Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.087445

REVIEW

AI-Native Security for 5G and 6G O-RAN: A Survey of Learning-Based Defense against False Base Stations, DDoS, and Signaling Storms

I Wayan Adi Juliawan Pawana[1,2], Vincent Abella[2], Eldridge Aaron Miole[2], Hoonyong Park[3] and Ilsun You[2,4,*]

[1]Department of Electrical Engineering, Udayana University, Badung, Indonesia
[2]Department of Cyber Security, Kookmin University, Seoul, Republic of Korea
[3]AUTOCRYPT Co., Ltd., Seoul, Republic of Korea
[4]KMU Global Research Center for ICT Convergence Security, Kookmin University, Seoul, Republic of Korea

*Corresponding Author: Ilsun You. Email: isyou@kookmin.ac.kr

Received: 16 June 2026; Accepted: 28 August 2026; Published: 28 September 2026

ABSTRACT: Fifth- and sixth-generation cellular networks remain exposed to a spectrum of threats ranging from the stealthy and targeted false base station (FBS) to the systemic distributed denial-of-service (DDoS) and signaling-storm attacks. The elevation of artificial intelligence (AI) to a native network capability by the International Telecommunication Union Radiocommunication Sector (ITU-R) framework for International Mobile Telecommunications-2030 (IMT-2030) has made learning-based defense the dominant response. This survey examines that response across the three threats jointly, advancing as a design position the view that effective defense is best served when AI is integrated into the radio access network (RAN) and core as a native function rather than appended as an external classifier. We organize the literature by five technique families, namely classical machine learning, deep learning, generative models, federated learning, and large language and agentic models, and for each we record not only how it detects an attack but where in the Open RAN (O-RAN) architecture it is deployed, mapping the surveyed methods onto the user equipment, the near-real-time and non-real-time RAN Intelligent Controller, and the 5G-core network function. Three observations organize the account: the three threats form a data-scarcity gradient that illuminates the progression of techniques; the deployment dimension, the question of where and how, is the one the primary literature most often omits and the one a practitioner most needs; and several threat-technique combinations that the architecture permits remain unexplored. Drawing on a semi-systematic review of 177 studies, the survey consolidates the datasets and reported performance that make learning feasible, examines the adversarial robustness of the defenses themselves, and sets out a forward agenda toward quantum-AI-native defense within the RAN.

KEYWORDS: 5G; 6G; open RAN; false base station; DDoS; signaling storm; AI-native security; federated learning; large language models

1 Introduction

In September 2025, the largest mobile operator in South Korea disclosed that several thousand of its subscribers had been compromised through a cluster of false base stations (FBS) operating undetected within its coverage area, an incident that resulted in unauthorized micropayments and prompted national concern regarding the integrity of cellular access [1,2]. This event was not anomalous. A separate prosecution in Norway concerned an offender who operated a vehicle-mounted, SMS-broadcasting FBS in populated

areas [3], and a device initially suspected to be an explosive in Paris was subsequently identified as an IMSI-catcher employed for fraudulent purposes [4]. These cases share a common root cause that has persisted from 2G systems through 5G deployments: a user device attaches to the base station presenting the strongest received signal, and the broadcast information governing that decision is neither authenticated nor integrity-protected prior to attachment. The false base station consequently remains one of the longest-standing unresolved vulnerabilities in cellular security.

False base stations represent one extreme of a broader threat landscape confronting 5G and 6G networks. An FBS attack is characteristically stealthy and targeted: it impersonates legitimate infrastructure in order to harvest subscriber identifiers, downgrade ciphering, inject spoofed emergency alerts, or overshadow legitimate transmissions directed at specific victims [5–8]. At the opposing extreme are volumetric and systemic threats that operate through resource exhaustion rather than concealment. Distributed denial-of-service (DDoS) attacks and signaling storms saturate the radio access network or the 5G core with connection requests or malformed control-plane traffic, thereby degrading or disabling service across an entire region [9,10]. An effective defensive capability for a contemporary network cannot be specialized to a single point on this spectrum. The same operational platform must be capable of identifying an individual rogue cell that interacts with a small number of devices and of absorbing large volumes of spurious requests arriving within a short interval.

The defensive paradigm available to address both extremes has changed substantially. Across the protocol stack, artificial intelligence (AI) and machine learning (ML) have advanced from research instruments to the principal means of detecting behaviors that rule-based systems fail to capture. More significantly, AI is increasingly being incorporated into the architecture of the next generation itself. The ITU-R Framework for IMT-2030 elevates "AI and Communication" to a standalone usage scenario for the first time in any generation of mobile standards and identifies ubiquitous intelligence as a defining capability of sixth-generation systems [11]. The Open RAN architecture operationalizes this objective through the RAN Intelligent Controller (RIC), which provides explicit deployment points, comprising near-real-time *xApps* and non-real-time *rApps*, at which learning models are intended to operate as first-class network functions rather than as external appliances [9,12,13]. The central defensive question for sixth-generation networks is therefore no longer whether AI is capable of detecting an attack, but rather where within the architecture a model should execute and how it should be integrated such that detection constitutes a native property of the network.

The existing literature is least developed on precisely this question, and two gaps recur. The first concerns depth of treatment. A substantial and expanding body of work demonstrates that a classifier, a deep neural network, or a generative model can identify an FBS, a DDoS flow, or a signaling anomaly with high reported accuracy; however, the majority of these studies conclude at the reporting of that result and leave the deployment dimension unaddressed, including which network function hosts the model, which protocol layer or plane supplies its input features, and what latency constraints the selected vantage point imposes. For a practitioner, the distinction between an xApp performing sub-second inference on live traffic within the near-real-time budget and an rApp conducting slower, above-one-second analytics over aggregated measurement reports determines whether a defense is operationally deployable. The second gap concerns breadth of coverage. Surveys typically address these threats in isolation, as exemplified by air-interface security surveys [14] and federated-learning intrusion-detection surveys [15] considered separately. This separation obscures the observation that FBS detection, DDoS detection, and signaling-storm detection constitute, at the level of the underlying learning problem, instances of a single task, namely anomaly detection under varying degrees of data scarcity. Articulating this shared structure, and demonstrating how

a limited set of AI techniques transfers across all three threats, constitutes a contribution that current surveys do not provide.

Table 1 positions this work relative to representative recent surveys along the dimensions relevant to an AI-native defense. Prior surveys exhibit strength on individual axes, including standardization pathways [16], optimization taxonomies [17], federated learning [15], and the adversarial threats introduced by AI itself [18]. To our knowledge, none jointly addresses the three threats under a common AI framework, organizes techniques by their O-RAN integration points, treats the datasets that render learning feasible, and characterizes the trajectory toward quantum-enhanced defense. The closest prior works are complementary rather than overlapping: the 5G-to-6G security survey [16] is broad but not technique-organized, the federated-learning survey [15] is single-technique, the review of machine-learning and transfer-learning strategies for intrusion detection in 5G and beyond [19] treats detection techniques and their datasets in depth but does not address where in the architecture a detector is deployed, and the adversarial-threat survey [18] concerns the security of AI rather than AI for defense.

Table 1: Comparison with representative recent surveys. ●: covered in depth; ◐: partial; ○: absent.

Survey	FBS	DDoS	Signaling	AI Technique Depth	O-RAN/Arch. Integration	Datasets Treated	Quantum Outlook
Yu et al. [14]	●	◐	◐	○	○	○	○
Yang et al. [16]	○	◐	○	◐	◐	○	●
Mustafovski et al. [17]	○	○	○	●	◐	○	○
Rezaei et al. [15]	○	◐	○	◐	◐	◐	○
Noor et al. [19]	○	●	◐	◐	○	●	◐
Altintaş et al. [18]	○	○	○	◐	◐	●	◐
This work	●	●	●	●	●	●	●

The principal contributions of this survey are as follows.

- **A cross-threat synthesis.** AI-based defenses against false base stations, DDoS, and signaling storms are surveyed jointly, establishing that they reduce to a common family of learning problems and that techniques validated on one threat transfer to the others.
- **An integration-oriented organization.** The AI techniques are organized according to a fixed analytical template, and each is associated with a deployment location within the O-RAN and AI-RAN architecture, comprising UE on-device execution, the near-real-time RIC *xApp*, the non-real-time RIC *rApp*, and the dedicated 5G-core network function. This organization relates the surveyed methods to their deployment context and enables comparison across an otherwise heterogeneous literature; where a deployment is the authors' inference rather than reported by the source, this is marked explicitly and the proportion of inferred assignments is reported.
- **A data-scarcity account of the field.** The three threats are shown to differ substantially in data availability; this gradient is advanced as an organizing lens for the observed progression of techniques from feature-based ML toward generative augmentation and quantum encoding, a progression that also reflects general advances in machine learning rather than data conditions alone; and the relevant datasets together with their evaluation pitfalls are consolidated.
- **A forward agenda toward quantum-AI-native defense.** The conditions under which quantum machine learning can realistically operate within the RAN are assessed in light of current hardware constraints, and concrete, evidence-supported open problems are identified, including threat–technique combinations that the literature has not yet examined.

The remainder of the paper is organized as follows, as summarized in Fig. 1. Section 2 presents the review methodology, research questions, and selection criteria. Section 3 establishes the 5G/6G and AI-RAN background and develops a taxonomy of the three threats. Section 4 surveys the AI-technique spine, comprising classical ML, deep learning, generative models, federated learning, and large language models, and maps each technique to its integration point. Section 5 synthesizes these results into an integration map and a comparative selection framework. Section 6 consolidates the datasets, benchmarks, and evaluation practices. Section 7 examines the security of the AI defenses themselves and the remaining open challenges. Section 8 sets out the future directions, centered on quantum-AI-native defense within the RAN. Section 9 concludes.

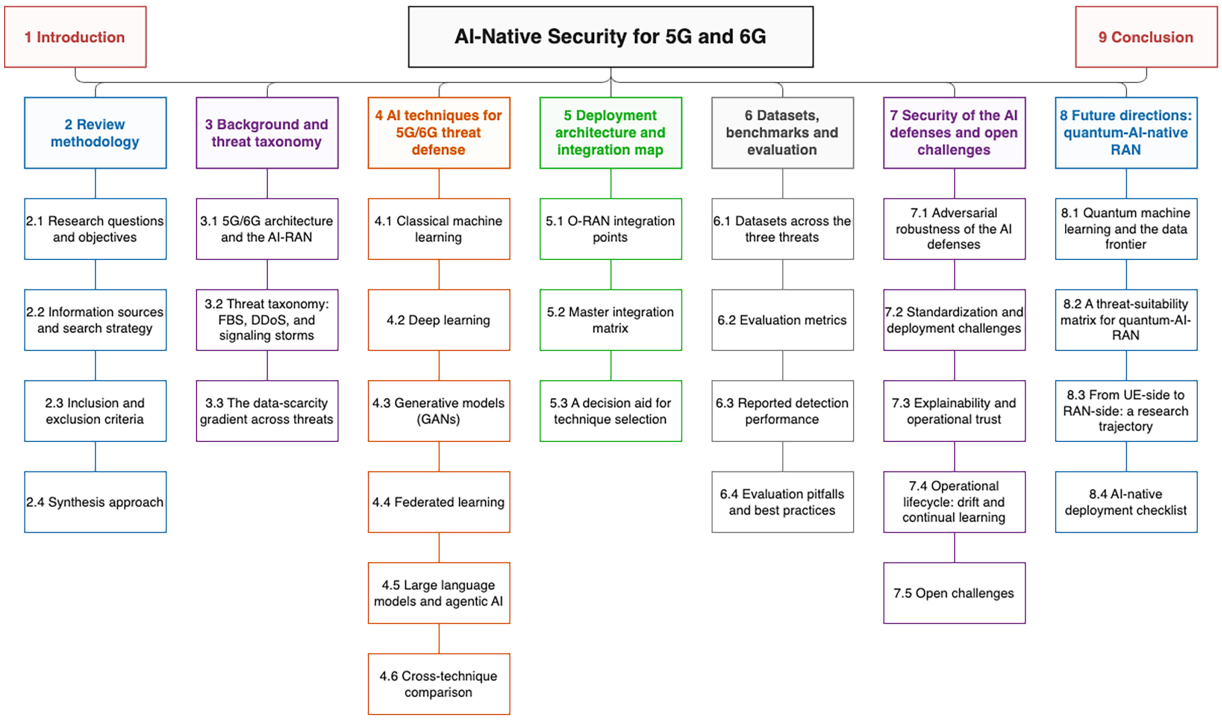

Figure 1: Structure of this survey.

2 Review Methodology

This survey follows a structured, semi-systematic protocol informed by the PRISMA guidelines. We describe it as semi-systematic because the AI-defense literature was identified through a reproducible database search, whereas the foundational threat-background literature was assembled by citation tracing (Section 2.2); we report this hybrid procedure explicitly rather than claim full PRISMA compliance. The protocol comprises four elements: the research questions that define the scope (Section 2.1), the information sources and search strategy used to identify candidate studies (Section 2.2), the inclusion and exclusion criteria applied during screening (Section 2.3), and the approach used to synthesize the retained studies (Section 2.4). The completed PRISMA 2020 checklist, reporting the location of each of the 27 items and identifying those that are not applicable to a review of engineering methods, is provided in Appendix A.

2.1 Research Questions and Objectives

The survey is organized around three research questions that correspond to the *what*, the *where and how*, and the *what next* of AI-based defense for 5G and 6G networks.

- **RQ1.** Which AI techniques are used to defend 5G and 6G networks against false base station, DDoS, and signaling-storm threats, and by what mechanism does each operate?
- **RQ2.** Where within the 5G/6G and O-RAN architecture do these techniques integrate, and how are their data sources, inference points, and latency constraints determined by that placement?
- **RQ3.** What challenges remain unresolved, and under what conditions can a quantum-AI-native defense within the RAN address the data-scale and data-scarcity frontier?

RQ1 governs the technique survey in Section 4; RQ2 governs the integration analysis in Section 5; RQ3 governs the challenges and future directions in Sections 7 and 8. Each research question is revisited and answered explicitly at the end of the corresponding section.

2.2 Information Sources and Search Strategy

The primary information source was the Scopus abstract and citation database, queried programmatically through the Publish or Perish tool to ensure reproducibility of the result sets. Scopus was selected for its coverage of the IEEE, ACM, Springer, Elsevier, and MDPI venues in which the relevant communications-security literature appears. The search was conducted in June 2026 and covered publications from January 2017 to June 2026, a window that captures the emergence of learning-based cellular defense while remaining contemporary with the O-RAN and IMT-2030 standardization activity.

The search strategy was decomposed into ten thematic query groups, each combining a technique facet, a threat facet, and a network facet with Boolean operators. The groups are summarized in Table 2. This decomposition was adopted in preference to a single broad query because the constituent literatures differ by orders of magnitude in volume: DDoS detection in cellular networks is a mature topic, whereas signaling-storm detection is a nascent one, and a single query tuned for recall on one would have suppressed the other.

Table 2: Thematic query groups and records identified (Scopus, January 2017–June 2026).

Group	Thematic Focus	Records
A	DDoS detection with AI in 5G/6G	162
B	Signaling-storm/RRC-flooding detection	14
C	GAN/generative augmentation for attack detection	269
D	Federated learning for intrusion detection	494
E	Classical ML/deep learning for FBS detection	26
F	Quantum machine learning for network security	432
G	AI-native 6G and O-RAN security framing	237
H	Adversarial robustness of ML-based defense	625
I	Datasets and benchmarks for intrusion detection	307
J	LLM and agentic AI for cellular security	1055
	Total identified	**3621**

The threat-background literature for Section 3, comprising the foundational false-base-station attack, countermeasure, and protocol-security studies, was assembled as a curated reference set rather than through the AI-keyword query groups, since these works predate the AI-defense literature and would not surface under technique-facet queries. The database search was additionally complemented by a hand-search of the 2024–2026 proceedings of major security and networking venues, including USENIX Security, NDSS, IEEE S&P, ACM CCS, INFOCOM, MobiCom, and SIGCOMM.

2.3 Inclusion and Exclusion Criteria

A record was included if it satisfied all of the following criteria: (i) it applied an AI, machine-learning, deep-learning, federated-learning, generative, or quantum-learning technique; (ii) the technique was directed at detecting or mitigating a false base station, DDoS, signaling-storm, or a closely related 5G/6G access- or core-network threat; (iii) it was a peer-reviewed publication or a citable preprint dated between 2017 and 2026; and (iv) it described the method in sufficient detail to permit classification by technique and deployment location.

A record was excluded if it satisfied any of the following criteria: (E1) it presented only a non-AI countermeasure; (E2) it addressed a non-cellular domain without transferable relevance; (E3) it lacked technical detail sufficient for classification; (E4) it constituted a superseded earlier version of an included work; or (E5) it was not available in English. Foundational threat-background references in Section 3 were exempt from inclusion criterion (i) and, correspondingly, from exclusion criterion (E1), as their role is to characterize the threats rather than to provide AI-based defenses.

2.4 Synthesis Approach

The retained records were de-duplicated across the ten query groups, reducing the 3621 identified records to 3141 unique records. Title and abstract screening against the criteria of Section 2.3 excluded the large majority as off-topic, non-AI, non-cellular, or insufficiently detailed for classification, and full-text eligibility assessment of the remainder retained the AI-defense studies that form the systematic component of the corpus. The flow is reported at the level of identified (3621), de-duplicated (480 duplicates removed), unique (3141), and included records. Together with the curated threat-background and standardization references identified by citation tracing, the cited corpus comprises 177 studies; the identification, de-duplication, and screening flow is shown in Fig. 2.

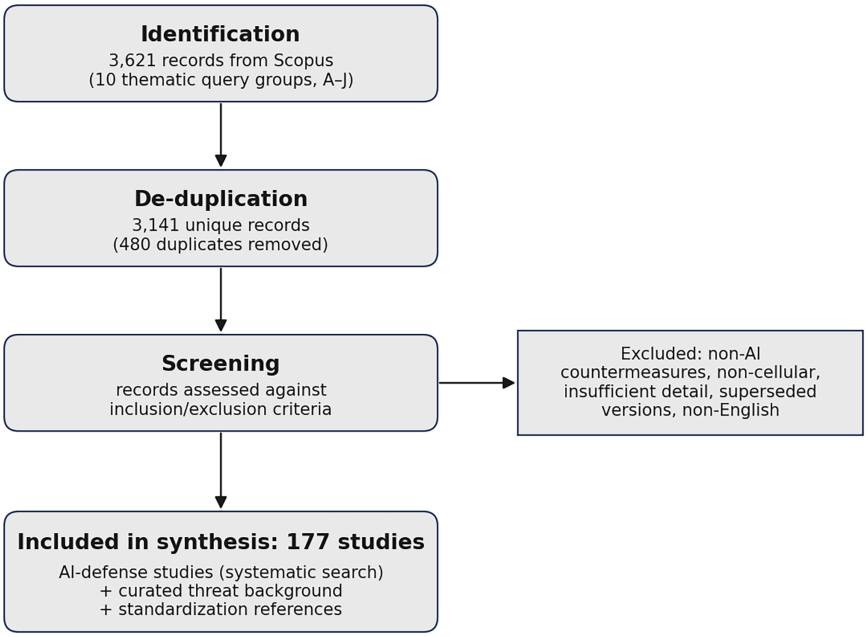

Figure 2: PRISMA-style flow of the systematic search. The 177-study corpus combines the AI-defense studies retained by the systematic Scopus search with a curated set of foundational threat-background and standardization references (Section 2.2).

Each included AI-defense study was then characterized along three axes: the AI technique family it employs, the threat it addresses, and the architectural location at which its model operates. To impose comparability across an otherwise heterogeneous literature, every technique surveyed in Section 4 is described using a single fixed template: operating principle, integration point (*where*), data-to-decision pipeline (*how*), threat fit, data regime, strengths and limitations, and representative studies. This template is the instrument through which the survey converts a collection of individual results into the integration map of Section 5. The classification additionally exposed which technique–threat combinations are densely populated and which are absent from the literature; the latter are reported as open directions in Section 8.

3 Background and Threat Taxonomy

This section establishes the technical context required for the remainder of the survey. Section 3.1 summarizes the elements of the 5G/6G architecture that determine where a defense can observe and act, including the O-RAN intelligent control loop that makes native AI deployment possible. Section 3.2 develops a taxonomy of the three threats considered in this work. Section 3.3 then advances the observation that organizes the rest of the paper: the three threats differ systematically in the availability of training data, and this gradient helps organize the trajectory of AI techniques surveyed in Section 4.

3.1 5G/6G Architecture and the AI-RAN Control Loop

A cellular network is conventionally partitioned into the user equipment (UE), the radio access network (RAN), and the core network. In 5G, the RAN consists of next-generation base stations (gNBs) and the core is the service-based 5G Core (5GC). Communication is further divided into a control plane, which carries signaling for connection management, mobility, and security, and a user plane, which carries subscriber traffic. The security procedures that protect a subscriber, including mutual authentication and key agreement and the subsequent ciphering of traffic, are specified in the 5G security architecture [20]. A defense can be positioned at any of these locations, and the choice is consequential: the UE observes the radio environment directly but in isolation, the gNB and RAN observe many devices but only their own cell, and the 5GC observes aggregated signaling but is removed from the air interface (Fig. 3).

