# 6G-XSec: Explainable Edge Security for Emerging OpenRAN Architectures

Haohuang Wen
The Ohio State University

Prakhar Sharma
SRI

Vinod Yegneswaran
SRI

Phillip Porras
SRI

Ashish Gehani
SRI

Zhiqiang Lin
The Ohio State University

## Abstract

The evolution from 5G to 6G cellular networks signifies a crucial advancement towards enhanced robustness and automation driven by the promise of ubiquitous Artificial Intelligence (AI) to overhaul network operations, commonly referred to as *AIOps*. However, 6G network operators also need to deal with evolving threats at the edge to ensure data integrity and availability. We introduce 6G-XSec, the first framework that seeks to automatically *monitor*, *analyze*, and *explain* anomalies and threats at the cellular network edge. Our framework enhances the emerging Open Radio Access Network (O-RAN) control plane with run-time analytic capabilities and explainability. A distinguishing aspect of our framework is the use of *expert referencing*, a coupling of lightweight unsupervised deep learning-based anomaly detection with large language models (LLMs) to first detect, analyze, and subsequently explain complicated real-world cellular threats and anomalies at run-time, based on enhanced security telemetry from the O-RAN data plane. We build a prototype 6G-XSec framework and evaluate it against 5 end-to-end cellular attacks from the literature, achieving 100% detection rate with our best model. We also propose effective LLM prompt templates for attack analysis and present qualitative results from 5 popular LLMs.

## CCS Concepts

• **Networks → Mobile networks**; **Programmable networks**;
• **Security and privacy → Mobile and wireless security**; **Intrusion/anomaly detection and malware mitigation**.

## Keywords

6G, OpenRAN, Anomaly Detection, Large Language Model

**ACM Reference Format:**
Haohuang Wen, Prakhar Sharma, Vinod Yegneswaran, Phillip Porras, Ashish Gehani, and Zhiqiang Lin. 2024. 6G-XSec: Explainable Edge Security for Emerging OpenRAN Architectures. In *The 23rd ACM Workshop on Hot Topics in Networks (HOTNETS '24), November 18–19, 2024, Irvine, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3696348.3696881

## 1 Introduction

Cellular networks are the backbone of modern wireless communication, impacting numerous sectors from transportation and entertainment to manufacturing and healthcare. In recent years, the cellular network standards and associated technologies have grown rapidly and gained widespread deployment. It is envisioned that future 6G networks will revolutionize from connected things to *connected intelligence* with ubiquitous Artificial Intelligence (AI) [44]. This transformative leap will enable advanced capabilities such as self-diagnosis, optimization, fault recovery, and threat mitigation, thereby enhancing network performance and trustworthiness.

**Opportunities Brought by OpenRAN.** The drive toward an intelligent 6G network is fueled by several factors, including network performance upgrades and the widespread integration of AI and machine learning (ML) technologies. A key revolution among these is the software-defined network architecture known as OpenRAN (O-RAN), which introduces unprecedented *programmability* to traditional cellular infrastructures. O-RAN transforms the previously monolithic cellular network into a disaggregated and interoperable mobile network architecture, as shown in Figure 1. This architecture incorporates principles from software-defined networks, allowing for modular design and the deployment of "plug-n-play" cellular control-plane applications (xApps) that perform dedicated tasks such as network monitoring and management. While much of the prior R&D has focused on network optimization and automation [20, 30, 41, 43, 51], we believe that the programmability of O-RAN also creates new opportunities in the security domain to enhance the resilience and trustworthiness of both public and private 6G networks.

*HOTNETS '24, November 18–19, 2024, Irvine, CA, USA*

ACM ISBN 979-8-4007-1272-2/24/11
https://doi.org/10.1145/3696348.3696881

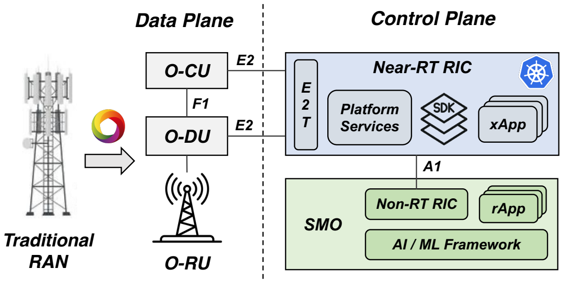


**Figure 1: O-RAN architecture for 5G / 6G cellular networks.**

**Case for Explainable 6G Edge Security.** In recent years, the mobile security community has uncovered numerous attack surfaces and exploits [15]. These vulnerabilities are inherited from prior generations (e.g., LTE), and can be readily exploited with commodity hardware [10] and open-source cellular stacks [9] to compromise the security and privacy of network infrastructures and users at the edge. Examples of attacks include denial or disruption of cellular services at base stations [19, 37, 38, 42], leakage of user's location and identity [25, 27, 32, 40], and network security downgrade [18]. To ensure security at the 6G edge, ideal solutions require both *analytic* capabilities and *explainability*. Effective cellular edge analytics refer to detection capabilities for not only current but also unseen and evolving threat patterns and variants. Moreover, explainability, i.e., the ability to explain *why* there is an attack, is also crucial for network operators to understand root causes and take corresponding actions, which also increases the network's trustworthiness.

**Technical Challenges.** While O-RAN provides a programmable way to realize novel control-plane applications, there are still high-level technical challenges. First, *Visibility*, or the ability to capture relevant threat modalities from network data, is the foremost challenge for every security application. However, the default O-RAN standard only provides limited support for security visibility, as evident by the existing service models [1, 6–8] that have lacked the focus and fidelity necessary to drive the development of practical security applications. Second, the *Analytic* capability is required to analyze sophisticated cellular network traffic and identify potential security threats. In the security domain, this is less studied compared to other networks and faces many challenges such as scalability [31, 55, 56, 59, 61]. Third, *Explainability* [58] is challenging due to the requirement of highly specialized domain insights and knowledge from massive 3GPP cellular specifications [13, 14], which is impossible to process manually.

