Digital Communications and Networks(DCN)

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/dcan

Towards secure intelligent O-RAN architecture: vulnerabilities, threats and promising technical solutions using LLMs

Mojdeh Karbalaee Motalleb[a], Chafika Benzaid[b], Tarik Taleb[*cd], Marcos Katz[b], Vahid Shah-Mansouri [a], Jaeho Kim[*d]

[a] *University of Tehran, Tehran, 14174-66191, Iran*
[b] *University of Oulu, Oulu, 90014, Finland*
[c] *Ruhr University Bochum, Bochum, Germany*
[d] *Sejong University, Seoul, 05006, Korea*

Abstract

The evolution of wireless communication systems will be fundamentally impacted by an open radio access network (O-RAN), a new concept defining an intelligent architecture with enhanced flexibility, openness, and the ability to slice services more efficiently. For all its promises and like any technological advancement, O-RAN is not without risks that need to be carefully assessed and properly addressed to accelerate its wide adoption in future mobile networks. In this paper, we present an in-depth security analysis of the O-RAN architecture, discussing the potential threats that may arise in different O-RAN architecture layers and their impact on the Confidentiality, Integrity, and Availability (CIA) triad. We also promote the potential of zero trust, Moving Target Defense (MTD), blockchain, and Large Language Models (LLM) technologies in fortifying O-RAN's security posture. Furthermore, we numerically demonstrate the effectiveness of MTD in empowering robust deep reinforcement learning methods for dynamic network slice admission control in the O-RAN architecture. Moreover, we examine the effect of Explainable AI (XAI) based on Large Language Models (LLM) in securing the system.

KEYWORDS:
Open radio access network, O-RAN security, Large language models, Blockchain, Moving target defense

1. Introduction

Wireless systems are becoming more capable but more complex in the next generation of cellular networks. Unlike previous generations, the next generation will be flexible, agile, modular, supporting heterogeneity in services, multiple technologies, and rapid deployment [1]. Radio Access Networks (RAN) performance is expected to be significantly improved with O-RAN, which combines and evolves the Cloud RAN (C-RAN) and virtual RAN (vRAN) to enable an open and flexible RAN. In the O-RAN architecture, the components of RANs are virtualized and decoupled, using compatible open interfaces developed for their interconnection. Moreover, the O-RAN's architecture utilizes Artificial Intelligence and Machine Learning (AI/ML) techniques to develop intelligent RAN layers, allowing to empower intelligent, data-driven closed-loop control for the RAN [2, 3, 4]. These features bring many benefits to the system, including reduced Capital Expenditures (CAPEX) and Operating Expenses (OPEX), increased agility and flexibility, and enhanced visibility and security.

For all its promises and like any technological advancement, O-RAN is not without risks that must be assessed and properly addressed to accelerate its widespread adoption in future mobile networks. Recent studies have shown that the O-RAN architecture introduces a new range of security challenges, driven by newly defined components and open interfaces, the use of open-source software, the disaggregation of hardware and software, and the reliance on cloud-native and AI technologies, among others [5]. Therefore, a comprehensive review of security aspects is necessary, considering potential risks, vulnerabilities, and applicable solutions. Such an investigation is crucial to strengthening O-RAN's security posture at this early stage of its development [6].

This paper explores security threats across the layers of the intelligent O-RAN architecture and proposes key technologies to mitigate them, emphasizing the need for proactive measures to secure next-generation networks. Unlike previous studies, such as [7], which focus on specific vulnerabilities and security methods for 5G, such as Zero Trust, our research examines a broader range of vulnerabilities in O-RAN and presents innovative solutions to secure both the near-Real-Time RAN Intelligent Controller (near-RT RIC) and the non-Real-Time RAN Intelligent Controller (non-RT RIC). These controllers integrate AI/ML methods for system automation, making it essential to safeguard AI/ML models against potential threats [8]. Moreover, the near-RT RIC and non-RT RIC incorporate third-party applications that leverage AI/ML techniques for resource allocation.

[1]Emails: [a] `mojdeh.karbalaee@ut.ac.ir`, `vmansouri@ut.ac.ir`
[b] `chafika.benzaid@oulu.fi`, `marcos.katz@oulu.fi`
[c] `tarik.taleb@ruhr-uni-bochum.de`
[d] `kimjh@sejong.ac.kr`.

We introduce a novel Moving Target Defense (MTD) technique to mitigate attacks on this system, demonstrating a significant reduction in adversarial attacks in the results.

In addition to traditional security mechanisms, we propose the novel use of Large Language Models (LLMs) to enhance the system's security. The LLM system can analyze data in real time and provide human-readable explanations to assist in detecting vulnerabilities. Using Explainable AI (XAI), the LLM model can identify significant changes in data patterns over time and alert the system to potential vulnerabilities.

Research contributions of this paper are listed as follows:

- An in-depth analysis of vulnerabilities and threats in the O-RAN architecture arising from the introduction of new technologies and common 5G RAN security issues.
- The proposal of four countermeasure approaches utilizing the zero trust concept, blockchain technology, LLM based XAI and the MTD paradigm.
- Considering the Confidentiality, Integrity, and Availability (CIA) table for the threats and approaches.
- Case studies and proof-of-concept demonstrations of MTD-based robust ML in O-RAN and LLM-based robust AI/ML in O-RAN, illustrating the effectiveness of MTD in enhancing the robustness of deep reinforcement learning models. We demonstrate that the secured MTD system significantly reduced the impact of adversarial attacks, with only a 21.5% decrease in admission rate (Fig. 4a) and a 21% decrease in Fig. 4b, compared to a 92% and 87% drop, respectively, in the absence of protection.

The remainder of this paper is as follows: Section II provides an overview of the O-RAN architecture, focusing on its key components: RAN, cloud, and management layers, along with ML and network slicing. Section III examines vulnerabilities and threats in the O-RAN architecture, analyzing their impact on CIA. Section IV explores emerging technologies such as zero-trust (ZT), blockchain, moving target defense (MTD), and LLMs to enhance O-RAN security. In Section V, we propose a novel MTD-based solution demonstrating its effectiveness in securing Deep Reinforcement Learning (DRL) against adversarial attacks in the Near-RT RIC. Additionally, we discuss the application of LLM-based Explainable AI (XAI) for detecting AI/ML attacks in O-RAN. Finally, conclusions are drawn in Section 6.

2. O-RAN Background

The O-RAN Alliance[2] has developed a novel RAN architecture to facilitate an open, intelligent, virtualized, and interoperable RAN, essential for cost-effective, next-generation wireless networks. This architecture integrates the advantages of C-RAN and vRAN, leveraging cloudification, centralization, and hardware-software decoupling to address vendor lock-in and proprietary issues via standard interfaces. O-RAN developed a multi-vendor ecosystem and embedded AI/ML for improved network intelligence.

The O-RAN architecture includes three components in the baseband side: the Radio Unit (O-RU), Distributed Unit (O-DU), and Central Unit (O-CU). The O-RU contains the Radio Frequency (RF) and low Physical (PHY) layers, while O-DU provides the functionalities of the high PHY, Medium Access Control (MAC), and Radio Link Control (RLC) layers. The Open Fronthaul (Open-FH) is the interface between the O-RU and the O-DU. The Open-FH interface includes a Control User Synchronization plane (CUS-plane) and a Management plane (M-plane). The O-CU is divided into two logical nodes the user plane (O-CU-UP) and the control plane (O-CU-CP). The O-CU-UP encompasses the Service Data Adaptation Protocol (SDAP), and the user plane part of the Packet Data Convergence Protocol (PDCP). The O-CU-CP hosts the Radio Resource Control (RRC) layer, and the control plane of the PDCP protocol [9]. Fig.1a illustrates O-RAN's architecture.