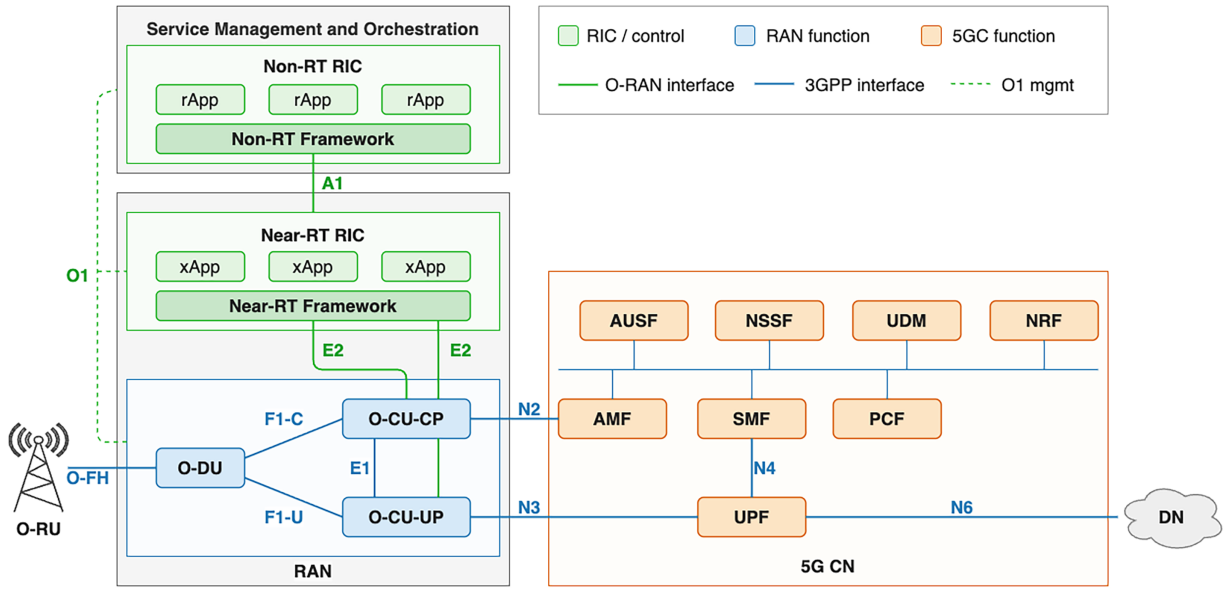

Figure 3: The 5G/6G and O-RAN architecture.

The decisive property for false base station attacks lies in the initial access procedure. Before a UE can authenticate, it must select a cell, and it does so by measuring the reference signals of nearby base stations and reading the unprotected system information broadcast on the downlink, governed by the idle-mode procedures of the standard [21]. The radio resource control (RRC) protocol then establishes a connection, after which non-access stratum (NAS) signaling with the core triggers authentication [22]. The structural vulnerability is that the information used for cell selection is transmitted, and acted upon, before any authentication has occurred. A device cannot, in general, distinguish a legitimate base station from an impostor at the moment it commits to attachment (Fig. 4).

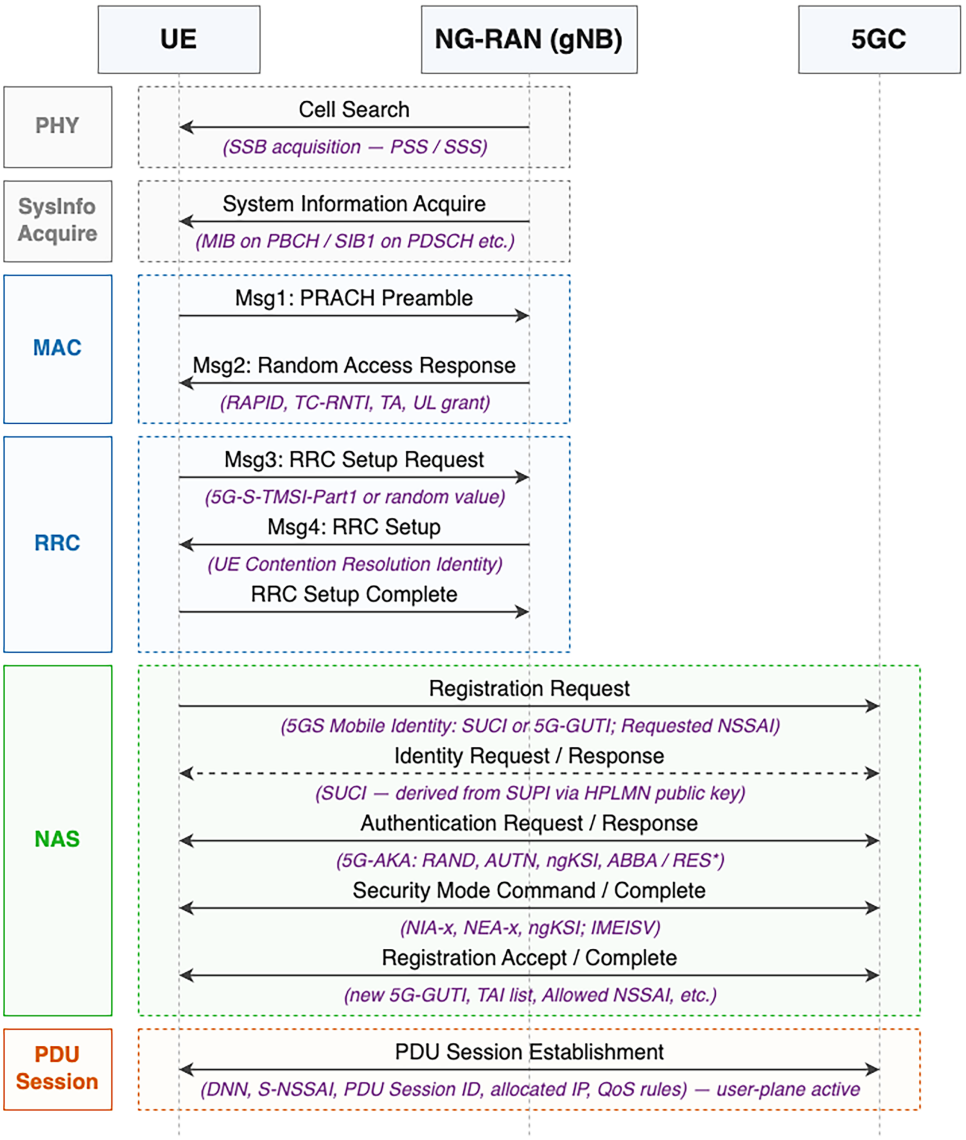

Figure 4: The initial-access procedure in 5G. The pre-authentication window (cell search and RRC connection setup, which precede NAS authentication) is the interval during which the UE reads and acts upon unauthenticated broadcast system information; this window is the structural basis of the FBS threat.

Formal and systematic analyses of the 4G and 5G control-plane procedures have repeatedly confirmed that this and related pre-authentication windows admit exploitable behaviors [23–27]. This

pre-authentication window is specific to the FBS, but the underlying architecture is not: the partition into UE, RAN, and core fixes the observation points for all three threats. The signaling storm and the distributed denial of service are observed not at the air interface but at the RAN aggregate and the core control plane.

The 6G vision, as expressed in the IMT-2030 framework, designates AI as a native capability of the network rather than an external adjunct [11]. The architectural mechanism that operationalizes this designation is the Open RAN (O-RAN) RAN Intelligent Controller (RIC), defined in the O-RAN Alliance architecture [28]. The RIC defines two control loops at different time scales. The non-real-time RIC hosts applications termed *rApps* and operates above the one-second time scale, making it suitable for training, analytics, and policy generation over aggregated data. The near-real-time RIC hosts applications termed *xApps* and operates between ten milliseconds and one second, making it suitable for inline detection and control actions on live traffic [9,12]. These two slots, together with the UE on-device location and the dedicated 5GC network function, constitute the set of deployment points to which Section 5 maps each AI technique. The significance of the RIC for the present survey is that it provides standardized, named locations at which a learning-based defense can reside as a first-class network function, a possibility that did not exist in earlier generations [13,29].

3.2 Threat Taxonomy: FBS, DDoS, and Signaling Storms

The three threats considered in this survey are summarized in Table 3 and detailed below. They are selected because they span the operational spectrum from concealed and targeted to overt and systemic, and because, despite this diversity, they reduce to a common learning problem, as Section 3.3 argues. We bound the scope to these three because each has a substantial AI-defense literature and together they cover that spectrum; closely related threats such as radio jamming, physical-layer eavesdropping, and GNSS or timing spoofing share methods with those surveyed here but fall outside the access- and core-network focus we adopt, and we note them only where a surveyed technique transfers to them.

Table 3: Taxonomy of the three threats considered in this survey.

Property	False Base Station	Signaling Storm	DDoS
Character	Stealthy, targeted	Volumetric, control-plane	Volumetric, systemic
Adversary	Passive or active impostor	Malicious/malfunctioning UEs	Botnet of compromised devices
Primary objective	Identity exposure, tracking, message injection, downgrade	Control-plane exhaustion	RAN/core exhaustion
Best observation point	UE radio environment; RAN aggregate reports	RAN/5GC control plane	RAN/5GC; network slice
Principal difficulty	Pre-authentication, unauthenticated broadcast	High volume, weak attributability	Malicious vs. legitimate surge at scale

False base stations

A false base station, also termed a rogue base station or, in its passive form, an IMSI-catcher, is an unauthorized transmitter that impersonates legitimate network infrastructure by broadcasting attractive

system information so that nearby devices select it [5,30]. The adversary may be passive, merely collecting identifiers and signaling, or active, engaging the victim device in protocol exchanges. The attack objectives form a spectrum of escalating capability. At the least intrusive end, the FBS exposes persistent or temporary subscriber identifiers, enabling identification and location tracking; commodity hardware has made such IMSI-catchers inexpensive to construct [31], and container-based methods have shown how rogue-cell attacks can be scaled [32]. Operator-deployed small cells expand the surface further, as the first systematic assessment of 4G LTE femtocells shows [33], and recent work refines IMSI-catching to minimize the service disruption that would otherwise betray it [34]. More capable attacks inject crafted messages, including spoofed public-warning and emergency alerts [7], exploit paging procedures to infer presence [35], or force a victim onto a less protected radio access technology through bidding-down [36]. A distinct and powerful class dispenses with the full rogue cell and instead overshadows legitimate signals at the physical layer, injecting or modifying messages while the victim remains connected to the genuine network [6,8]. Beyond the rogue cell itself, a family of protocol-level attacks exploits weaknesses in the authentication and session procedures, including impersonation in 4G networks [37], the Ghost Telephonist hijack of circuit-switched fallback [38], key-reinstallation and replay attacks on the NAS layer [39,40], and privacy threats in the authentication and key agreement protocol that enable activity monitoring [41]. Systematic studies have catalogued practical attacks on the privacy and availability of 4G and 5G access [42,43], device identification from the attach request [44], and the leakage of application identity from encrypted traffic through fingerprinting [45,46]. The 5G subscription concealed identifier mitigates the simplest identity-exposure attacks by encrypting the long-term identifier [47], but backward compatibility and the unauthenticated broadcast surface ensure that the threat persists in practice, as the real-world incidents of Section 1 demonstrate. The standardization response is itself ongoing: the 3GPP study on 5G security enhancements against FBS catalogues candidate countermeasures, such as digital signing of system information, together with their deployment costs and residual limitations [48].

Conventional false base station detection

Before the learning-based methods that are the subject of this survey, a substantial body of work approached FBS detection through hand-crafted rules, signatures, and measurement heuristics, and this work establishes the baseline that AI methods seek to surpass. Crowdsourced and distributed systems aggregate measurements from many devices or sensors to locate anomalous cells, as in FBS-Radar [49] and the city-wide SeaGlass sensor network [50]. Device-side applications and catcher-catchers flag suspicious cells from locally observable indicators [51,52], and device-centric protocol monitoring extends this to richer signaling features [53]. Specification-based methods detect misbehavior by checking observed traffic against the expected protocol [54], radio-frequency fingerprinting attributes transmissions to forensic signatures [55], and multi-step approaches combine indicators to detect both the rogue station and the attacks it enables [56]. Further proposals address detection during handover [57], defensive link routing [58], and the verifiable authentication of public-warning messages [59], while measurement studies assess how widely the 5G security enhancements that would blunt these attacks are actually deployed and detectable in practice [60,61]. The common limitation of these approaches, their dependence on fixed rules and anticipated signatures, is precisely what motivates the learning-based detection surveyed in Section 4.

Distributed denial of service

DDoS attacks against cellular networks aim to exhaust finite resources at the RAN or the core so that legitimate service is degraded or denied. The expansion of the attack surface in 5G is driven by two factors: the proliferation of massive machine-type and IoT devices, which can be conscripted into botnets, and the introduction of network slicing, which multiplies the logical targets that must each be protected [10]. Unlike the FBS, the DDoS adversary does not conceal its presence; the defensive difficulty arises instead

from distinguishing malicious volume from legitimate surges and from the scale at which detection must operate. The availability of compromised user terminals as attack origins has been documented in operational 5G traces [62], and location-aware variants demonstrate that denial of service can be directed with spatial precision rather than applied indiscriminately [63].

Signaling storms

A signaling storm is a control-plane denial-of-service attack in which a population of devices, whether malicious or malfunctioning, generates a surge of signaling procedures that overwhelms the RAN or the 5GC, even when the user-plane traffic volume remains modest. Characteristic instances include floods of RRC connection requests, abuse of the random-access procedure, and exploitation of the RRC inactivity timer to maximize signaling churn [9,64]; sub-use cases specific to the disaggregated O-RAN architecture have been characterized together with ML-based detection [65]. Related control-plane attacks include random-access failure through timing-advance misalignment [66], energy-efficient jamming of the downlink control channel [67], relay of unprotected RRC broadcast messages [68], and covert denial of service over the O-RAN open fronthaul [69]; in the wireless-blackhole attack a fake base station strands users by manipulating RRC broadcast [70]. Systematic fuzzing of the RAN–core interface has exposed many of the input-validation flaws these attacks exploit [71]. The signaling storm occupies an intermediate position between the FBS and the DDoS. Like the DDoS, it is volumetric and systemic; like the FBS, it is difficult to attribute, because the offending devices continually change temporary identifiers and frequently abandon procedures before authentication completes, leaving the network unable to bind the activity to an authenticated subscriber [72]. This combination of high volume and weak attributability makes the signaling storm the most analytically demanding of the three for a rule-based defense, and a natural candidate for learning-based detection.

Table 4 complements the conceptual taxonomy by classifying the primary-literature attacks discussed above according to their threat and attack class, thereby consolidating the substantial body of studies cited in this subsection into a single comparative view.

Table 4: Representative primary-literature attack and vulnerability studies by threat, organized by attack class.

Threat	Attack/Vulnerability Class	Representative Studies
False base station	Identity exposure, tracking (IMSI-catching)	[30–32,34]
	Message injection; physical-layer overshadowing	[6–8]
	Protocol/AKA-layer impersonation and replay	[37,38,41]
	Systematic attack surveys; traffic fingerprinting	[42,43,45]
Signaling storm	RRC/RACH floods, inactivity-timer abuse	[9,64,65]
	Control-plane variants; RAN–core fuzzing	[66,67,69,71]
DDoS	Expanded 5G surface (massive IoT, slicing)	[10]
	UE-as-attacker; location-aware denial of service	[62,63]

3.3 The Data-Scarcity Gradient across Threats

The three threats differ not only in mechanism but, more consequentially for an AI-based defense, in the availability of the labeled data on which learning depends. This difference is systematic, and it forms a gradient that the remainder of the survey treats as its organizing principle (Fig. 5).

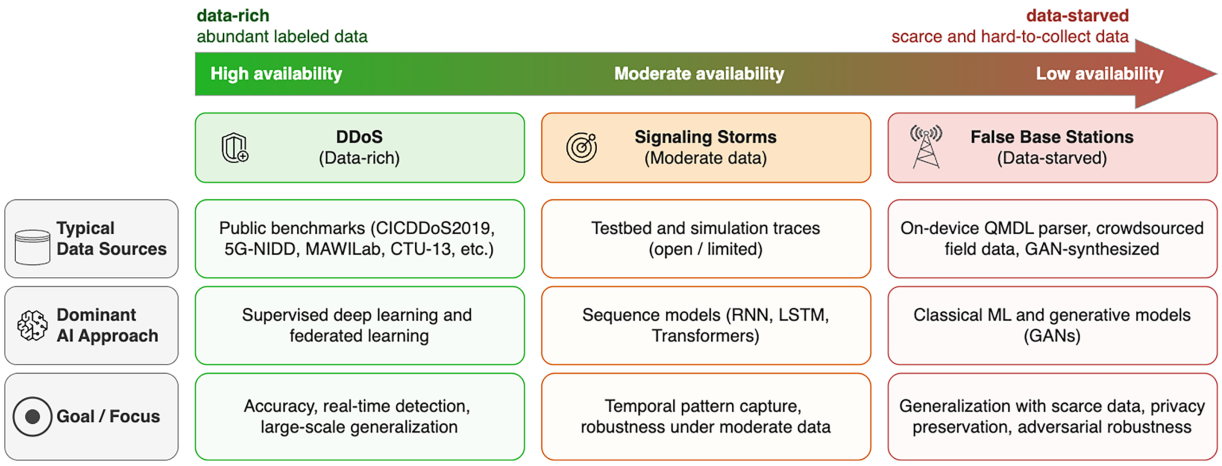

Figure 5: The data-scarcity gradient across the three threats. Data availability decreases from DDoS to signaling storms to false base stations, and the dominant AI response shifts accordingly, motivating the technique ordering of Section 4.

At one end, DDoS detection is comparatively data-rich. Public benchmark datasets of denial-of-service traffic exist and have been extended to the cellular setting, and the volumetric nature of the attack means that abundant positive examples can be generated in a testbed without difficulty. A supervised classifier can therefore be trained directly, and much of the DDoS literature surveyed in Section 4 proceeds on this assumption. Signaling-storm detection occupies the middle of the gradient. Realistic data is scarcer, because reproducing control-plane storms requires a faithful RAN or O-RAN testbed rather than a generic traffic generator, and consequently the available evidence rests largely on simulation and on a small number of testbed studies [9,72].

False base station detection sits at the data-starved extreme. There is no public corpus of real FBS attacks at scale, both because such attacks are sporadic and because capturing them requires privileged access to device-level radio measurements that are not ordinarily exposed. The field has responded in two characteristic ways, each of which prefigures a technique surveyed later. The first response is to build the data-collection capability that does not otherwise exist, by extracting and labeling the cellular diagnostic information available on the device; the MODI dataset, which captures NAS and RRC attack and normal traffic, is an instance of this approach [73]. The second response is to manufacture the missing data synthetically, using a generative adversarial network to produce realistic FBS attack samples where real ones cannot be obtained [74].

This gradient is the organizing lens of the technique survey. It offers one account of the order in which the five technique families of Section 4 are presented: where data is abundant, supervised methods, classical or deep, can be trained directly; as data becomes scarcer, the field turns first to generative augmentation, which manufactures training examples, then to federated learning, which pools distributed and privacy-sensitive data across operators without centralizing it, and finally to large language models, whose broad pretraining substitutes for task-specific labels at the data-scarce extreme. We advance the gradient as a useful axis for comparison rather than as a sole cause. The same progression also reflects general advances in machine learning, including greater representational capacity and available compute, that hold independently of data availability; the alignment between data conditions and technique adoption is therefore a correlation we observe, not a causal mechanism we establish. With that caveat, the gradient organizes the survey and motivates the quantum-enhanced methods of Section 8, whose advocates argue that quantum feature

encoding is most valuable in the high-dimensional, low-sample regime that the data-starved end of this gradient represents.

4 AI Techniques for 5G/6G Threat Defense

This section surveys the five families of AI technique that constitute the defensive repertoire against the three threats: classical machine learning (Section 4.1), deep learning (Section 4.2), generative models (Section 4.3), federated learning (Section 4.4), and large language models and agentic AI (Section 4.5). Section 4.6 then compares them directly. The families are presented in an order that broadly follows the data-scarcity gradient of Section 3.3, as illustrated in Fig. 6: classical methods, which require clean labeled features; deep methods, which require large volumes of data; generative methods, which manufacture data that is otherwise unavailable; federated methods, which pool data that cannot be centralized; and large language models, which substitute broad pretraining for task-specific data and so operate at the data-scarce end. Each family is described using a single fixed template, comprising its operating principle, its integration point within the architecture (*where*), its data-to-decision pipeline (*how*), the threat to which it is best suited, its data regime, its strengths and limitations, and the representative studies that instantiate it. This uniform treatment is what permits the cross-technique comparison of Section 4.6 and the integration map of Section 5; it also distinguishes a structured survey from a catalogue, because the *where* and *how* fields situate each method in a concrete deployment context rather than leaving it as an abstract result.