**Envisioned Framework.** We describe our vision of an AI-driven and explainable edge security framework, namely 6G-XSEC, for the O-RAN architecture. It involves novel data-plane and control-plane extensions (e.g., xApps), enabling the network with advanced capabilities to monitor, analyze, and explain runtime anomalies and threats. Our preliminary results suggest that this framework not only achieves 100% detection rate for 5 types of unseen cellular attacks but also explains why these threats deviate from benign traffic. Looking forward, we envision our framework will enable lower-skilled and private cellular operators to rapidly detect, diagnose, and recover from runtime faults and attacks, driving toward the goal of explainable AIOps in future-generation networks.

**Contributions.** Our paper makes the following contributions:

- Design of a novel explainable edge security framework based on the cellular O-RAN architecture.
- Development of xApps employing unsupervised anomaly detection and LLM-based expert referencing to detect and analyze emerging threats and anomalies at run-time.
- Demonstration of a prototype system on an O-RAN compliant cellular network testbed.
- Preliminary evaluation of unsupervised threat detection with practical cellular traces and 5 attacks, showing that it can effectively detect unseen cellular threat patterns.
- Evaluation of the expert referencing approach using 5 popular LLMs against 5 attacks.

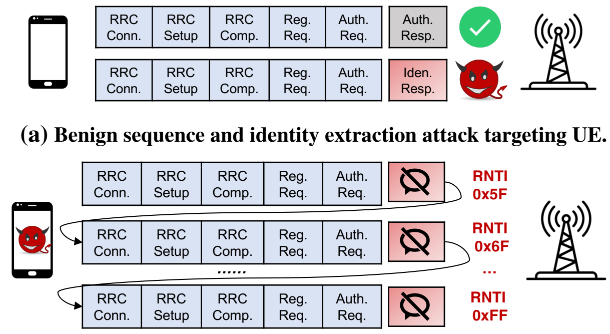
## 2 Relevant Background

### 2.1 OpenRAN Architecture

A typical cellular network is broken down into several entities. User Equipment (*UE*) represents end-user devices. The Radio Access Network (*RAN*), also commonly known as a base station, or a gNodeB in a 5G context, serves as an access point to relay user traffic. *Core Network* is the mobile network backend connecting users to the external Internet.

**O-RAN Data Plane.** As shown in Figure 1, the OpenRAN (O-RAN) architecture [4] follows 3GPP's functional split [11] to divide the monolithic RAN design into logical components: the Radio Unit (O-RU), Distributed Unit (O-DU), and Central Unit (O-CU). The O-RU is the typical radio hardware deployed in the front-haul network, responsible for handling layer-1 (L1) physical radio signals from nearby user equipment. The O-DU is a logical component hosted at the network edge, managing layer-2 (L2) functions such as Media Access Control (MAC). The O-CU handles layer-3 (L3) control protocols like the Radio Resource Control (RRC) [14]. The O-CU connects to core network (CN) functionalities, such as the Access and Mobility Management Function (AMF) and the User Plane Function (UPF) of the 5G core. The O-RAN data plane components communicate through standard and open interfaces, such as F1 connecting O-DUs and O-CUs [12].

**O-RAN Control Plane.** The control logic of O-RAN is separated from the data plane based on the Software-Defined Networking (SDN) principles. The O-RAN control functions are



**(a) Benign sequence and identity extraction attack targeting UE.**

**(b) RAN Denial-of-Service (DoS) attack from UE.**

**Figure 2: Illustration of two distinct attacks exploiting unprotected cellular messages between a UE and RAN.**

implemented in Near Real-time RAN Intelligent Controllers (*nRT-RICs*), which act as control service proxies connected to the RAN via the standard E2 interface [5]. The nRT-RIC hosts various modular xApps that implement customized network management functions such as monitoring, analytics, and control [4]. The interactions between the control and data planes are defined by four basic E2 primitives: *report, insert, control,* and *policy*. To interact, each xApp needs to define E2 Service Models (E2SMs) as function-specific protocols by using these E2 primitives, based on the generic E2 Application Protocol (E2AP) [7]. For instance, the O-RAN Alliance has demonstrated several exemplar E2SMs for key performance monitoring and network slicing management [1, 8]. O-RAN also defines different latency requirements. The control loop of a nRT-RIC completes within 10ms to 1s, while time-insensitive tasks, e.g., ML model training, are handled within the Service Management and Orchestration (SMO) framework by rApps on non-real-time RICs.

## 2.2 Cellular Threat Landscape

**Threat Model.** We focus on a specific threat model within the cellular network domain. In this model, adversaries at the network edge attempt to compromise devices and infrastructure through the open wireless interface. These adversaries exploit vulnerabilities in cellular protocol standards by transmitting, flooding, and hijacking unprotected protocol messages over the air, while still adhering to cryptographic assumptions [15, 32, 37, 38, 48, 62]. This threat model is highly practical because adversaries can use readily available software-defined radios (SDRs) that work with open-source cellular software stacks, allowing them to easily program and execute malicious attack logic as a rogue UE or an adversarial relay in the victim's network [9, 33]. On the other hand, we assume the internal components and communication in the networks, including the base stations, the core network, and the O-RAN control plane, are trusted.

**Anomalies Exhibited by Attacks.** To detect and counter the aforementioned attacks, the natural solution is to inspect anomalies from the cellular traffic, as the attack traffic likely involves highly unusual traces and payloads. In the later framework descriptions, we utilize this key principle to design our attack detection mechanisms. In Figure 2, we illustrate two concrete attacks and discuss below how they manifest deviated anomalies from normal cellular traffic.

