL (Q-learning)	Ultra-dense B5G	Strategic honeypot placement to maximise attacker engagement	Simulation-only; limited real-world testing
P19	[41]	Au, Pe	IoT intrusion detection (roadmap)	Conceptual DL/ML framework	IoT–5G convergence	Roadmap unifying AI intrusion techniques for IoT–5G	Conceptual only; no implemented model
P20	[42]	Au, Pe, OL	Encrypted DNS (DoH) threat detection	DFL + incremental learning (SVM/RF/LR/DT)	B5G privacy-preserving detection	Detects DoH tunnelling while preserving privacy	Simulation-only; possible real-time overhead
P21	[43]	Au, Ac	DDoS mitigation + QoS	DRL (DQN)	5G SDN-enabled networks	Balances DDoS defence with QoS for benign users	Simulated traffic; limited hybrid-attack evaluation

ID	Ref	Agentic features	Security focus	AI model / architecture	5G context	Key contribution (short)	Main limitation
P22	[44]	Au, Ac, OL	DDoS mitigation (V2X)	DQN	5G V2X	RL defence for vehicular edge DDoS mitigation	Synthetic data; needs real-world validation

Taken together, the studies reflect four dominant technical paradigms: federated learning (FL) for privacy-preserving and distributed security analytics (e.g., slice protection and anomaly detection), deep reinforcement learning (DRL) for adaptive response under changing attack conditions (e.g., DDoS or jamming), multi-agent reinforcement learning (MARL) for coordinated decision-making in distributed environments (e.g., SDN-IoT or V2X settings), and hybrid approaches that integrate FL with deep learning, DRL, or game-theoretic formulations to strengthen robustness and strategic behaviour. Notably, although these paradigms align well with the autonomy and adaptivity requirements of 5G security, most evaluations remain simulation-heavy, with comparatively limited evidence of large-scale deployment feasibility, cross-domain interoperability, and adversarial robustness under real operational constraints.

3.2. Descriptive Analysis of Reviewed Studies

This section presents a descriptive analysis of the 22 included studies, focusing on publication trends and how research attention has evolved over time.

1) Temporal Distribution

Figure 2 illustrates the year-by-year publication distribution of the included papers and highlights a clear upward trend in scholarly interest in agentic AI for 5G security. The temporal pattern suggests that the topic has moved from an emerging niche to a more established research direction within a relatively short period, aligning with the broader maturation of 5G deployments and the parallel shift toward more autonomous, software-driven network operations.

Figure 2 derivation: publication counts extracted from 22 studies meeting inclusion criteria, showing year-by-year distribution from 2020-2025 database searches. Thus, the first one appeared in 2020, right around the time 5G started rolling out commercially. After that, the number of publications began to increase: 2020 had 2, then 1 in 2021,

followed by 4 in 2022, 5 in 2023, peaked at 7 in 2024, and finally 3 in 2025. That peak in 2024 suggests researchers are now more involved in applying Agentic AI to security systems for next-generation networks.

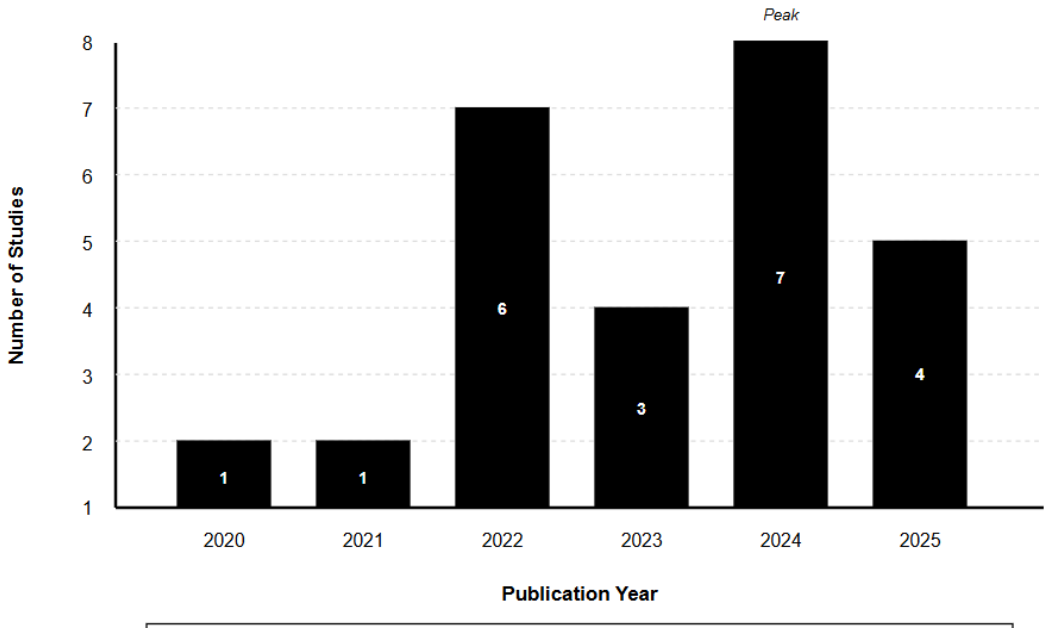

Figure 2. Number of Studies by Publication Year

2) AI Models and Architectures

Figure 3 summarises the dominant AI paradigms adopted across the reviewed studies and groups them into four categories: Federated Learning (FL), Deep Reinforcement Learning (DRL), Multi-Agent Reinforcement Learning (MARL), and hybrid architectures that combine FL with complementary techniques such as game theory or deep learning. This categorisation provides a high-level view of how researchers are operationalising "agentic" behaviour in 5G security settings—either through distributed training and privacy-preserving collaboration (FL), adaptive sequential decision-making (DRL), coordinated multi-entity control (MARL), or integrated frameworks that aim to balance multiple requirements such as robustness, privacy, and strategic response.