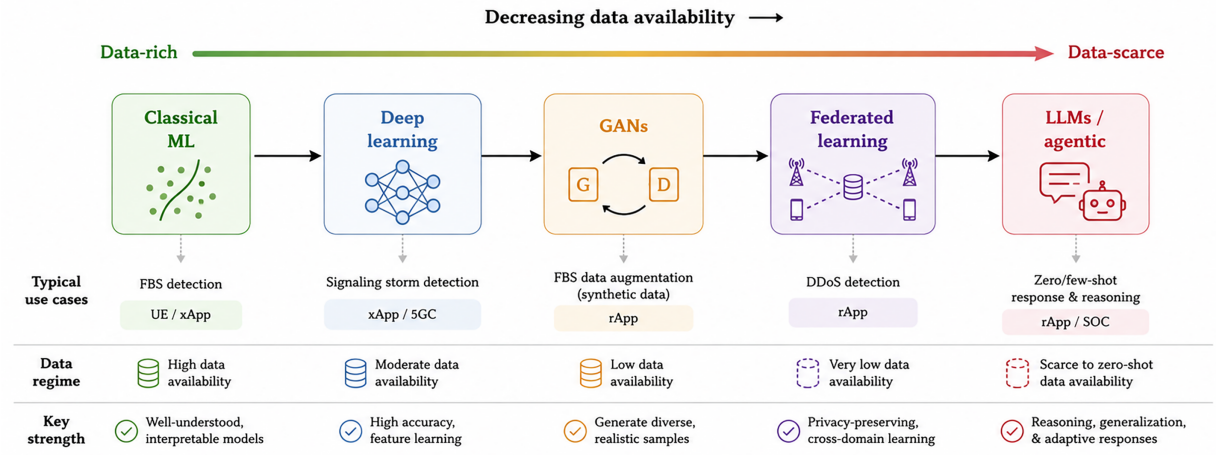

Figure 6: The AI-technique spine arranged along the data-scarcity gradient of Section 3.3, with large language models added at the data-scarce end via pretraining.

4.1 Classical Machine Learning

Principle. Classical machine learning denotes the family of shallow, non-deep models, including decision trees, random forests, support vector machines, gradient-boosted ensembles such as XGBoost, and k-nearest-neighbor classifiers, applied to features that an engineer has explicitly designed. The model does not learn a representation; it learns a decision boundary over a representation that is supplied to it. For cellular defense, the engineering of those features from radio and protocol measurements is where most of the difficulty, and most of the value, resides.

Where. Because classical models are computationally inexpensive at inference time, they admit the widest range of deployment locations of any family considered here. A trained random forest can execute on the UE itself, consuming the device's own radio measurements, or within the near-real-time RIC as an xApp

that scores aggregated measurement reports from many devices. The training step, by contrast, is performed offline. This combination makes classical ML the natural occupant of the on-device and near-real-time positions in the integration map of Section 5.

How. The pipeline begins with the collection of measurements that the device or the network already produces. For false base station detection, the canonical features are derived from the reference signal received power and related radio measurements, supplemented by fields read from the broadcast system information; from these, a classifier learns to separate the radio signature of a legitimate cell from that of an impostor [75]. The features are extracted, normalized, and presented to the trained model, which emits a per-sample or per-cell verdict. The interpretability of the resulting decision, traceable to specific feature values, is an operational advantage when an analyst must justify a response.

Threat fit. Classical ML is best suited to false base station detection, the data-starved end of the gradient. The reason is instructive: although labeled FBS data is scarce in aggregate, each individual sample is rich in engineered features, and shallow models generalize from comparatively few examples where a deep network would overfit. A network-side 5G function for FBS detection built on a machine-learning classifier demonstrates the approach at the RAN [76], while ensemble methods combining temporal and graph features illustrate its extension to more structured signals [77]. The same family also serves the data-rich DDoS setting, where feature-selection methods identify the small set of traffic attributes that most efficiently separate attack from background [78,79].

Data regime. Classical methods are the appropriate choice when labeled data is limited but informative features can be hand-designed. They degrade when the discriminating structure is not expressible in engineered features, and they do not adapt to attack variants that fall outside the feature design. Their dependence on expert feature engineering is simultaneously their strength in the low-data regime and their limitation as attacks evolve.

Strengths and limitations. The strengths are low inference cost, suitability for on-device and near-real-time deployment, interpretability, and effectiveness with modest training data. The limitations are brittleness to distribution shift, a ceiling on accuracy imposed by the quality of the hand-designed features, and limited capacity to detect novel attacks whose signatures were not anticipated. Several recent studies address the brittleness directly by combining multiple classifiers into robust ensembles, including for attack detection within Open RAN [80,81].

Representative studies. RSRP-based feature classifiers for FBS detection [75]; a machine-learning 5G network function for FBS detection [76]; ensemble and temporal-graph approaches [77,80]; rogue base station detection by machine learning in industrial settings [82,83]; hybrid feature-selection classifiers for 5G DDoS and for the protection of 5G network slices [78,84]; and lightweight two-stage classifiers that balance detection accuracy against operational efficiency [85]. In-device detection of control-plane threats from the temporal and spatial consistency of broadcast messages [86] and latency-anomaly detection of SIMBox fraud at the cellular edge [87] extend the family to the device and the edge.

4.2 Deep Learning

Principle. Deep learning replaces hand-designed features with representations learned directly from data. Three architectural families recur in cellular defense. Convolutional networks extract spatial structure from signal or image-like representations; recurrent networks and their gated variants, together with transformers, model the temporal structure of signaling and traffic sequences; and autoencoders learn a compressed model of normal behavior whose reconstruction error serves as an unsupervised anomaly score. The common premise is that the discriminating structure is too complex to specify by hand and must be learned.

Where. Deep models are more demanding at both training and inference than classical models, which shifts their natural deployment locations. Training is performed offline or within the non-real-time RIC. Inference may occur within the near-real-time RIC as an xApp when the model is compact enough to meet the sub-second budget, or at the 5GC control-plane boundary where signaling sequences are observed. Lightweight and quantized variants are an active concern precisely because the near-real-time budget constrains model size.

How. For signaling-storm detection, the input is a sequence of control-plane events, and a recurrent or attention-based model learns the temporal signature that distinguishes a storm from legitimate high load, a pipeline shown in Fig. 7. A temporal-attention LSTM applied to next-generation application-protocol signaling at the 5GC boundary exemplifies this pipeline [88]; adaptive-threshold and extreme-value-theory detectors for RRC signaling storms in Open RAN provide the strong statistical baselines against which such sequence models are measured [89,90]. For DDoS, the input is flow or packet-level traffic, and convolutional, transformer, or hybrid models classify it; transformer-based detectors operating on cellular intrusion datasets illustrate the current state of the art [91,92], including dual-branch transformers over multivariate 5G user-plane key performance indicators [93], and convolutional and mixture-of-experts architectures have been applied to 5G intrusion detection [94]. Where labels are scarce, autoencoders trained only on normal traffic can flag deviations from it without requiring attack examples [95,96].

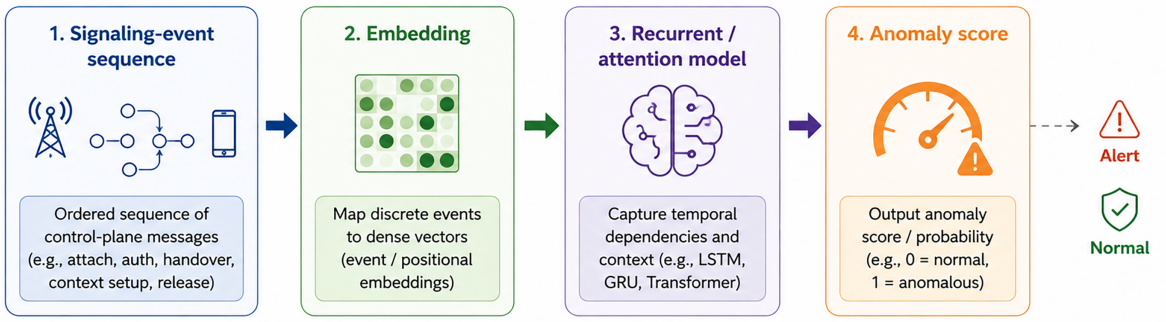

Figure 7: A representative deep-learning detection pipeline for control-plane signaling.

Threat fit. Deep learning is best suited to the signaling storm, whose defining feature is temporal: the attack manifests as an abnormal sequence of control-plane procedures rather than as a static per-sample signature, and sequence models are designed for exactly this structure. It is equally effective against DDoS, where the abundance of traffic data satisfies its appetite for examples. It is least naturally suited to false base station detection, where the scarcity of data invites overfitting unless augmentation or transfer learning is applied; the one prominent deep-learning FBS result in this survey is a quantum-enhanced model treated separately in Section 8 [97].

Data regime. Deep methods are data-hungry. They excel where data is abundant, as in DDoS, and where temporal structure provides additional signal, as in signaling storms. In the data-starved FBS regime they require either the generative augmentation of Section 4.3 or explicit transfer learning to remain viable.

Strengths and limitations. The strengths are high accuracy, the elimination of manual feature engineering, and a native capacity to model temporal and spatial structure. The limitations are a dependence on large training sets, a computational cost that can conflict with the near-real-time latency budget, and reduced interpretability, which has prompted a parallel line of work coupling deep detectors with explainability methods [98].

Representative studies. Temporal-attention LSTM for 5GC signaling anomalies [88]; adaptive-threshold and extreme-value-theory detectors for RRC storms [89,90]; transformer, spatiotemporal, and hybrid CNN–LSTM models for cellular DDoS [78,91,92]; autoencoder-based unsupervised anomaly detection [95,96]; and explainable real-time DDoS detection [98].

4.3 Generative Models

Principle. A generative adversarial network learns the distribution of a dataset by training a generator to produce samples that a discriminator cannot distinguish from real ones. In cellular defense the objective is rarely generation for its own sake; it is augmentation. By synthesizing realistic attack samples, a GAN compensates for the scarcity and class imbalance that otherwise cripple a supervised detector. This makes the generative family the field's direct response to the data-starved end of the gradient.

Where. Data synthesis is a training-time activity, not an inline one. A GAN therefore resides offline or within the non-real-time RIC as an rApp, where it manufactures or rebalances the training corpus. The detector that consumes the synthetic data is then deployed wherever its own family dictates, whether on the UE, in an xApp, or at the core. The generative model and the deployed detector occupy different locations in the integration map, a separation that Section 5 makes explicit.

How. The pipeline, depicted in Fig. 8, trains a GAN on the limited real samples that do exist, generates additional synthetic attack samples, and augments the training set of a downstream classifier with them. For false base station detection, a GAN has been used to create realistic FBS attack data in 5G where real captures are unavailable [74]. For DDoS, a conditional tabular GAN synthesizes 5G-core traffic to train a detector [99], broader studies compare GAN- and diffusion-based augmentation for intrusion detection and pair synthesis with explainability [100,101], and generative synthesis is combined with oversampling to correct class imbalance in Open RAN intrusion detection [102].

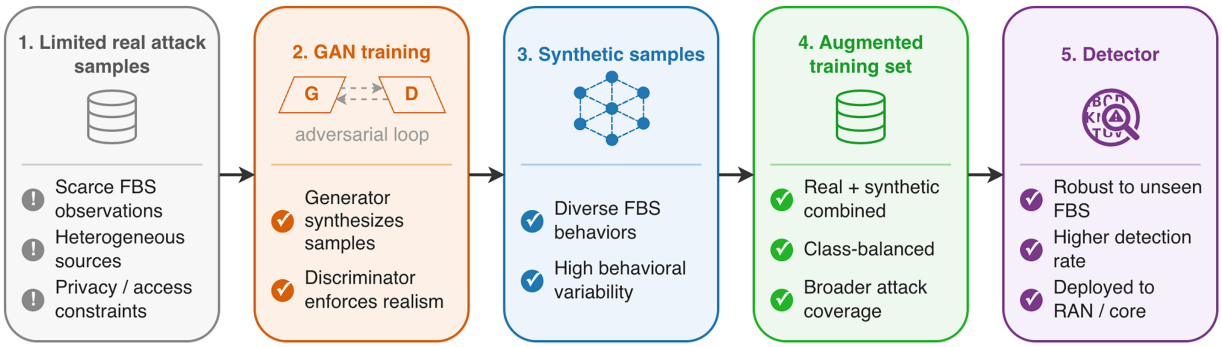

Figure 8: GAN-based data augmentation as the field's response to FBS data scarcity.

Threat fit. Generative augmentation is best suited to false base station detection, where data is scarcest and the alternative, building an on-device collection pipeline such as MODI, is costly. It also serves DDoS and intrusion detection wherever class imbalance suppresses the detection rate of minority attack classes [103]. It is least relevant to the data-rich settings where sufficient real examples already exist.

Data regime. In cellular defense the generative family is deployed for the data regime it addresses: scarcity and imbalance. It requires a seed of real samples from which to learn, and its value diminishes as that seed grows, since abundant real data renders synthesis unnecessary.

Strengths and limitations. The strengths are the direct mitigation of data scarcity and class imbalance and a measurable improvement in minority-class detection. The limitations are the training instability

characteristic of adversarial objectives, the difficulty of verifying that synthetic samples are realistic rather than artifactual, and a dual-use hazard: the same generative capability that augments a defender's training set can be turned to craft adversarial attacks against detectors, a risk examined in Section 7 [104,105].

Representative studies. GAN synthesis of FBS attack data [74]; conditional tabular GAN for 5GC DDoS traffic [99]; comparative GAN and diffusion augmentation for 5G intrusion detection [100]; conditional GAN dataset synthesis [103]; synthesis coupled with explainable detection [101]; and a systematic account of generative models and large language models across the 6G architecture, which situates defensive augmentation within the wider set of generative network functions [106].

4.4 Federated Learning

Principle. Federated learning trains a shared model across many nodes without moving their data to a central location. Each node trains locally on its own data and transmits only model updates, which a coordinator aggregates into a global model that is redistributed. The motivation in cellular defense is twofold: the data that would most improve detection is distributed across base stations and operators, and it is privacy-sensitive and often legally constrained, so it cannot be pooled directly. Federated learning is therefore the response to a data regime in which data is not absent but unshareable.

Where. Federated learning is intrinsically distributed. The non-real-time RIC is the natural coordinator, orchestrating training rounds across gNBs or across operators, while the resulting global model is deployed downstream to xApps, network functions, or devices. The technique thus spans the integration map rather than occupying a single point, and its defining infrastructural requirement is the aggregation and communication fabric that connects the nodes.

How. Each participating node, whether a base station, an edge server, or an operator domain, trains a local detector on its own traffic; the updates are aggregated, often with secure-aggregation or privacy-preserving protocols, into a global model, an arrangement shown in Fig. 9. For DDoS detection across 5G-core virtualized network functions, a federated scheme combining gradient boosting and deep learning has been demonstrated [107], and further studies extend federated DDoS detection to cyber-physical and edge settings and to public benchmark traffic [108–110]. Transfer-learning variants address the non-identical data distributions across nodes [111]. Federated graph neural networks and mixture-of-experts detectors accommodate the heterogeneity of distributed 5G traffic [112,113], and outlier-aware federated autoencoders detect control-plane anomalies, such as those in the packet-forwarding control protocol of the 5G core, under limited labels [114].

Threat fit. Federated learning is best suited to DDoS, whose detection benefits most from a cross-domain view that no single operator possesses, and whose traffic is both abundant and privacy-sensitive. It applies equally to the broader intrusion-detection problem in 5G and beyond [15,115]. Its application to false base station detection across operators is, notably, absent from the surveyed literature, a gap that appears as an empty cell of the technique–threat coverage matrix presented in Section 5 and is carried into the open challenges of Section 7.5.

Data regime. Federated learning addresses the regime in which data exists in quantity but is distributed and unshareable. It does not manufacture data, as a GAN does, nor does it reduce the data requirement, as a classical model does; it relaxes the constraint that the data be centralized.

Strengths and limitations. The strengths are privacy preservation, cross-operator collaboration, and scalability to many nodes. The limitations are communication overhead, sensitivity to non-identically distributed data across nodes, and a distinctive security exposure: because training is delegated to untrusted

nodes, federated learning is vulnerable to model-poisoning, which has motivated adversarially resilient and verifiable-privacy variants examined in Section 7 [116,117].

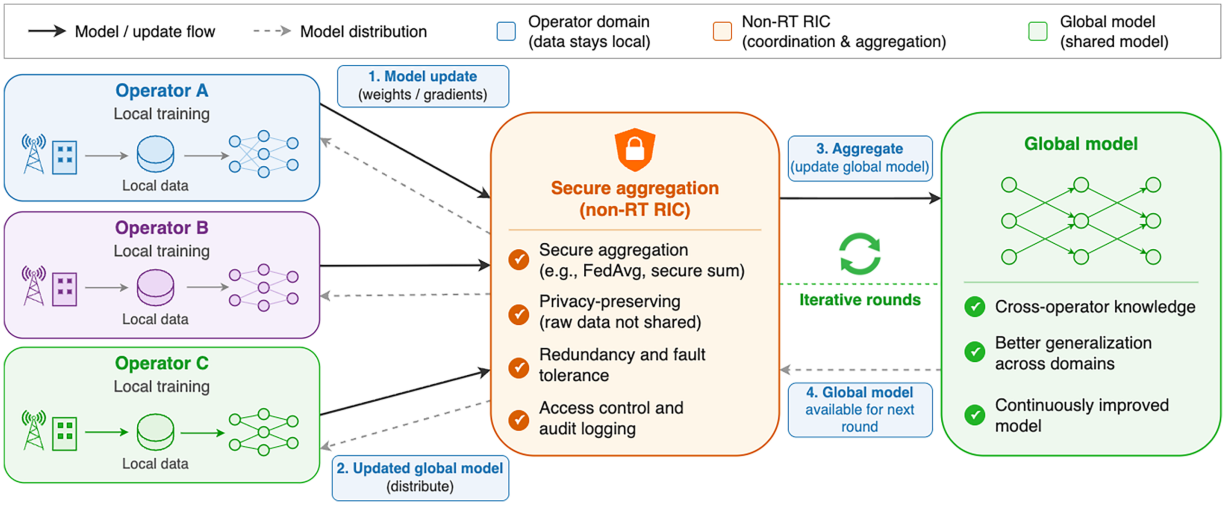

Figure 9: Federated learning across operator and RAN domains, coordinated by the non-real-time RIC.

Representative studies. Federated gradient-boosting and deep learning for 5GC DDoS [107,118]; federated DDoS detection for cyber-physical and edge networks [108,109]; federated detection on public DDoS benchmarks [110]; federated transfer learning for non-IID 5G data [111]; and surveys of federated intrusion detection for 5G and beyond [15].

4.5 Large Language Models and Agentic AI

Principle. Large language models (LLMs) are large transformer networks pretrained on broad corpora; they are applied to security through their capacity for few-shot and zero-shot reasoning over text and text-serialized data and, in agentic configurations, for multi-step planning and action. Two uses recur: as detectors that classify or reason about serialized traffic, logs, and specifications, and as agents that close the loop from detection to response. They are distinguished from the deep models of Section 4.2 by pretrained knowledge that reduces the need for task-specific labeled data, inverting rather than satisfying the data assumption of the earlier families.

Where. LLMs are the most computationally demanding family surveyed. Inference latency and resource cost place them at the non-real-time RIC (rApp), at the edge or cloud, or in the security-operations layer, not on the inline near-real-time path. Local, on-premise deployment is an active concern where 5G configuration data is too sensitive to transmit to a cloud-hosted model [119], and agentic frameworks reside at the non-real-time RIC, orchestrating detection-and-response loops [120–122].

How. For detection, traffic, flows, or logs are serialized to text and presented to the LLM, which classifies or flags anomalies, frequently with a natural-language explanation [123–127]. A second pattern applies LLMs to specifications and protocols, parsing 3GPP documents to extract security requirements and detect vulnerabilities [128,129]. Retrieval-augmented and knowledge-graph variants ground the model in cyber-threat intelligence to limit hallucination [130,131]. Agentic frameworks add a planning-and-action loop that proposes and executes mitigations [120,121]. These two roles, detection and agentic response, are illustrated in Fig. 10.

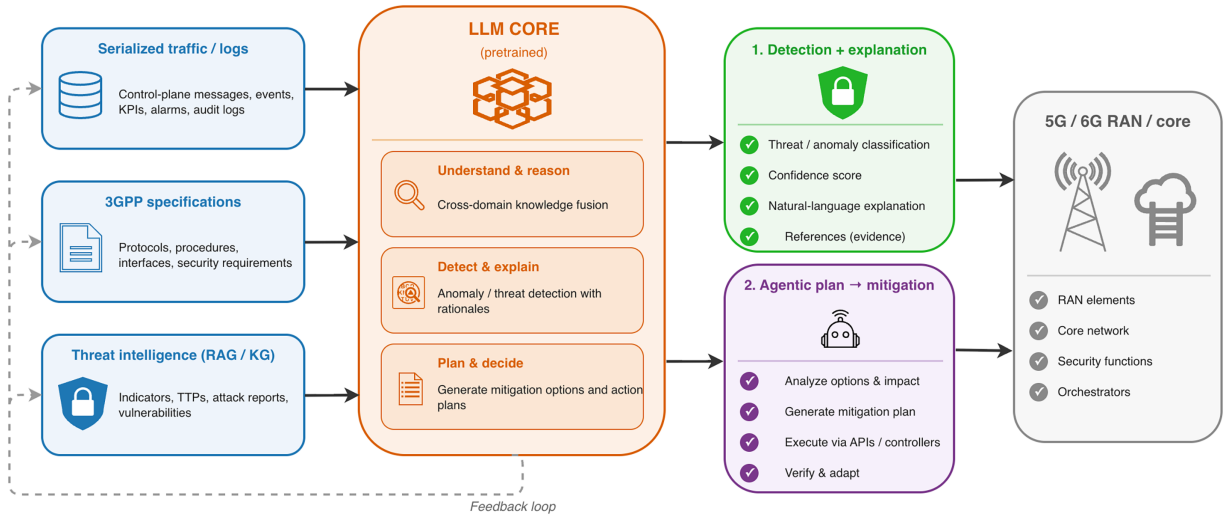

Figure 10: The two roles of large language models in 5G/6G defense. Serialized traffic and logs, 3GPP specifications, and retrieval-grounded threat intelligence are processed by a pretrained LLM that both detects and explains threats and, in agentic configurations, plans and executes mitigations against the RAN or core, closing the loop. The latency profile sites the model at the non-real-time RIC or in the security-operations layer.