The O-RAN architecture also includes a management part which comprises Service Management and Orchestration (SMO), Near Real-Time RAN Intelligent Controllers (RICs), and O-Clouds blocks. SMO includes functions such as Non-Real-Time RIC. Generally, the near-RT and non-RT RIC are responsible for AI/ML methods and making the system more intelligent. The AI/ML technologies plays a crucial role in the resource allocation within RAN systems. In the O-RAN system, near-RT RICs are functions that provide near real-time control and optimization of network resources through the E2 interface. This includes xApplications (xApps), which are third-party applications that run by leveraging the modules and capabilities of a system for functionalities such as resource allocation.

The O-Cloud platform, known as a cloud computing platform, hosts O-RAN architecture components depicted in Fig. 1b [10]. The RAN network functions can be deployed as Virtualized Network Functions (VNFs) on Virtual Machines (VMs) or as Cloud-native Network Functions (CNFs) in containers. The O-Cloud platform supports these options with its virtualization layer, which includes operating systems, hypervisors, and container engines. Additionally, the O-RAN ecosystem supports and interfaces with bare-metal, hardware-based RAN functions. The SMO system connects to the O-Cloud via the O2 interface, enabling efficient resource and workload management [11].

In the following, we provide a concise overview of the key techniques and features employed within the O-RAN system, enhancing its flexibility and performance.

2.1. Network Slicing in O-RAN

Network slicing, essential for 5G revenue, dynamically creates customized virtual networks on shared infrastructure, integrating network functions and resources across RAN, transport, and core networks to meet specific service needs. RAN slicing involves the isolation of Physical Resource Blocks (PRBs) and specific Virtual Network Functions (VNFs) such as MAC, RLC in the O-DU, and PDCP, SDAP in the O-CU for various services as illustrated in Figure 1 of [2]. In addition, core slicing virtualizes and isolates nodes like UPF and AMF, catering to the specific needs of each service. Finally, transport slicing creates dedicated pathways across the shared underlay network, ensuring guaranteed performance for these diverse service connections. By working together, RAN, core, and transport slicing unlock the full potential of 5G networks. O-RAN's virtualization and intelligence are key to advancing RAN slicing, essential for end-to-end network services [2, 12].

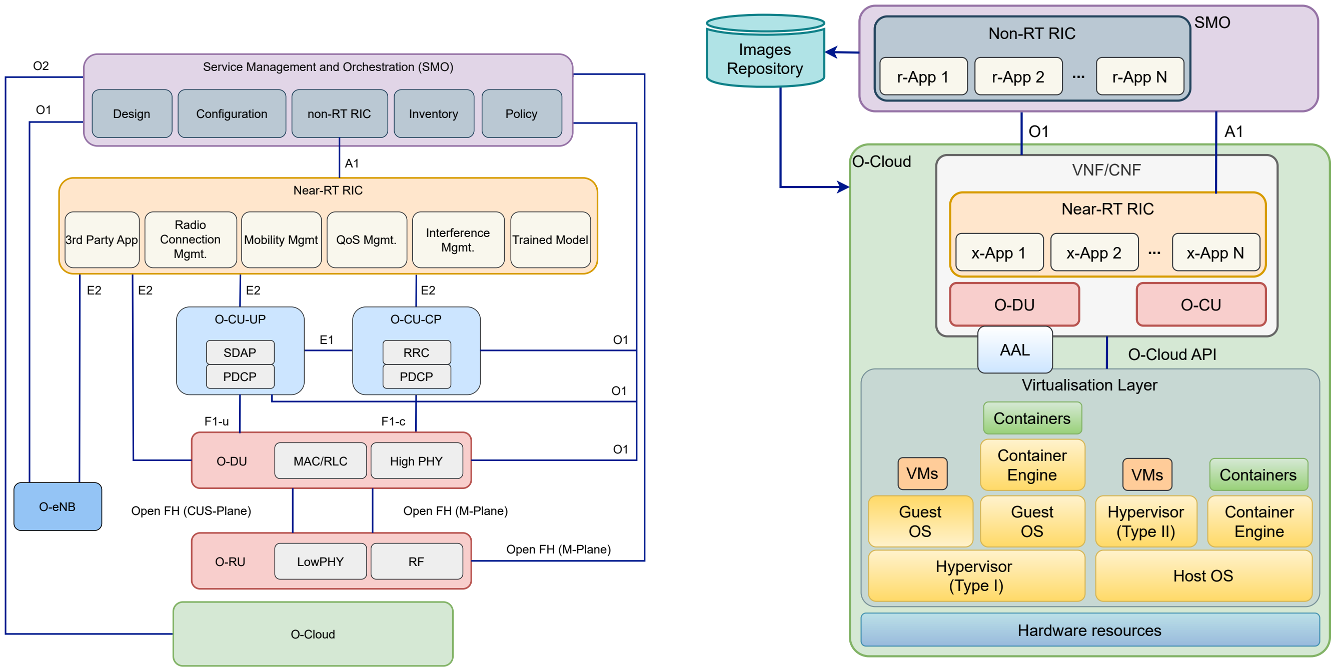
2.2. Radio Intelligent Controller (RIC)

The Near-RT and Non-RT RICs are essential for O-RAN system management, serving as an open hosting platform and optimizing RAN functions. The RIC consists of Near-RT RIC and Non-RT RIC, facilitating intelligent RAN optimization on near-real-time (10 – 1000 msec) and non-real-time (greater than 1s) scales, respectively. The Near-RT RIC uses xApps for real-time RAN control via E2 interfaces with O-RAN components, while the Non-RT RIC employs rApps for broader RAN optimization and is linked to the Near-RT RIC through the A1 interface for policy and AI/ML model management. The near-RT RIC and non-RT RIC are vital components responsible for the AI/ML workflow in the O-RAN architecture[11, 13, 1].

2.3. ML aspect in O-RAN

The O-RAN architecture incorporates AI/ML to add intelligence across its RAN layers, a move seen as pivotal for highly autonomous RAN functions that improve service quality and lower OPEX. AI/ML is expected to be instrumental in a range of RAN use cases, from resource allocation to anomaly detection and cybersecurity. Subsequently, we will outline potential ML techniques applicable to O-RAN and detail the general ML lifecycle.

[2] https://www.o-ran.org



(a)

(b)

Fig. 1. (a) The O-RAN high-level architecture with components and interfaces, (b) The O-Cloud architecture, which is a set of computing resources and virtualization infrastructure.

2.3.1. *ML techniques*

In the O-RAN system, various ML techniques are utilized: (1) supervised learning for model training with labeled data and subsequent prediction on new data; (2) unsupervised learning to find patterns in unlabeled data; (3) Reinforcement learning (RL) and Deep RL (DRL) for learning optimal actions through interaction with the environment; and (4) Federated Learning (FL) for privacy-preserving collaborative model training across distributed entities without data exchange, using a central server to aggregate local model updates. In addition, LLMs can also be incorporated to enhance communication performance and the decision-making processes by analyzing and generating human-like text, providing valuable insights within the O-RAN architecture. Moreover, integrating LLMs with existing ML methods can significantly improve the system's overall intelligence and efficiency.