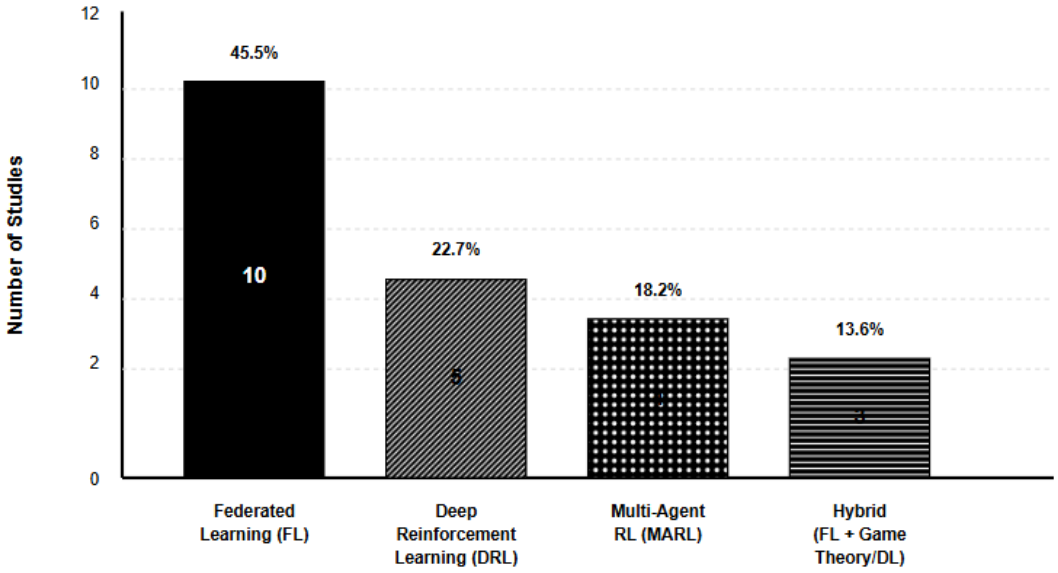

Figure 3. Number of AI approaches in reviewed studies

Figure 3 derivation: Studies were classified according to the primary AI paradigm described in their methodology sections. Where a study employed multiple approaches (e.g., FL combined with DRL or game-theoretic components), it was counted in each relevant category. The distribution indicates a strong preference for distributed and decentralised learning paradigms, reflecting the architectural realities of 5G and B5G environments where data and control are often dispersed across edge nodes, slices, and heterogeneous domains. Federated Learning emerged as the most frequently adopted approach, appearing in 10 studies, suggesting that privacy preservation and distributed model training are central design priorities for many 5G security solutions. Deep Reinforcement Learning followed with 5 studies, highlighting the relevance of adaptive decision-making for dynamic threat response (e.g., DDoS mitigation or policy selection). Multi-Agent Reinforcement Learning was observed in 4 studies, reinforcing the need for coordinated autonomy in distributed settings such as SDN-enabled domains or vehicular networks. Finally, 3 studies used hybrid architectures (e.g., FL integrated with game theory or deep learning), reflecting efforts to overcome the limitations of single-paradigm approaches and further underscoring the overall shift toward decentralised intelligence rather than purely centralised security control.

3) Security Focus Area

Figure 4 shows the percentages of the area of focus in the papers regarding security. Network slicing, 32% of the studies focused on interslice isolation, resource allocation fairness and slice-specific throughput detection. Intrusion detection was prioritized in 27% of the studies, focusing on high-throughput 5G traffic, while the other 18% focused on DDoS mitigation mechanisms to preserve quality of service. Anomaly detection and specialised applications such as jamming and encrypted traffic detection accounted for 14% and 9%, respectively. The strong focus on network slicing intrusion detection reflects 5G-specific attack surfaces, with features such as slice isolation and disaggregated RAN interfaces absent in legacy networks. Figure 4 derivation: Secure focus extracted from study objective and experimental scenarios, classified by primary application area.

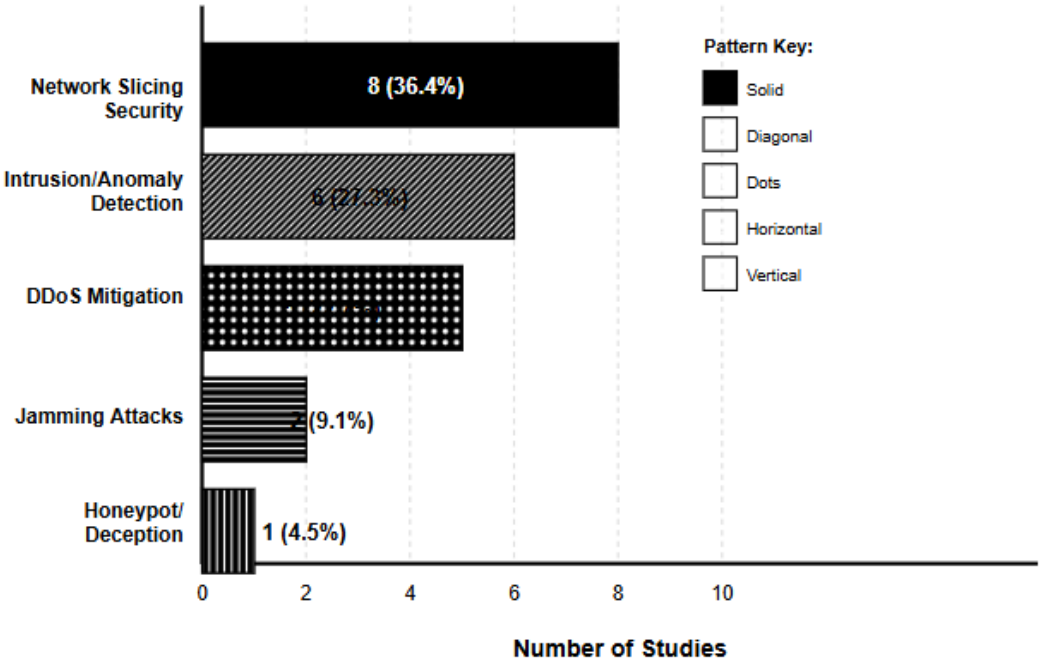

Figure 4. Security focus distribution

4) Agentic Characteristics

Figure 5 shows that agentic characteristics are widely distributed across the studies. The selection criteria ensured that all selected studies had the core characteristics of Agentic AI. Autonomy was the principle of all the papers, so it was universal with 100 %. Adaptivity was a very close second, as all the agents were able to modify defences based on environmental feedback. Goal orientation was explicitly defined in [26] as maximising

detection accuracy and minimising false positives. [25] shows how environmental perception varied from simple traffic volume monitoring to multidimensional state representations incorporating slice health and historical attack patterns [32]. Two studies [31], [40] demonstrated weaker environmental perception, relying primarily on pre-trained models with limited runtime adaptation, though they still qualify as agentic due to their autonomous decision-making capabilities. Figure 5 derivation: Studies evaluated against four agentic characteristics identified from methodology description and architectural specifications.

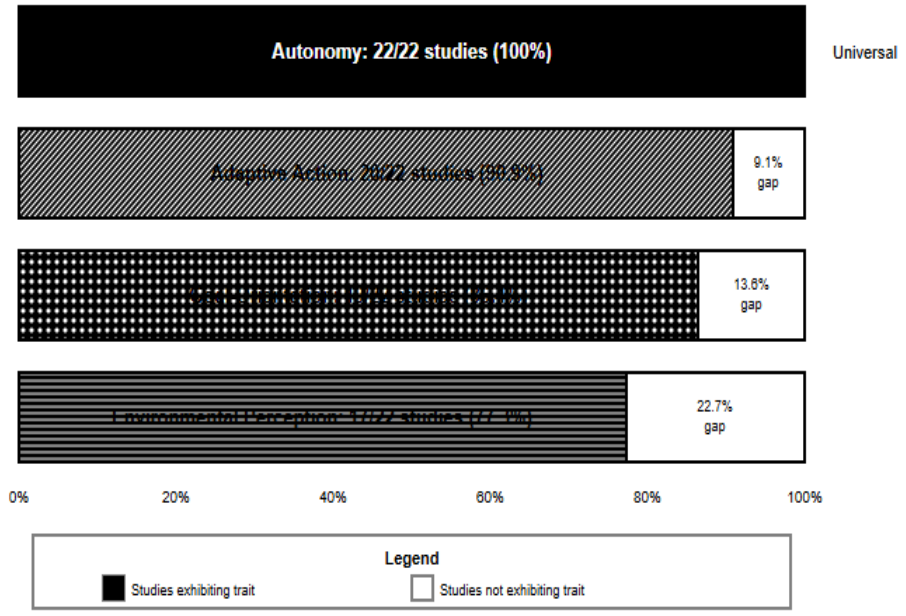

Figure 5. Prevalence of Agentic Features Across Studies

5) Deployment Contexts

Figure 6 shows the different deployment contexts identified across the studies. For the deployment and network layers, the core 5G architecture accounted for 68.2%, with specific attention to beyond 5G and Open RAN at 31.8%, reflecting anticipation of future network evolution. Diversity in AI applications across specialised fields such as vehicular V2X networks covered by [35], [44] and ultra-dense deployment [38] demonstrates the agentic AI applicability from automation to intelligent transport systems. Figure 6 derivation: Deployment context extracted from experimental setups and target infrastructure specifications.