- **Univariate Anomalies**. Figure 2a illustrates a benign cellular trace along with another representing an identity extraction attack conducted by Man-in-The-Middle (MiTM) attackers [32, 40]. In the attack, the adversary overwrites the downlink authentication request message to maliciously ask the victim UE to transmit its identifier in plain text. Since the identifier is bound to a specific user's SIM, this step allows tracking of the victim's location. This sequence exhibits an *out-of-order message sequence* from normal traffic where the UE typically responds to the authentication request with the corresponding response payload.
- **Multivariate Anomalies**. Figure 2b describes a Denial-of-Service (DoS) attack targeting the RAN from a malicious UE. The UE establishes multiple fabricated RRC connections at the authentication stage, which consumes the RAN's resources and prevents other legitimate UEs from connecting [37, 38]. This attack differs from benign traffic in that the network observes malicious patterns with both abnormal message sequences and device parameters. In this instance, the RAN is flooded with a rapid succession of uncompleted UE connection requests from a stream of unique temporary identifiers (RNTI), and thus these variables jointly constitute the attack pattern deviated from normal traffic.

## 3 6G-XSec Framework Overview

A design overview of the 6G-XSec framework architecture is shown in Figure 3. First, the RAN data plane is extended to collect security telemetry and report through the O-RAN standard E2 interface to the nRT-RIC (§3.1). Next, the telemetry is analyzed by an unsupervised anomaly detector xApp (MobiWatch) to identify deviated traffic patterns (§3.2). Finally, the anomalous traces are processed by an LLM-based expert (§3.3) to generate insights for threat explanation. The following describes each of these components.

## 3.1 Security Telemetry Collection

**Telemetry Definition.** Data telemetry is the foundation for any data-driven AI techniques. However, existing service models in the O-RAN standards only provide coarse-grained

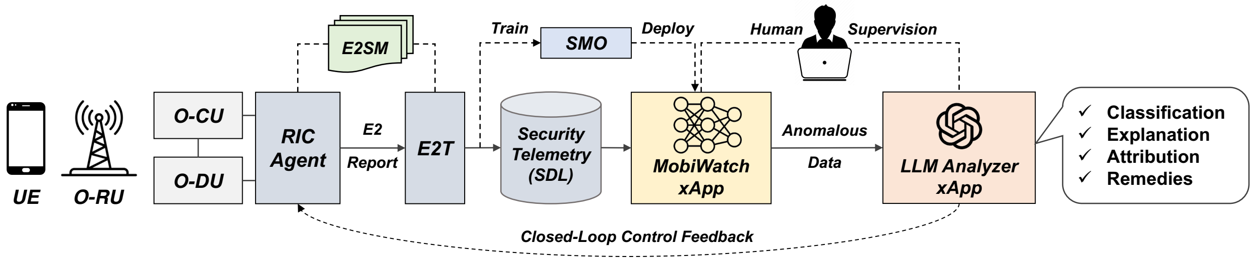


**Figure 3: Overview of the 6G-XSEC framework. Security telemetry is extracted via E2 interfaces from the cellular data plane. The collected telemetry is used for either model training or inference. The MOBIWATCH xApp and LLM xApp provide capabilities for anomalous traffic detection as well as anomaly explanation and remediation.**

security telemetry that mainly focuses on network performance monitoring [1], which is thus inadequate for monitoring and detecting sophisticated cellular threats at the edge [32, 37, 38, 48, 62]. To address this limitation, we adopt the design of MOBIFLOW security telemetry from our prior work [60], which extends the cellular data plane to collect fine-grained states and statistics at the cellular protocol level. The collected telemetry is formulated as multivariate time-series data entries [17], where each telemetry entry $x_i$ is collected at each control message transmission:

$$x_i^T = [t_i, m_i, p_{1i}, p_{2i}, p_{3i} ..., p_{ki}]$$

where $m_i$ represents the message, and $p_{ki}$ denotes the UE-specific parameter from a defined parameter set $p_{ki} \in K$ collected at timestamp $t_i$. Therefore, a time series collected from the RAN spanning $M$ messages is denoted by:

$$\mathcal{T} = \{x_1, x_2, x_3, ..., x_M\}$$

**Telemetry Collection.** We provide an example list of MOBIFLOW telemetry in Table 1. The message $m_i$ could be collected from RRC and NAS protocol packets [13, 14]. The parameter set $K$ could include UE-specific identifiers and state parameters. The listed telemetry is readily extracted from standard RAN interfaces, such as F1 Application Protocol (F1AP) [12] and NG Application Protocol (NGAP) [11]. By instrumenting these interfaces or parsing the *pcap* streams, the RIC agent at the RAN data plane extracts, encodes, and reports the telemetry to the nRT-RIC. More precisely, the telemetry report is based on an established subscription between an xApp and the RAN, and the telemetry is eventually decoded and handled by the custom xApp logic. To comply with the E2SM standard, we extend the E2SM-KPM [1] service model from the O-RAN reference implementation. This service model enables the RIC agent to report security telemetry via the E2 report operation [7] per time interval, where the telemetry can be encoded as *(key, value)* data. Upon receiving the security telemetry, the xApp stores it in the *Shared Data Layer* (SDL) which is a centralized database that can be accessed by other nRT-RIC services and xApps.

| Category | Telemetry | Description |
|---|---|---|
| Message | RRC Message | Uplink / Downlink Radio Resource Control (RRC) protocol message [14] |
| | NAS Message | Uplink / Downlink Non-Access-Stratum (NAS) protocol message [13] |
| Identifier | RNTI | Radio Network Temporary Identifier |
| | S-TMSI | Temporary Mobile Subscriber Identity |
| | SUPI | Subscription Permanent Identifier |
| State | Cipher_alg | Ciphering algorithm employed by the UE |
| | Integrity_alg | Integrity algorithm employed by the UE |
| | Establish_cause | RRC establishment cause from the UE |

