# BTS-Band: An Explainable AI Detection Solution for Base Transceiver Station Resource Depletion Attack in O-RAN

Assrar Maamary
*Concordia University*
*assrar.maamary@mail.concordia.ca*

Hyame Assem Alameddine
*Ericsson Research*
*hyame.a.alameddine@ericsson.com*

Chadi Assi
*Concordia University*
*chadi.assi@concordia.ca*

Mourad Debbabi
*Concordia University*
*mourad.debbabi@concordia.ca*

***Abstract*—Incorporating intelligent controllers in Open Radio Access Networks (O-RAN) enables the development of multi-purpose applications to enhance the RAN's performance and bolster its security against new and existing attacks. The Base Transceiver Station (BTS) Resource Depletion (BTS-RD) attack is an example of the existing attacks, which aims at exhausting the BTS resources. To detect this attack in an O-RAN disaggregated setting, we propose BTS-Band, a novel BTS-RD O-RAN compliant detection solution that leverages a Bidirectional Long Short Term Memory (BiLSTM) AutoEncoder (AE) (BiLSTM-AE) complemented with the SHapley Additive exPlanation (SHAP) to further explain its detection results. We test the efficiency of BTS-Band in detecting three different variations of the BTS-RD attack that we emulate using a disaggregated Fifth Generation (5G) testbed based on the open source OpenAirInterface (OAI) project. Our experimental results show that the BTS-Band is able to detect BTS-RD variations with an average F1-score of 92.4% through leveraging statistical features that capture the signaling between O-RAN components. Moreover, the detection explanations produced by SHAP demonstrate that the BTS-RD attack variations exhibit different signatures successfully captured by the BTS-Band.**

***Index Terms*—O-RAN, 5G, BTS resource depletion attack, Anomaly detection, Machine learning.**

## I. Introduction

The transition to an open Radio Access Network (RAN) has recently gained momentum, with O-RAN being an important initiative in the field, spearheaded by O-RAN ALLIANCE [1]. With this architecture, openness, intelligence, and disaggregation are key concepts promoted for the Fifth Generation (5G) and beyond networks. These concepts revolutionize the RAN by breaking the long-standing vendor lock-in, further modularizing the RAN components, and enabling intelligent automation of RAN management [1]. O-RAN architecture, however, remains susceptible to BTS-RD attack, an attack which existed in previous generations and which exploits the lack of integrity check in Radio Resource Control (RRC) messages communicated between the RAN and the User Equipment (UE) [2]. These exchanged control plane messages involve allocating some RAN resources to the UE in the early steps of the registration procedure without verifying its identity (i.e., before authenticating it). To perform this attack, attackers can exploit one or multiple UEs to endlessly restart the initial registration procedure without properly completing it while continuously changing the UE's identity [3]. The attack can manifest in different variations, depending on the message within the registration procedure, after which attackers decide to stop responding to the network. Each variation triggers a specific timer or network behavior to release hanging connections. However, attackers spawn a large number of fake connections in a short time period (i.e., less than or equal to the timer) to ensure the depletion of RAN resources before the timer expiry and the release of malicious connections [3].

This underscores the role of a timely detection solution to report such attacks in their early stages. In fact, the authors of [4] proposed a rule-based detection approach that relies on predefined thresholds to decide whether counters for malicious events are significant enough to report a BTS-RD attack. However, such a solution does not account for the intricate patterns of benign network traffic and requires human expertise and intervention to adjust the thresholds upon changing network conditions. Other work in the literature attempted to determine malicious connection's source based on identifying the UE's location and comparing it to a presumably benign connection [5]. Such approach may not be efficient in the case of malicious UE mobility. These works do not study the impact of multiple BTS-RD variations in ORAN components under a disaggregated RAN architecture and fall short in presenting an intelligent detection solution to detect BTS-RD attack variations despite changing network conditions.

To fill these gaps, we study BTS-RD attack variations on O-RAN components and leverage the O-RAN architecture benefitting from intelligent radio management and open interfaces, to monitor radio signaling and extract related features for the BTS-RD attack detection. For that, we build BTS-Band, a novel eXplainable Artificial Intelligence (XAI) BTS-RD attack detection solution tailored to capture temporal dependencies between O-RAN components' signaling messages to efficiently detect different variations of the BTS-RD attack. Additionally, to increase the trust in its detection results BTS-Band leverages the SHapley Additive exPlanations (SHAP) [6], a game theoretical approach to explain its detection results.