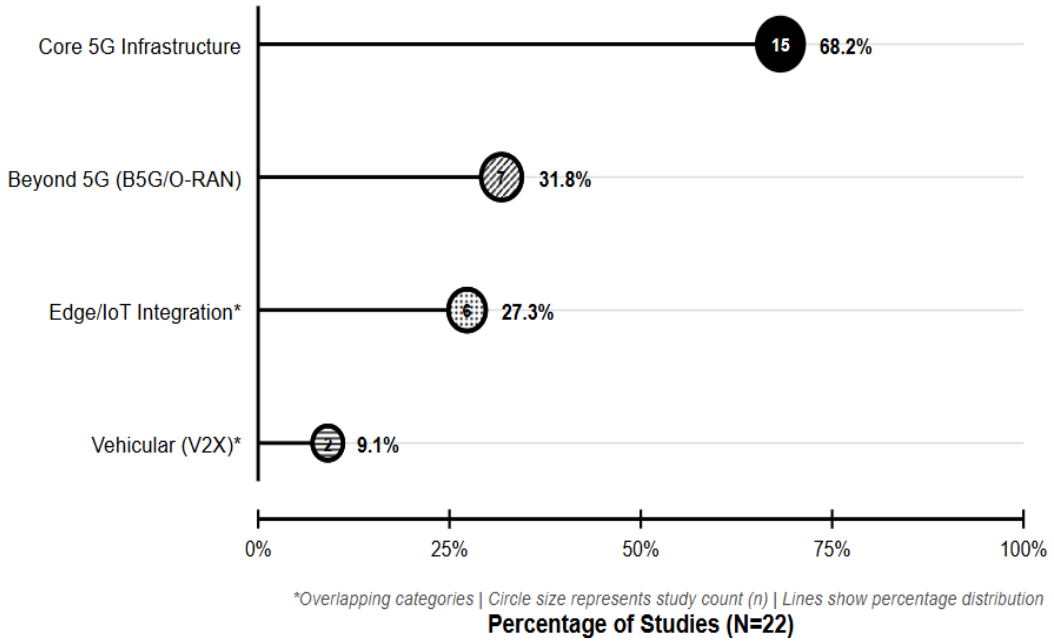

Figure 6. Deployment Context network layers

6) Evaluation Methodologies

Figure 7 shows the different percentages of the paper distribution across the studies, highlighting a critical validation gap. 86.4% of the papers are simulations, with 9.1% from network Emulation and 4.5% with limited real-world testing. This large percentage highlights the critical gap between theoretical and operational deployment, which represents a fundamental limitation affecting generality. The near-complete absence of operational validation fundamentally constrains claims about real-world performance. Figure 7 derivation: Evaluation methodology classified from experimental design sections: simulations (synthetic traffic, controlled environments), emulations (software network replication), and real-world (operational infrastructure deployment).

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

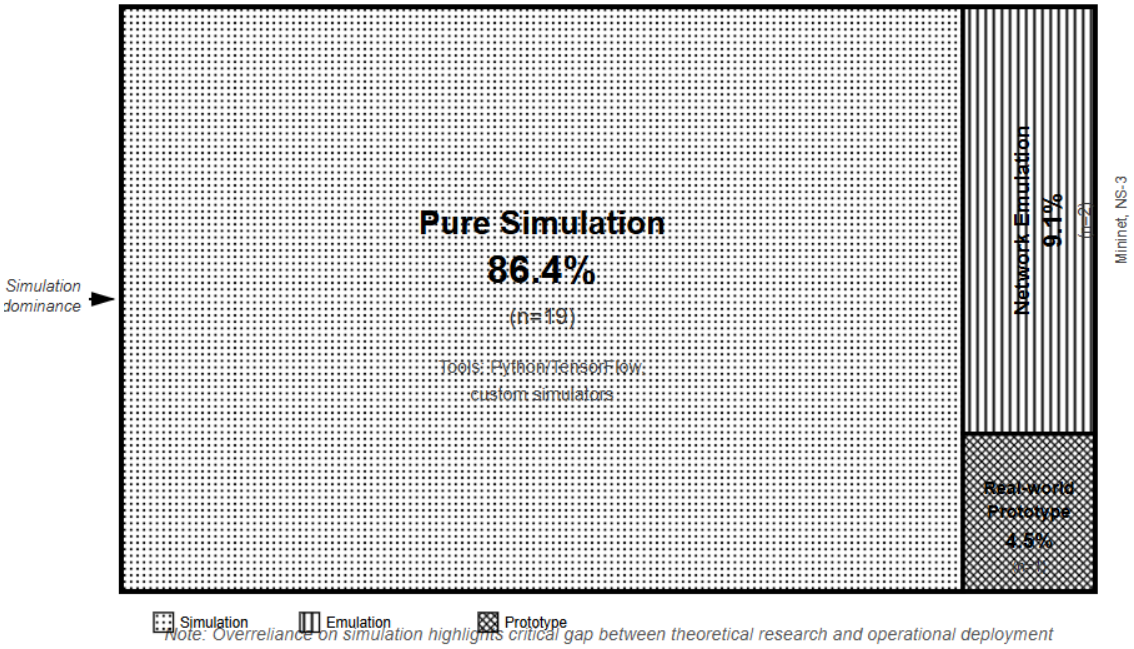

Figure 7. Evaluation Methodologies

The next section discusses the research questions and how they align with the review's findings.

3.3. What are the specific application areas and security threats that agentic AI addresses in 5G networks?

According to the studies, agentic AI was applied in four dominant areas: (1) network intrusion and anomaly detection, (2) DDoS detection and mitigation, (3) network slicing security, and (4) physical layer and specialised threats. A critical analysis of the papers reveals inconsistencies between performance reports and real-world feasibility.

1) Intrusion and Anomaly Detection

Anomaly detection has been widely explored across several studies, including [23], [24], [34], which identify unusual network behaviours. At the same time [26]–[28], [41], [42] are intrusion detection systems that address unauthorised access. Despite being tested in simulated, controlled, and simplified threat models, it still achieves detection accuracies of 94.8% to 97.3%. The critical limitations of these studies include the use of legacy datasets, such as using a 2009 benchmark predicting virtualised network function, slice isolation violation and Open RAN exploitation. DRL-based IDS [26] achieved 96.7% accuracy on the NSL-KDD dataset. [27] made use of federated learning for OpenRAN to

address disaggregated RAN interfaces but lacked O-RAN Alliance validation. Furthermore, [28] used a GRU-based federated DRL approach to achieve 97.2% accuracy in intrusion detection, but acknowledged the need for simulated traffic evaluations. Despite all this, no study looked at the adversarial machine learning attacks as a whole, that is, adversarial examples, model inversions and membership inference. Modern attacks leverage crafted inputs to exploit learned detection policies, yet this critical dimension remains untested.

Due to these limitations, certain practical situations pose challenges, such as an IDS trained on NSL-KDD failing to recognise attacks targeting slice orchestration APIs or containerised network functions. This is because these types of attack patterns did not exist in pre-5G datasets. Another critical issue is how a false positive in an industrial 5G system could trigger an emergency shutdown. This would halt production lines, resulting in millions of losses, yet no study has validated false-positive rates under operational industrial conditions.

2) DDoS Mitigation

DDoS mitigation represented the second most addressed area, with four studies [25], [36], [43], [44] focused on autonomous attack suppression. Reinforcement learning agents demonstrated impressive reductions in attack impact (89% to 96.7%). However, in the studies, simplified threat models dominated single-vector volumetric flooding with predictable signatures, ignoring modern vector botnets that exploit volumetric, protocol, and application-layer vulnerabilities simultaneously.