In O-RAN architecture, Non-RT RIC and Near-RT RIC are responsible for AI/ML techniques, where they can play the role of ML training host and/or ML model host/actor [13]. The ML training host VNF trains models within the Non-RT RIC, while the ML model host/actor VNF, for inference, may reside in either Non-RT or Near-RT RIC. In RL, Near-RT RIC conducts online training and inference, while Non-RT RIC is for offline training and Near-RT RIC for inference. FL uses Non-RT RIC as the central server and Near-RT RIC for distributed training.

2.3.2. *ML Life Cycle Procedure*

Despite the variety of ML techniques supported and the deployment scenarios considered for placing the ML training hosts and ML model hosts/actors, a general ML lifecycle in the O-RAN architecture can be described as follows (See Fig. 2) [1, 13]:Firstly, the ML Designer, deployed the model (stage 1 and 2). The data is selected for training (stage 3) and fed into the ML model during the training and inference stages. The data are typically collected over E2, O1, and A1, from O-CU, O-DU, and RICs (stage 8). The collected data are prepared in the RICs to fit the ML models by performing data pre-processing operations, including dataset balancing, normalization, and removing noise, among others. The ML model goes first through the training process, where

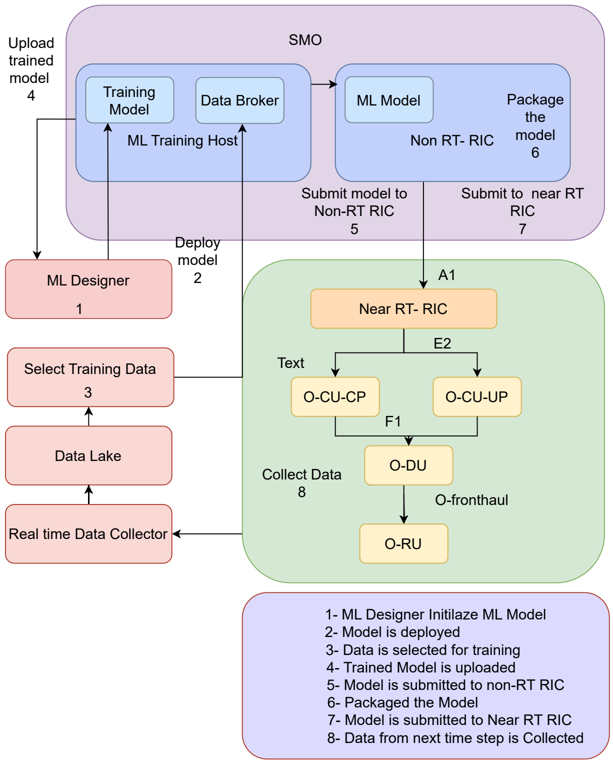


Fig. 2. ML Model Life Cycle in the O-RAN Architecture.

the ML designer or SMO/Non-RT RIC will select and implement the ML algorithm to train in the ML training host. The trained model is then uploaded (stage 4) and validated to ensure its reliability and accuracy. Once the model is validated, it is stored and published in the SMO/Non-RT RIC catalog (stage 5). After a model has been validated (stage 6), it can be deployed and executed (stage 7).

3. Vulnerabilities and Threats in O-RAN Architecture

The openness and disaggregation of the O-RAN architecture facilitate compliance with security standards and enable improved security agility, adaptability, and resiliency for future mobile networks. In addition to these benefits, the O-RAN architecture introduces the potential for an increased attack surface [14]. The O-RAN Alliance's Security Work Group 11 focuses on securing O-RAN, but their measures are insufficient, particularly against malicious AI/ML methods. Therefore, additional security perspectives are necessary. This section discusses key vulnerabilities and threats to O-RAN, including the new security issues of O-RAN technologies.

3.1. O-RAN System Vulnerabilities

As previously discussed, the O-RAN system comprises three different sides (radio, management, cloud), each with its own vulnerabilities tied to their respective roles and functions. This section delves into the vulnerabilities inherent to the different sides of the O-RAN architecture.

3.1.1. O-RU/O-DU and Open-FH Vulnerabilities

In radio communication, the O-RAN architecture and other RAN generations have inherent vulnerabilities. This section outlines these vulnerabilities, particularly focusing on O-RAN. One key threat is the False Base Station (FBS) attack, where an attacker poses as a legitimate base station to execute a Man-in-The-Middle (MiTM) attack. Three FBS attack scenarios on an O-RU include hijacking fronthaul, recruiting a standalone O-RU, and gaining unauthorized physical access. These attacks can compromise both O-RAN and other RAN systems [14, 15, 16].

There are several risks associated with FBSs in the network, including stealing subscriber information, altering and redirecting transmitted data, and compromising subscriber privacy. The FBS attacks may help in penetrating O-DU and beyond in the CN and launching DoS attacks to cause loss of service or degradation of its performance.

Given that the O-DU and O-RU can be from different vendors, they may have varying security levels. The O-DU's role in managing traffic between the management system and the O-RU increases the risk of unauthorized access to other systems, such as RICs, via the Open-FH interface. An unprotected Open-FH interface can also enable MiTM attacks, allowing data tampering, disclosure, and DoS attacks. For instance, an unauthorized device on the Open-FH Ethernet L1 interface could launch a flooding attack, causing unavailability or performance degradation of legitimate network elements.

3.1.2. Near-RT RIC Vulnerabilities

Through standardized interfaces and hardware support, the Near-RT RIC provides a safe and reliable platform for hosting xApps. The xApps are independent of the Near-RT RIC and may be supplied by a third-party vendor. The Near-RT RIC and xApps can be sources of different security threats [14].

A malicious or compromised xApp has the potential to negatively impact the service delivery for a subscriber, a group of subscribers, or a specific geographic area by manipulating data collected from E2 nodes (i.e., O-DU, O-CU-CP and O-CU-UP) and A1 interface. It introduces also the risk of obtaining unauthorized access to E2 nodes and Near-RT RIC, exploiting the RAN functions and engendering harmful effects to the overall system. Leakage of sensitive data (e.g., UE identification and location) is another menace that could stem from malicious/compromised xApps. The disclosure of sensitive information will not only pose privacy violation issues but may also lead to the launch of other attacks, such as impersonation and UE tracking attacks. The xApps cannot operate independently from the components of the Near-RT RIC. They need to interact with these components to access their functionalities. For instance, they communicate with the App Manager during registration and the Sub Manager to subscribe to data from E2 nodes. Due to this communication, a malicious xApp can affect other components of Near-RT RIC too.

This could happen by exploiting shared resources, manipulating control messages, disrupting event processing, compromising security credentials, introducing hidden logic bombs, or exfiltrating sensitive data through communication channels within the framework. Additionally, resources such as CPU and RAM limits can be specified in the xApp descriptors to prevent resource exhaustion, which is enforced by Kubernetes. Hence, a malicious xApp can use more resources than it needs.

The indefinite functional split between Near-RT RIC and E2 nodes, which depends on the available xApps and the capabilities of E2 nodes, may result in conflicts between decisions taken by the Near-RT RIC and the E2 nodes. Moreover, developing multiple xApps with overlapping objectives within the same RAN may lead to conflicting actions between xApps. Those conflicts can degrade the system's performance or may cause a Denial-of-Service (DoS) attack intentionally or unintentionally in the O-RAN architecture.