To provide a holistic overview of our research, we outline our contributions as follows:

- We propose *BTS-Band*, a novel approach for BTS-RD attack detection with XAI. BTS-Band is composed of a *BTS-Stalker* module that monitors RAN signaling and a *BTS-BodyGuard* module that leverages features provided

978-3-903176-75-1 ©2025 IFIP

by the BTS-Stalker for BTS-RD attack detection augmented with results explainability using SHAP. To the best of our knowledge, we are the first to present an XAI solution for BTS-RD attack detection.
- Using the OpenAirInterface (OAI) open-source project [7], we setup an O-RAN compliant 5G testbed and emulate three different variations of the BTS-RD attack while modifying their intensities providing an extensive study on their potential impact in an O-RAN setup.
- Our experimental results show that the BTS-Band achieves an average F1-score of 92.4% across all BTS-RD attack variations while detecting most of them with F1-scores exceeding 94%.

The remainder of this article is structured as follows: we first review related work (Section II) and explain the UE registration procedure (Section III). Then we elaborate on the BTS-RD attack (Section IV), the proposed BTS-Band solution (Section V), and the testbed with attack emulation (Section VI). Finally, we analyze attack impacts (Section VII), present experimental results (Section VIII), and describe BTS-BodyGuard deployment (Section IX) before concluding (Section X).

## II. Related Work

While [3] provides a general overview of the BTS-RD attack in 4G systems, and [2] confirms its persistence in 5G, both studies focus on the overarching concept of the attack, where an adversary floods the RAN with registration requests using random UE identifiers to exhaust its resources. From a practical perspective, the study in [8] demonstrates the feasibility of only one variation of the BTS-RD attack in 5G systems. None of the aforementioned works addresses the need for a BTS-RD detection solution. In contrast, the work in [5] focused on distinguishing malicious RRC connections impersonating a specific UE from legitimate connections initiated by the actual victim UE. Their detection solution depends on physical and channel features to create a fingerprint distinguishing transmitting UEs, assuming attackers have procured the 5G Temporary Mobile Subscriber Identity (5G-TMSI) of a UE and used it to deny the UE access to the network by exploiting various RRC messages. Authors of [4] present a rule-based detection solution for the BTS-RD attack. Their solution depends on predefined thresholds, which, if violated, will trigger a BTS-RD attack detection alert. Such a solution fails to adapt to dynamic network patterns where changing network conditions can occur, hence making the threshold value obsolete. This requires human expertise to continuously adjust the threshold in order to limit false positives and negatives. Accordingly, these shortcomings motivated our aforementioned contributions.

## III. Background

We focus on the UE's initial registration in O-RAN, which attackers exploit for the BTS-RD attack. The process begins with the UE sending a preamble for uplink synchronization [9], followed by the RAN granting access via a *Random Access (RA) Response* message (msg) ( Figure 1). The UE then sends an *RRC Setup Request*, starting the RRC connection [9], which the RAN monitors using an inactivity timer [10].

The O-RAN Distributed Unit (O-DU) forwards the *RRC Setup Request* to the O-RAN Central Unit Control Plane (O-CU-CP), and if accepted, Signaling Radio Bearer 1 (SRB1) is established for UE–RAN connection. The O-DU then delivers the *RRC Setup* msg to the UE, after which the UE enters the connected state and can use the Cell Radio Network Temporary Identifier (C-RNTI). Finally, the UE sends an *RRC Setup Complete* msg including its identity and a Registration Request (i.e., includes the UE's Subscription Concealed Identifier (SUCI)). The O-CU-CP, then, forwards this *Registration Request* to the Access Mobility Function (AMF) after which two possible scenarios of interest can occur:

*1)* ***Successful Registration****:* If the SUCI is valid among other conditions [11], the AMF sends the UE an *Authentication Request* containing an authentication challenge and starts the T3560 timer [12]. The UE must reply with an *Authentication Response* before the timer expires, otherwise its connection is released. If successful, both parties proceed with security mode establishment, followed by Registration *Accept/Complete* msgs, finalizing the 5G registration [11].

*2)* ***Unsuccessful Registration****:* In cases where the SUCI is invalid, or for other specific reasons (e.g., unsupported requested network behavior or network congestion) [11], the AMF may send a *Registration Reject*, after which the UE's RRC connection is released.