**Table 1: An example list of MOBIFLOW security telemetry collected from the cellular data plane.**

## 3.2 Unsupervised Anomaly Detection

**Motivation.** Typical machine learning techniques for detecting security threats are broadly classified into either supervised or unsupervised approaches. Particularly in the cellular domain, one major challenge for training a supervised classifier lies in the scarcity of well-labeled datasets in the wild. While there are a few public datasets available [26, 31, 47, 54], they are collected from LTE networks or do not involve control-layer traffic. Another major challenge is obtaining adversarial data samples (e.g., attack traces). Due to these reasons, we adopt an unsupervised approach, which has been widely used in security intrusion or anomaly detection [29, 49]. From the attack descriptions in §2.2, cellular attacks typically exhibit certain levels of *group anomalies* [50] (e.g., out-of-order message sequences in Figure 2a), and thus we hypothesize that these outliers are distinguishable from benign traffic. Therefore, our objective is to train an unsupervised neural network only on readily available benign cellular traffic data, through which the model learns the latent representations and is capable of estimating how to distinguish unknown data deviations from the benign data distributions.

**Formulation.** Our problem setup is similar to prior multivariate outlier detection tasks [17, 36, 57]. The unsupervised anomaly detection xApp consumes the multivariate time series $\mathcal{T}$ as the input. We further use a sliding window of size $N$ to convert the time-series data into sequences $S = \{S_1, ..., S_{M-N+1}\}$, where each sequence $S_i = \{x_i, ..., x_{i+N-1}\}$. All categorical variables within each sequence $S$ is one-hot encoded. The objective is to learn a mapping function $\phi : S \mapsto Z$ where $Z \in \mathbb{R}^{M-N+1}$. Therefore, when given an unseen sequence $\hat{S}_n = \{\hat{x}_n, ..., \hat{x}_{n+N-1}\}$, the model should output an anomaly score measuring the extent to which the given sample deviates from the trained data. Based on a predefined threshold, a final decision $y \in \{0, 1\}$ is produced where 1 indicates the given sequence is anomalous.

**Approaches.** We describe the unsupervised deep learning approaches to detect anomalous behaviors indicative of faults or security threats from the telemetry. In this context, we explore various unsupervised techniques for outlier detection [29, 36, 46, 49, 63]. We train these models with normal data (devoid of any attacks) and test them on unseen data that has a mixture of normal and abnormal flows.

- **Autoencoders:** Autoencoders [53] are neural network architectures that compress an input vector to a lower dimensional representation and use that to create the original input vector, formulated as $\hat{S} = f_{AE}(S)$. In doing so, they fit their internal parameters to minimize the reconstruction error between the input and the reconstructed output, thereby learning to represent the input data distribution. Any outlier sequences $\hat{S}$ when passed through the model will exhibit a large reconstruction error (e.g., mean square error), thus indicating an outlier.
- **Sequence Modeling:** Owing to the fact that our security telemetry captures the cellular protocol message flows, they can be naturally represented as sequence data. We fit a Long Short Term Memory model (LSTM) [35] model to this representation. The model is trained on normal flow sequences, to predict the next telemetry, denoted by $\hat{x}_{i+N} = f_{LSTM}(x_i, ...x_{i+N-1})$, depending on the window size $N$. If the actual telemetry $x_{i+N}$ deviates from the model's predicted output $\hat{x}_{i+N}$, it is thus considered an outlier.

**Deployment.** These unsupervised anomaly detection models are integrated into an xApp, namely MobiWatch in the nRT-RIC, as shown in Figure 3. Through the internal communication interfaces (e.g., the message routing mechanisms in the O-RAN reference implementation [3]), the xApp can access the SDL for the dynamic security telemetry from the networks. Regarding the model training and deployment, it can be conducted either in an offline fashion or in the Service Management and Orchestration (SMO) infrastructure on the O-RAN control plane [52]. Incoming telemetry for real-time inference is done with the deployed models in the xApp, which outputs the decision of whether a given input sequence is anomalous and requires further analysis.

## 3.3 LLM-based Expert Referencing

LLMs have exhibited exceptional text summarization, comprehension, and deduction abilities [22] including network management and incidence analysis [34, 39]. Motivated by these examples, we investigate how to leverage LLM's deduction capabilities for explainable security [58], specifically to generate domain-specific insights including *what* are the anomalies (classification), *why* the cellular sequence is anomalous (explainability), *who* is responsible (attribution), and *how* to mitigate this threat (remediation).

Figure 3 depicts the integration of a control-plane xApp for LLM-based expert referencing. Once MobiWatch flags a cellular sequence as anomalous, the sequence (plus its context window) is passed to the LLM xApp. The motivation for chaining a lightweight anomaly detector and the LLM analyzer is that LLMs are prohibitively large and expensive to deploy and invoke in a cellular network as it may introduce a high volume of cellular traffic in practice. The MobiWatch xApp serves as a pre-filter that extracts only the meaningful and novel network incidents that require further analysis. Additionally, the results from MobiWatch and LLM could be cross-compared to ensure the decisions are indeed reliable. Upon receiving the sequences, the xApp constructs inputs based on prompt templates and accesses the LLMs through RESTful web APIs from either a pre-trained LLM (e.g., GPT-4 [16]) or a locally fine-tuned model. However, LLM's decision might not be accurate, also known as the hallucination problem [64]. As such, human supervision is required in cases such as when the LLM and the anomaly detector generate contradictory results.