Threat fit. LLMs suit two regimes. The first is the data-scarce and zero-day regime: pretrained knowledge enables zero-shot and few-shot detection where labeled examples are absent, aligning LLMs with the FBS and specification-analysis end of the gradient [128,132]. The second is the response-and-operations regime that the other families do not address, namely triaging alerts, explaining decisions, and closing the loop to mitigation [120,131]; autonomous threat-hunting agents already demonstrated within 5G networks indicate the operational regime that agentic LLM configurations are being positioned to occupy [133]. Across the three threats they have been applied to 5G/6G anomaly and intrusion detection [123,130,134], to O-RAN-sited security [135,136], and to 5G protocol and vulnerability analysis relevant to FBS and signaling [128,129]. The same transformer machinery is concurrently being adopted for 6G network functions beyond security, which is what makes its co-location with the defense pipeline architecturally plausible [137].

Data regime. LLMs invert the data assumption of the supervised families: pretraining substitutes for task-specific labels, making them most valuable where labels are scarce. The cost is a different scarcity, namely compute and, for fine-tuning, domain data, together with a dependence on grounding to control hallucination.

Strengths and limitations. The strengths are few- and zero-shot capability, native explainability, the ability to reason over heterogeneous text and specifications, and a path to agentic response. The limitations are an inference cost incompatible with the near-real-time budget, hallucination and the consequent need for retrieval grounding, sensitivity to prompt and serialization, privacy concerns that motivate local deployment [119], and a large attack surface of their own, including prompt injection and adversarial inputs, which is an instance of the security-of-AI problem of Section 7.

Representative studies. Anomaly and intrusion detection in 5G/6G [123–125,130,134]; O-RAN-sited and agentic frameworks [120,121,132,135,136]; 5G protocol and vulnerability analysis [119,128,129]; and explainable, retrieval-grounded LLM detection [126,127,131]. Foundational work on adapting LLMs to networking tasks [138] and a survey of the dual role of LLMs in network security [139] frame the broader trajectory.

4.6 Cross-Technique Comparison

The five families are not competitors for a single role but complementary instruments selected by the data regime and the deployment constraint. Their relationship to the data-scarcity gradient is direct. Classical machine learning is the instrument of last resort when labeled data is scarce but features are informative, and it is the only family that deploys comfortably on the device itself. Deep learning is the instrument of choice when data is abundant or when temporal structure carries the signal, at the cost of a computational budget that strains the near-real-time slot. Generative modeling is not a detector at all but a manufacturer of the data the other detectors need, active when scarcity or imbalance would otherwise defeat them. Federated learning is an orchestration discipline rather than a model architecture, invoked when the obstacle is not the quantity of data but the impossibility of centralizing it. Large language models invert the data premise altogether, substituting broad pretraining for task-specific labels and extending the repertoire beyond detection to explanation and agentic response, at a computational cost that confines them to the non-real-time layer.

Table 5 summarizes the families along the dimensions that govern their selection. Two columns deserve emphasis because they are the ones most often omitted from the primary literature and most needed by a practitioner. The deployment column records *where* each family is sited, and the latency column records the constraint that placement imposes; together they convert an accuracy result into a deployability judgment. The pattern across the table is a central organizing observation of this survey: each family aligns with the data conditions of its target threat, and the five together trace the same broad progression toward the data-starved regime that Section 3.3 described among the threats. We present this correspondence as an organizing structure rather than a causal law; it is what the integration map of Section 5 formalizes.

Table 5: Comparison of the five AI-technique families. The lead-threat row records the authors' analytical judgement of each family's primary target.

Dimension	Classical ML	Deep Learning	Generative (GAN)	Federated Learning	LLMs/Agentic
Lead threat	FBS	Signaling storm	FBS (synthesis)	DDoS	Zero-day/specs
Role	Detector	Detector	Data augmenter	Orchestrator	Detector + responder
Deployment (*where*)	UE/xApp	xApp/5GC	rApp/offline	rApp (spans nodes)	rApp/SOC/edge
Latency profile	Low (inline)	Medium–high	Training-time	Training (rounds)	High (offline)
Data regime	Scarce, feature-rich	Data-hungry	Scarce/imbalanced	Distributed	Pretrained; few-shot
Interpretability	High	Low	n/a (augmenter)	Inherited	High (native)
Principal limitation	Feature brittleness	Compute vs. latency	Instability; dual-use	Poisoning; non-IID	Cost; hallucination

Table 6 maps each family to the representative studies surveyed in Sections 4.1–4.5, recording the threat each addresses and its typical deployment slot, thereby enabling a systematic comparison of the studies discussed across the five subsections.

Answer to RQ1

The AI techniques defending 5G/6G networks against FBS, DDoS, and signaling storms reduce to five families whose mechanisms are, respectively, decision-boundary learning over engineered features,

representation learning over raw or sequential data, distributional learning for data synthesis, distributed training without data centralization, and pretrained language modeling that enables few-shot reasoning and agentic response. Each operates by the pipeline detailed above, and each is matched to a threat by the data regime that threat imposes.

Table 6: Representative studies by AI-technique family, with threat focus and typical deployment slot.

Family	Representative Studies	Threat Focus	Typical Slot
Classical ML	[75–77,80]	FBS; DDoS	UE/xApp/5GC
Deep learning	[88,91,93,95]	Signaling; DDoS	xApp/5GC
Generative (GAN)	[74,99,100,102]	FBS; DDoS (imbalance)	rApp/offline
Federated	[107,108,112,114]	DDoS	rApp (spans nodes)
LLMs/agentic	[120,123,126,128]	Zero-day; DDoS; specs	rApp/SOC

5 Deployment Architecture and Integration Map

The preceding section established what each AI technique does and to which threat it is suited. This section answers the question that the primary literature most often leaves open: where in the 5G/6G architecture each technique should be deployed, and how that placement is determined. Section 5.1 specifies the four deployment points that the O-RAN architecture makes available. Section 5.2 consolidates the survey into a master integration matrix that maps technique and threat onto those points. Section 5.3 reduces the matrix to a decision aid that a practitioner can apply directly.

5.1 O-RAN Integration Points

The O-RAN architecture exposes four locations at which a learning-based defense can be sited, distinguished by what each can observe and by the latency budget each imposes. These four points, shown in Fig. 11, are the coordinate system of the integration map.

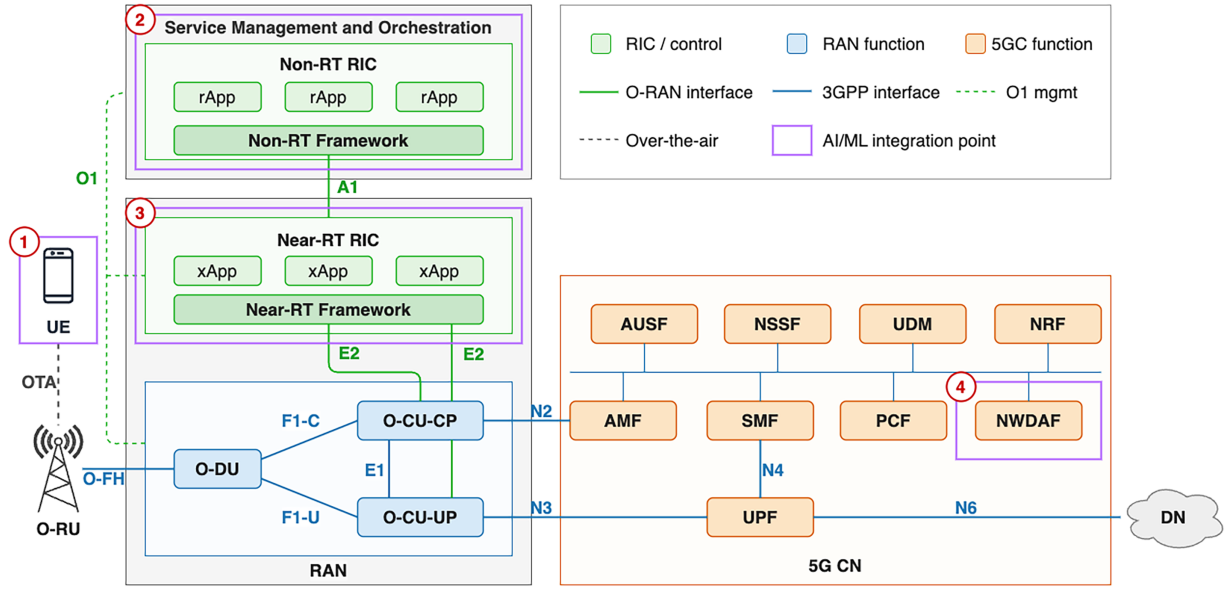

Figure 11: The four AI/ML deployment points that constitute the coordinate system of the integration map.

The **UE on-device** position places the model on the handset or module itself. Its vantage is the local radio environment, observed directly but in isolation from other devices, and its latency budget is set only by the device's own resources. It is the natural home of lightweight classical models for false base station detection, which operate on the reference-signal measurements the device already collects. The principal obstacle at this position is the acquisition of labeled device-level data, which the MODI on-device collection pipeline was constructed to address [73].

The **near-real-time RIC**, hosting *xApps*, operates on a budget between ten milliseconds and one second and observes live traffic and signaling aggregated across a cell or group of cells. It is the position for inline detection that must act on traffic as it arrives, including real-time DDoS and jamming detection; the latency of machine-learning random-access DDoS detection at this position has been measured directly, confirming its feasibility within the budget [140], and self-adaptive jamming detection has been demonstrated as an xApp [141]. Real-time frameworks for securing the O-RAN random-access channel against DDoS [142] and modular xApp-based defense frameworks embedded in the near-real-time RIC [143] further populate this position. O-RAN-compliant layer-3 attack-detection services [144], cross-layer real-time detection distributed across the Open RAN [145], and near-real-time RIC intrusion detection targeting rogue-base-station downgrade attacks [146] demonstrate layer-3 and false-base-station detection at this slot, while explainable anomaly detection addresses its operational-trust requirement [147].

The **non-real-time RIC**, hosting *rApps*, operates above the one-second budget and is the position for activities that are inherently not inline: model training, federated-learning orchestration, generative augmentation, and analytics over aggregated reports. Runtime-safety supervision of the near-real-time control loop is itself sited here [12]. Large-language-model analytics and agentic response frameworks also reside at this layer, their latency profile placing them alongside training and analytics rather than on the inline path [120,135].

The **5GC network function** position places the model within a dedicated core-network function, where it observes control-plane signaling aggregated across the network but is removed from the air interface. It is the position for network-side detection of threats whose signature appears in core signaling, including network-side false base station detection and signaling-based DDoS. A native 5G network function for FBS detection using machine learning [76], a real-time implementation of such a function [148], a signaling-DDoS detection system for the 5G standalone core [10], predictive DDoS detection built on the 5G-core Network Data Analytics Function [149], provenance-graph attack detection and attribution in the 5G core [150], and deep-learning intrusion detection for roaming traffic in the cloud-native core [151] instantiate this position.

5.2 Master Integration Matrix

Table 7 is the central artifact of this survey. It arranges the four deployment points against the three threats and records, in each cell, the technique family that the surveyed evidence places there together with a representative study. The matrix is read as a deployment guide: a defender selects a row by the operational constraint, that is, the deployment point and the latency budget it imposes, and a column by the threat, and the cell names the technique and an instance of its use.

Three features of the matrix carry the argument of the paper. First, the populated cells are not distributed uniformly, but neither do they form a simple diagonal; they follow from where each threat is observable and the latency within which it must be addressed. The FBS is the threat addressed at the widest range of deployment points, appearing in all four, because it is observable both in the device's local radio environment and in aggregated RAN and core reports. The near-real-time RIC and the 5GC network function are populated for all three threats, since every threat leaves a signature in live traffic

or in aggregated control-plane signaling; the UE position, by contrast, is populated only for the FBS, the single threat with a device-local signature; and the non-real-time position holds the training-time techniques (generative augmentation for the FBS and federated training for DDoS), while the signaling storm, which in the surveyed corpus has neither a device-local nor a training-time defense, is confined to the two control-plane rows. Second, the latency budget of each row constrains the admissible techniques independently of the threat, which is why a single threat may appear with different techniques at different positions. Third, some technique–threat combinations that the architecture permits are simply unpopulated. Because this deployment matrix is indexed by deployment point rather than by technique, the clearest example, the absence of any federated approach to cross-operator false base station detection, is a technique–threat gap that does not correspond to a single cell of the present matrix; we therefore record it in the technique–threat coverage matrix of Table 8.

Table 7: Master integration matrix: deployment point × threat, with technique family and a representative study. A dagger (†) marks a cell whose deployment slot is our inference from the technique's latency and data profile rather than one the cited source states; 3 of the 10 populated cells (30%) are so marked.

Deployment Point (Latency)	False Base Station	Signaling Storm	DDoS
UE on-device (local)	Classical ML on RSRP features [75]; on-device data via MODI [73]	—	—
Near-RT RIC/xApp (10 ms–1 s)	Rogue-BS/L3 detection xApp [29,144,146]	Inline RRC-storm/service-model/ cross-layer detection [65,72,89,145]	Real-time ML DDoS/ RACH/jamming [140–142]
Non-RT RIC/rApp (>1 s)	GAN augmentation of FBS data [74]†	—	Federated DDoS training [107]†
5GC network function (core)	Native NF for FBS detection [76,148]	Temporal-attention LSTM at 5GC [88]†	NWDAF/SA-core DDoS detection [10,149–151]

Table 8: Technique–threat coverage matrix over the surveyed corpus. ●: an established line of work; ◐: emerging or partial (few studies, or via a related technique); ○: absent from the surveyed literature. Empty (○) cells are the unexplored combinations.

Technique Family	False Base Station	Signaling Storm	DDoS
Classical ML	●	◐	●
Deep learning	◐	●	●
Generative (GAN)	●	○	●
Federated learning	○	○	●
LLMs/agentic	◐	◐	●
Quantum learning	◐	○	○

The deployment matrix above is indexed by *where* a defense runs. To make the technique–threat gaps discussed above explicit, Table 8 re-presents the same corpus indexed by *technique family* against threat. In this view its empty cells are exact: the absence of federated false base station detection, of generative and federated signaling-storm defense, and of any quantum treatment of the signaling storm or of DDoS,

denotes technique–threat combinations that the architecture permits but that the surveyed literature does not yet occupy.

5.3 A Decision Aid for Technique Selection

The matrix can be reduced to a procedure. Given a threat to defend and the dominant operational constraint, the following decision logic, summarized in Fig. 12, selects a technique family and a deployment point.

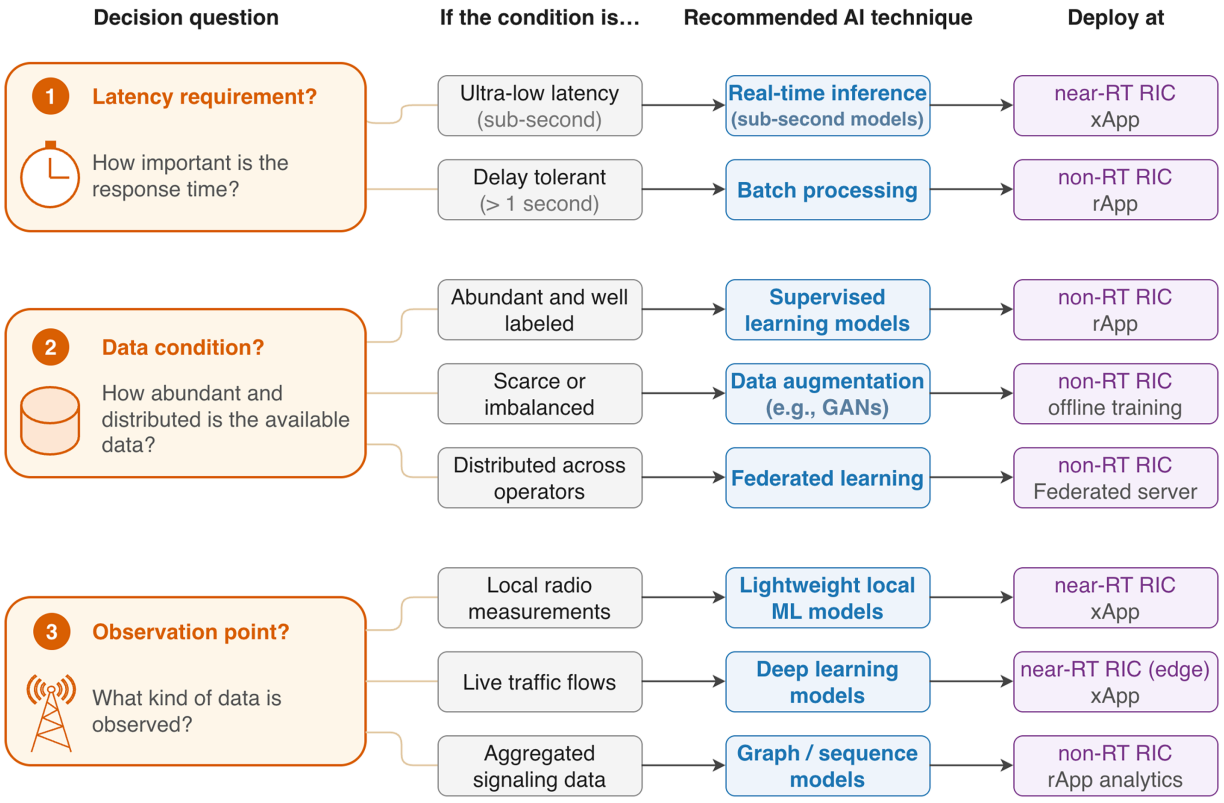

Figure 12: Decision aid mapping a (threat, constraint) pair to an AI technique and deployment point.

The first branch is the latency constraint, because it is the most restrictive. If detection must act on traffic inline, within the sub-second budget, the deployment point is the near-real-time RIC and the admissible techniques are the compact ones: classical models and lightweight deep models that meet the budget. If detection may operate over aggregated data above the one-second budget, the non-real-time RIC and the 5GC network function become available, and the heavier deep and federated methods are admissible.

The second branch is the data condition. If labeled data for the target threat is abundant, as for DDoS, a supervised deep or classical detector is trained directly. If data is scarce or imbalanced, as for FBS, the procedure inserts a generative-augmentation stage at the non-real-time RIC before the detector is trained. If data is distributed across operators and cannot be centralized, the procedure adopts federated learning, orchestrated from the non-real-time RIC.

The third branch is the observation point dictated by the threat. A threat visible only in the device's local radio environment, such as the radio signature of an FBS, favors the UE position; a threat visible in aggregated control-plane signaling, such as a signaling storm or a network-side view of an FBS, favors the

5GC network function; a threat visible in live traffic, such as a DDoS flood, favors the near-real-time RIC. Where a threat is observable at more than one point, as an FBS is at both the UE and the aggregated RAN view, the choice reverts to the latency and data branches.

Answer to RQ2

AI-based defenses integrate into the 5G/6G architecture at four O-RAN deployment points: the UE on-device position, the near-real-time RIC xApp, the non-real-time RIC rApp, and the 5GC network function. The placement of a given defense is determined jointly by the latency budget it must meet, the data condition of its target threat, and the point at which that threat is observable. The master integration matrix of Table 7 records the resulting assignments, and Fig. 12 reduces them to a selection procedure at the level of technique family and deployment slot. The further specification of model, features, retraining cadence, and packaging is a system-design task beyond the scope of this survey; the procedure is a synthesis instrument for comparing defenses, not a deployment prescription.

6 Datasets, Benchmarks and Evaluation

A survey of learning-based defense is incomplete without an account of the data on which the learning depends and the practices by which it is evaluated. This section consolidates that account. Section 6.1 inventories the datasets available for the three threats and shows that their distribution reproduces the data-scarcity gradient of Section 3.3. Section 6.2 reviews the metrics by which detectors are evaluated. Section 6.3 consolidates the detection performance reported across the surveyed methods. Section 6.4 identifies the evaluation pitfalls that recur in the surveyed literature and states the practices that mitigate them.