The lack of proper isolation between an xApp and the other Near-RT RIC components may be a source of serious security breaches. In fact, with the recent trend to evolve VNFs into CNFs, complete isolation between co-hosted CNFs is hard to realize due to the lack of strong hardware isolation in the emerging cloud-native platforms (e.g., Kubernetes). Thus, an xApp with compromised isolation can be exploited to escalate the privilege granted to it, carry out shared resource exhaustion attacks, steal secrets and sensitive information from memory, and conduct DoS attacks against co-hosted xApps and the Near-RT RIC platform.

3.1.3. SMO Vulnerabilities

SMO security is critical because a vulnerability can allow attacks on O-RAN components and lateral movement within the network. Weak authentication and authorization can let attackers access and alter SMO data, control O-RAN components, and steal sensitive information. For example, unauthorized access to Non-RT RIC via SMO can lead to UE tracking or issuing false policies to Near-RT RIC. Additionally, SMO and Non-RT RIC are susceptible to DoS attacks, which can impair network monitoring and control functions. The security concerns for rApps in Non-RT RIC are similar to those for xApps [14].

3.2. O-Cloud Vulnerabilities

The O-Cloud platform in O-RAN architecture faces common cloud security risks, including software flaws, valid account access, and lack of interface authentication. Malicious actors can exploit VMs and containers running O-RAN components, leading to privilege escalation, malware contamination, unauthorized deployment of VMs/containers, root server access, and system destruction. They can also access and manipulate sensitive data. Deploying vulnerable VMs/containers risks DoS attacks on shared resources, which can be economically damaging if turned into an EDoS attack. Supply chain attacks can inject malicious code or extract private keys from VM/container images. Additionally, an unprotected O2 interface between O-Cloud and SMO is vulnerable to MiTM attacks, allowing tampering and disclosure of services and requests.

3.3. Open Source Code Vulnerabilities

Open-source software is crucial for building the software-based O-RAN architecture, used in both cloud infrastructure and O-RAN components. It accelerates development, promotes vendor independence, and reduces costs. However, it also poses security challenges. The open source code allows attackers to find and exploit vulnerabilities.

Without an accurate, up-to-date inventory of open-source codes and dependencies, managing and mitigating high-risk vulnerabilities becomes difficult due to the volume, variety, and lack of standard naming conventions.

3.4. ML System Vulnerabilities

Integrating ML techniques into O-RAN enhances autonomous RAN functions but also introduces significant security challenges. ML models are vulnerable to adversarial attacks that manipulate decisions, compromise model integrity, or reveal private information. Attacks include altering training datasets, injecting fake data during online learning, or crafting inputs to deceive models during operation. Collaborative learning methods like FL face model poisoning attacks, where malicious agents tamper with local model parameters to compromise the global model. FL is also susceptible to inference attacks, allowing attackers to deduce private training data using local model parameters [5, 17].

Based on accessibility, attacks on ML models can be categorized into white-box, black-box, and gray-box attacks [17]. Indeed, the adversarial attack is considered as a white box, gray box, or black box when the attacker can have full, partial, or no access to the training data and the targeted model's parameters and architecture, respectively. The white-box attack is deemed less realistic due to the assumption of an attacker with full knowledge, which is hard to achieve in real-world scenarios.

3.5. Threats against 5G Radio Networks

Common threats to traditional RAN architectures are also applicable to O-RAN architecture. This includes (i) jamming attacks, which consist of blocking radio signals; for example by introducing intentional interference in the communication channels; (ii) sniffing attacks, which focus on observing and collecting data packets with the purpose of extracting sensitive information (e.g., UE location and cell configuration) as well as using the extract information to craft new attacks; and (iii) spoofing attacks, which refer to creating a fake signal that is hard to distinguish from the actual signal, allowing an attacker to impersonate a base station, cause a DoS, or bypass physical-layer signal authentication [17], among others.

3.6. Physical Threats

Physical threats, though not unique to O-RAN, are crucial to understanding its vulnerabilities. The physical infrastructure, including cell sites and data centers, faces risks from unauthorized access, power outages, natural disasters, and hardware failures. Intruders can sabotage hardware or alter settings to provoke DoS, inject malware, or access other network components. Natural disasters like snow, floods, earthquakes, and lightning can damage physical components. Lack of proper procedures for hardware failures and power outages increases the risk of unavailability. Physical security is more challenging in O-RAN due to the higher number of cell sites, data centers, and vendors.

Table 1 summarizes the main security threats discussed above, highlighting their impact on the CIA triad. Note that the threats marked with the (✓) sign affect a CIA principle, while those marked with (x) do not. Moreover, (✓) and (x) indicate whether the potential mitigation of vulnerabilities through Zero Trust (ZT), Blockchain (BC), Moving Target Defense (MTD), and LLM investigated in Section 4 is applicable or not, respectively.

4. Security Solutions in O-RAN

There are different possible solutions for security threats and vulnerabilities [18]. This section discusses several key emerging technologies that can be leveraged to improve the security of the O-RAN architecture.

4.1. Zero Trust

Zero trust (ZT) is a valuable security model for enhancing O-RAN security. Based on "never trust, always verify," it assumes breaches can occur anytime from internal or external threats. ZT principles include continuous identification and authentication, enforcing least-privilege access, maintaining risk-based policies, checking communication channels, and continuous security monitoring. Implementing ZT protects the entire O-RAN architecture, from hardware to applications. AI/ML techniques and Security-as-a-Service (SECaaS) enable ZT by allowing instant threat identification and automated security adjustments [19, 20].

4.2. Blockchain

Blockchain (BC) is a promising solution for securing O-RAN architecture with a zero trust mindset. Its features of decentralization, immutability, transparency, auditability, and smart contract auto-execution support various security controls in O-RAN. These controls include privacy-enhanced identity management, mutual authentication, dynamic access control, integrity and non-repudiation of data and software, and secure resource sharing. For example, in AI security, blockchain can ensure the integrity and provenance of data in a ML pipeline and protect against poisoning attacks on FL models [18, 19, 21, 22, 23].

4.3. MTD

MTD has recently emerged as an effective approach to enable proactive security. The core principle of MTD is to constantly and dynamically modify the configuration of the network and services to increase uncertainty and complexity for attackers. In fact, the dynamicity introduced by MTD reduces the attacker's opportunities to gather useful information on vulnerabilities of the target environment, preventing their exploitation. To this end, different MTD techniques can be applied, which are broadly categorized into *shuffling* (e.g., network topology, VMs/containers placement), *diversity* (e.g., in underlying technology used to implement or run a service), and *redundancy* (e.g., by providing multiple replicas of a network component or service). In O-RAN, the MTD approach can be used to prevent intrusions, mitigate DoS attacks, and increase the robustness of ML models to adversarial attacks (Table 1), among others. For example, the resiliency of ML models can be strengthened by continuously changing the ML algorithm, the features used for its training, or the model's parameters [17]. Moreover, to determine whether we have resources to allocate to UE, we can use the AI/ML method for the admission control system. This AI/ML system can be protected using MTD by considering different AI/ML training models with different configurations that are chosen randomly by MTD.