# 4 Preliminary Evaluation

One of the major challenges in cellular security research is the scarcity of open traffic datasets, especially adversarial samples with anomalous or attack traces. As a proof-of-concept, we have built a 5G network testbed for experimentation and data collection as 6G standards are currently under development [2]. In support of open science, we have released the source code and datasets[1] to reproduce our system and results.

**Testbed Setup.** We have built an end-to-end 5G standalone (SA) network with open-sourced software and commodity hardware platforms. Specifically, we use OpenAirInterface (OAI) [9] as our gNodeB (CU and DU) and core network implementations, and a commodity software-defined radio (SDR) Ettus USRP B210 as the RU front-end [10]. The OAI CU is extended with an E2 RIC agent that extracts security telemetry and handles communication with the nRT-RIC's E2 interface. We adopt the O-RAN Software community (OSC)'s nRT-RIC reference implementation [3], which allows us to

[1] Our source code and datasets are available at https://github.com/5GSEC/MobiWatch

| Dataset | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Benign | Autoencoder | 93.23% | 93.23% | N/A | N/A |
| | LSTM | 91.15% | 91.15% | N/A | N/A |
| Attack | Autoencoder | 100% | 100% | 100% | 100% |
| | LSTM | 95.00% | 88.68% | 100% | 94.00% |

**Table 2: Summary of detection performance of the two deep learning models on benign and attack datasets.**

realize our architecture in Figure 3. The whole testbed setup is deployed in a standalone Ubuntu 20.04 desktop with 12 Intel i7 CPU cores and 32GB RAM.

**Dataset Collection.** We collect both *benign* and *attack* datasets from our testbed. To ensure the diversity of the benign cellular traffic, we first collected traffic from four different commodity 5G smartphone models including Google Pixel 5 and 6, Samsung Galaxy A22, and A53. In addition, we leverage COLOSSEUM, a software-defined large-scale 5G network emulator capable of generating traffic in various scenarios [21], and we deployed our test network with open-sourced UEs from OAI. To collect data, we instrument the F1AP and NGAP interface to obtain *pcap* streams, which are further parsed into MOBIFLOW security telemetry formats. In summary, we collected 2.5MB of *pcap* files for benign traffic, constituting over 100 UE sessions with the network. For the attack dataset, we implemented 5 cellular attacks by inserting malicious logic into the OAI's UE stack [32, 37, 38, 40]. To ensure no real-world devices are affected, we conducted the attack dataset collection on COLOSSEUM as well.

**Dataset Labeling.** We developed the following two principles to label our datasets. First, all traces from benign datasets are naturally considered benign. Second, the attack datasets contain a mixture of benign and malicious sequences as each attack occurs at a certain point within a network session. We manually identify and label each malicious telemetry entry $x_i \in \mathcal{T}$, and consider any sequences that involve $x_i$ to be malicious, i.e., $\{S_{i-N+1}, ..., S_i\}$ for window size $N$.

## 4.1 Can Unsupervised Deep Learning Models Detect Cellular Attacks?

We evaluate whether a lightweight unsupervised deep learning model described in §3.2 can effectively detect deviated patterns that represent anomalies and threats in cellular networks. We use the benign dataset to train our models, by leveraging categorical features in the security telemetry described in Table 1 including the control messages and device identifiers such as UE's RNTI and TMSI. After training, we select a 99% percentile threshold among the reconstruction errors for anomaly detection, assuming 1% outliers within the training set caused by network noise. Table 2 summarizes the detection performance from both the Autoencoder and the LSTM models. Our results show that both models deployed at MOBIWATCH correctly classified all attack sequences as anomalous, indicating 100% recall or no false negatives. However, we observe a small portion of false positives (<10%) in both the benign (through cross-validation) and the attack datasets. The major causes are unusual message sequences and network interference (e.g., RRC message retransmissions). Since the dataset we have collected in this paper is relatively small, we intend to expand our dataset and evaluation methodology as important future work.

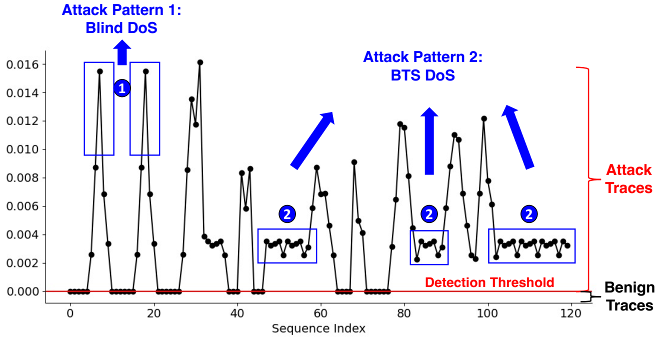


**Figure 4: Visualization of our Autoencoder's reconstruction errors of attack dataset sequences. ① and ② respectively highlight similar anomaly patterns for attack events within the same category (①: Blind DoS attack and ②: BTS DoS attack).**

To further facilitate analysis of our results, we visualize the reconstruction errors of the attack sequences from the Autoencoder model in Figure 4. In the figure, all data points above the threshold bar are considered outliers, and those under the bar are considered benign. Interestingly, we observe similar patterns among attack events of the same attack type. For instance, ① and ② represent the Blind DoS and BTS DoS attacks [38], respectively, in which different attack instances of the same type exhibit highly similar group anomaly patterns with respect to the reconstruction errors. This observation also holds for the remaining attack types. As a result, this feature is potentially useful for training a supervised attack classifier to recognize and cluster events of different attack types based on their reconstruction error patterns.

> **Takeaway:** Our preliminary results show that lightweight unsupervised anomaly detection models can detect unseen threat patterns by training only on benign cellular traffic.