6.1 Datasets Across the Three Threats

The datasets available for the three threats differ in kind, not merely in size, and the difference aligns with the gradient established earlier. Table 9 summarizes the principal datasets and their character.

For DDoS, public benchmark datasets exist and are widely used. General-purpose denial-of-service corpora such as CICDDoS2019 provide labeled attack traffic that federated and deep detectors are trained and evaluated on [110], while datasets curated for the cellular setting, including 5G-NIDD, supply traffic representative of a 5G network and have become a reference benchmark for transformer-based detectors [91] and AI-driven 5G intrusion detection [152]. The cellular DDoS picture is further strengthened by open datasets derived from operational and testbed networks, including traces in which user terminals themselves act as attack origins [62] and O-RAN testbed measurements that combine user mobility with DDoS and DoS attack traffic [153]. DDoS is, in short, data-rich.

Table 9: Principal datasets and data sources for the three threats.

Dataset/Source	Threat	Type	Public	Representative Use
CICDDoS2019	DDoS	Emulated, labeled	Yes	Federated DDoS detection [110]
5G-NIDD	DDoS/ intrusion	Testbed, 5G-specific	Yes	Transformer detector [91]
Operational 5G DDoS traces	DDoS	Real operational	Yes	UE-as-attacker analysis [62]

(Continued)

Table 9 (continued)

Dataset/Source	Threat	Type	Public	Representative Use
O-RAN testbed (OpenIreland)	DDoS/ DoS	Real testbed, O-RAN	Yes	Mobility + attack traces [153]
Testbed/simulation traces	Signaling storm	Simulated/testbed	Limited	O-RAN storm detection [72]
MODI	FBS	Real on-device (NAS/RRC)	Yes	On-device collection [73]
GAN-synthesized FBS data	FBS	Synthetic	n/a	Augmented training [74]

For signaling storms, the situation is materially weaker. Reproducing a control-plane storm requires a faithful RAN or core testbed rather than a generic traffic generator, and as a consequence the available evidence rests largely on simulation and on a small number of testbed studies rather than on shared public corpora [9,72]. There is no widely adopted public signaling-storm benchmark comparable to those for DDoS, which constrains the comparability of results across studies.

For false base station detection, the data problem is most acute, and the field's two responses define the dataset landscape. The first response collects real device-level data where none was previously available: the MODI dataset captures NAS and RRC attack and normal traffic from on-device diagnostics, providing one of the few real labeled corpora for the threat [73]. The second response manufactures data synthetically, using a generative adversarial network to produce realistic FBS attack samples for training [74]. The reliance on collection pipelines and synthesis, rather than on established public benchmarks, is itself the signature of the data-starved regime.

6.2 Evaluation Metrics

Detectors in this field are evaluated predominantly with the standard classification metrics: accuracy, precision, recall, the F1 score, and the area under the receiver-operating-characteristic curve. For the imbalanced data that characterizes attack detection, where benign samples vastly outnumber attacks, accuracy alone is misleading, because a detector that never raises an alarm can attain high accuracy while detecting nothing. Precision, recall, the F1 score, and the area under the precision-recall curve are the informative metrics in this regime, and the false-positive rate is operationally decisive, since each false alarm imposes a cost on the analysts or automated responders that must adjudicate it.

A metric that the deployment perspective of this survey makes essential, yet that the primary literature reports only intermittently, is detection latency. For a detector intended for the near-real-time RIC, an accuracy figure unaccompanied by an inference-latency figure is insufficient to establish deployability, because the defining constraint of that position is the sub-second budget. The studies that report latency directly, such as the measurement of machine-learning random-access DDoS detection in Open RAN [140], are therefore disproportionately useful, and the broader omission of latency reporting is a gap in evaluation practice rather than merely in presentation.

6.3 Reported Detection Performance

Table 10 consolidates the detection performance reported by a representative set of the surveyed studies, together with the dataset or setting on which it was obtained and the deployment slot, the last classified as stated by the source or inferred by us; that the majority are inferred is itself a measure of how often the primary literature leaves the deployment point unspecified. The table is offered with a strong caveat that the

following subsection then develops: the values are as reported by the original authors, on heterogeneous datasets, threat models, and evaluation protocols, and they are *not* directly comparable. Its purpose is not to rank methods but to convey the order of magnitude of reported results and, more importantly, their spread.

Table 10: Representative reported detection performance. Values are as reported by the original authors on heterogeneous datasets and protocols and are *not* directly comparable. Entries for which the source does not state a single numeric value are summarized qualitatively. Slot per Section 5. A dagger (†) marks a deployment slot we inferred from the technique's latency and data profile rather than one the source states; 12 of the 20 slot assignments (60%) are so marked, quantifying how often the primary literature leaves the deployment point unspecified.

Study	Threat	Technique	Setting/Dataset	Reported Result (Value Where Stated)	Slot
[76]	FBS	Classical ML	5G-core NF, 7 classifiers	MLP best of seven	5GC-NF
[81]	FBS	ML anomaly	Static vs. mobile	100% static; 65%–76% mobile (recovered by retraining)	UE
[54]	FBS	Spec.-based	Testbed	≈98% accuracy, low overhead	5GC-NF†
[154]	FBS	One-class SVM	Interference-rich ISAC	Robust recall (anomaly-prioritized)	UE†
[97]	FBS	Quantum NN	B5G, quanvolutional	≈90% accuracy	UE
[10]	Signaling/DDoS	Random forest	5G SA core	≈98% accuracy	5GC-NF
[89]	Signaling	Threshold-based	O-RAN testbed	Detect ≤90 ms; 60 ms mitigation window	xApp
[90]	Signaling	Adaptive (EVT)	O-RAN, complex load	>93% acc/prec/recall, low latency	xApp
[72]	Signaling	ML	O-RAN service model	precision ≈0.99	xApp
[88]	Signaling	Attention LSTM	5GC NGAP boundary	99.8% acc; 99.85% F1	5GC-NF†
[78]	DDoS	CNN-LSTM	5G IoT, multi/binary	≈99.99% accuracy	xApp†
[155]	DDoS	CNN-LSTM (JA4)	SDN-5G	99.98% acc; 100% recall	xApp†
[91]	DDoS	Transformer	5G-NIDD (realistic)	≈99.8% multi-class accuracy	5GC-NF†
[140]	DDoS	XGBoost	O-RAN RACH	>97% prec/recall, <MAC timer	xApp
[107]	DDoS	Federated	5GC VNFs	High accuracy, low FPR	rApp†
[156]	DDoS	Federated ensemble	NCSRD-DS-5GDDos, SMOTE-Tomek	≈98% acc/prec/recall/F1	rApp†
[116]	DDoS (adv.)	Robust FL	5G-NIDD, poisoned	+20%–70% adversarial acc. vs. standard FL	rApp†
[157]	Intrusion	Quantum ML	Noisy quantum HW	F1 degrades with noise/depth (NISQ)	rApp†
[123]	DDoS/intrusion	LLM	5G-NIDD, leave-one-attack-out	≈76% recall on held-out attack	rApp†
[128]	5G protocol	LLM (GPT-4)	3GPP 5G specifications	Zero-shot vulnerability detection	rApp†

That spread is itself a finding. Where a study evaluates on an idealized or static setting, reported accuracy clusters near the ceiling: a random-forest detector reports figures at or above 98% for signaling-DDoS on the 5G core [10], a hybrid CNN–LSTM detector reports near-99.99% for 5G-IoT DDoS [78], and a specification-based FBS detector reports 98% [54]. Where a study confronts distribution shift rather than a matched train–test split, the numbers fall and become more informative: an FBS anomaly detector reports 100% accuracy in static conditions but 65%–76% under mobility before dynamic retraining recovers it [81], and an out-of-distribution evaluation that holds an entire attack type out of training reports roughly 76% recall on the unseen attack [123]. High accuracy on a realistic dataset does not by itself establish robustness to such shift: a transformer reports approximately 99.8% multi-class accuracy on the realistic 5G-NIDD dataset under a matched split [91], a near-ceiling figure that an in-distribution evaluation can yield even on representative data. The studies that report latency rather than only accuracy are the most operationally useful: RRC-storm detection within 90 ms leaving a 60 ms mitigation window [89], random-access DDoS classification faster than the MAC contention-resolution timer [140], and microsecond-scale runtime verification compatible with the near-real-time budget [12].

6.4 Evaluation Pitfalls and Best Practices

Several evaluation pitfalls recur across the surveyed literature, and each admits a corresponding best practice.

The first is the accuracy paradox under class imbalance, already noted: high accuracy on imbalanced data can conceal a near-total failure to detect the minority attack class. The corresponding practice is to report precision, recall, F1, and precision-recall area, and to state the class balance of the evaluation set explicitly.

The second is the simulation-to-reality gap, most pronounced for signaling storms, whose evidence base rests largely on simulation and testbed studies for want of a public benchmark, but present wherever results are obtained on simulated or testbed data. A detector that performs well on synthetic traffic may not transfer to an operational network whose traffic distribution differs. The corresponding practice is to validate on real or operational data where it exists, and otherwise to characterize the fidelity of the simulation and to treat the result as provisional.

The third is specific to generative augmentation: when a detector is both trained and evaluated on data synthesized by the same generative model, the evaluation risks circularity, measuring the detector's fit to the generator rather than to the real threat. The corresponding practice is to reserve real data for evaluation even when synthetic data is used for training, and comparative studies of generative and diffusion-based augmentation provide a template for assessing synthesis quality independently [100,103].

The fourth concerns federated and quantum methods, whose evaluations carry additional confounders. Federated results depend on the degree of non-identical distribution across nodes, which should be reported rather than assumed benign; quantum results obtained on simulators may not survive execution on physical, noisy hardware, a discrepancy that benchmarking on actual quantum processors has begun to quantify [157,158]. The corresponding practice is to report the operating conditions, the data heterogeneity for federated methods and the hardware and noise model for quantum methods, under which a result was obtained.

The overarching observation is that the field lacks a standardized, cross-threat benchmark that would permit detectors to be compared on common ground, and that the absence is most damaging precisely where data is scarcest. The consolidation of datasets and the explicit reporting of imbalance, latency, real-data validation, and operating conditions are the practices through which individual results can be made comparable in the interim. These evaluation gaps are revisited among the open challenges of Section 7.

7 Security of the AI Defenses and Open Challenges

The preceding sections treated AI as the defender. This section inverts the perspective and treats AI as a new attack surface. An AI-native network does not merely add a detector to an existing system; it introduces a learning component that is itself a target, and the security of that component is a precondition for the security it is meant to provide. Section 7.1 examines the adversarial robustness of the AI defenses directly. Section 7.2 addresses the standardization and assurance challenges that AI-native deployment raises. Section 7.3 considers explainability and operational trust. Section 7.4 addresses the operational lifecycle and concept drift. Section 7.5 consolidates the open challenges that motivate the future directions of Section 8.

7.1 Adversarial Robustness of the AI Defenses

When detection is performed by a learned model, the model becomes a target, and the IMT-2030 designation of AI as a native network capability accordingly enlarges the attack surface rather than merely strengthening the defense [18]. Two classes of attack against the defenses themselves are salient, together forming the dual attack surface depicted in Fig. 13.

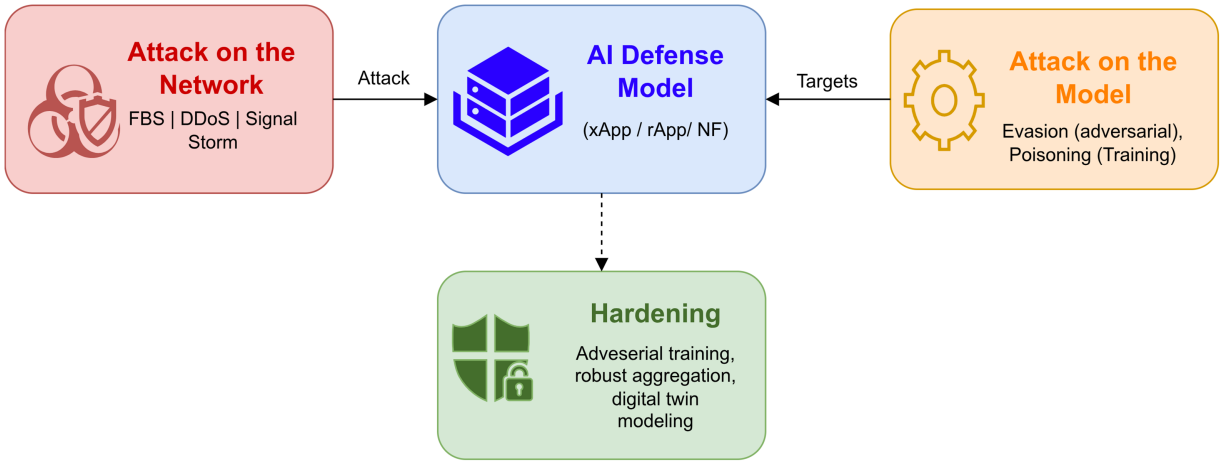

Figure 13: The dual attack surface introduced by an AI-native defense.

The first is evasion, in which an adversary perturbs its traffic or signals so that a detector misclassifies an attack as benign. Because the generative models of Section 4.3 can learn the distribution a detector expects, the same capability that augments training data can be redirected to craft adversarial inputs that evade detection. Studies of generative-adversarial attacks against 5G intrusion-detection systems demonstrate this dual use concretely, and some pair the attack with a corresponding recovery procedure that restores detector performance after the attack is identified [104,105]. Adversarial machine learning has been demonstrated directly against the learning components of Open RAN, including attacks on the deep-reinforcement-learning xApps that drive the RIC [159,160] and a system-level analysis of the adversarial threats to O-RAN intelligence [161], and the threat has been surveyed for 6G networks as a whole [162]. The defensive response is to harden detectors against perturbation, for example through adversarially optimized representations that preserve class separation under attack [163], adversarial training within adaptive detectors [164,165], and digital-twin monitoring that compares a model's live output against a trusted twin to detect and localize adversarial manipulation within the RIC [166].

The second is poisoning, in which an adversary corrupts the training process rather than the inference input. The federated learning of Section 4.4 is distinctively exposed here, because it delegates training to nodes that the coordinator does not control, so that a malicious node can submit crafted updates that degrade or backdoor the global model. The defensive response has two strands: adversarially resilient aggregation that tolerates a fraction of malicious updates [116], and verifiable-privacy mechanisms, such as zero-knowledge proofs over federated updates, that allow the integrity of a contribution to be checked without exposing its data [117]. More broadly, adaptive defensive frameworks that reason about the evolving threat to the AI components themselves are beginning to appear [167].

The implication for the field is that adversarial robustness cannot be an afterthought appended to a deployed detector. In an AI-native network, the robustness of the learning component is part of the security property being claimed, and a defense whose own model is trivially evaded or poisoned provides a false assurance more dangerous than none. This consideration recurs as an open challenge in Section 7.5.

7.2 Standardization and Deployment Challenges

The transition from a research result to a deployed network function encounters a set of challenges that are organizational and procedural as much as technical. The foremost is assurance. A safety-critical telecommunications network is subject to certification, and a learned model whose behavior is defined by training data rather than by an auditable specification is difficult to certify by the methods developed for deterministic functions. This tension underlies a recurring objection to AI-native design, namely that security functions ought to be kept separate from learning components precisely so that they can be assured.

The objection is real, and we do not claim it is settled. We observe, however, that the standardized direction points toward AI-native design rather than away from it: the IMT-2030 framework places AI among the native capabilities of 6G [11], which suggests that, if that direction is followed, assurance will need to be achieved within an AI-native paradigm rather than by avoiding it. The emerging mechanisms reflect this. Runtime-safety supervision of the near-real-time control loop, in which a separate component performs predictive verification and fail-safe enforcement over the actions of a learning-based controller, offers assurance without requiring the learning component itself to be statically certified [12]. Validation infrastructure specific to the open and AI-native RAN, such as cyber-range frameworks for O-RAN and 6G security testing, provides the environment in which such mechanisms can be exercised before deployment [168]. Concretely, the relevant standardization activity spans 3GPP working group SA3, which defines the 5G security architecture [20], studies dedicated enhancements against false base stations [48], and underpins the core's Network Data Analytics Function on which security analytics can be built [149], and the O-RAN Alliance, whose security working group (WG11) maintains the requirements and threat model for the open RAN [169] against which the defenses surveyed here will ultimately be certified; work toward embedding security directly in the RIC illustrates this direction [143]. The standardization of AI security for the RAN through these bodies remains in progress, and the gap between the pace of research and the pace of standardization is itself a deployment challenge.

7.3 Explainability and Operational Trust

A detector that a security analyst cannot interrogate is difficult to act upon, and the deep models that achieve the highest accuracy are the least transparent. The operational consequence is that an unexplained alarm is costly to adjudicate and an unexplained non-alarm is impossible to audit. This has motivated a line of work that couples detection with explainability, attaching to each decision an account of the features or inputs responsible for it. Explainable real-time DDoS detection [98], graph neural networks paired with attribution methods such as SHAP [170], and explainable detection in 6G health and IoT settings [171]

illustrate the direction. For an AI-native defense intended to operate as a network function with minimal human supervision, explainability is not a presentational nicety but a requirement for the trust that delegation of authority to the model presupposes.

7.4 The Operational Lifecycle: Drift and Continual Learning

A consideration that the surveyed literature addresses only in passing, yet that the deployment perspective of this survey makes central, is the operational lifecycle of a learning-based defense. The detectors of Section 4 are typically presented as static artifacts, trained once and evaluated once, but an AI-native network function must contend with non-stationarity: legitimate traffic evolves and adversaries adapt, so a detector's accuracy degrades over time through concept drift. The studies that confront this directly are instructive. An FBS anomaly detector whose accuracy falls from 100% to 65%–76% under mobility recovers it through dynamic retraining [81], and adaptive-threshold and transfer-learning methods are explicit responses to distribution shift [90,111]. Sustaining a deployed defense therefore requires the apparatus of machine-learning operations: monitoring for drift, continual or periodic retraining, model versioning, and the ability to roll back to a previous model when a new one regresses or is poisoned. In the O-RAN setting this apparatus has a natural home, since the non-real-time RIC, whose rApps already perform training and analytics, is the locus for retraining and version management, while runtime-safety supervision provides a rollback trigger [12]. Treating drift, continual learning, and the model lifecycle as first-class concerns of AI-native defense, rather than as deployment afterthoughts, is among the clearest gaps between the research literature and operational practice.

7.5 Open Challenges

The preceding analysis, together with the evaluation gaps of Section 6, yields a set of open challenges that the field must address for AI-native defense to mature.

- **Robustness as a default.** Adversarial robustness against evasion and poisoning is presently an add-on rather than a design requirement, yet in an AI-native network the integrity of the learning component is part of the security property claimed. Robustness must become a default of the defense, not an optional hardening step.
- **Standardized cross-threat benchmarks.** The absence of a shared benchmark spanning FBS, signaling storms, and DDoS prevents detectors from being compared on common ground and is most damaging where data is scarcest. A standardized, cross-threat benchmark with real or high-fidelity data is a prerequisite for cumulative progress.
- **The latency–accuracy–explainability trilemma.** The near-real-time slot rewards compact, fast models; accuracy rewards large ones; operational trust rewards explainable ones. Reconciling the three within a single deployable detector remains unresolved and is sharpened by the deployment perspective of Section 5.
- **Assurance of learning components.** Certifying a model whose behavior derives from data rather than specification is unresolved, and runtime supervision is a partial rather than a complete answer.
- **Unexplored threat–technique combinations.** The empty cells of the technique–threat coverage matrix (Table 8) in Section 5, foremost the absence of federated false base station detection across operators, indicate defenses that the architecture permits but the literature has not yet attempted.

The last of these challenges is the most generative, because it points to specific defenses that could be built. It, together with the data-scale frontier, is the subject of the future directions in Section 8, where the role of quantum-enhanced learning is assessed and the answer to RQ3 is given.

8 Future Directions

The data-scarcity gradient that has organized this survey has an endpoint. Classical models economize on data, generative models manufacture it, and federated models pool it; but each of these remains a classical response to the data problem. This section considers the direction the field takes when the classical responses are exhausted, namely the introduction of quantum machine learning into the RAN, and it does so without overstating the maturity of that direction. This direction is one of several the survey identifies rather than its focus, examined here alongside large-language-model and agentic defenses and the technique–threat combinations the literature has not yet attempted. Section 8.1 locates quantum learning at the data frontier and states the hardware constraint that governs it. Section 8.2 introduces a threat-suitability matrix that determines, for each threat, whether a quantum-RAN defense is appropriate. Section 8.3 traces the concrete research trajectory from existing UE-side work to a RAN-side proposal. Section 8.4 offers a deployment checklist that consolidates the survey's practical guidance. The section closes with the answer to RQ3.

8.1 Quantum Machine Learning and the Data Frontier

Quantum machine learning encodes classical data into the state of a quantum system, embedding it in a feature space whose dimension grows exponentially in the number of qubits. The argument advanced for its relevance to security is that this encoding is most valuable precisely in the regime where classical methods struggle: the high-dimensional, low-sample regime that the data-starved end of the gradient represents. Quantum support vector machines, quantum convolutional networks, and quanvolutional networks have all been proposed for intrusion and threat detection on this basis [172,173], and quantum machine learning has been positioned more broadly for 6G network intelligence with explicit attention to its own adversarial exposure [174]. That exposure is not hypothetical: poisoning attacks against QML-driven classifiers in O-RAN have already been demonstrated [175].