The approach in [25], while based on DRL, achieved sub-50ms decision latency in SDN environments but considered only 100-node networks with synthetic attack traffic. Scalability to ultra-large 5G deployments is unexamined. Similarly, the vehicular V2X DDoS mitigation in [44] reported response times of 8.3ms, meeting ultra-low latency requirements in controlled environments without vehicle mobility, handover disruptions, or channel degradation. Multi-agent DDoS detection [36] achieved 34% latency improvement but incurred 120-180ms of coordination overhead, incompatible with sub-10ms URLLC requirements. It is worth noting, however, that none of these studies investigated reward signal manipulation, in which adversaries craft traffic to train the agent to identify malicious traffic as legitimate. All these limitations, however, become a problem when applied correctly, that is, in a 5G smart city, a sophisticated botnet could

probe the DRL defence to learn the pattern and mimic a legitimate emergency service. This means the RL agent then learns and classifies these attacks as emergency service while maintaining apparent QoS metrics.

3) Network Slicing Security

Network slicing security was another application area specific to 5G, where agentic systems addressed problems such as isolation violations, resource exhaustion, and inter-slice attacks [24], [30], [31], [34], [35], [38]. This is a highly critical application domain, since slice isolation underpins multi-tenancy in 5G and is crucial for regulatory compliance in industries that demand strict data separation. However, overly soiled models contradict operational reality where slices share physical infrastructure, control channels and orchestrate functions.

The federated learning approach to slice security in [24] achieved 95.8% accuracy in detecting isolation violations. It, however, assumed independent slice action, which is unrealistic for actual 5G. The game-theoretic strategic defence in [38] was an elaborate strategy that combined two-layer federated learning with mean-field game theory and attained a 43% improvement in slice compromise prevention against adaptive adversaries. Unfortunately, it required a 2.7-second equilibrium computation, which is unacceptable for real-time breach containment. Scalability validation limited to 100 nodes raises questions about networks that may host hundreds of slices. Despite studies examining interslice security, none evaluated lateral movement scenarios in which compromising one slice enables attacks on adjacent slices via shared infrastructure. Furthermore, they seem to lack privacy-preserving decision mechanisms and transparent governance protocols, which are essential for cross-tenant collaboration [45][46].

4) Physical-Layer and Specialised Threats

There are limited physical-layer threats, including jamming, eavesdropping, and signal manipulation. One study addressed jamming [29] using federated DRL. It achieved a success rate of 91.7%, although it modeled jammers as static adversaries with predictable patterns. This totally disregarded reactive jamming that adaptively adjusts interference. Furthermore, it did manage to raise questions about reactive jamming adjusting within milliseconds; anti-jamming coordination incurred 340ms overhead.

Furthermore, the study did involve other socialised applications, such as strategic deployment [40], which unfortunately remained simulation-centric and did not address effects in resource-constrained environments, including scalability issues. [42] evaluated encrypted traffic detection, but unfortunately, it was evaluated against synthetic traffic rather than operational captures using realistic encryption protocols. QoS optimisation by [32], [35], [43] omitted simultaneous active threat mitigation, leaving questions about performance under combined constraints. Table 4 shows the summary of AI models vs the threat types and performance.

Table 4: AI models vs Threat Types and Performance

AI Model	DDos	Intrusion	Slicing	Anomaly	Physical	Performance
DQN	[25], [43], [44]	[26]	-	-	-	89-96.8% accuracy/impact
MARL	[36]	-	[35]	-	-	34% latency improvement
Federated Learning	-	[27], [28]	[24], [30], [31], [34], [35], [38]	[23]	[29], [42]	94.8-97.3% accuracy
Hybrid	-	[28]	[32],[38]	-	[29]	+8.6-43% over single-paradigm

3.4. Which AI algorithms and architectural models are currently deployed for 5G threat detection and mitigation, and what are their comparative strengths and limitations?

The reviewed studies addressed diverse algorithmic and architectural combinations with different trade-offs between performance, scalability, privacy and responsiveness.

1) Single-Agent Approaches using Deep Q-Networks

Single-agent architectures [25], [26], [40], [43] relied on placing agents at centralised control points, offering complete visibility but consequently introducing critical vulnerabilities. The algorithm has a variety of strengths, such as DQN demonstrating effectiveness across SDN-based DDoS mitigation, wireless intrusion detection, QoS-preserving DDoS mitigation and V2X security. It handles high-dimensional state spaces

and generalises to unseen attacks with 83-91% effectiveness, while also learning from experience. DQN converges within 150-200 episodes, yielding high detection accuracy and 89-94% reductions in attack impact. However, despite all these, it does have weaknesses, such as being unacceptable for production areas with daily merging variants. The only way to learn is to retrain the model completely. Vulnerability to diverse manipulation, which enables attacks to craft traffic exploiting learned Q value estimates.

The SDN-based DRL mitigation in [25], for example, demonstrated sub-50ms decision latency for 100-node networks; scalability beyond this modest scale was not validated. Centralised architectures introduce fundamental scalability limits by creating a processing bottleneck in which a single agent must make all security decisions for entire networks. Single points of failure are more critical concerns. An example of these would be a DQN controller managing 5G core protection that has to handle security decisions across all network slices simultaneously. During sudden traffic spikes or synchronised assaults on multiple slices, that sole system gets overwhelmed. When malicious data skewing its decision scores is used to hack the defence setup, the whole defence setup collapses, putting every user at risk.

2) Multi-Agent Systems Using MARL

The distributed agents coordinated across network elements [32], [35], [36], [39], offering resilience and eliminating single points of failure but introducing coordination complexity. MARL demonstrates strong capabilities, achieving 34% latency reduction [36] and graceful degradation [35] in distributed decision-making. This does eliminate centralised bottlenecks. The agents, however, slowed down their communication, taking 120 to 180 ms, which is beyond the 10 ms speed needed. Training complexity also increases when dealing with multiple-agent environments, as each environment evolves as the others learn. MARL setup introduces new limitations, such as scaling tested up to just 50 agents, so how it works when the number of agents increases. Byzantine vulnerability emerged as a critical concern bad actors could mess things up by messing up their strategies, but none of the studies tried using defences against such. Crashes were treated like chance events, not smart attacks aimed at causing maximum chaos. Running this setup takes way more effort than handling one agent alone, needing teamwork across machines, aligning decisions, plus tracking inter-agent actions that pop out of nowhere. How this would work is in a 5G network spread across 200 edge nodes

using MARL agents, a skilled hacker might take over 10 key units placed where data comes together. Instead of cooperating, these Byzantine agents send fake strategy changes when syncing, slowly weakening overall threat spotting ability but also creating specific gaps on purpose. Because sync delays last between 120 and 180 milliseconds, the system can't react fast enough to breaches that target those created weaknesses.