4.4. Large Language Models

The deployment of Large Language Models (LLMs) within O-RAN networks can significantly enhance cybersecurity measures by capitalizing on their exceptional data processing and pattern recognition capabilities. In the context of O-RAN, where a diverse array of virtualized network functions operates across open interfaces, LLMs can meticulously monitor and analyze network traffic and system logs. This enables the early detection of anomalous behaviors that could signal a security breach, such as unusual login patterns or unexpected changes in data flow, which are critical in the multi-vendor O-RAN environment.

LLMs can dynamically adjust security policies for each O-RAN network slice by analyzing data to make smart access choices, fine-tune encryption, and improve intrusion detection, resulting in personalized security. We can fine-tune the LLM system for specific tasks according to our requirements for the next generation of RAN system [24, 25]. For instance, we can fine-tune the LLM system to analyze the data and diagnosis to early warnings.

Let us consider a specific scenario: in the event of a sudden surge in traffic indicating a potential DDoS attack within a network slice, an LLM equipped with real-time analytics can autonomously adjust traffic

Table 1
Impact of threats and vulnerabilities in O-RAN system on Confidentiality (**C**), Integrity (**I**) and Availability (**A**); and the Potential Mitigation of Vulnerabilities through Zero Trust (**ZT**), Blockchain (**BC**), Moving Target Defense (**MTD**), Large Language Model (**LLM**).

Threats and Vulnerabilities	**C**	**I**	**A**	**ZT**	**BC**	**MTD**	**LLM**
Conflicts among xApps or rApps	x	x	✓	x	x	✓	✓
Accessing a misconfigured x/rApps	✓	x	x	x	✓	x	✓
Altering Data through malicious x/rApps attacks	✓	✓	x	✓	✓	x	✓
Conflicts between Near-RT RIC and O-gNB/eNB	x	x	✓	x	x	✓	✓
FBS attacks on O-RU	✓	✓	x	✓	✓	x	x
Eavesdropping on air interfaces	✓	x	x	x	x	x	x
Accessing the O-RU/DU/CU and degrading the O-RAN's performance	x	x	✓	x	x	✓	x
MiTM attack from the Open-FH over M-plane or CUS-plane	✓	✓	x	✓	✓	x	x
Misconfiguration, lack of isolation and security in the O-Cloud	✓	✓	✓	✓	✓	✓	✓
Open-source code vulnerabilities	✓	✓	✓	✓	✓	✓	✓
Adversarial attacks against ML	✓	✓	✓	✓	✓	✓	✓
Jamming attacks	x	x	✓	x	x	✓	✓
Spoofing attacks	✓	✓	✓	✓	✓	✓	✓
Physical threats	✓	✓	✓	x	x	x	x

rules and resource allocations to mitigate the threat. This proactive approach not only ensures uninterrupted service, but also enhances overall security by continuously monitoring for vulnerabilities and updating configurations. In the realm of O-RAN, where AI/ML-driven solutions are paramount, LLMs can also contribute to the secure orchestration of network elements by generating and updating security configurations and orchestrating responses to threats in collaboration with the SMO framework. This not only streamlines the management of complex O-RAN architectures, but also fortifies them against sophisticated cyber threats, ensuring the network's integrity and the trust of its users.

By integrating LLMs into the O-RAN security strategy, network operators can leverage the full potential of AI to maintain a robust, adaptive, and intelligent defense system, keeping pace with the evolving cyber-security landscape while supporting the continuous growth and innovation inherent to O-RAN networks.

In addition, LLMs can be used to enhance XAI systems in O-RAN by providing human-like explanations for the decisions and predictions made by various AI/ML components. As a result, XAI reduces the risk of false positives and improves the accuracy of AI / ML security [26, 27]. When systems or operators understand the reasoning behind AI decisions, they can fine-tune the system to be more precise, leading to better detection of genuine threats and fewer mistakes. In other words, XAI with the help of LLMs not only makes AI more transparent but also smarter and more reliable when it comes to keeping the network safe [28].

4.5. *Effect of Security Solutions on different Vulnerabilities*

This section examines the impact of ZT, BC, MTD, and LLM on the vulnerabilities listed in Table 1. Conflicts among xApps or rApps, and between Near-RT RIC and O-gNB/eNB, and accessing the O-RU/DU/CU and degrading the O-RAN's performance can be prevented and resolved by implementing the MTD method, which constantly changes the configuration and environment of the system. Moreover, MTD could potentially mitigate jamming attacks by dynamically changing frequencies or communication patterns.

BC can prevent misconfigured x/rApps from being accessed by ensuring that configurations are recorded immutably, making misconfigurations easier to detect. When malicious x/rApps attacks alter data, BC ensures data integrity, while Zero Trust prevents unauthorized access, mitigating risk. In order to prevent FBS attacks on O-RU and MiTM attacks from the Open-FH over M-plane or CUS-plane, BC could provide a secure and transparent method for firmware distribution and communication channel, while Zero Trust could prevent unauthorized access to the system.

Misconfiguration, open-source code vulnerabilities, and adversarial attacks against machine learning can be secured by employing Blockchain for immutable logging and verification, Zero Trust for rigorous access control and continuous authentication, and MTD to dynamically alter the system's attack surface, complicating potential exploitation efforts. Moreover, LLM can help in detecting many threats shown in Table 1 such as adversarial attacks against AI/ML using XAI, open-source code vulnerabilities, jamming and spoofing using various analyses, and pattern recognition techniques.

5. Secure O-RAN Case Studies

In this section, we investigate two case studies: MTD-based Robust ML and LLM-based XAI Robust AI/ML in O-RAN. MTD and LLM-based XAI were preferred over ZT and blockchain for securing the O-RAN architecture with ML techniques due to their flexibility and efficiency. MTD provides dynamic protection against evolving threats, while LLM-based XAI enables real-time anomaly detection and explainability. In contrast, ZT and blockchain face challenges related to scalability, complexity, and performance, making them less suitable for the high-performance needs of O-RAN.

The first study explores the application of the MTD approach in enhancing deep reinforcement learning methods for dynamic network slice admission control within the O-RAN architecture. The second study focuses on the use of an LLM XAI system for diagnosing and explaining aberrant behavior.

5.1. *MTD-based Robust ML in O-RAN*

This section presents a practical study, corroborating the capabilities of the MTD approach in empowering robust DRL methods for dynamic network slice admission control in the O-RAN architecture [8]. While AI/ML is essential in the O-RAN for functions such as resource allocation and network slicing, its security is vital to ensure the reliability of 5G and 6G networks. Therefore, MTD is chosen for the study due to its agility in reconfiguring ML systems within O-RAN, effectively disrupting attack vectors and fortifying against the complex threats of future wireless networks.

5.1.1. System Scenario

We consider a scenario of service admission control, as shown in Fig. 3, in which we have two different services in the O-RAN architecture. In order to provide a service requirement, a specific amount of resources is needed. Each service is assigned to its slices based on the network slicing technique in the O-RAN architecture. Each slice contains VNFs in the O-DU and O-CU layers.

In this study, we implement a simulation for the O-RAN architecture by considering the O-DU and O-CU as specific VNFs with memory requirements. For simplicity, we assume that O-DU and O-CU use the same processors. Additionally, in the near-RT RIC, the AI/ML models are trained to solve the resource allocation problem. This model is implemented as an xApp within the system. We suppose that the system has enough CPU and storage resources while it has restricted memory resources. We consider a dynamic resource allocation model for VNFs of O-DU and O-CU slices for service admission control problems. Our goal is to maximize the total service admission rate. We suppose that services have the same priority in this system model. In this service, we assume the system is dynamic, and in each time slot, we have service requests from the two services that arrive following a Poisson process. Additionally, we assume that these two services have a service departure rate that has an exponential distribution.