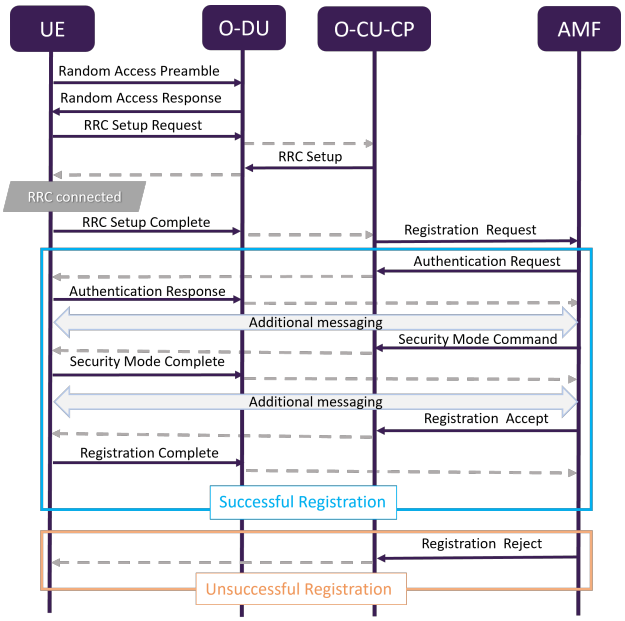
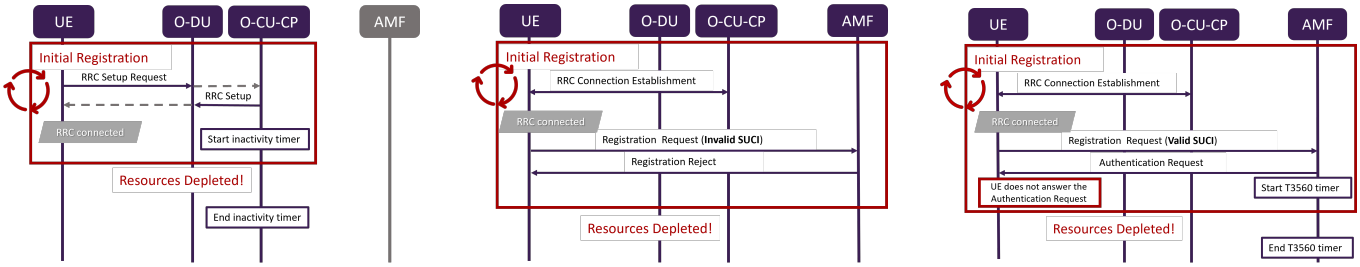



Fig. 1: Initial registration procedure [10], [11]

## IV. BTS-RD Attack

As early messages of the initial registration procedure are unprotected [2], a UE can acquire some RAN resources before verifying its identity. Consequently, attackers can exploit this vulnerability by acquiring RAN resources, such as in the RRC




(a) RRC-Storm (b) Invalid SUCI Flooding (c) Valid SUCI Flooding

Fig. 2: BTS-RD attack variations

Setup msg of the registration procedure, and repeatedly restarting the RA procedure using different CRNTIs and randomized UE identities (i.e., 39-bit random value) [3]. The number of RRC connections that the attacker needs to cause a BTS-RD attack, resulting in a Denial of Service (DoS) depends on the RAN's capacity to host RRC connections. In this section, we detail several variations the BTS-RD attack can manifest in, while explaining the feasibility of each.

*1)* ***RRC-Storm****:* Attackers continuously initiate the RA procedure, completing it up to the *RRC Setup Request* and then restarting it. By holding these RRC connections, they can exhaust RAN resources, denying legitimate UEs access to the network [8]. Here, the difference between the number of *RRC Setup* msgs and *RRC Setup Complete* msgs becomes key to detect the attack in its early stages. Further, the RAN releases each malicious connection when its inactivity timer expires (Figure 2a), making the number of Context Release Request messages from the O-DU important for the detection.