3) Federated Learning Systems

Federated multi-agent systems [23], [24], [27]–[29], [42] integrated distributed decision-making with privacy-preserving collaborative learning, aligning naturally with 5G multi-tenant architecture. Federated learning systems have strengths, such as lower privacy risks since each group trains locally and then combines results later, with detection accuracies of 95.2%-97.3%. Knowledge moves between groups, yet original data stays hidden at all times. Decentralised operation allows real-time local security decisions while periodically aggregating learned models; however, it has weaknesses, such as model merging that can take a few minutes or longer, which clashes with how quickly threats need responses. Instead of speeding things up, privacy tweaks slow down training by 15-35%, making it harder to keep both data safe and responses quick, especially when there is no clear way to set those controls. Because individual systems update faster than the main model integrates them, devices may run on outdated versions when attacks evolve rapidly, potentially missing new threat indicators. A weakness in the system stems from a Byzantine vulnerability. Hackers could send fake model updates that compromise overall results or introduce hidden flaws, but so far, no research has tested robust fixes such as Krum, Bulyan, or cryptographic verification. This misses a real need in shared 5G setups where attackers might intentionally hijack users. Tests were limited to 100 nodes, which makes it unclear whether these methods can handle live networks with many more participants. Coordination at that scale has not yet been proven. A use case where the vulnerabilities can be addressed is a federated learning setup that secures network slices for 50 business clients. It appears safe at first glance, yet a clever hacker breaches only 5 of those (that is, 1 in 10). They could slip altered models into the mix when updates are combined. Instead of breaking everything, these tainted inputs slowly create hidden gaps that allow certain threats to go unnoticed, such as unauthorised access to banking systems. Though general performance still looks solid on paper, defences against targeted intrusions drop sharply, from catching nearly all such attempts to fewer than 6 out of 10 over time, even as standard checks show no red flags.

4) Hybrid Architectures Combining Multiple Paradigms

Complex hybrid designs [28], [29], [38] merge complementary techniques (FL + DRL, FL + game theory) to achieve superior performance. Fewer errors were observed with hybrid strategies, up to 43% fewer than with a single-paradigm approach. Federated DRL for IIoT intrusion detection [28] hit 97.2% right calls by merging private learning with quick-thinking controls. One advanced method is the Two-layer FL with mean-field game theory [38], which detects compromised zones much more frequently through clear back-and-forth threat tracking. However, despite the great improvements, it has costs and complexity. Computation took 15–40% longer compared to basic approaches. Game-theoretic equilibrium computation took 2.7 seconds, which may be too slow for quick responses. Managing the system requires knowledge of several AI areas, such as federated learning, reinforcement learning, and strategic modelling, while also troubleshooting unexpected tool conflicts and balancing conflicting goals across different design levels. Integrating different approaches, however, is challenging; most hybrid methods appear random rather than being built step by step from the task at hand. No paper provided clear blueprints for the smart blending of techniques or explained when certain mixes make sense. Key questions remain unanswered: when does federated learning substantially improve reinforcement learning? When should we use game-style thinking instead of straight RL alone? Missing solid theory means real-world builders have no reliable rules to follow while crafting combined setups Table 5 shows a summary of the learning approaches against performance and operational characteristics

Table 5: Learning Approaches vs. Performance and Operational Characteristics

Approach	Accuracy	Latency	Max Scale	Adversarial Testing?	Real Deployment?	Key Limitation
Single-AgenT DQN	89-96.8%	<50ms decision	100 nodes	No	No	Single point of failure
MARL	78-96%	120-180ms coord	50 agents	No	No	Coordination overhead vs URzLC
Federated Learning	94.8-97.3%	Minutes-hours agg	100 nodes	No	No	Byzantine vulnerability

Hybrid	97.2% (+43% gain)	2.7s equilibrium	100 nodes	Partial [37]	No	15-40% overhead, multi-domain expertise

3.5. What empirical, methodological, and operational gaps persist in the current literature regarding scalability, adversarial robustness, and real-world deployment feasibility?

The analysis reviews systematic limitations that constrain the readiness of agentic AI for operational 5G deployment.

1) The Simulation-Reality Gap

Most research leaned heavily on simulations - 86.4% used them alone, while 9.1% mixed in network emulation instead. Only 4.5% attempted any real-world checks. Because of this, findings may not generalize outside controlled settings, a major drawback when applying results more broadly. Simulated setups typically employ basic threat scenarios; attack methods are often routine or fixed. Network settings remain unchanged during tests, whereas data traffic is generated rather than drawn from real-world breaches. Instead of adjusting on the fly to bypass security, attackers adhere to preset rules. Because of these shortcuts, results such as 88–98% detection rates or 89–96.7% less damage may reflect best-case outcomes. Real-world performance would likely fall short. Furthermore, datasets are getting outdated: Most hacking detection research still relies on old standards like NSL-KDD or UNSW-NB15 - benchmarks made before 5G even existed, so they do not cover risks tied to virtual functions, broken slice barriers, or hacked split radio systems. Because of this gap, tools built using earlier threats will not detect risks unique to 5G systems.

2) Adversarial Robustness: The Missing Dimension

Across all 22 studies, adversarial robustness testing was virtually absent. Hackers now use sophisticated data tricks to fool smart systems, disrupt shared learning, or manipulate feedback loops. Even though these defences are becoming sharper, none of the works tested defences against fake inputs, stolen model leaks, guessing private information, or corrupted updates to detect break-ins, stop floods, or detect anomalous behaviour. One paper [38] used a strategy-style setup but still ran only pretend trials that

took about 2.5 seconds to balance. It does have a critical vulnerability. Self-driving tech might handle common attacks just fine, yet crash hard when tricked by clever tweaks that target its decision-making. These hacks are a significant gap in demonstrating that these systems can remain reliable in hostile 5G networks.

An example would be someone trying to break in might test a DRL-powered defence against floods of fake traffic. By watching how it reacts, like checking what gets rewarded, they shape harmful data flows that look like urgent help requests. Instead of blocking them, the system begins allowing them through because stopping them incurs penalties. Over time, it is tricked into sustaining harmful traffic without appearing faulty. If no one tests its robustness to adversarial attacks, this flaw will not surface until it is already being exploited.

3) Scalability: The Unvalidated Frontier

With a limited validation scope, the largest tests covered only 100 nodes or 50 agents. Real-world 5G deployments may require coordination among thousands of units distributed across large areas, including densely populated urban areas. So systems that work well in lab conditions (like 100-node trials) could break down when pushed to real-life size because:

a) Chatting slows way down when more agents join - each new one adds extra hassle that stacks up fast
b) As systems spread out, agreement takes longer due to tangled communication paths
c) Attacks from tricky insiders messing with team systems when things get big
d) Edge resource constraints limiting agent computational capacity

Scaling up, however, causes delays: systems with 50 agents show a lag of 120–180ms, Competing objectives: Federated learning preserves data privacy, which is essential when multiple groups collaborate, but it slows model updates, sometimes taking minutes or even hours. That delay conflicts with rapidly evolving cyber threats that require immediate responses [47],[48]. Differential privacy mechanisms further reduce learning speed by 15-35%. So far, no research has mapped out how these factors conflict or offered clear rules for choosing settings that balance secrecy, precision, and rapid responses. When agents make rapid local decisions based on outdated data, and updates

arrive only after delays, they may overlook emerging attack styles. This gap arises because learning occurs gradually in the background rather than in real time. The risk of this delay depends on how rapidly threats change, yet no one knows exactly where the tipping point lies. Missing subtle shifts becomes more likely if changes outpace model refreshes.

4) Lack of Benchmarking Standards

Without fixed 5G safety rules, it is hard to judge which study performs better. Varying datasets, such as NSL-KDD, UNSW-NB15, or synthetic traffic, disrupt consistency. Network sizes differ, from a few machines to nearly a hundred. Some tests assess one type of attack; others examine multiple threats simultaneously. Additionally, researchers employ different measures of success: one uses accuracy, another uses F1-score. At the same time, someone else tracks the amount of damage avoided. Furthermore, no 5G-focused test data exists, and public sources do not cover unique risks such as broken slice separation, hacked O-RAN links, compromised virtual functions, or mmWave signal tampering. Because of this gap, results cannot be independently verified; without consistent reference points, real improvement is difficult to measure.