Suppose we have a tuple that represents the required resources for VNF m in the O-DU or O-CU ($m^{\mathfrak{z}}$, $\mathfrak{z} \in c, d$) within slice s, denoted as $\bar{\psi} s^{m^{\mathfrak{z}}} = \{\psi^{m^{\mathfrak{z}}}_{\mathsf{C},s}, \psi^{m^{\mathfrak{z}}}_{\mathsf{S},s}, \psi^{m^{\mathfrak{z}}}_{\mathsf{M},s}\}$. Here, $\psi^{m}_{\mathsf{C},s}$, $\psi^{m}_{\mathsf{S},s}$, $\psi^{m}_{\mathsf{B},s}$, and $\psi^{m}_{\mathsf{M},s}$ indicate the required amounts of CPU, storage, bandwidth, and memory, respectively, for the VNFs of the O-DU (d) or O-CU (c). Assume there are N data centers designated for the VNFs of the O-DU and O-CU. Each data center n possesses a memory resource capacity denoted as χ^{n}_{s}. Assume $x_{m^{\mathfrak{z}}_{s},n} \in 0, 1$ is a binary variable indicating whether the VNF $m^{\mathfrak{z}}_{s}$ in layer O-DU/O-CU ($\mathfrak{z} \in c, d$) within slice s is being hosted by data center n. In this system model, we aim to maximize the service admission rate ($\sum_{n=1}^{N} \sum_{m_s=1}^{M_s} x_{m_s,n}$) with the constraint that $x_{m_s,n}$ is a binary variable. Additionally, $\sum_{s=1}^{S} \sum_{m_s=1}^{M_s} x_{m_s,n} \bar{\psi}^{\mathfrak{z},tot}_{\mathsf{M},s} \leq \chi^{n}_{\mathsf{M},s} \quad \forall n$, meaning that the total memory used by the VNFs hosted on server n must not exceed the server's total memory. Hence the main problem is

$$\max_{\boldsymbol{X},\boldsymbol{M}} \quad \sum_{n=1}^{N} \sum_{m_s=1}^{M_s} x_{m_s,n} \tag{1a}$$

$$\text{subject to} \quad \sum_{s=1}^{S} \sum_{m_s=1}^{M_s} x_{m_s,n} \bar{\psi}^{\mathfrak{z},tot}_{\mathsf{M},s} \leq \chi^{n}_{\mathsf{M},s} \quad \forall n \tag{1b}$$

$$x_{m_s,n} \in \{0, 1\} \quad \forall n, \forall s, \forall m_s \tag{1c}$$

This problem was modeled and solved in Python using the PPO model which is a DRL method.

5.1.2. Proposed Service Admission Algorithm

To solve this service admission control problem, we consider a DRL method that is implemented in the Near-RT RIC. Moreover, we assume the memory is quantized [29]. Therefore, we have discrete action and space. The DRL method adopted is Proximal Policy Optimization (PPO); an actor-critic method. Two models have been developed in the Actor-Critic system, namely: the Actor and the Critic. The Actor decides to take which action, and it updates the policy network for the selected agent. The Critic corresponds to the value function. During updating the Actor, the Critic modifies the network parameters for the value function. In the DRL models, we need to consider three aspects to solve the optimization problem, namely state, action, and reward. In this system, the state is the remaining memory we have in each time step, appended to the service arrival rate for two services which are random variables with a Poisson distribution, while the actions are the service admission for the two services that we considered. Moreover, the reward is the function of the service admission rate and the remaining memory. A reward is a huge negative number if the remaining memory is less than zero.

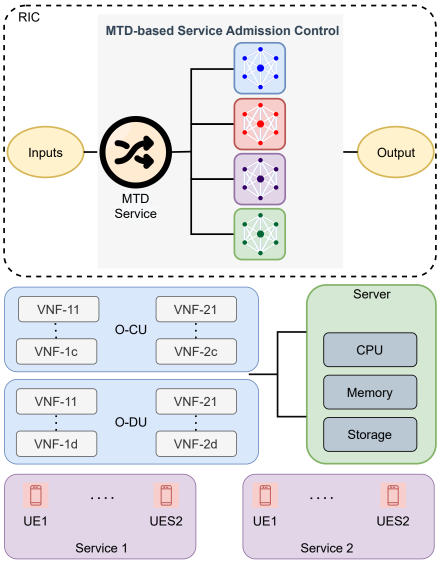


Fig. 3. MTD-based dynamic VNF placement scenario based on service request.

5.1.3. Attack Model

This section describes a malicious adversarial attack on the proposed PPO method. We consider a black-box poisoning attack against the PPO-based DRL agent. To this end, we use a weak adversary attack as in [30] to attack the system. Suppose the attacker determines to attack the time step t, it generates an arbitrary state $\hat{s}_t$ and the associated reward function $\hat{r}(\hat{s}_t, .)$. When the agent observes the altered state $\hat{s}_t$, it applies action a_t and observes $\hat{r}(\hat{s}_t, a_t)$, rather than $r(s_t, a_t)$.

Therefore, we assume that in each time step, the state of the system, which is the remaining memory and the service arrival rate of two services, is perturbed. In our simulations, we altered the service arrival rates of two services and converted them to the uniform random variable between zero and the service arrival rate. Therefore, we blocked part of service arrival rates in these simulations based on the weak adversary attack in [30].

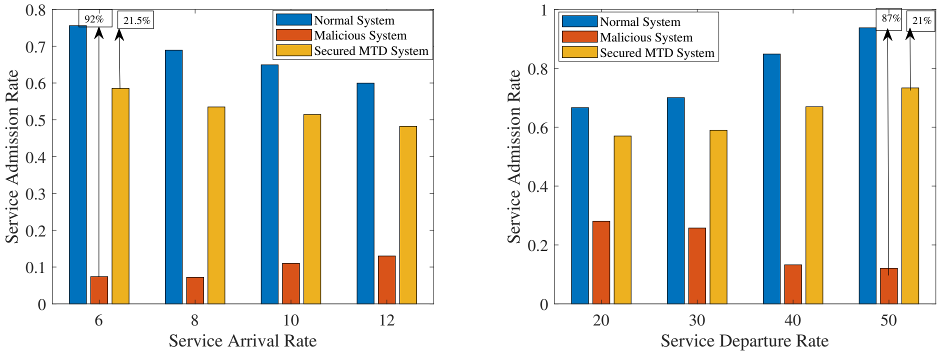
5.1.4. MTD technique

To tackle the adversarial attack issue, we adopt the MTD approach, where the defender has multiple configurations for the ML models. In this scenario, as shown in Fig 3, we use four different PPO models with varying configurations for learning. We assume that the adversarial attacker can randomly affects one of these models during the training. After the models are trained, a random model is selected among the four models to run each input and returns the output generated by that model. Thanks to the dynamicity introduced by the proposed MTD method, attackers will have less impact on the system because they attack one of the models and do not know which model is selected.