*2)* ***Invalid SUCI Flooding (Invalid-SUCI)****:* Attackers perform the initial registration up to the *Registration Request*, using a SUCI that does not match any valid UE. The AMF rejects the request, but the RAN still takes time to release the hanging resources. Attackers do not wait for the AMF's rejection (Figure 2b) and continuously repeat the attack to congest the RAN and exhaust its resources. Accordingly, the number of *Registration Reject* msgs received by the O-CU becomes an attack indicator, as rejections will occur far more frequently than under normal network conditions.

*3)* ***Valid SUCI Flooding (Valid-SUCI)****:* Attackers send a valid SUCI in the *Registration Request* triggering the AMF to proceed with an *Authentication Request*. As the Third Generation Partnership (3GPP) standard does not mandate the core network functions to check the freshness of a received SUCI, attackers can reuse previously observed SUCIs in the registration request to trick the AMF into starting authentication (a SUCI replay attack) [13]. To cause a DoS, they ignore the received *Authentication Request* and repeat the procedure, leaving connections hanging (Figure 2c). These hanging connections are only released when the AMF T3560 timer expires [12], delaying resource recovery and potentially exhausting RAN resources. This behavior shows the importance of tracking the number of sent *Authentication Request* msgs and comparing it with the number of received *Authentication Response* msgs, which should normally match.

## V. BTS-Band

To efficiently detect BTS-RD attack, we present in this section our novel BTS-Band solution, comprising a BTS-Stalker module and a BTS-BodyGuard module (Figure 3).

### A. BTS-Stalker: A BTS Monitoring Solution

This module ingests network traffic and computes performance metrics from both O-CU and O-DU via the following units:

*1)* ***Data Collection Unit****:* Uses Tshark to monitor RAN traffic and capture key protocols associated with the initial registration procedure and the disconnection of UEs from the network. Captured traffic is then stored in PCAP files.

*2)* ***Data Extraction Unit****:* Pre-processes collected PCAP files using Pyshark to extract relevant information associated with messages transmitted between the O-CU and the O-DU (i.e., registration and deregistration msgs).

*3)* ***Data Segmentation Unit****:* Counts RRC messages (e.g., setup request, setup complete) from pre-processed packets while keeping their temporal order. For that, a sliding window approach is adopted to segment packet streams into overlapping intervals where the window size (i.e., set to 3 seconds) defines the duration for each segment, and the window slide (i.e., set to 1 second) is the progression step through the data, dictating the start of a new window. This setting allows capturing dependent sequences and reveals the temporal proximity of messages. For instance, the sampled packets within a window can capture the entirety or most of the registration messages. Furthermore, the 1 second slide provides temporal granularity, revealing which messages appear in close temporal proximity (e.g., messages of the RRC connection establishment). The calculated counters, saved in CSV files, reflect RAN conditions and are used as statistical features to detect BTS-RD attacks.

### B. BTS-BodyGuard: A BTS-RD Attack Detection Solution

The BTS-BodyGuard module hosts an unsupervised Machine Learning (ML) model, namely a BiLSTM-AE, that leverages counters collected from the BTS-Stalker to detect BTS-RD attacks. It encompasses the following units.