The constraint that governs any near-term deployment is the state of the hardware. Current quantum processors are noisy and limited in qubit count, and results obtained on idealized simulators do not necessarily survive execution on physical devices. As Table 11 records, benchmarking on noisy quantum computers and on a superconducting processor establishes that quantum inference is presently neither fast nor reliable enough for inline operation [157,158], and its adversarial robustness remains underexplored [176]. The consequence for the integration map of Section 5 is specific and decisive: a quantum-enhanced defense cannot occupy the near-real-time xApp slot, whose budget is sub-second, and must instead reside in the non-real-time rApp slot, where its latency is tolerable and its role is analytic rather than inline. Any realistic proposal for quantum-AI-native defense in this decade is therefore a proposal for a non-real-time rApp.

Table 11: Representative quantum machine-learning studies relevant to a RAN-sited defense, and their state of evidence.

Study	Role/Method	State of Evidence
[97]	Quanvolutional NN, UE-side FBS detection	≈90% accuracy on the simulator
[157]	QML intrusion detection on noisy hardware	F1 degrades with noise and circuit depth (NISQ)
[158]	QNN robustness on a superconducting processor	Robustness gap between simulator and real hardware
[175]	Poisoning attack on QML in O-RAN	Attack demonstrated against QML-driven classifiers
[176]	SoK on adversarial robustness of QML	Security under attack remains underexplored

8.2 A Threat-Suitability Matrix for Quantum-AI-RAN

Rather than assert that quantum learning benefits the RAN in general, which the evidence does not support, this survey provides an analytical instrument for determining where it might. A candidate threat is suitable for a quantum-RAN defense to the extent that it satisfies four criteria: the threat must be observable at the RAN, so that a RAN-sited model has the data; its feature space must be high-dimensional, so that quantum encoding has something to exploit; its data must be scarce, so that the quantum few-shot advantage is relevant rather than redundant; and it must tolerate the latency of a non-real-time rApp, so that the hardware constraint of Section 8.1 is respected. Table 12 scores the three threats against these criteria.

Table 12: Threat-suitability matrix for a quantum-AI-native RAN defense (non-real-time rApp). +: favorable; ~: conditional; −: unfavorable.

Criterion	FBS	Signaling Storm	DDoS
Observable at RAN (aggregated reports)	+	+	+
High-dimensional feature space	+	+	+
Data scarce (few-shot advantage relevant)	+	~	−
Tolerates non-real-time (rApp) latency	+	~	−

The matrix is offered as a tool rather than a verdict, and it is deliberately threat-agnostic: it does not designate a single threat as the target of quantum-RAN defense but exposes the conditions under which any threat would qualify. The +, ~, and − entries are the authors' qualitative assessment against the four criteria, offered to be contested rather than presented as measured quantities. Across its rows, it indicates that the data-starved, latency-tolerant profile scores most favorably and the data-rich or latency-critical profiles least, while leaving the determination for any specific deployment to its own operational constraints.

8.3 From UE-Side to RAN-Side: A Research Trajectory

The proposal of a quantum rApp is not speculative in the abstract; it has a concrete point of departure in existing work. A quantum-enhanced detector based on a quanvolutional neural network has already been demonstrated for false base station detection in beyond-5G networks, sited on the user equipment, where it operates on the device's own radio observations [97]. This UE-side result is a proof of concept that quantum encoding can serve FBS detection at all.

The natural generalization follows the logic of the integration map. The UE observes its local radio environment in isolation, whereas the RAN, through the RIC, observes the measurement reports of many devices at once. Aggregating reports across many devices produces a feature space of substantially higher dimension than any single device commands, which is precisely the condition under which quantum encoding is argued to be advantageous. A RAN-side quantum rApp that consumes aggregated multi-device

measurement reports is therefore the principled extension of the UE-side proof of concept, and it occupies the rApp slot that the latency constraint of Section 8.1 prescribes. The broader interest in quantum methods for next-generation network security, spanning vehicular and cyber-physical settings, indicates that the supporting techniques are developing in parallel [172,173,177]. This trajectory, from a demonstrated UE-side detector to a proposed RAN-side rApp is among the most concrete of the open directions, and it extends the quantum-learning row of the technique–threat coverage matrix (Table 8), whose only surveyed entry is the UE-side proof of concept above, to the RAN-side position that no study yet occupies. Such a defense would also complement the standardized direction: the 3GPP study on enhancements against false base stations frames the same problem from the specification side [48].

Quantum learning is not the only frontier. Beyond the detection and analysis roles surveyed in Section 4.5, agentic LLMs that close the loop from detection to autonomous mitigation are an emerging direction [120,121], as are LLM-centric adaptive-defense frameworks now appearing in adjacent domains [167]. Their integration into the RAN control loops, and their own substantial attack surface of prompt injection and adversarial inputs, will require dedicated study.

8.4 AI-Native Deployment Checklist

The practical guidance distributed across this survey can be consolidated into a checklist for taking a learning-based defense from a research result to an AI-native network function. The checklist is intended to make the deployment dimension, the dimension the primary literature most often omits, explicit and routine.

1. **Select the deployment slot** from the latency budget and the observation point of the threat: UE, near-real-time xApp, non-real-time rApp, or 5GC network function (Section 5).
2. **Characterize the data regime** of the target threat and insert the corresponding remedy: direct supervised training where data is abundant, generative augmentation where it is scarce, federated orchestration where it is distributed and unshareable (Section 4).
3. **Report latency alongside accuracy**, since an accuracy figure without an inference-latency figure does not establish deployability at a latency-constrained slot (Section 6).
4. **Validate on real or operational data**, reserving real data for evaluation even when synthetic data is used for training, and characterizing simulation fidelity where real data is unavailable (Section 6).
5. **Harden against evasion and poisoning** before deployment, treating the robustness of the learning component as part of the security property claimed rather than as an optional addition (Section 7).
6. **Attach an explanation** to each decision sufficient for an analyst to adjudicate an alarm and audit a non-alarm (Section 7).
7. **Provide runtime-safety supervision** so that the actions of a learning-based controller can be verified and fail-safe enforcement applied without static certification of the model (Section 7).
8. **State the operating conditions** under which results hold, including data heterogeneity for federated methods and hardware and noise model for quantum methods (Section 6).

Answer to RQ3

The open challenges remaining for AI-native defense are the establishment of robustness and assurance as defaults, the construction of standardized cross-threat benchmarks, the reconciliation of latency, accuracy, and explainability, and the exploration of the threat–technique combinations the architecture permits but the literature has not attempted. A quantum-AI-native defense addresses the data-scale and data-scarcity frontier under specific conditions, not in general: because current quantum hardware precludes inline operation, such a defense is realistic only as a non-real-time rApp, and it is suitable only for threats that are data-scarce,

high-dimensional, RAN-observable, and latency-tolerant. The false base station, whose UE-side quantum detection has already been demonstrated, satisfies these conditions most closely, and its generalization to a RAN-side quantum rApp consuming aggregated measurement reports is the most concrete realization of the frontier this survey identifies.

9 Conclusion

This survey has argued, as a design position rather than a surveyed fact, that effective defense of 5G and 6G networks against false base stations, distributed denial of service, and signaling storms is best served by artificial intelligence that is native to the network rather than appended to it, and that the question worth answering is no longer whether a learning model can detect an attack but where in the architecture it should run and how it should be integrated. We distinguish this position from the evidence: the literature establishes that learning-based detection works and that the O-RAN architecture provides native deployment points, but it does not yet establish that native integration outperforms external detection, and we have not claimed otherwise. From this premise the paper has made four contributions. It has surveyed the AI defenses against the three threats jointly, showing that they reduce to a common family of learning problems and that techniques validated on one threat transfer to the others. It has mapped each technique family to a concrete deployment point in the O-RAN architecture, converting method descriptions into deployment guidance. It has shown that the three threats form a data-scarcity gradient that correlates with, and helps organize, the observed progression of techniques. And it has set out a forward agenda toward quantum-enhanced defense, qualified by the hardware constraints that bound it.

The organizing observation has been that the threats and the techniques are aligned by data. DDoS is data-rich and is met by supervised deep and federated methods; the signaling storm is data-scarce and is met by sequence models over a thin evidence base; the false base station is data-starved and has driven the field toward on-device collection, generative synthesis, and ultimately quantum encoding. The progression of AI techniques from classical machine learning through deep learning, generative augmentation, and federated learning to large language models is, we have argued, usefully read along the axis of worsening data conditions, even as it also reflects general advances in machine learning; the integration map of Section 5 records where each response is deployed. The three research questions are answered in their respective sections: the techniques and their mechanisms in Section 4, their architectural integration in Section 5, and the open challenges and the conditional role of quantum learning in Sections 7 and 8.

Two findings of the survey are likely to outlast the specific methods it catalogues. The first is methodological: the deployment dimension, recorded as the where and how of each technique, is the dimension the primary literature most often omits and the one a practitioner most needs, and a survey that supplies it converts a body of accuracy results into an account of what can actually be fielded. The second is diagnostic: the empty cells of the technique–threat coverage matrix (Table 8), foremost the absence of federated false base station detection across operators and the still-nascent state of quantum defense at the RAN, are not omissions of this survey but gaps in the field, and they identify where the next defenses can be built. As the IMT-2030 framework moves artificial intelligence from an option to a native capability of the network, the discipline of asking not only whether a defense works but where it lives and how it is assured will determine whether AI-native security becomes a property of deployed networks or remains a property of papers.

Acknowledgement: This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT).

Funding Statement: This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00437252, Development of anti-sniffing technology for mobile communication and AirGap environments).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, I Wayan Adi Juliawan Pawana and Ilsun You; methodology, I Wayan Adi Juliawan Pawana; validation, I Wayan Adi Juliawan Pawana and Vincent Abella; formal analysis, I Wayan Adi Juliawan Pawana; investigation, I Wayan Adi Juliawan Pawana and Vincent Abella; data curation, Vincent Abella and Hoonyong Park; writing—original draft preparation, I Wayan Adi Juliawan Pawana; writing—review and editing, Vincent Abella, Eldridge Aaron Miole, Hoonyong Park, and Ilsun You; visualization, Vincent Abella and Eldridge Aaron Miole; supervision, Ilsun You; project administration, Ilsun You; funding acquisition, Ilsun You. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as aguest editor and editorial board of this journal, Ilsun You had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no conflicts of interest.

Appendix A PRISMA 2020 Checklist

Table A1: PRISMA 2020 checklist: location of each item in this manuscript.

Item	Topic	Location in this Manuscript
Title		
1	Title	Title page; Section 2
Abstract		
2	Abstract	Abstract
Introduction		
3	Rationale	Section 1
4	Objectives	Section 2.1
Methods		
5	Eligibility criteria	Section 2.3
6	Information sources	Section 2.2
7	Search strategy	Section 2.2; Table 2
8	Selection process	Section 2.4
9	Data collection process	Section 2.4
10a	Data items: outcomes	Sections 6.2 and 6.3
10b	Data items: other variables	Sections 2.4, 5.2 and 6.1
11	Study risk-of-bias assessment	Not applicable; Section 6.4
12	Effect measures	Not applicable
13a	Synthesis: eligibility for each synthesis	Section 2.4
13b	Synthesis: data preparation	Section 2.4
13c	Synthesis: tabulation and visual display	Sections 4.6, 5.2 and 6.3

(Continued)

Table A1 (continued)

Item	Topic	Location in this Manuscript
13d	Synthesis: methods of synthesis	Section 2.4
13e	Synthesis: heterogeneity	Not applicable
13f	Synthesis: sensitivity analyses	Not applicable
14	Reporting bias assessment	Section 6.4
15	Certainty assessment	Not applicable
Results		
16a	Study selection	Section 2.4; Fig. 2
16b	Studies excluded	Section 2.3; Fig. 2
17	Study characteristics	Sections 4, 5.2 and 6.1
18	Risk of bias in studies	Not applicable
19	Results of individual studies	Sections 6.3 and 4
20a	Results of syntheses: characteristics	Sections 4.6, 5.2 and 5.3
20b	Results of syntheses: statistical results	Not applicable
20c	Results of syntheses: heterogeneity	Not applicable
20d	Results of syntheses: sensitivity	Not applicable
21	Reporting biases	Section 6.4
22	Certainty of evidence	Not applicable
Discussion		
23a	General interpretation	Sections 5, 7 and 9
23b	Limitations of the evidence	Sections 6.4 and 7
23c	Limitations of the review processes	Section 2
23d	Implications	Sections 8 and 8.4
Other information		
24a	Registration	Not applicable
24b	Protocol	Section 2
24c	Amendments	Not applicable
25	Support	Back Matter
26	Competing interests	Back Matter
27	Availability of data, code and other materials	Back Matter; Section 2.2; Table 2

References

1. Priezkalns E. Rogue base stations linked to KT micropayments fraud; 2 Chinese nationals arrested. Commsrisk. 2025 [cited 2026 May 21]. Available from: https://commsrisk.com/4-rogue-base-stations-linked-to-kt-micropayments-fraud-2-chinese-nationals-arrested/.
2. Korea JoongAng Daily. Two days after Denial, KT admits that 5561 subscribers were hacked. Korea JoongAng Daily. 2025 [cited 2026 May 21]. Available from: https://koreajoongangdaily.joins.com/news/2025-09-11/business/industry/Two-days-after-denial-KT-admits-that-5561-subscribers-were-hacked/2397128.
3. Priezkalns E. 3 years in Prison for Fraudster who drove SMS-blasting IMSI-catcher around norway. Commsrisk. 2024 [cited 2026 May 21]. Available from: https://commsrisk.com/3-years-in-prison-for-fraudster-who-drove-sms-blasting-imsi-catcher-around-norway/.

4. Priezkalns E. Paris IMSI-catcher mistaken for Bomb was actually used for Health Insurance SMS phishing scam. Commsrisk. 2023 [cited 2026 May 21]. Available from: https://commsrisk.com/paris-imsi-catcher-mistaken-for-bomb-was-actually-used-for-health-insurance-sms-phishing-scam/.
5. Shaik A, Borgaonkar R, Park S, Seifert JP. On the impact of rogue base stations in 4G/LTE self organizing networks. In: Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks; 2018 Jun 18–20; Stockholm, Sweden. p. 75–86. doi:10.1145/3212480.3212497.
6. Yang H, Bae S, Son M, Kim H, Kim SM, Kim Y. Hiding in plain signal: physical signal overshadowing attack on LTE. In: Proceedings of the 28th USENIX Security Symposium; 2019 Aug 14–16; Santa Clara, CA, USA. Berkeley, CA, USA: USENIX Association; 2019. p. 55–72.
7. Lee G, Lee J, Lee J, Im Y, Hollingsworth M, Wustrow E, et al. This is your president speaking: spoofing alerts in 4G LTE networks. In: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services; 2019 Jun 17–21; Seoul, Republic of Korea. p. 404–16. doi:10.1145/3307334.3326082.
8. Erni S, Kotuliak M, Leu P, Roeschlin M, Capkun S. AdaptOver: adaptive overshadowing attacks in cellular networks. In: Proceedings of the 28th Annual International Conference on Mobile Computing and Networking; 2022 Oct 17–21; Sydney, Australia. p. 743–55. doi:10.1145/3495243.3560525.
9. Tabiban A, Alameddine HA, Salahuddin MA, Boutaba R. Signaling storm in O-RAN: challenges and research opportunities. IEEE Commun Mag. 2024;62(6):58–64. doi:10.1109/mcom.002.2300317.
10. Park S, Cho B, Kim D, You I. Machine learning based signaling DDoS detection system for 5G stand alone core network. Appl Sci. 2022;12(23):12456. doi:10.3390/app122312456.
11. ITU-R. Recommendation ITU-R M.2160-0 (11/2023)—framework and overall objectives of the future development of IMT for 2030 and beyond. 2023 [cited 2026 May 21]. Available from: https://www.itu.int/rec/R-REC-M.2160-0-202311-I.
12. Jawad MA, Munna MMH, Kabir AH, Antu NH, Tulona RF. A runtime safety copilot for AI-native O-RAN: predictive verification and fail-safe enforcement in near-RT RIC control loops. IEEE Access. 2026;14:63106–20. doi:10.1109/access.2026.3686132.
13. Sousa M, Saraiva T, Mata L, Cilínio M, Vieira P. NEXA-NET: advancing knowledge-driven autonomy in AI-native 6G and beyond wireless networks. IEEE Wirel Commun. 2026:1–10. doi:10.1109/mwc.2026.3671234.
14. Yu C, Chen S, Wang F, Wei Z. Improving 4G/5G air interface security: a survey of existing attacks on different LTE layers. Comput Netw. 2021;201:108532. doi:10.1016/j.comnet.2021.108532.
15. Rezaei H, Taheri R, Nowroozi E, Hajizadeh M, Shiaeles S, Bauschert T. A survey on security and privacy in federated learning-based intrusion detection systems for 5G and beyond networks. IEEE Open J Commun Soc. 2026;7:253–300. doi:10.1109/ojcoms.2025.3644477.
16. Yang M, Qu Y, Ranbaduge T, Thapa C, Sultan NH, Ding M, et al. From 5G to 6G: a survey on security, privacy, and standardization pathways. ACM Comput Surv. 2026;58(8):1–38. doi:10.1145/3785467.
17. Mustafovski R, Marinova G, Qehaja B, Hajrizi E, Gagica S, Guliashki V. AI-driven network optimization for the 5G-to-6G transition: a taxonomy-based survey and reference framework. Future Internet. 2026;18(3):155. doi:10.3390/fi18030155.
18. Altintaş M, Karhan SN, Tok YE, Toprak AG, Mercan ÖB. When beneficial intelligence turns hostile: a survey of adversarial threats in AI-native 6G. IEEE Open J Commun Soc. 2026;7:3468–511. doi:10.1109/ojcoms.2026.3678511.
19. Noor K, Imoize AL, Li CT, Weng CY. A review of machine learning and transfer learning strategies for intrusion detection systems in 5G and beyond. Mathematics. 2025;13(7):1088. doi:10.3390/math13071088.
20. 3GPP. TS 33.501 security architecture and procedures for 5G system (Release 19). 2026 [cited 2026 May 21]. Available from: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3169.
21. 3GPP. TS 38.304 NR; User Equipment (UE) procedures in idle mode and in RRC inactive state (Release 19). 3GPP; 2026 [cited 2026 May 21]. Available from: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3192.

22. 3GPP. TS 38.331 NR; radio resource control (RRC) protocol specification (Release 19). 2026 [cited 2026 May 21]. Available from: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197.
23. Hussain SR, Chowdhury O, Mehnaz S, Bertino E. LTEInspector: a systematic approach for adversarial testing of 4G LTE. In: Proceedings of the 2018 Network and Distributed System Security Symposium; 2018 Feb 18–21; San Diego, CA, USA. doi:10.14722/ndss.2018.23313.
24. Hussain SR, Echeverria M, Karim I, Chowdhury O, Bertino E. 5GReasoner: a property-directed security and privacy analysis framework for 5G cellular network protocol. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security; 2019 Nov 11–15; London, UK. p. 669–84. doi:10.1145/3319535.3354263.
25. Ludant N, Vomvas M, Noubir G. Unprotected 4G/5G control procedures at low layers considered dangerous. arXiv:2403.06717. 2024.
26. Ishtiaq AA, Das SSS, Rashid SM, Ranjbar A, Tu K, Wu T, et al. Hermes: unlocking security analysis of cellular network protocols by synthesizing finite state machines from natural language specifications. In: Proceedings of the 33rd USENIX Security Symposium; 2024 Aug 14–16; Philadelphia, PA, USA. Berkeley, CA, USA: USENIX Association; 2024. p. 4445–62.
27. Rahman MM, Karim I, Bertino E. CellularLint: a systematic approach to identify inconsistent behavior in cellular network specifications. In: Proceedings of the 33rd USENIX Conference on Security Symposium; 2024 Aug 14–16; Philadelphia, PA, USA. Berkeley, CA, USA: USENIX Association; 2024. p. 5215–32.
28. O-RAN ALLIANCE. O-RAN architecture description (O-RAN.WG1.OAD). O-RAN ALLIANCE working group 1, technical specification. 2024 [cited 2026 May 21]. Available from: https://specifications.o-ran.org/download?id=641.
29. Huang JH, Cheng SM, Kaliski R, Hung CF. Developing xApps for rogue base station detection in SDR-enabled O-RAN. In: Proceedings of the IEEE INFOCOM 2023—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); 2023 May 10–13; Hoboken, NJ, USA. p. 1–6. doi:10.1109/infocomwkshps57453.2023.10225868.
30. Dabrowski A, Pianta N, Klepp T, Mulazzani M, Weippl E. IMSI-catch me if you can: IMSI-catcher-catchers. In: Proceedings of the 30th Annual Computer Security Applications Conference; 2014 Dec 8–12; New Orleans, LA, USA. p. 246–55. doi:10.1145/2664243.2664272.
31. Mjølsnes SF, Olimid RF. Easy 4G/LTE IMSI catchers for non-programmers. In: Rak J, Bay J, Kotenko I, Popyack L, Skormin V, Szczypiorski K, editors. Computer network security. Cham, Switzerland: Springer; 2017. p. 235–46. doi:10.1007/978-3-319-65127-9_19.
32. Heish WL, Hong BK, Cheng SM. Toward large-scale rogue base station attacks using container-based virtualization. In: Proceedings of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall); 2019 Sep 22–25; Honolulu, HI, USA. p. 1–6. doi:10.1109/vtcfall.2019.8891312.
33. Yang Y, Zhang Y, Wan T, Duan H, Chang D, Li Y, et al. Small cell, big risk: a security assessment of 4G LTE femtocells in the wild. In: Proceedings of the 2026 Network and Distributed System Security Symposium; 2026 Feb 23–27; San Diego, CA, USA.
34. Paci A, Bologna G, Palamà I, Bianchi G. FlashCatch: minimizing disruption in IMSI catcher operations. In: Proceedings of the 18th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2025 Jun 30–Jul 3; Arlington, VA, USA. p. 124–35. doi:10.1145/3734477.3734705.
35. Hussain SR, Echeverria M, Chowdhury O, Li N, Bertino E. Privacy attacks to the 4G and 5G cellular paging protocols using side channel information. In: Proceedings of the 2019 Network and Distributed System Security Symposium; 2019 Feb 24–27; San Diego, CA, USA. doi:10.14722/ndss.2019.23442.
36. Karakoc B, Fürste N, Rupprecht D, Kohls K. Never let me down again: bidding-down attacks and mitigations in 5G and 4G. In: Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2023 May 29–Jun 1; Guildford, UK. p. 97–108. doi:10.1145/3558482.3581774.