## 4.2 Can LLMs Explain Cellular Anomalies?

Following §3.3, we studied whether LLMs can explain why a cellular sequence is benign or anomalous - akin to a security analyst. We start with existing web-based LLMs (e.g., ChatGPT-4o [16]) and zero-shot prompting strategies (i.e., no example provided). Table 3 presents the overall performance,

| Attack / Trace | Chat GPT-4o | Gemini | Copilot | Llama3 | Claude 3 Sonnet |
|---|---|---|---|---|---|
| BTS DoS [38] | ✓ | ✓ | ✓ | ✗ | ✗ |
| Blind DoS [38] | ✓ | ✗ | ✗ | ✓ | ✗ |
| Uplink ID Extr [32] | ✗ | ✗ | ✗ | ✗ | ✓ |
| Downlink ID Extr [40] | ✓ | ✓ | ✗ | ✓ | ✓ |
| Null Cipher & Int. [37] | ✓ | ✓ | ✗ | ✓ | ✓ |
| Benign Sequence 1 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Benign Sequence 2 | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 3: Evaluation results from different LLMs. ✓ indicates the LLM correctly classifies the attack or benign sequence, and ✗ indicates the LLM produced wrong decisions for that trace.**

in which we tested five baseline LLMs and manually verified if they can provide correct explanations for each attack. To avoid bias, we also tested with two other benign sequences. Surprisingly, without any examples, all existing LLMs have demonstrated capabilities to analyze cellular protocol data and distinguish between anomalous and benign traces with accurate explanations. Among these baselines, ChatGPT-4o performs relatively well as it only missed one attack in total.

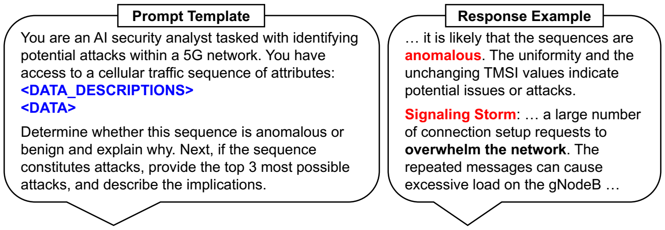
To better illustrate the response quality, Figure 5 presents the prompt template and an example response from ChatGPT-4o. This response shows that the LLM can accurately identify a potential signaling storm (equivalent to BTS DoS [38] in Figure 2b) based on the repeated RRC message patterns. For more complicated attacks such as Blind DoS [38], some baseline LLMs could still extract implicit relations (e.g., replayed TMSI numbers in different UE sessions) and thus correctly suggest the attack. Most LLMs fall short in detecting uplink identity extraction attack [32] as it exhibits standard compliant protocol traces and is thus challenging to distinguish from normal traces. To further evaluate the stability of LLM's reasoning, we conducted repeated experiments on ChatGPT-4o, and observed consistent results as in Table 3.

**Takeaway:** Existing baseline LLM models are promising to classify and explain complex cellular anomalies and attacks with only zero-shot prompting.

## 5 Conclusion, Discussion, and Future Work

We presented an end-to-end framework based on O-RAN to illustrate how AI could drive cellular networks toward automatic and explainable threat detection and mitigation. The following discusses the current limitations, as well as the challenges and opportunities for 6G and OpenRAN networks.

**Limitations.** As a defense framework deployed on the network side, MobiWatch faces challenges in handling certain types of cellular threats. These include downlink attacks that drop protocol messages and rogue base stations that directly communicate with user devices. Another limitation is the difficulty in distinguishing between anomalous events and actual attacks. For example, the transmission of a plain-text user identity could indicate either an identity extraction attack [32, 40] or just an unusual network event. This ambiguity can lead to false positives or false negatives, highlighting the need for further studies to better monitor and understand these threats in real-world networks.



**Prompt Template**

You are an AI security analyst tasked with identifying potential attacks within a 5G network. You have access to a cellular traffic sequence of attributes:
<DATA_DESCRIPTIONS>
<DATA>

Determine whether this sequence is anomalous or benign and explain why. Next, if the sequence constitutes attacks, provide the top 3 most possible attacks, and describe the implications.

**Response Example**

... it is likely that the sequences are anomalous. The uniformity and the unchanging TMSI values indicate potential issues or attacks.

Signaling Storm: ... a large number of connection setup requests to **overwhelm the network**. The repeated messages can cause excessive load on the gNodeB ...

**Figure 5: Prompt template and response from ChatGPT-4o showing an accurate analysis of a BTS DoS attack event (Figure 2b) based on cellular security telemetry.**

**Automated Network Responses.** Our preliminary work has demonstrated that foundation models such as GPT-4 are capable of generating remedies when provided with abnormal or attack traces. This highlights the potential for future cellular networks to perform self-management and control, such as security countermeasure deployment and fault recovery through data plane control primitives such as dApps [28]. The O-RAN E2SM's RAN Control specification [8] defines a set of actions that could be incorporated into the AI pipeline. Reinforcement learning approaches [51] may also be adopted to ensure the network maintains healthy states.

**Specialized LLM for 6G.** While we show that baseline LLMs can provide accurate reasoning for cellular network data, these models are known to suffer from various weaknesses such as hallucination [64]. To address these issues, one could adopt Retrieval-Augmented Generation (RAG) [45] to augment prompts or adapt the pre-trained LLMs to specialized cellular domains (e.g., through fine-tuning), by leveraging accurate cellular protocol knowledge (e.g., those from 3GPP specifications [12]). Within this domain, prior work has used such documentation to guide security testing [23, 24].

**Zero-Trust O-RAN Architecture.** The adoption of O-RAN architecture in future cellular networks also opens new attack vectors. Unprotected O-RAN interfaces and services [5] could be potentially exploited to compromise network availability and user privacy. Malicious adversaries may poison the AI models with malicious telemetry to drive the network into abnormal states. As such, a zero-trust and resilient O-RAN architecture is necessary to prevent these threats in advance.

## Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This research was supported in part by NSF awards ITE-2226443, ITE-2326882 and CNS-2112471.

## References

[1] 2023. O-RAN.WG3.E2SM-R003-v04.00: O-RAN E2 Service Model (E2SM) KPM.

[2] 2024. 6G standardization – an overview of timeline and high-level technology principles. https://www.ericsson.com/en/blog/2024/3/6g-standardization-timeline-and-technology-principles.

[3] 2024. O-RAN Software Community. https://wiki.o-ran-sc.org/display/ORAN.

[4] 2024. O-RAN.WG1.OAD-R003-v11.00: O-RAN Architecture Description.

[5] 2024. O-RAN.WG3.E2AP-R003-v05.00: O-RAN E2 Application Protocol (E2AP).

[6] 2024. O-RAN.WG3.E2SM-CCC-R003-v03.00: O-RAN E2 Service Model (E2SM) Cell Configuration and Control.

[7] 2024. O-RAN.WG3.E2SM-R003-v05.00: O-RAN E2 Service Model (E2SM).

[8] 2024. O-RAN.WG3.E2SM-RC-R003-v05.00: O-RAN E2 Service Model (E2SM), RAN Control.

[9] 2024. oai / openairinterface5G. https://gitlab.eurecom.fr/oai/openairinterface5g.

[10] 2024. USRP Software Defined Radio (SDR). https://www.ettus.com/products/.

[11] 3GPP. 2024. NG-RAN Architecture description. http://www.3gpp.org/DynaReport/38401.htm.

[12] 3GPP. 2024. NG-RAN F1 Application Protocol (F1AP). http://www.3gpp.org/DynaReport/38473.htm.

[13] 3GPP. 2024. Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS). http://www.3gpp.org/DynaReport/24301.htm.

[14] 3GPP. 2024. Radio Resource Control (RRC). http://www.3gpp.org/DynaReport/38331.htm.

[15] 3GPP. 2024. Security architecture and procedures for 5G System. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3169.

[16] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[17] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3395–3404.

[18] Evangelos Bitsikas and Christina Pöpper. 2021. Don't Hand It Over: Vulnerabilities in the Handover Procedure of Cellular Telecommunications. In *Annual Computer Security Applications Conference*. 900–915.

[19] Evangelos Bitsikas and Christina Pöpper. 2022. You have been warned: Abusing 5G's Warning and Emergency Systems. In *Proceedings of the 38th Annual Computer Security Applications Conference*. 561–575.

[20] Leonardo Bonati, Salvatore D'Oro, Michele Polese, Stefano Basagni, and Tommaso Melodia. 2021. Intelligence and learning in O-RAN for data-driven NextG cellular networks. *IEEE Communications Magazine* 59, 10 (2021), 21–27.

[21] Leonardo Bonati, Pedram Johari, Michele Polese, Salvatore D'Oro, Subhramoy Mohanti, Miead Tehrani-Moayyed, Davide Villa, Shweta Shrivastava, Chinenye Tassie, Kurt Yoder, et al. 2021. Colosseum: Large-scale wireless experimentation through hardware-in-the-loop network emulation. In *2021 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 105–113.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[23] Yi Chen, Di Tang, Yepeng Yao, Mingming Zha, XiaoFeng Wang, Xiaozhong Liu, Haixu Tang, and Dongfang Zhao. 2022. Seeing the Forest for the Trees: Understanding Security Hazards in the {3GPP} Ecosystem through Intelligent Analysis on Change Requests. In *31st USENIX Security Symposium (USENIX Security 22)*. 17–34.

[24] Yi Chen, Yepeng Yao, XiaoFeng Wang, Dandan Xu, Chang Yue, Xiaozhong Liu, Kai Chen, Haixu Tang, and Baoxu Liu. 2021. Bookworm Game: Automatic Discovery of LTE Vulnerabilities Through Documentation Analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1197–1214.

[25] Merlin Chlosta, David Rupprecht, Christina Pöpper, and Thorsten Holz. 2021. 5G SUCI-Catchers: Still catching them all? (2021).

[26] Cooper Coldwell, Denver Conger, Edward Goodell, Brendan Jacobson, Bryton Petersen, Damon Spencer, Matthew Anderson, and Matthew Sgambati. 2022. Machine learning 5g attack detection in programmable logic. In *2022 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 1365–1370.

[27] Adrian Dabrowski, Nicola Pianta, Thomas Klepp, Martin Mulazzani, and Edgar Weippl. 2014. IMSI-catch me if you can: IMSI-catcher-catchers. In *Proceedings of the 30th annual computer security applications Conference*. 246–255.

[28] Salvatore D'Oro, Michele Polese, Leonardo Bonati, Hai Cheng, and Tommaso Melodia. 2022. dapps: Distributed applications for real-time inference and control in o-ran. *IEEE Communications Magazine* 60, 11 (2022), 52–58.

[29] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 1285–1298.

[30] Salvatore D'Oro, Leonardo Bonati, Michele Polese, and Tommaso Melodia. 2022. Orchestran: Network automation through orchestrated intelligence in the open ran. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 270–279.

[31] Mitziu Echeverria, Zeeshan Ahmed, Bincheng Wang, M Fareed Arif, Syed Rafiul Hussain, and Omar Chowdhury. 2021. PHOENIX: Device-Centric Cellular Network Protocol Monitoring using Runtime Verification. *arXiv e-prints* (2021), arXiv–2101.

[32] Simon Erni, Martin Kotuliak, Patrick Leu, Marc Röschlin, and Srdjan Capkun. 2022. AdaptOver: adaptive overshadowing attacks in cellular networks. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 743–755.