In this scenario, we delve into the O-RAN near-RT RIC architecture, specifically employing the AI/ML approach, notably the PPO model, for resource allocation. The RIC layer, constituting the new AI/ML controller within the O-RAN system, plays a pivotal role in service admission control and resource allocation. As elucidated in the O-RAN white papers, RL methods find implementation within the near-RT RIC for the resource allocation. In this context, we explore the integration of MTD for fortifying the system. To accomplish this, we trained four distinct models, each configured as an individual xApp in the near RT RIC.

5.1.5. Performance Results

Here, we consider two different services with varying memory requirements for the admission control problem (1). To evaluate the efficiency of the PPO-based dynamic service admission control solution



(a) **(b)**

Fig. 4. Service admission rate vs. (a) mean service arrival rate and (b) mean service departure rate.

and the effectiveness of the proposed MTD method in withstanding adversarial attacks against DRL, we consider three scenarios, as shown in Fig. 4. In the first scenario, we have a normal system without any attack. The system is trained using the PPO model to admit services based on their resource requirements. The system is implemented in Python, considering two different services with distinct requirements. At each time step, a varying number of requests arrive from these two services, and we solve equation (1) using the PPO model, implemented via the Stable-Baselines3 library in Python. In the second scenario, the system is under attack while using a single PPO model. In this case, the attacker manipulates the system state, specifically the remaining memory, and alters its values. In the third scenario, we employ the proposed MTD technique with four PPO models. These four models are implemented by varying hyperparameters, including the discount factor, batch size, learning rate, and others. We assume that the attacker targets one of these PPO models. At each step in the MTD system, one of the models is selected for the admission control task.

For the three scenarios, the average service admission rate is measured in terms of the mean service arrival rate and the mean service departure rate. Fig. 4a and Fig. 4b report the comparative results. It is observed that the service admission rate of the system decreases with the increase of the service arrival rate, which is attributed to the limited available resources. Furthermore, as the service departure rate increased, the service admission rate increased due to the release of memory. We can also notice a significant enhancement in the system's performance under adversarial attacks after using the MTD technique. Fig. 4a shows that the secured MTD system experienced only 21.5% lower admission rate under adversarial attack, compared to 92% drop-in admission rate when the system is not secured. Similar observations hold true in Fig. 4b, where we can see that the secured MTD system limited the attacker's impact to 21% decrease in the admission rate, compared to 87% without protection from adversarial attacks.

5.2. *LLM-based XAI Robust AI/ML in O-RAN*

In a previous scenario, the AI/ML component responsible for service admission control was managed using the PPO model. We assumed a weak adversarial attack was in play. To diagnose and explain this unusual behavior, an LLM XAI system could take action. For example, the LLM could analyze the model's decision-making process and generate a plain-language report: "The service admission model has rejected 15 devices in the last 15 minutes, a significant difference from its normal pattern of one rejection per 15 minutes."

The LLM system employs XAI techniques to identify the malicious model. Using the Isolation Forest technique, an unsupervised ML algorithm for anomaly detection, the system can detect outlier data based on features such as mean and variance. The LLM then explains these anomalies in a human-readable format. This insight enables the O-RAN system to quickly recognize malicious interference with the PPO model. An immediate investigation is recommended to confirm the nature of the detected anomaly and take steps to remove that model from the system.

By leveraging the capabilities of the LLM-based XAI system, network operators can gain a deeper understanding of the underlying issues affecting AI/ML-driven service admission control. This will ensure that the integrity and security of the O-RAN system are maintained.

5.2.1. *System Scenario*

In this system scenario, we show how the LLM-based XAI system can analyze the output data coming from the models (which can be the service admission rate) and translate it into human-readable language to help the mobile operators to detect any attack to any trained models of the MTD system. This represents an advanced MTD system that integrates the LLM model and XAI to analyze and clarify attacks, subsequently removing the affected model from the MTD system. Suppose one of the four models is targeted in an attack. When the system selects this xApp, the data pattern for service admission differs from that of other xApps (i.e., service admission is notably lower for this specific xApp compared to others). The LLM system can analyze the data pattern, identify the attacked model based on the pattern, describe it in human-readable language, and then request action, which could be performed by either the system operator or the SMO, to remove the specific xApp from the O-RAN system [31].

5.2.2. *Analyzing the system using LLM based on XAI*

We studied Fig. 4-a (where service arrival rate is 12) whenever one of the 4 trained models was attacked. We used GPT-4's data analyst with isolation forest to spot unusual patterns in the outputs of these four models over time. We provided the data to GPT-4 for the detection of malicious activity within the system. The service admission rates for models x1, x2, and x4 were similar, averaging around 60%, whereas model x3 averaged approximately 15%. We analyzed it using LLM based on XAI. The LLM based XAI used the Isolation Forest algorithm to analyze whether there is any anomaly detection in our system.

The results reveal significant differences and potential issues among the series analyzed. Series x1 and x4 display consistent values with moderate variation typical of time-series data. Series x2 shows higher peaks (e.g., 63) and slightly more variability, which seems contextually normal. In contrast, series x3 stands out with consistently lower and less varied values. Identified as an anomaly by the Isolation Forest algorithm, x3 exhibits significantly lower mean and variance compared to x1, x2, and x4. This deviation suggests poisoning attack or any error in the system. Further investigation, including system log reviews, configuration checks, or security audits, is essential to identify and address potential malicious activity or technical faults in x3.

6. Conclusion

In this paper, we first examined the O-RAN architecture, focusing on the integration of network slicing and ML techniques within this system. We then conducted a detailed analysis of key vulnerabilities and threats affecting O-RAN, including risks associated with the RAN, O-Cloud, open-source code, ML, radio networks, and physical security. To address these challenges, we explored four promising approaches: the ZT concept, blockchain technology, LLMs, and MTD paradigms.

We also considered the CIA framework to evaluate both the attacks and the proposed approaches. Additionally, we presented a proof of concept demonstrating the effectiveness of MTD in enhancing the resilience of DRL models against adversarial poisoning attacks.

Furthermore, we examined a service admission control system within the O-RAN architecture and addressed it using a PPO model. Three scenarios were analyzed: a normal system, a system under attack, and an MTD-enabled system with four PPO models operating under attack. Our findings demonstrate that the MTD approach significantly improves the system's reliability

We studied the impact of LLM-based Explainable AI (XAI) in detecting attacks within the O-RAN AI/ML system to enhance the MTD technique. Using ChatGPT-4o, we analyzed data to identify malicious activity. The LLM successfully detected an attack and issued a warning to isolate the affected system from the MTD framework .

6.1. Limitations and Future Research Directions

While the four proposed approaches offer significant benefits, securing O-RAN still faces several challenges: (i) maintaining continuous risk monitoring without impacting network performance for ZT, (ii) addressing scalability, performance, and privacy issues in blockchain, (iii) developing MTD strategies that balance security, performance, and cost, and (iv) leveraging LLMs to automate tasks, enhance explainable AI (XAI), and reduce risks in AI/ML systems. Moreover, a key limitation of LLM-based XAI system is its dependence on accurate anomaly detection, which can be affected by false positives. Additionally, evolving threat patterns may reduce the model's ability to adapt in real-time.

In addition, To enhance system security through MTD, it is crucial to deploy and train multiple models, despite inherent limitations. Future MTD strategies should focus on developing an optimal selection mechanism based on model probability. For example, if an XAI-based model is identified as an anomaly-prone model, its selection probability within the MTD system can be progressively reduced with each iteration until it is eventually excluded from the model pool.