5) Responsible AI and Ethical Considerations

Privacy-safe choices in shared setups: According to [45], when different groups operate joint systems, such as federated or agent-based networks, they require built-in tools to protect private information and clear decision-making rules. Network slicing methods do not include such protections, even though collaboration across tenants could leak confidential information or allow outsiders to infer usage patterns. In addition, some security tools operate independently, shutting down components or halting data flow without explanation. These choices stay hidden because there is no clear way for people to see how they were made. You cannot check if the logic is right when you do not get an explanation. No visibility means doubts grow fast, especially where rules demand answers. When no one understands the call, confidence drops significantly. Last but not least, fairness issues arise with the use of AI. No study has checked whether smart AI performs unevenly across different renters, priority levels, or locations. Skewed data or rewards might hurt certain users, even when overall scores look good.

3.6. Conceptual Framework for Agentic AI in 5G Threat Detection and Mitigation

This review builds a Conceptual Framework for Agentic AI in 5G security, drawing on insights from 22 real-world studies shown in Figure 8. It shows how self-driven threat response operates through a loop of intelligence split into four parts: Perception, Decision, Action, and Feedback. These layers rely on three key supports that run across them all: keeping data private, handling growth smoothly, and resisting attacks effectively.

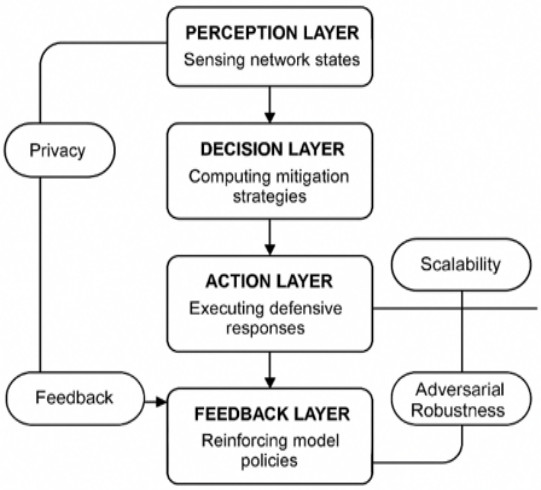

Figure 8. Conceptual Framework of Agentic AI for 5G Threat Detection and Mitigation.

The perceptual layer of agents monitors network state through data-flow inspection, abnormal behaviour detection, and usage tracking at the RAN, MEC, and Core levels. The Decision Layer not only reacts but also enhances protection schemes through trial-and-error learning, shared control settings, and OR teamwork among several agents, based on safety objectives and live requests. The Action Layer intervenes at the proper moment. For instance, this may involve blocking suspicious traffic, reallocating resources, or isolating the affected slices using software-defined networking, virtual functions, or local processing. Ultimately, the Feedback Layer concludes the self-running cycle by monitoring the performance of previously executed actions and updating agent tactics via reward-driven training.

For the architecture to work, enablers are required. These enablers span all layers of the architecture, each with its own tasks and confirmations. These enablers are privacy preservation, scalability, robustness against Byzantine agents, robustness against adversarial machine learning attacks, and intelligent evasion strategies. This framework

provides a unified structure for analysing how autonomous agents perceive threats, decide on responses, execute actions, and continuously adapt in adversarial 5G environments.

A 5G setup powering smart factories faces a coordinated attack: first, massive fake traffic floods the factory's IoT network; next, evasive data patterns evade security alarms; then, hackers try to slip into nearby business zones. Framework response as follow.

1. Perception: Sensors on edge devices spot weird traffic like signs of DDoS attacks or notice an industrial IoT network slowing down, while also catching cases where one network probes another against rules
2. Decision: Agents spread work together using a robust method that handles faulty inputs, enabling them to spot dangers, assess how serious each is, since the IIoT network runs vital factory systems, and pick actions that reduce harm while keeping service quality steady.
3. Action: Edge agents handle quick tasks like slowing down traffic fast (under 10ms), keeping network slices separate so threats do not spread, or teaming up to block DDoS attacks across different spots; meanwhile, central agents tweak resources to keep industrial IoT performance safe and stable
4. Feedback: Agents notice how attacks change, like DDoS moving to the application layer or traffic tweaking its stealth tactics, they adjust their responses on the fly using real-time learning, keep user data safe with built-in noise techniques, yet still spread key warnings among connected teams

Decision logs make choices clear so people can understand why systems get isolated. Fairness checks ensure that low-priority tenants still receive adequate security coverage. Privacy safeguards prevent data leakage between tenants when systems work together. This example shows how the new proposed framework tackles key missing pieces: stronger defence against changing threats, support for large-scale systems through decentralised control, better trade-offs between data privacy and quick decisions using on-device learning, and clearer accountability via open models, unbiased outcomes, and secure handling, all overlooked in today's simulated studies yet critical when putting solutions into real-world use.

3.7. Limitations

1) Methodological Limitations

The literature review methodology may limit the systematic findings. There are four databases on IEEE Xplore, ACM, SpringerLink, and ScienceDirect; searches were limited to English-language publications, thus excluding potentially relevant publications from other venues and languages. Only 22 studies on agentic AI and 5G security met the inclusion criteria. This small pool of studies is not an exhaustive list of all relevant work. Instead, it shows that the field is quite young. While the PRISMA 2009 framework helped ensure the review's meticulous methodology, it may have omitted grey literature, preprints, or new evidence not yet published in peer-reviewed fora. Publication bias likely affected the sample, as studies with positive results are more likely to be published than those with negative or less conclusive results, potentially making the efficacy of agentic AI appear better than it actually is.

2) Study-Level Limitations

The studies exhibited serious limitations that constrain generalizability and practical applicability. Almost 20 out of 22 studies relied mainly on simulation environments, very few on emulation and minimal real-life deployment validation. There are concerns about performance issues. This is due to actual network complexity, the complexity of real attacks, and operational constraints that are missing in simulations. Most evaluation methods used older intrusion detection datasets rather than newer ones. They used the NSL-KDD and UNSW-NB15 datasets or synthetic attacks rather than the 5G dataset, which captures attacks on slicing or functions. Since synthetic attacks may not accurately reflect real-world adversarial behavior, they may overestimate actual detection performance.

Most-cited works have serious limitations in their scalability analyses. In fact, the most extensive evaluations performed involve (at most) 100 nodes or 50 agents. Hence, there remains considerable uncertainty regarding performance in ultra-large 5G deployments involving thousands of coordinating agents across large geographic areas. Adversarial robustness has hardly received any attention. No study has quantitatively checked the resilience against Byzantine agents, adversarial machine learning attacks and clever evasion attacks aimed at the designed algorithms. The absence of standard evaluation metrics for agentic AI in 5G security studies hinders comparisons and the evaluation of

relative effectiveness. These considerations of energy efficiency and computational overhead become important in battery-powered edge devices and resource-constrained IoT gateways. However, these studies were not conducted. It creates uncertainty regarding deployment in resource-constrained environments.

3) Scope and Coverage Limitations

According to the review, there were clear deficiencies regarding the threats and uses. Physical-level attacks, such as jamming and eavesdropping on communication signals, can compromise the wireless link. However, there is limited focus on them. In fact, the only paper that deals with jamming is [6]. The work relied heavily on core 5G network measures rather than addressing them. Real-world risks such as app-layer DDoS and complex, combined attacks are increasing rapidly but have not been systematically studied. They barely touched critical installations that require custom-made defences, such as special setups, satellite-connected 5G, and drone systems.