[33] Ismael Gomez-Miguelez, Andres Garcia-Saavedra, Paul D Sutton, Pablo Serrano, Cristina Cano, and Doug J Leith. 2016. srsLTE: An open-source platform for LTE evolution and experimentation. In *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*. 25–32.

[34] Pouya Hamadanian, Behnaz Arzani, Sadjad Fouladi, Siva Kesava Reddy Kakarla, Rodrigo Fonseca, Denizcan Billor, Ahmad Cheema, Edet Nkposong, and Ranveer Chandra. 2023. A Holistic View of AI-driven Network Incident Management. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 180–188.

[35] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[36] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.

[37] Syed Rafiul Hussain, Mitziu Echeverria, Imtiaz Karim, Omar Chowdhury, and Elisa Bertino. 2019. 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*. 669–684.

[38] Hongil Kim, Jiho Lee, Eunkyu Lee, and Yongdae Kim. 2019. Touching the untouchables: Dynamic security analysis of the LTE control plane. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1153–1168.

[39] Manikanta Kotaru. 2023. Adapting Foundation Models for Operator Data Analytics. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 172–179.

[40] Martin Kotuliak, Simon Erni, Patrick Leu, Marc Roeschlin, and Srdjan Čapkun. 2022. LTrack: Stealthy Tracking of Mobile Phones in LTE. In *31st USENIX Security Symposium (USENIX Security 22)*. 1291–1306.

[41] Andrea Lacava, Michele Polese, Rajarajan Sivaraj, Rahul Soundrarajan, Bhawani Shanker Bhati, Tarunjeet Singh, Tommaso Zugno, Francesca Cuomo, and Tommaso Melodia. 2023. Programmable and customized intelligence for traffic steering in 5g networks using open ran architectures. *IEEE Transactions on Mobile Computing* (2023).

[42] Gyuhong Lee, Jihoon Lee, Jinsung Lee, Youngbin Im, Max Hollingsworth, Eric Wustrow, Dirk Grunwald, and Sangtae Ha. 2019. 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*. 404–416.

[43] Hoejoo Lee, Youngcheol Jang, Juhwan Song, and Hunje Yeon. 2021. O-RAN AI/ML workflow implementation of personalized network optimization via reinforcement learning. In *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 1–6.

[44] Khaled B Letaief, Wei Chen, Yuanming Shi, Jun Zhang, and Ying-Jun Angela Zhang. 2019. The roadmap to 6G: AI empowered wireless networks. *IEEE communications magazine* 57, 8 (2019), 84–90.

[45] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[46] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*. Springer, 703–716.

[47] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. 2016. Mobileinsight: Extracting and analyzing cellular network information on smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 202–215.

[48] Norbert Ludant and Guevara Noubir. 2021. SigUnder: a stealthy 5G low power attack and defenses. In *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. 250–260.

[49] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. 2018. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. In *Network and Distributed Systems Security (NDSS) Symposium*.

[50] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)* 54, 2 (2021), 1–38.

[51] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. 2022. ColO-RAN: Developing machine learning-based xApps for open RAN closed-loop control on programmable experimental platforms. *IEEE Transactions on Mobile Computing* 22, 10 (2022), 5787–5800.

[52] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. 2022. Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. *arXiv preprint arXiv:2202.01032* (2022).

[53] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. *Biometrika* 71 (1986), 599–607.

[54] Sehan Samarakoon, Yushan Siriwardhana, Pawani Porambage, Madhusanka Liyanage, Sang-Yoon Chang, Jinoh Kim, Jonghyun Kim, and Mika Ylianttila. 2022. 5g-nidd: A comprehensive network intrusion detection dataset generated over 5g wireless network. *arXiv preprint arXiv:2212.01298* (2022).

[55] Alessio Scalingi, Salvatore D'Oro, Francesco Restuccia, Tommaso Melodia, Domenico Giustiniano, et al. 2024. Det-RAN: Data-Driven Cross-Layer Real-Time Attack Detection in 5G Open RANs. In *IEEE International Conference on Computer Communications*. 1–10.

[56] Zhaowei Tan, Jinghao Zhao, Boyan Ding, and Songwu Lu. 2023. Cell-DAM: User-Space, Rootless Detection and Mitigation for 5G Data Plane. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1601–1619.

[57] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. 2022. TranAD: deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment* 15, 6 (2022), 1201–1214.

[58] Luca Vigano and Daniele Magazzeni. 2020. Explainable security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 293–300.

[59] Haohuang Wen, Phillip Porras, Vinod Yegneswaran, Ashish Gehani, and Zhiqiang Lin. 2024. 5G-Spector: An O-RAN Compliant Layer-3 Cellular Attack Detection Service. In *Proceedings of the 31st Annual Network and Distributed System Security Symposium (NDSS'24)*. San Diego, CA.

[60] Haohuang Wen, Phillip Porras, Vinod Yegneswaran, and Zhiqiang Lin. 2022. A fine-grained telemetry stream for security services in 5g open radio access networks. In *Proceedings of the 1st International Workshop on Emerging Topics in Wireless*. 18–23.

[61] Haohuang Wen, Phillip Porras, Vinod Yegneswaran, and Zhiqiang Lin. 2023. Thwarting Smartphone SMS Attacks at the Radio Interface Layer. In *30th Annual Network and Distributed System Security Symposium, NDSS*.

[62] Hojoon Yang, Sangwook Bae, Mincheol Son, Hongil Kim, Song Min Kim, and Yongdae Kim. 2019. Hiding in Plain Signal: Physical Signal Overshadowing Attack on {LTE}. In *28th USENIX Security Symposium (USENIX Security 19)*. 55–72.

[63] Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. 2019. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 1409–1416.

[64] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).