7. Acknowledgements

Tarik Taleb and Jaeho Kim are co-corresponding authors. Prof. Jaeho Kim was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Information Technology Research Center (ITRC) grant funded by the Korea government(IITP-2025-RS-2021-II211816) and by the Technology Innovation Program (RS-2022-00154678) funded By the Ministry of Trade, Industry & Energy (MOTIE, South Korea). In addition, this work was partly conducted at ICTFICIAL Oy, Finland. It was partly funded by the European Union's HORIZON-JUSNS-2023 HE research and innovation program (6G Path project, Grant No. 101139172) and the European Union's Horizon Europe research and innovation programme HORIZON-JU-SNS-2022 under the RIGOUROUS project (GrantNo. 101095933). The views expressed are solely those of the authors, and the European Commission is not responsible for any use of this information.

References

[1] M. Polese, L. Bonati, S. D'oro, S. Basagni, T. Melodia, Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges, IEEE Communications Surveys & Tutorials 25 (2) (2023) 1376–1411.

[2] M. K. Motalleb, V. Shah-Mansouri, S. Parsaeefard, O. L. A. López, Resource allocation in an open ran system using network slicing, IEEE Transactions on Network and Service Management 20 (1) (2022) 471–485.

[3] S. Nouri, M. K. Motalleb, V. Shah-Mansouri, S. P. Shariatpanahi, Semi-supervised learning approach for efficient resource allocation with network slicing in O-RAN, arXiv preprint arXiv:2401.08861.

[4] T. Taleb, C. Benzaïd, R. A. Addad, K. Samdanis, AI/ML for beyond 5G systems: Concepts, technology enablers & solutions, Computer Networks 237 (2023) 110044.

[5] D. Mimran, R. Bitton, Y. Kfir, E. Klevansky, O. Brodt, H. Lehmann, Y. Elovici, A. Shabtai, Evaluating the Security of Open Radio Access Networks, arXiv preprint arXiv:2201.06080.

[6] H. Park, T.-H. Nguyen, L. Park, An investigation on open-ran specifications: Use cases, security threats, requirements, discussions., CMES-Computer Modeling in Engineering & Sciences (2024) 141 (1).

[7] H. A. Kholidy, A. Karam, J. Sidoran, M. A. Rahman, M. Mahmoud, M. Badr, M. Mahmud, A. F. Sayed, Toward zero trust security in 5G open architecture network slices, in: MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM), IEEE, 2022, pp. 577–582.

[8] M. K. Motalleb, C. Benzaïd, T. Taleb, V. Shah-Mansouri, Moving target defense based secured network slicing system in the O-RAN architecture, in: GLOBECOM 2023-2023 IEEE Global Communications Conference, IEEE, 2023, pp. 6358–6363.

[9] M. K. Motalleb, V. Shah-Mansouri, S. N. Naghadeh, Joint power allocation and network slicing in an open ran system, arXiv preprint arXiv:1911.01904 (2019).

[10] O-RAN Security Focus Group (SFG) Study on Security for O-Cloud v01.00, `https://www.o-ran.org/o-ran-resources` (2022).

[11] O-RAN Alliance Working Group 1, O-RAN-Architecture-Description-v06.00, `https://www.o-ran.org/o-ran-resources` (2022).

[12] A. Javadpour, F. Ja'fari, T. Taleb, C. Benzaïd, Reinforcement learning-based slice isolation against DDOS attacks in beyond 5G networks, IEEE Transactions on Network and Service Management 20 (3) (2023) 3930–3946.

[13] ORAN ALLIANCE Working Group 2 Study AI/ML Workflow Description and Requirements v01.03, `https://www.o-ran.org/o-ran-resources` (2021).

[14] ORAN ALLIANCE- Security Focus Group (SFG), O-RAN Security Threat Modeling and Remediation Analysis v03.00, `https://www.o-ran.org/o-ran-resources` (2022).

[15] D. Dik, M. S. Berger, Open-ran fronthaul transport security architecture and implementation, IEEE Access 11 (2023) 46185–46203.

[16] J. Groen, S. D'Oro, U. Demir, L. Bonati, M. Polese, T. Melodia, K. Chowdhury, Implementing and evaluating security in O-RAN: Interfaces, intelligence, and platforms, IEEE Network 2024.

[17] C. Benzaïd, T. Taleb, AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler?, IEEE Network 34 (6) (2020) 140 – 147.

[18] C. Benzaid, T. Taleb, M. Z. Farooqi, Trust in 5G and Beyond Networks, IEEE Network Magazine 35 (3) (2021) 212 – 222.

[19] C. Benzaid, T. Taleb, J. Song, AI-based autonomic and scalable security management architecture for secure network slicing in B5G, IEEE Network 36 (6) (2022) 165–174.

[20] H. Jiang, H. Chang, S. Mukherjee, J. Van der Merwe, Oztrust: An o-ran zero-trust security system, in: 2023 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), IEEE, 2023, pp. 129–134.

[21] S. K. Poorazad, C. Benzaïd, T. Taleb, Blockchain and deep learning-based IDS for securing SDN-Enabled industrial iot environments, in: GLOBECOM 2023-2023 IEEE Global Communications Conference, IEEE, 2023, pp. 2760–2765.

[22] X. Wang, B. Wang, Y. Wu, Z. Ning, S. Guo, F. R. Yu, A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability, IEEE Communications Surveys & Tutorials 2024.

[23] L. Giupponi, F. Wilhelmi, Blockchain-enabled network sharing for O-RAN, arXiv preprint arXiv:2107.02005.

[24] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, K. Huang, Pushing large language models to the 6g edge: Vision, challenges, and opportunities, arXiv preprint arXiv:2309.16739.

[25] T. Senevirathna, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, S. Wang, A Survey on XAI for 5G and Beyond Security: Technical Aspects, Challenges and Research Directions, IEEE Communications Surveys & Tutorials (2024).

[26] E. Cambria, L. Malandri, F. Mercorio, N. Nobani, A. Seveso, XAI meets llms: A survey of the relation between explainable AI and large language models, arXiv preprint arXiv:2407.15248 (2024).

[27] X. Wu, H. Zhao, Y. Zhu, Y. Shi, F. Yang, T. Liu, X. Zhai, W. Yao, J. Li, M. Du, et al., Usable XAI: 10 strategies towards exploiting explainability in the LLM era, arXiv preprint arXiv:2403.08946 (2024).

[28] T. Datta, J. P. Dickerson, Who's thinking? a push for human-centered evaluation of LLMs using the XAI playbook, arXiv preprint arXiv:2303.06223 (2023).

[29] A. Javadpour, F. Ja'fari, T. Taleb, C. Benzaïd, Enhancing 5G network slicing: Slice isolation via actor-critic reinforcement learning with optimal graph features, in: GLOBECOM 2023-2023 IEEE Global Communications Conference, IEEE, 2023, pp. 31–37.

[30] T. Wu, Y. Yang, S. Du, L. Wang, On Reinforcement Learning with Adversarial Corruption and its Application to Block MDP, in: International Conference on Machine Learning, PMLR, 2021, pp. 11296–11306.

[31] A. J. Dave, T. N. Nguyen, R. B. Vilim, Integrating LLMs for explainable fault diagnosis in complex systems, arXiv preprint arXiv:2402.06695.