37. Rupprecht D, Kohls K, Holz T, Poepper C. IMP4GT: IMPersonation attacks in 4G NeTworks. In: Proceedings of the 2020 Network and Distributed System Security Symposium; 2020 Feb 23–26; San Diego, CA, USA. doi:10.14722/ndss.2020.24283.
38. Zheng Y, Huang L, Shan H, Li J, Yang Q, Xu W. Ghost telephonist impersonates you: vulnerability in 4G LTE CS fallback. In: Proceedings of the 2017 IEEE Conference on Communications and Network Security (CNS); 2017 Oct 9–11; Las Vegas, NV, USA. p. 1–9. doi:10.1109/cns.2017.8228629.
39. Fan W, Shi B, Peng C. NReplay: 5G key reinstallation attack based on NAS layer vulnerabilities. In: Proceedings of the MILCOM 2024—2024 IEEE Military Communications Conference (MILCOM); 2024 Oct 28–Nov 1; Washington, DC, USA. p. 1088–93. doi:10.1109/milcom61039.2024.10773741.
40. Raza MT, Guo Y, Lu S, Anwar FM. On key reinstallation attacks over 4G LTE control-plane: feasibility and negative impact. In: Proceedings of the Annual Computer Security Applications Conference; 2021 Dec 6–10; Virtual Event. p. 877–86. doi:10.1145/3485832.3485833.
41. Borgaonkar R, Hirschi L, Park S, Shaik A. New privacy threat on 3G, 4G, and upcoming 5G AKA protocols. Proc Priv Enhancing Technol. 2019;2019(3):108–27. doi:10.2478/popets-2019-0039.
42. Shaik A, Borgaonkar R, Asokan N, Niemi V, Seifert JP. Practical attacks against privacy and availability in 4G/LTE mobile communication systems. In: Proceedings of the 2016 Network and Distributed System Security Symposium; 2016 Feb 21–24; San Diego, CA, USA. doi:10.14722/ndss.2016.23236.
43. Shaik A, Borgaonkar R, Park S, Seifert JP. New vulnerabilities in 4G and 5G cellular access network protocols: exposing device capabilities. In: Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks; 2019 May 15–17; Miami, FL, USA. p. 221–31. doi:10.1145/3317549.3319728.
44. Fraunholz D, Schorghofer-Vrinssen R, Konig H, Zahoransky R. Show me your attach request and I'll tell you who you are: practical fingerprinting attacks in 4G and 5G mobile networks. In: Proceedings of the 2022 IEEE Conference on Dependable and Secure Computing (DSC); 2022 Jun 22–24; Edinburgh, UK. p. 1–8. doi:10.1109/dsc54232.2022.9888899.
45. Kohls K, Rupprecht D, Holz T, Pöpper C. Lost traffic encryption: fingerprinting LTE/4G traffic on layer two. In: Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks; 2019 May 15–17; Miami, FL, USA. p. 249–60. doi:10.1145/3317549.3323416.
46. Zhai L, Qiao Z, Wang Z, Wei D. Identify what you are doing: smartphone apps fingerprinting on cellular network traffic. In: Proceedings of the 2021 IEEE Symposium on Computers and Communications (ISCC); 2021 Sep 5–8; Athens, Greece. p. 1–7. doi:10.1109/iscc53001.2021.9631415.
47. Norrman K, Näslund M, Dubrova E. Protecting IMSI and user privacy in 5G networks. In: Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications; 2016 Jun 18–20; Xi'an, China. doi:10.4108/eai.18-6-2016.2264114.
48. 3GPP. TR 33.809 study on 5G security enhancements against false base stations (FBS). 3GPP Technical Report (Release 18), SA WG3. 2024 [cited 2026 May 21]. Available from: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3539.
49. Li Z, Wang W, Wilson C, Chen J, Qian C, Jung T, et al. FBS-radar: uncovering fake base stations at scale in the wild. In: Proceedings of the 2017 Network and Distributed System Security Symposium; 2017 Feb 26–Mar 1; San Diego, CA, USA. doi:10.14722/ndss.2017.23098.
50. Ney P, Smith I, Cadamuro G, Kohno T. SeaGlass: enabling city-wide IMSI-catcher detection. Proc Priv Enhancing Technol. 2017;2017(3):39–56. doi:10.1515/popets-2017-0027.
51. CellularPrivacy. Android IMSI-Catcher Detector (AIMSICD). 2024 [cited 2024 Nov 30]. Available from: https://github.com/CellularPrivacy/Android-IMSI-Catcher-Detector.
52. Park S, Shaik A, Borgaonkar R, Seifert JP. Anatomy of commercial IMSI catchers and detectors. In: Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society; 2019 Nov 11; London, UK. p. 74–86. doi:10.1145/3338498.3358649.
53. Echeverria M, Ahmed Z, Wang B, Arif MF, Hussain SR. PHOENIX: device-centric cellular network protocol monitoring using runtime verification. In: Proceedings of the 2021 Network and Distributed System Security Symposium; 2021 Feb 21–25; Virtual Event.

54. Park H, Astillo PVB, Ko Y, Park Y, Kim T, You I. SMDFbs: specification-based misbehavior detection for false base stations. Sensors. 2023;23(23):9504. doi:10.3390/s23239504.
55. Zhuang Z, Ji X, Zhang T, Zhang J, Xu W, Li Z, et al. FBSleuth: fake base station forensics via radio frequency fingerprinting. In: Proceedings of the 2018 on Asia Conference on Computer and Communications Security; 2018 Jun 4–8; Incheon, Republic of Korea. p. 261–72. doi:10.1145/3196494.3196521.
56. Mubasshir KS, Karim I, Bertino E. Gotta detect 'Em all: fake base station and multi-step attack detection in cellular networks. In: Proceedings of the 34th USENIX Security Symposium; 2025 Aug 13–15; Seattle, WA, USA. Berkeley, CA, USA: USENIX Association; 2025. p. 5365–84.
57. Li X, Zheng K, Guo S, Ma X. Precheck sequence based false base station detection during handover: a physical layer security scheme. arXiv:2307.01396. 2023.
58. Purification S, Kim J, Kim J, Chang SY. Fake base station detection and link routing defense. Electronics. 2024;13(17):3474. doi:10.3390/electronics13173474.
59. Purification S, Chang SY. Verifiable alerts for 4G/5G public warning system. In: Proceedings of the 2025 IEEE Conference on Communications and Network Security (CNS); 2025 Sep 8–11; Avignon, France. p. 1–9. doi:10.1109/cns66487.2025.11195018.
60. Nie S, Zhang Y, Wan T, Duan H, Li S. Measuring the deployment of 5G security enhancement. In: Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2022 May 16–19; San Antonio, TX, USA. p. 169–74. doi:10.1145/3507657.3528559.
61. Tucker T, Bennett N, Kotuliak M, Erni S, Capkun S, Butler K, et al. Detecting IMSI-catchers by characterizing identity exposing messages in cellular traffic. In: Proceedings of the 2025 Network and Distributed System Security Symposium; 2025 Feb 24–28; San Diego, CA, USA. doi:10.14722/ndss.2025.241115.
62. Christopoulou M, Garos A, Vekraki A, Santorinaios D, Koufos I, Karamitsiani S, et al. User terminals as attackers: an open dataset analysis of DDoS attacks in 5G networks. In: Proceedings of the 2024 IEEE Conference on Standards for Communications and Networking (CSCN); 2024 Nov 25–27; Belgrade, Serbia. p. 301–7. doi:10.1109/cscn63874.2024.10849694.
63. Erni S, Kotuliak M, Martinovic I. GLaDoS: location-aware denial-of-service of cellular networks. In: Proceedings of the 34th USENIX Security Symposium; 2025 Aug 13–15; Seattle, WA, USA. Berkeley, CA, USA: USENIX Association; 2025.
64. Shabbir M, Kandeepan S, Rowe W, Al-Hourani A. Detection of RRC inactivity timer based signalling storm attack on 5G and B5G networks. Comput Netw. 2026;282:112286. doi:10.1016/j.comnet.2026.112286.
65. Mayhoub S, Chatzimiltis S, Alimohammadi H, He Z, Abdulkareem SA, Shojafar M, et al. A new sub-use case for signaling storm attack in open RAN and an ML-based detection approach. In: Proceedings of the 2024 IEEE Conference on Standards for Communications and Networking (CSCN); 2024 Nov 25–27; Belgrade, Serbia. p. 308–13. doi:10.1109/cscn63874.2024.10849726.
66. Kwao E, Lee J, Park J, Hong B, Kim T, Bang I. Random access failure attack on cellular networks: forcing timing advance misalignment. In: Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2024 May 27–29; Seoul, Republic of Korea. p. 248–53. doi:10.1145/3643833.3656125.
67. Ashik MH, Hossain M. ReaperPulse: a targeted energy-efficient control channel jamming in 5G. In: Proceedings of the 2025 ACM Workshop on Wireless Security and Machine Learning; 2025 Jul 3; Arlington, VA, USA. p. 2–7. doi:10.1145/3733965.3733971.
68. Dong Y, Wan T, Wu T, Hussain SR. Evaluating time-bounded defense against RRC relay in 5G broadcast messages. In: Proceedings of the 18th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2025 Jun 30–Jul 3; Arlington, VA, USA. p. 236–41. doi:10.1145/3734477.3734718.
69. Lin W, Li Z, Chen B, Liu J, Cheng RG, Zhang F. 5G-muffler: covert DoS attacks over open fronthaul interface of O-RAN 5G network. In: Proceedings of the IEEE INFOCOM 2025—IEEE Conference on Computer Communications; 2025 May 19–22; London, UK. p. 1–10. doi:10.1109/infocom55648.2025.11044568.
70. Chang SY, Purification S. Securing cellular availability: the wireless blackhole threat and defense. In: Proceedings of the 2025 IEEE Conference on Communications and Network Security (CNS); 2025 Sep 8–11; Avignon, France. p. 1–9. doi:10.1109/cns66487.2025.11195014.

71. Bennett N, Zhu W, Simon B, Kennedy R, Enck W, Traynor P, et al. RANsacked: a domain-informed approach for fuzzing LTE and 5G RAN-core interfaces. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security; 2024 Oct 14–18; Salt Lake City, UT, USA. p. 2027–41. doi:10.1145/3658644.3670320.
72. Neto FHC, Da Silva PRB, Alves EC, Lima JPSH, Paiva TWP. A novel service model framework for signaling storm detection in open RAN. In: Proceedings of the IEEE INFOCOM 2025—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); 2025 May 19; London, UK. p. 1–6. doi:10.1109/infocomwkshps65812.2025.11152770.
73. Abella V. MODI dataset: NAS and RRC attack and normal captures for LTE and 5G. 2026 [cited 2026 May 22]. Available from: https://github.com/roastedbeans/modi-dataset.
74. Park Y, Pawana IWAJ, Kim B, You I. Harnessing GAN to create realistic FBS (false base station) attack data in 5G. In: Proceedings of the 2024 IEEE International Symposium on Consumer Technology (ISCT); 2024 Aug 13–16; Bali, Indonesia. p. 404–8. doi:10.1109/isct62336.2024.10791271.
75. Nakarmi PK, Sternby J, Ullah I. Applying machine learning on RSRP-based features for false base station detection. In: Proceedings of the 17th International Conference on Availability, Reliability and Security; 2022 Aug 23–26; Vienna, Austria. p. 1–7. doi:10.1145/3538969.3543787.
76. Park H, Astillo PVB, Kim T, You I. 5G native network function for false base station detection using machine learning technique. Inf Syst Front. 2025;27(6):2279–94. doi:10.1007/s10796-025-10614-1.
77. Sun S, Abualhaol I, Poitau G, Esswie A, Repeta M. An ensemble approach for fake base station detection using temporal graph analysis and anomaly detection. In: Proceedings of the 2024 Wireless Telecommunications Symposium (WTS); 2024 Apr 10–12; Oakland, CA, USA. p. 1–6. doi:10.1109/wts60164.2024.10536680.
78. Sadhwani S, Mathur A, Muthalagu R, Pawar PM. 5G-SIID: an intelligent hybrid DDoS intrusion detector for 5G IoT networks. Int J Mach Learn Cybern. 2025;16(2):1243–63. doi:10.1007/s13042-024-02332-y.
79. Ji S, Mishra AK. Enhanced machine learning framework for DDoS attack endurance in vehicles security for 5G networks. In: Proceedings of the 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI); 2025 Apr 4–6; Greater Noida, India. p. 1516–21. doi:10.1109/iccsai64074.2025.11063900.
80. Kirana MC, Cherkaoui S. Robust ensemble model for attack detection in open RAN. In: Proceedings of the GLOBECOM 2025—2025 IEEE Global Communications Conference; 2025 Dec 8–12; Taipei, Taiwan. p. 4000–5. doi:10.1109/globecom59602.2025.11431904.
81. Islam A, Purification S, Chang SY. Anomaly detection against fake base station threats using machine learning. J Cybersecur Priv. 2025;5(4):94. doi:10.3390/jcp5040094.
82. Liu IH, Chen HH, Tang BH, Li JS. Rogue base station detection in industrial Internet of Things. Sens Mater. 2025;37(4):1605. doi:10.18494/sam5279.
83. Jin J, Lian C, Xu M. Rogue base station detection using a machine learning approach. In: Proceedings of the 2019 28th Wireless and Optical Communications Conference (WOCC); 2019 May 9–10; Beijing, China. p. 1–5. doi:10.1109/wocc.2019.8770554.
84. Ebadinezhad S, Hussein IK. Securing 5G network slicing: a hybrid machine learning approach for DDoS threat detection and performance evaluation. In: Trends in sustainable computing and machine intelligence. Cham, Switzerland: Springer Nature; 2026. p. 12–29. doi:10.1007/978-3-032-13177-5_2.
85. Tahori K, Fatani IFE, Moughit M. Conditional counter-inspection with curriculum-biased experts for lightweight 5G intrusion detection. Future Internet. 2026;18(3):116. doi:10.3390/fi18030116.
86. Wu T, Al Ishtiaq A, Yang T, Dong Y, Tu K, Song Z, et al. Guardians of the air: in-device detection of 5G control-plane threats. In: Proceedings of the 2026 IEEE Symposium on Security and Privacy (SP); 2026 May 18–21; San Francisco, CA, USA. p. 2759–78. doi:10.1109/sp63933.2026.00204.
87. Kouam AJ, Viana AC, Martins P, Adjih C, Tchana A. SigN: SIMBox activity detection through latency anomalies at the cellular edge. In: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security; 2025 Aug 25–29; Hanoi, Vietnam. p. 1442–58. doi:10.1145/3708821.3733902.
88. Feng S, Cui B, Chang S, Jiang M. Temporal attention LSTM network for NGAP anomaly detection in 5GC boundary. Comput Model Eng Sci. 2025;144(2):2567–90. doi:10.32604/cmes.2025.067326.

89. Nguyen DK, El Malki R, Rebecchi F. RRC signaling storm detection in O-RAN. In: Proceedings of the 2025 IEEE Symposium on Computers and Communications (ISCC); 2025 Jul 2–5; Bologna, Italy. p. 1–7. doi:10.1109/iscc65549.2025.11326128.
90. Nguyen DK, El Malki R, Rebecchi F, Knopp R, Önen M. Beyond static thresholds: adaptive RRC signaling storm detection with extreme value theory. In: Proceedings of the 2025 International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM); 2025 Oct 27–31; Barcelona, Spain. p. 360–8. doi:10.1109/mswim67937.2025.11309091.
91. Harshdeep K, Sumalatha K, Mathur R. DeepTransIDS: transformer-based deep learning model for detecting DDoS attacks on 5G NIDD. Results Eng. 2025;26:104826. doi:10.1016/j.rineng.2025.104826.
92. Wu G. A spatiotemporal transformer framework for robust threat detection in 6G networks. Internet Technol Lett. 2025;8(3):e70017. doi:10.1002/itl2.70017.
93. Tian Z, Kalidindi RV, Gurusamy M. Dual-branch transformer for anomaly-based intrusion detection from multivariate KPIs in the 5G user plane. In: Proceedings of the 2025 IEEE Conference on Communications and Network Security (CNS); 2025 Sep 8–11; Avignon, France. p. 1–9. doi:10.1109/cns66487.2025.11194999.
94. Ilias L, Doukas G, Lamprou V, Ntanos C, Askounis D. Convolutional neural networks and mixture of experts for intrusion detection in 5G networks and beyond. Front Artif Intell. 2026;8:1708953. doi:10.3389/frai.2025.1708953.
95. Islam A, Chang SY, Kim J, Kim J. Anomaly detection in 5G using variational autoencoders. In: Proceedings of the 2024 Silicon Valley Cybersecurity Conference (SVCC); 2024 Jun 17–19; Sejong, Republic of Korea. p. 1–6. doi:10.1109/svcc61185.2024.10637312.
96. Ilias L, Palmos S, Doukas G, Blika A, Kiokes G, Ntanos C, et al. Convolutional autoencoders coupled with hypernetworks for recognizing attacks in 5G networks and beyond. IEEE Open J Commun Soc. 2025;6:7885–98. doi:10.1109/ojcoms.2025.3612124.
97. Pawana IWAJ, Kwon H, Linawati, You I. Quantum-enhanced detection of false base stations in beyond 5G networks using quanvolutional neural network. In: Proceedings of the 2025 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS); 2025 Oct 9–11; Bali, Indonesia. p. 207–12. doi:10.1109/icsgteis68532.2025.11284521.
98. Albashayreh A, Al-Sharaeh S, Tashtoush Y, Zahariev P. Enhancing 5G network security: a deep learning framework for real-time DDoS detection and explainable threat analysis. IEEE Access. 2025;13:168217–35. doi:10.1109/access.2025.3613386.
99. Kim YS, Kim YE, Kim H. A model training method for DDoS detection using CTGAN under 5GC traffic. Comput Syst Sci Eng. 2023;47(1):1125–47. doi:10.32604/csse.2023.039550.
100. Pamuditha P, Hettiarachchi S, Gunarathne Y, Moremada C, Sandeepa C, Liyanage M. Enhancing intrusion detection in 5G networks: a comparative study of generative adversarial and diffusion networks. In: Proceedings of the 2025 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom); 2025 Jun 23–26; Chisinau, Republic of Moldova. p. 1–6. doi:10.1109/blackseacom65655.2025.11193953.
101. Hossain MJ, Alam K, Fahad Monir M, Mozammal Hoque M, Ahmed T. Explainable AI meets synthetic data: a deep learning framework for detecting network intrusion in NextG network infrastructure. IEEE Access. 2025;13:114979–5001. doi:10.1109/access.2025.3585783.
102. Amachaghi EN, Abdulkareem SA, Foh CH, Mi D, Shojafar M. Improving intrusion detection in O-RAN with synthetic data generation: a GAN and SMOTE approach. Telemat Inf Rep. 2025;20:100269. doi:10.1016/j.teler.2025.100269.
103. Hani A, El-Emam NN. Enhanced intrusion detection systems dataset synthesis using conditional generative adversarial networks with the adaptive whale optimization algorithm. Int J Adv Soft Comput Appl. 2026;18(1):263–85. doi:10.15849/ijasca.v18i1.71.
104. Alslman Y, Alkasassbeh M, Abdel-Rahman MJ. Breaking and healing: GAN-based adversarial attacks and post-adversarial recovery for 5G IDSs. IEEE Access. 2025;13:132109–25. doi:10.1109/access.2025.3587605.
105. Shameli R, Rajkumar S. High-speed threat detection in 5G SDN with particle swarm optimizer integrated GRU-driven generative adversarial network. Sci Rep. 2025;15(1):10025. doi:10.1038/s41598-025-95011-z.