3.8. Implications

1) Theoretical Implications

The review in question presents relevant theoretical insights at the intersection of robotics, cybersecurity and telecommunications. The fact that agentic systems have been successfully developed and used to address several 5G security challenges suggests that the concept generalises from robotics and cooperative multi-agent systems to adversarial security problems. Afterward, clear theory gaps emerge when systems join attacks at a single location, and attacks make independent choices to establish an equilibrium that balances teamwork, faster attacks, and self-protective data. Combining federated learning with reinforcement learning can enable private data methods and support informed decision-making. Instead of keeping these, blending them opens new ground, especially where privacy rules, teamwork-based updates, and step-by-step decisions meet under pressure. Such compromises mean that new research should develop multi-robot or agent methods that anticipate hostility, rather than tweaking team-based models designed for safe settings, and that these methods can be based on a combination of federated learning and reinforcement learning.

2) Practical Implications

Agentic AI gives telecom providers and cybersecurity teams practical guidance for securing live 5G networks without exposing customer data. The evidence shows that multi-agent, distributed setups handle network stress better and scale more smoothly than centralised models. Although these results appear promising, it is important to note that most were conducted in simulated environments. This implies that there is a need to obtain real-world testbed results before proceeding further and pursuing full-scale adoption.

3) Policy Implications

This review has highlighted issues that regulators and standards bodies must address quickly as agentic AI begins to run within 5G networks. Policymakers should design clear rules that allow organisations to share threat insights securely, using privacy-preserving techniques such as federated learning, and to define legal liability in the event of system malfunctions. Regulators of multi-agent systems should require pre-deployment testing to account for rogue or faulty nodes. The system must be checked to ensure it is not vulnerable to manipulations through data or collusion. Accountability guidelines must address errors such as blocking the wrong user and failing to detect an attack, making them traceable and auditable, perhaps by using explainable AI tools. Currently, there are no common testing standards. As such, 3GPP, ITU, and ETSI need to develop evaluation frameworks focused on 5G security. The frameworks should use realistic threat datasets and performance measures extending beyond mere accuracy. Many IoT devices have limited energy and processing potential. Hence, regulations should ensure that power use is continuously and visibly reported. In addition, they should encourage the adoption of lightweight AI models that will not burden the network.

3.9. Future Research Directions

While agentic progress continues, future research must shift from simulation-based studies to real-world testing that demonstrates behaviour in practice. Access to near-production network environments through collaborations between operators and research institutions would expose gaps hidden in short laboratory experiments. Strengthening resilience against adversarial manipulation remains equally important, particularly as multi-agent systems introduce opportunities for poisoned rewards, disrupted coordination, or malicious nodes. This indicates a need for further research on

robust learning mechanisms, game-theoretic attacker-defender models, and behavioural anomaly detection that monitors agents themselves. Increasing the number of agents and nodes also requires a different approach, such as hierarchical coordination, lightweight consensus methods, hybrid single-agent/multi-agent modes, and online learning that can manage device traffic. In addition, future systems should consider current security studies, such as zero trust, and how to integrate them. Energy efficiency remains a critical challenge since most 5G devices and edge nodes cannot support heavy computation; therefore, work on model compression, low-precision inference, knowledge distillation, and cloud–edge hybrid designs must be validated on real 5G hardware.

4. CONCLUSION

This review examined 22 recent studies on agentic AI for detecting and mitigating threats in 5G networks. Most research indicates that federated learning performs best when multiple users share security configurations, with nearly half of the studies focusing on this setting. Reinforcement learning enables systems to respond more quickly and detect hazards with 94.8%- 97.3% accuracy. When using several AI agents together, they can coordinate across locations but introduce delays of 120-180 milliseconds, which is too slow for ultra-fast connections. Mixing different methods boosts performance by up to 43%, though processing requirements increase by 15-40%. Despite advances in algorithms, key problems remain. Most tests rely on simulations (about 86%), making results difficult to apply in real-world settings. When it comes to facing challenging enemies or sophisticated attacks, almost no testing has been conducted. Systems have not been tested beyond 100 nodes, so we are uncertain how they would perform in large networks with thousands of nodes. Also, balancing privacy and rapid response remains challenging; sharing data safely slows things down when quick action's needed. The proposed Conceptual Framework runs on a loop sense, decide, act, learn, with a strong focus on privacy, room to grow, and defence against attacks. It helps turn experimental AI into real-world systems. Future steps would include testing it in live 5G networks, conducting rigorous attack simulations, demonstrating its performance across thousands of units, developing 5G-specific benchmarks, and incorporating ethical AI practices. With 5G spreading into key systems like automated factories, driverless cars, telemedicine, and connected urban areas, agentic AI gives network providers a way to build self-repairing systems that spot and block complex hacks in moments all while keeping user data

private; yet getting it ready for real use means closing known weaknesses by running tough field tests, checking resistance to malicious inputs, and proving performance at scale so this tech can actually deliver on securing future networks from ever-changing digital dangers.

ACKNOWLEDGMENT

Acknowledgement to all the reviewers who helped make this manuscript robust with their insightful comments. Special mention to the database that provided access to the papers used in this study.

REFERENCES

[1] S. Sharma, I. You, and M. Atiquzzaman, "Security, Privacy and Reliability in 5G Networks," *IEEE Commun. Surv. Tutorials*, 2020, doi: 10.1109/COMST.2020.2991754.

[2] Z. Zhang *et al.*, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Veh. Technol. Mag.*, 2020, doi: 10.1109/MVT.2019.2952109.

[3] N. Panwar and S. Sharma, "5G Enabled Internet of Things: Security and Privacy Issues," *IEEE Sens. J.*, 2021, doi: 10.1109/JSEN.2021.3050832.

[4] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2020.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[6] M. Wooldridge, *An Introduction to MultiAgent Systems*. Wiley, 2009.

[7] I. B. M. Research, "Agentic AI and Autonomous Systems White Paper," 2025.

[8] M. Eriksson and J. Holmberg, "Towards Autonomous Networks with Agentic AI," *Ericsson Technol. Rev.*, 2025.

[9] S. Chishakwe, N. Moyo, B. M. Ndlovu, and S. Dube, "Intrusion Detection System for IoT environments using Machine Learning Techniques," *2022 1st Zimbabwe Conf. Inf. Commun. Technol. ZCICT 2022*, pp. 1–7, 2022, doi: 10.1109/ZCICT55726.2022.10045992.

[10] S. Kaur, N. Gupta, and R. Kumar, "A Systematic Review of AI and ML Techniques for Cybersecurity," *Comput. Secur.*, 2023.

[11] L. U. Khan, I. Yaqoob, M. Imran, and Z. Khan, "AI and Machine Learning for 5G Security: A Comprehensive Survey," *IEEE Commun. Surv. Tutorials*, 2023.

[12] M. Nasralla and R. Hassan, "A Systematic Literature Review on 5G Security Challenges and Machine Learning-Based Solutions," *IEEE Access*, 2022.

[13] S. Hussain, M. H. Rehman, and R. Sana, "Deep Learning in Intrusion Detection Systems: A Systematic Review," *ACM Comput. Surv.*, 2023.

[14] A. Ferdowsi and W. Saad, "A Comprehensive Survey on Machine Learning for Wireless Security," *IEEE Commun. Surv. Tutorials*, 2019.

[15] S. Alghamdi, "Security Challenges in 5G Networks: A Survey," *Sensors*, vol. 21, no. 24, 2021.

[16] M.-T. Nguyen, "A Survey of Reinforcement Learning Applications in Communication Networks," *IEEE Access*, 2020.

[17] N. Kaur, N. Gupta, and R. Kumar, "A Systematic Review of AI and Machine Learning Techniques for Cybersecurity," *Comput. Secur.*, 2023, doi: 10.1016/j.cose.2023.103248.

[18] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *BMJ*, vol. 339, no. 7716, pp. 332–336, 2009, doi: 10.1136/bmj.b2535.