106. Nurakhov Y, Aibagarov S, Kassymbek N, Mukhanbet A, Kumalakov B, Imankulov T. The impact of generative AI on 6G network architecture and service. Electronics. 2026;15(7):1345. doi:10.3390/electronics15071345.
107. Maiga AA, Ataro E, Githinji S. XGBoost and deep learning based-federated learning for DDoS attack detection in 5G core network VNFs. In: Proceedings of the 2024 6th International Conference on Computer Communication and the Internet (ICCCI); 2024 Jun 14–16; Tokyo, Japan. p. 128–33. doi:10.1109/iccci62159.2024.10674312.
108. Munaweera P, Prasad S, Hewa T, Siriwardhana Y, Ylianttila M. Federated learning-powered DDoS attack detection for securing cyber physical systems in 5G and beyond networks. In: Proceedings of the 14th International Conference on the Internet of Things; 2024 Nov 19–22; Oulu, Finland. p. 273–8. doi:10.1145/3703790.3703822.
109. Awasthi A, Mishra S, Vediya P, Miranka H, Battula RB, Gopalani D. FL-MultiDDoSNet: a visual federated fusion for securing future networks against DDoS attacks. In: Proceedings of the 2025 IEEE Future Networks World Forum (FNWF); 2025 Nov 10–12; Bengaluru, India. p. 1–6. doi:10.1109/fnwf66845.2025.11317246.
110. Alsarhan A, Barhoush M, Khassawneh B, Al-Essa M, Aljaidi M, Al-Na'amneh Q. Deep learning utilization for DDoS attack detection with federated learning: a case study on the CICDDoS2019 dataset. Eng Technol Appl Sci Res. 2026;16(1):31203–8. doi:10.48084/etasr.14119.
111. Bellmunt A, Otero B, Rodríguez E, Masip-Bruin X. Federated transfer learning-based intrusion detection system in 5G networks. Expert Syst Appl. 2026;305:130868. doi:10.1016/j.eswa.2025.130868.
112. Maduranga M, Nandan D, Vidanage B, Neththikumara S, Kulasekara D, Narasinghe NM. A federated graph neural network approach for intrusion detection in 5G network. In: Proceedings of the 2025 9th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI); 2025 Nov 19–21; Colombo, Sri Lanka. p. 1–6. doi:10.1109/slaai-icai68534.2025.11318484.
113. Ilias L, Doukas G, Lamprou V, Mouzakitis S, Ntanos C, Askounis D. A federated learning-based network intrusion detection system for 5G and IoT using mixture of experts. Electronics. 2026;15(5):1057. doi:10.3390/electronics15051057.
114. Wang X, Liao X, Cao Z, Min S. 5G-FedGuard: outlier-aware federated autoencoder for PFCP anomaly detection in 5G core networks. In: Proceedings of the 2025 International Conference on Future Communications and Networks (FCN); 2025 Jul 13–16; Istanbul, Turkey. p. 1–6. doi:10.1109/fcn66513.2025.11296734.
115. Alam A, Umer A, Ullah I, Alsayat A. AI-enabled cybersecurity framework for future 5G wireless infrastructures. Sci Rep. 2026;16(1):7055. doi:10.1038/s41598-026-37444-8.
116. Zafar S, Legg P, White J, Salman A. Adversarially resilient federated learning for heterogeneous edge nodes in 5G networks with non-IID data. Internet Things. 2026;37:101919. doi:10.1016/j.iot.2026.101919.
117. Reis MJCS. ZK-FLGuard: verifiable privacy via zero-knowledge proofs in federated anomaly detection for 5G edge-IoT systems. Int J Comput Theory Eng. 2026;18(1):27–37. doi:10.7763/ijcte.2026.v18.1386.
118. Suresh Sanu A, Chitra M. A novel federated approach to model aggregation for global model in cooperative wireless communication. Cybern Syst. 2025;2:1–31. doi:10.1080/01969722.2025.2551007.
119. Bene P, Bernardini A, Sagratella L, Maunero N, Settembre M. Optimizing local LLM deployment for 5G CVE classification avoiding external data exposure. In: Proceedings of the 2025 IEEE Conference on Communications and Network Security (CNS); 2025 Sep 8–11; Avignon, France. p. 1–3. doi:10.1109/cns66487.2025.11195016.
120. Wen H, Sharma P, Yegneswaran V, Gehani A, Porras P, Lin Z. MobiLLM: an agentic AI framework for closed-loop threat mitigation in 6G open RANs. In: Proceedings of the MILCOM 2025—2025 IEEE Military Communications Conference (MILCOM); 2025 Oct 20–24; Washington, DC, USA. p. 1–6. doi:10.1109/milcom64451.2025.11310651.
121. Chatzimiltis S, Mashhadi MB, Shojafar M, Debbah M, Tafazolli R. Agentic AI for 6G: a new paradigm for autonomous RAN security compliance. IEEE Comm Stand Mag. 2026:1–9. doi:10.1109/mcomstd.2026.3681685.
122. Chatzistefanidis I, Leone A, Yaghoubian A, Irazabal M, Nassim S, Bariah L, et al. MX-AI: agentic observability and control platform for open and AI-RAN. arXiv:2508.09197. 2025.
123. Atsilmis E, Uzunidis D, Karkazis P. Large language models for out-of-distribution attack detection in 6G networks. In: Proceedings of the 2025 IEEE Conference on Network Function Virtualization and Software-Defined Networking (NFV-SDN); 2025 Nov 10–12; Natal, Brazil. p. 1–6. doi:10.1109/nfv-sdn66355.2025.11349498.

124. Abasi AK, Aloqaily M, Guizani M. Anomaly detection in 6G networks using large language models (LLMs). In: Proceedings of the 2025 International Wireless Communications and Mobile Computing (IWCMC); 2025 Jun 23–27; Venice, Italy. p. 1466–71. doi:10.1109/iwcmc65282.2025.11059535.
125. Bani Melhem S, Golec M, Alwarafy A, Khamayseh Y. LENS: lightweight and explainable LLM-based APT detection at the edge for 6G security. IEEE Access. 2025;13:172402–15. doi:10.1109/access.2025.3616235.
126. Houssel PRB, Layeghy S, Singh P, Portmann M. eX-NIDS: a framework for explainable network intrusion detection leveraging large language models. Comput Electr Eng. 2026;129:110826. doi:10.1016/j.compeleceng.2025.110826.
127. Kim YC, Lee C, Yoon Y. Payload-aware intrusion detection with CMAE and large language models. ACM Trans Priv Secur. 2026;29(1):1–30. doi:10.1145/3769682.
128. Shahriar A, Hisham SJ, Rahman KMA, Islam R, Hossain MS, Hwang RH, et al. 5GPT: 5G vulnerability detection by combining zero-shot capabilities of GPT-4 with domain aware strategies through prompt engineering. IEEE Trans Inf Forensics Secur. 2025;20:7045–60. doi:10.1109/tifs.2025.3586480.
129. Shan Q, Cui J. LLM guided automated security analysis for 5G UE control plane protocols. In: Proceedings of the 2025 IEEE 25th International Conference on Communication Technology (ICCT); 2025 Oct 17–19; Chengdu, China. p. 1879–83. doi:10.1109/icct67417.2025.11374120.
130. Wang G, Hong X, He X, Liu Y, Hong Y, Chen M, et al. KG-LLM: a framework for intrusion detection in 6G industrial Internet of Things. IEEE Netw. 2026;40(3):152–9. doi:10.1109/mnet.2026.3659223.
131. Blefari F, Cosentino C, Pironti FA, Furfaro A, Marozzo F. CyberRAG: an agentic RAG cyber attack classification and reporting tool. Future Gener Comput Syst. 2026;176:108186. doi:10.1016/j.future.2025.108186.
132. Albaseer A, Hamood M, Al-Sabri R, Abdallah M, Al-Fuqaha A. A zero-touch O-RAN framework for federated few-shot IDS with LLM-oracle verification. In: Proceedings of the GLOBECOM 2025—2025 IEEE Global Communications Conference; 2025 Dec 8–12; Taipei, Taiwan. p. 3170–5. doi:10.1109/globecom59602.2025.11432014.
133. Alnfiai MM. AI-powered cyber resilience: a reinforcement learning approach for automated threat hunting in 5G networks. EURASIP J Wirel Commun Netw. 2025;2025(1):68. doi:10.1186/s13638-025-02497-2.
134. Rezaei H, Taheri R, Shojafar M. FedLLMGuard: a federated large language model for anomaly detection in 5G networks. Comput Netw. 2025;269:111473. doi:10.1016/j.comnet.2025.111473.
135. Moore J, Abdalla AS, Khanal P, Marojevic V. Integrated LLM-based intrusion detection with secure slicing xApp for securing O-RAN-enabled wireless network deployments. In: Proceedings of the 2025 IEEE International Conference on Communications Workshops (ICC Workshops); 2025 Jun 8–12; Montreal, QC, Canada. p. 274–9. doi:10.1109/iccworkshops67674.2025.11162166.
136. Chatzimiltis S, Shojafar M, Mashhadi MB, Tafazolli R. AI-on-RAN for cyber defense: an XAI-LLM framework for interpretable anomaly detection. IEEE Trans Netw Sci Eng. 2026;13:3301–19. doi:10.1109/tnse.2025.3629983.
137. Cortés-Polo D, Calle-Cancho J, Paoletti ME, Haut JM. Transformer-based network for enhanced wireless service differentiation in 6G. Expert Syst Appl. 2026;300:130379. doi:10.1016/j.eswa.2025.130379.
138. Wu D, Wang X, Qiao Y, Wang Z, Jiang J, Cui S, et al. NetLLM: adapting large language models for networking. In: Proceedings of the ACM SIGCOMM 2024 Conference; 2024 Aug 4–8; Sydney, Australia. p. 661–78.
139. Liu H, Xue J, Zhao S, Liu Y, Lu Z. The dual role of large language models in network security: survey and research trends. In: Proceedings of the 2025 ACM Workshop on Wireless Security and Machine Learning; 2025 Jun 16–20; Chicago, IL, USA. p. 20–5. doi:10.1145/3733965.3733972.
140. Branco da Silva PR, Henriques Sales de Lima JP, Costa Alves E, Farfan WS, Coutinho VA, do Prado Paiva TW, et al. Evaluation of the latency of machine learning random access DDoS detection in open RAN. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking; 2024 Sep 30–Oct 4; Washington, DC, USA. p. 2306–11. doi:10.1145/3636534.3701546.
141. Rahman MH, Hossen MS, Stephenson NH, Shah VK, Da Silva A. SAJD: self-adaptive jamming attack detection in AI/ML integrated 5G O-RAN networks. In: Proceedings of the MILCOM 2025—2025 IEEE Military Communications Conference (MILCOM); 2025 Oct 6–10; Los Angeles, CA, USA. p. 1130–5. doi:10.1109/milcom64451.2025.11310373.

142. Aksu N, Saridas I, Gülen U, Fuladi R, Tuna ÖF, Basaran ST. 3PS-RAN: a real-time framework for securing the O-RAN RACH against DDoS attacks toward NextG. IEEE Access. 2026;14:64949–65. doi:10.1109/access.2026.3687602.
143. Kakani PK, Balannagari MRC, Tammen M, Schotten HD. Securing O-RAN: a modular xApp-based defense framework in the near-RT RIC. In: Proceedings of the 2025 IEEE Future Networks World Forum (FNWF); 2025 Nov 10–12; Bangalore, India. p. 1–6. doi:10.1109/fnwf66845.2025.11317551.
144. Wen H, Porras P, Yegneswaran V, Gehani A, Lin Z. 5G-spector: an O-RAN compliant layer-3 cellular attack detection service. In: Proceedings of the 2024 Network and Distributed System Security Symposium; 2024 Feb 26–Mar 1; San Diego, CA, USA.
145. Scalingi A, D'Oro S, Restuccia F, Melodia T, Giustiniano D. Det-RAN: data-driven cross-layer real-time attack detection in 5G open RANs. In: Proceedings of the IEEE INFOCOM 2024—IEEE Conference on Computer Communications; 2024 May 20–23; Vancouver, BC, Canada. p. 41–50. doi:10.1109/infocom52122.2024.10621223.
146. Dimou S, Noubir G. ARGOS: anomaly recognition and guarding through O-RAN sensing. arXiv:2506.06916. 2025.
147. Sun C, Pawar U, Khoja M, Foukas X, Marina MK, Radunovic B. SpotLight: accurate, explainable and efficient anomaly detection for open RAN. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking; 2024 Sep 30–Oct 4; Washington, DC, USA. p. 923–37. doi:10.1145/3636534.3649380.
148. Son D, Park Y, Kim B, You I. A study on the implementation of a network function for real-time false base station detection for the next generation mobile communication environment. J Wirel Mob Netw Ubiquitous Comput Dependable Appl. 2024;15(1):184–201. doi:10.58346/jowua.2024.i1.013.
149. Xylouris G, Vekraki A, Christopoulou M, Kourtis MA, Markakis EK, Trakadas P. Advancing predictive security for consumer applications in beyond 5G/6G networks with annotated datasets. IEEE Trans Consum Electron. 2025;71(2):5108–18. doi:10.1109/tce.2025.3567151.
150. Pacherkar HS, Yan G. PROV5GC: hardening 5G core network security with attack detection and attribution based on provenance graphs. In: Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2024 May 27–29; Seoul, Republic of Korea. p. 254–64. doi:10.1145/3643833.3656129.
151. Wayan I, Abella V, Lastre J, Ko Y, You I. Enhancing roaming security in cloud-native 5G core network through deep learning-based intrusion detection system. Comput Model Eng Sci. 2025;145(2):2733–60. doi:10.32604/cmes.2025.072611.
152. Singh S. Advancing network security in 5G: leveraging the 5G-NIDD dataset for intrusion detection and mitigation. In: Proceedings of the 2025 IEEE 12th International Conference on Cyber Security and Cloud Computing (CSCloud); 2025 Nov 7–9; New York, NY, USA. p. 1–6. doi:10.1109/cscloud66326.2025.00055.
153. Xavier BM, Dzaferagic M, Martinello M, Ruffini M. Performance measurement dataset for open RAN with user mobility and security threats. Comput Netw. 2024;253:110710. doi:10.1016/j.comnet.2024.110710.
154. Karahan SN, Janjua MB, Yazıcı İ. Fake base station detection in interference-rich ISAC networks using machine learning approaches. In: Proceedings of the 2025 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT); 2025 Nov 17–19; Sakhir, Bahrain. p. 1–6. doi:10.1109/3ict68299.2025.11442106.
155. Durmuş Ö, Atasoy F, Türkoğlu M. Comparative detection of DDoS attacks on software-defined 5G network data using deep neural networks and machine learning methods. IEEE Access. 2025;13:194657–76. doi:10.1109/access.2025.3631396.
156. Lakshmi V, Rajkumar S. Hybrid ensemble federated learning using SMOTE-Tomek for efficient DDoS detection on constrained edge devices over 5G networks. Results Eng. 2025;28:107601. doi:10.1016/j.rineng.2025.107601.
157. Cirillo F, Esposito C, Taek Seo J. Benchmarking quantum machine learning methods for intrusion detection on noisy quantum computers. Quantum Mach Intell. 2026;8(1):27. doi:10.1007/s42484-026-00379-4.
158. Zhang HF, Chen ZY, Wang P, Guo LL, Wang TL, Yang XY, et al. Experimental robustness benchmarking of quantum neural networks on a superconducting quantum processor. arXiv:2505.16714. 2025.
159. Hassan T, Meneghello F, Restuccia F. AdvO-RAN: adversarial deep reinforcement learning in AI-driven open radio access networks. In: Proceedings of the Twenty-Sixth International Symposium on Theory, Algorithmic

Foundations, and Protocol Design for Mobile Networks and Mobile Computing; 2025 Oct 27–30; Houston, TX, USA. p. 231–40. doi:10.1145/3704413.3764425.
160. Ergu YA, Nguyen VL, Hwang RH, Lin YD, Cho CY, Yang HK. Unmasking vulnerabilities: adversarial attacks against DRL-based resource allocation in O-RAN. In: Proceedings of the ICC 2024—IEEE International Conference on Communications; 2024 Jun 9–13; Denver, CO, USA. p. 2378–83. doi:10.1109/icc51166.2024.10623131.
161. Chiejina A, Kim B, Chowhdury K, Shah VK. System-level analysis of adversarial attacks and defenses on intelligence in O-RAN based cellular networks. In: Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks; 2024 May 27–29; Seoul, Republic of Korea. p. 237–47. doi:10.1145/3643833.3656119.
162. Keerthi S, Anitha M, Rao A, Advaith PN, Nair AR, Manjunath A. Adversarial attacks on 6G networks—a survey. In: Proceedings of the 2025 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER); 2025 Oct 17–18; Mangalore, India. p. 464–9. doi:10.1109/discover66922.2025.11258985.
163. Latif S, Ahmad J, Boulila W, Djenouri D. Adversarially optimized multi-space prototypical network for intrusion detection in 5G-enabled IoT systems. Trans Mach Learn Comm Netw. 2026;4:575–90. doi:10.1109/tmlcn.2026.3673694.
164. Neha BT. Adaptive intrusion detection system leveraging dynamic neural models with adversarial learning for 5G/6G networks. In: Proceedings of the 2025 4th International Conference on Computer Technologies (ICCTech); 2025 Feb 20–23; Kuala Lumpur, Malaysia. p. 103–7. doi:10.1109/icctech66294.2025.00028.
165. Karatapu HRT, Rahaman SU, Patchipulusu S. Exposing and mitigating adversarial AI threats in IoT/5G intrusion detection systems using random forest and feature reduction. In: Proceedings of the 2025 IEEE International Carnahan Conference on Security Technology (ICCST); 2025 Oct 13–17; San Antonio, TX, USA. p. 1–8. doi:10.1109/iccst63435.2025.11293874.
166. Ahmad Soleymani S, Eslamnejad M, Shojafar M, Tafazolli R. Detecting and mitigating adversarial machine learning in open RAN using digital twin. IEEE Wirel Commun Lett. 2025;14(12):3972–6. doi:10.1109/lwc.2025.3609661.
167. Mao J, Wei Z, Li B, Zhang R, Hu X, Yang L. Cognitive security for AI-native IIoT: an LLM-centric adaptive defense framework. In: Proceedings of the 2026 International Conference on Computing, Networking and Communications (ICNC); 2026 Feb 16–19; Maui, HI, USA. p. 7–11. doi:10.1109/icnc68183.2026.11416932.
168. Chaskos E, Kolokotronis N, Shiaeles S. A next-generation cyber-range framework for O-RAN and 6G security validation. Future Internet. 2026;18(1):29. doi:10.3390/fi18010029.
169. O-RAN ALLIANCE. O-RAN security requirements and analysis specification (O-RAN.WG11). O-RAN ALLIANCE Working Group 11 (Security), Technical Specification. 2024 [cited 2026 May 21]. Available from: https://specifications.o-ran.org/download?id=772.
170. Azeez SD, Ilyas M, Ahmed SR. Graph neural networks with multi-head attention and SHAP-based explainability for robust, interpretable, and high-throughput intrusion detection in 5G-enabled software defined networks. Comput Mater Contin. 2026;87(3):1–10. doi:10.32604/cmc.2026.074930.
171. Kaur N, Gupta L. Explainable AI assisted IoMT security in future 6G networks. Future Internet. 2025;17(5):226. doi:10.3390/fi17050226.
172. Palanivel R, Muthulakshmi P, Dey S, Mehta S, Sapkale P. Role of QML in 6G integrated vehicular networks. In: Quantum computing and machine learning for 6G. Hoboken, NJ, USA: John Wiley & Sons, Ltd.; 2026. p. 327–48.
173. Swathi BH, Arvind R, Benvin S, Gupta R, Maranan R, Jayanthi R. Securing 5G-enabled cyber-physical systems: an optimized reflection equivariant quantum neural network approach to DDoS attack detection. In: Proceedings of the 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS); 2025 Apr 17–19; Namakkal, India. p. 606–12. doi:10.1109/icssas66150.2025.11080985.
174. Nguyen VL, Nguyen LH, Hwang RH, Canberk B, Duong TQ. Quantum machine learning for 6G network intelligence and adversarial threats. IEEE Comm Stand Mag. 2025;9(3):40–8. doi:10.1109/mcomstd.2025.3575261.
175. Ergu YA, Nguyen VL, Lin PC, Hwang RH. Q-poison: quantum adversarial attacks against QML-driven interference classification in O-RAN. In: Proceedings of the 2025 IEEE International Conference on Machine Learning for

Communication and Networking (ICMLCN); 2025 May 12–15; Barcelona, Spain. p. 1–6. doi:10.1109/icmlcn64995.2025.11140504.

176. Nowmi SR, Lopez J, Imon MMA, Pouryousef S, Rahman MS. SoK: critical evaluation of quantum machine learning for adversarial robustness. arXiv:2511.14989. 2026.
177. Alhumaima RS, Al-Karawi Y, Al-Raweshidy H. Cybersecure synchronisation of entangled quantum neural networks with reconfigurable intelligent surface and quantum key distribution for 6G holographic communications. IET Quantum Commun. 2026;7:e70032. doi:10.1049/qtc2.70032.