[19] D. and L. Moher Alessandro and Tetzlaff, Jennifer and Altman, Douglas G., "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *PLoS Med.*, vol. 6, no. 7, p. e1000097, 2009.

[20] J. McGowan, M. Sampson, and D. M. Salzwedel, "PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement," *J. Clin. Epidemiol.*, vol. 75, pp. 40–46, 2016, doi: 10.1016/j.jclinepi.2016.01.021.

[21] B. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. Boca Raton, FL: CRC Press, 2015.

[22] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, p. b2700, 2009, doi: 10.1136/bmj.b2700.

[23] P. F. Saura, J. M. Bernabé Murcia, E. G. de la Calera Molina, A. Molina Zarca, J. Bernal Bernabé, and A. F. Skarmeta Gómez, "Federated Network Intelligence Orchestration for Scalable and Automated FL-Based Anomaly Detection in B5G Networks," *Comput. Mater. Contin.*, vol. 80, no. 1, pp. 163–193, 2024, doi: https://doi.org/10.32604/cmc.2024.051307.

[24] R. Moreira *et al.*, "An intelligent native network slicing security architecture empowered by federated learning," *Futur. Gener. Comput. Syst.*, vol. 163, p. 107537, 2025, doi: https://doi.org/10.1016/j.future.2024.107537.

[25] J. Á. Fernández-Carrasco, L. Segurola-Gil, F. Zola, and R. Orduna-Urrutia, "Security and 5G: Attack mitigation using Reinforcement Learning in SDN networks," in *2022 IEEE Future Networks World Forum (FNWF)*, 2022, pp. 622–627, doi: 10.1109/FNWF55208.2022.00114.

[26] A. Arun, C. H. Basha, M. J. Rani, V. Jamuna, P. Vijayakumar, and K. J. Kumar, "Deep Reinforcement Learning-Based Intrusion Detection System for Next-Generation Wireless Networks," in *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2025, pp. 283–291, doi: 10.1109/ICIRCA65293.2025.11089597.

[27] F. Alalyan, M. Awad, W. Jaafar, and R. Langar, "Secure Distributed Federated Learning for Cyberattacks Detection in B5G Open Radio," vol. 6, no. November 2024, pp. 3067–3081, 2025.

[28] A. Kaur, "Intrusion Detection Approach for Industrial Internet of Things Traffic Using Deep Recurrent," *IEEE Trans. Artif. Intell.*, vol. 6, no. 1, pp. 37–50, 2025, doi: 10.1109/TAI.2024.3443787.

[29] H. Sharma, N. Kumar, and R. Tekchandani, "Mitigating Jamming Attack in 5G Heterogeneous Networks : A Federated Deep Reinforcement Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2439–2452, 2023, doi: 10.1109/TVT.2022.3212966.

[30] K. Bedda, Z. Fadlullah, and M. M. Fouda, "Efficient Wireless Network Slicing in 5G Networks : An Asynchronous Federated Learning Approach," *2022 IEEE Int. Conf. Internet Things Intell. Syst.*, pp. 285–289, 2022, doi: 10.1109/IoTaIS56727.2022.9976007.

[31] A. Thantharate, R. Paropkari, V. Walunj, C. Beard, and P. Kankariya, "Secure5G : A Deep Learning Framework Towards a Secure Network Slicing in 5G and Beyond," *2020 10th Annu. Comput. Commun. Work. Conf.*, pp. 852–857, 2020, doi: 10.1109/CCWC47524.2020.9031158.

[32] E. Ossongo, M. Esseghir, and L. M. Ieee, "A multi-agent Federated Reinforcement Learning-based optimization of quality of service in various LoRa network slices," 2023.

[33] L. Pu *et al.*, "Federated Learning-based Heterogeneous Load Prediction and Slicing for 5G Systems and Beyond," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 166–172, doi: 10.1109/GLOBECOM48099.2022.10000870.

[34] S. Wijethilaka and M. Liyanage, "A Federated Learning Approach for Improving Security in Network Slicing."

[35] Y. Cui, H. Shi, R. Wang, and P. He, "Multi-Agent Reinforcement Learning for Slicing Resource Allocation in Vehicular Networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 2005–2016, 2024, doi: 10.1109/TITS.2023.3314929.

[36] D. K. Dake, "DDoS and Flash Event Detection in Higher Bandwidth SDN-IoT using Multiagent Reinforcement Learning," pp. 16–20, 2021, doi: 10.1109/ICCMA53594.2021.00011.

[37] F. Alalyan, B. Bousalem, W. Jaafar, and R. Langar, "Secure peer-to-peer federated learning for efficient cyberattacks detection in 5G and beyond networks," *ICC 2024 - IEEE International Conference on Communications*. IEEE, pp. 1752–1757, 2024.

[38] H. Sedjelmaci and A. Boualouache, "When Two-Layer Federated Learning and Mean-Field Game Meet 5G and Beyond Security : Cooperative Defense Systems for 5G and Beyond Network Slicing."

[39] M. Kiely *et al.*, "Exploring the Efficacy of Multi-Agent Reinforcement Learning for Autonomous Cyber Defence : A CAGE Challenge 4 Perspective," no. 2024, 2025.

[40] P. Radoglou-Grammatikis *et al.*, "Strategic Honeypot Deployment in Ultra-Dense Beyond 5G Networks: A Reinforcement Learning Approach," *IEEE Trans. Emerg. Top. Comput.*, vol. 12, no. 2, pp. 643–655, 2024, doi: 10.1109/TETC.2022.3184112.

[41] A. Villafranca, K. M. Thant, and I. Tasic, "AI-Enabled IoT Intrusion Detection : Unified Conceptual Framework and Research Roadmap," pp. 1–38, 2025.

[42] D. Cajaraville-aboy *et al.*, "CO-DEFEND : Continuous Decentralized Federated Learning for Secure DoH-Based Threat Detection," pp. 1–17.

[43] S. Khozam, G. Blanc, S. Tixeuil, and E. Totel, "DDoS Mitigation while Preserving QoS: A Deep Reinforcement Learning-Based Approach," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, 2024, pp. 369–374, doi: 10.1109/NetSoft60951.2024.10588889.

[44] B. Bousalem, M. A. Sakka, V. F. Silva, W. Jaafar, A. Ben Letaifa, and R. Langar, "DDoS Attacks Mitigation in 5G-V2X Networks: A Reinforcement Learning-Based Approach," in *2023 19th International Conference on Network and Service Management (CNSM)*, 2023, pp. 1–5, doi: 10.23919/CNSM59352.2023.10327917.

[45] 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. Informatics Educ.*, vol. 9, no. 1, 2025, doi: 10.20961/ijie.v9i1.100723.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[46] J. Mutengeni, A. Musasa, and B. Mutunhu, "Local Area Network Based Collaboration Using Distributed Computing," *2022 1st Zimbabwe Conf. Inf. Commun. Technol. ZCICT 2022*, pp. 1–7, 2022, doi: 10.1109/ZCICT55726.2022.10045858.

[47] P. Vareta, H. Muzenda, T. Nyamupaguma, Y. Dube, and B. Ndlovu, "The Rise of Quantum Computing and Its Impact on Cybersecurity," *Indones. J. Comput. Sci.*, vol. 14, no. 6, 2025, doi: 10.33022/ijcs.v14i6.5040.

[48] P. Maitireni, V. Ncube, B. Ndlovu, and T. Sibanda, "Quantum Computing Cryptography : A Systematic Review of Innovations , Applications , Challenges , and Algorithms," vol. 7, no. 4, pp. 3668–3710, 2025, doi: 10.63158/journalisi.v7i4.1331.