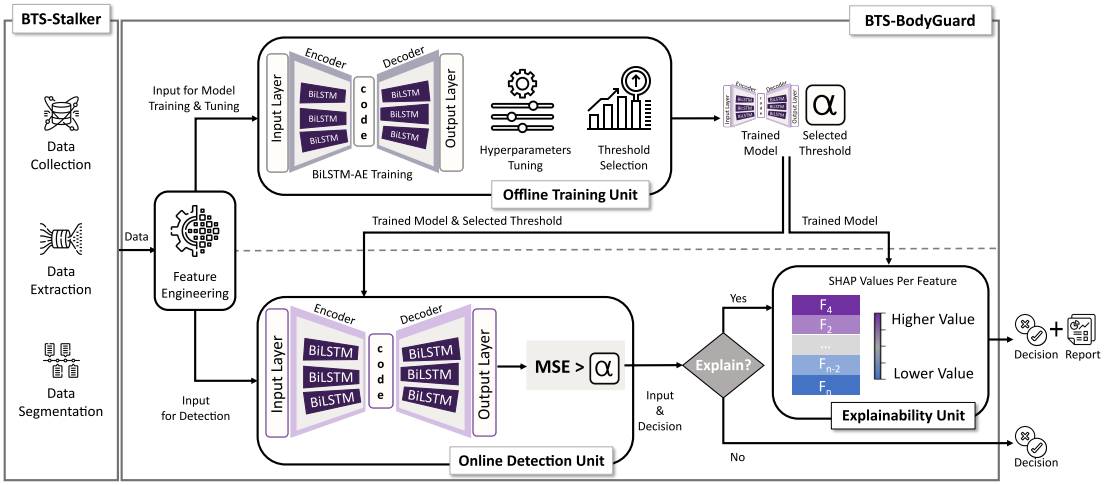



Fig. 3: An overview of the BTS-Band architecture

*1) **Feature Engineering Unit**:* Prepares timestamped counters from BTS-Stalker for the BiLSTM-AE model by indexing them in temporal order, normalizing with Min-Max scaling, and shaping them into sequences. It removes low-variance and highly correlated features to avoid redundancy, retaining 15 key features (Table I) that best capture normal network behavior and highlight patterns linked to BTS-RD attacks. This ensures relevant input for effective anomaly detection.

*2) **Offline Training Unit**:* Is responsible for training, validating, and testing a BiLSTM-AE model tailored for BTS-RD attack detection. Given that 5G procedures mostly follow a strict sequential flow of signaling messages, BiLSTM-AE becomes a well-suited option as it combines the strength of AEs in unsupervised anomaly detection with BiLSTM's ability to capture temporal dependencies in both forward and backward directions. This design allows the model to learn a compact representation of normal network behavior from benign data, making it effective even when labeled attack data is scarce [14], and enabling it to distinguish anomalies such as BTS-RD attacks from legitimate traffic. The BiLSTM-AE reconstructs inputs based on benign training data. Such reproduction may be erroneous. In this work, we use the Mean Squared Error (MSE) to quantify the reconstruction error. If the reconstruction error exceeds a preset threshold $\alpha$, the input is flagged as anomalous. Otherwise, it is benign. The threshold $\alpha$ is selected to maximize the F1-score as explained later in Section VIII-B. This unit can be used to retrain the model and re-select the threshold $\alpha$ whenever the operator deems suitable. Finally, the retrained model is passed to the online detection unit with the selected $\alpha$.

*3) **Online Detection Unit**:* Identifies BTS-RD attacks using the trained model from the offline training unit and input sequences from the feature engineering unit. It reconstructs the input, computes the MSE between input and output, and compares it to the threshold $\alpha$. This comparison drives BTS-BodyGuard's decision. If the evaluated sequence is deemed malicious, an alert is raised. The network operator may also request an explanation, provided by the explainability unit described hereafter.

*4) **Explainability Unit**:* Adds transparency to the BTS-BodyGuard's classification decisions. We use SHAP with the KernelExplainer implementation. SHAP is known for its robust game-theoretic foundation [6]. It requires a benign dataset to establish a baseline expectation for the model's prediction in the absence of specific input features. With this baseline, the explainer uses the trained model received from the offline training unit and an input sequence of interest (i.e., obtained either from the offline training unit or the online detection unit) to generate instance-level explanations by computing the given instance's SHAP values. These values quantify each feature's contribution (at a given timestep) to the gap between the model's prediction and the baseline. Higher SHAP values mean stronger influence on the detection decision.

## VI. Network Emulation and Data Collection

To detect the BTS-RD attack in its different variations, we need a suitable dataset. To the best of our knowledge, there is no publicly available 5G dataset reproduced in a disaggregated setup that provides BTS-RD attack features closely aligned with the RRC and network protocols [14]. Also, we want to evaluate the impact of the attack variations on the RAN under different intensity levels associated with each variation. Therefore, we leverage the open source OAI project encompassing a 5G RAN and UEs (version 2024.w43) along with a core network (version v2.1.0), to deploy a testbed (Figure 4) in a virtual containerized environment [7]. We configure a virtual machine with 16 CPU cores, 32 GB of RAM, and Ubuntu 22.04 to host the testbed. The latter does not implement the Open Fronthaul split adopted by O-RAN [1], thus, the O-DU will be referred to as DU. This does not affect the attack implementation and detection.

### A. Benign Network Emulation

We emulate normal traffic with 45 UEs performing RRC-based operations supported by OAI at the time of the writing (i.e., registration and deregistration). To account for daily

TABLE I: Features used by BTS-BodyGuard

| Feature | Tag | Collected At | Definition |
|---|---|---|---|
| Transmitted RAR | F1 | O-DU | The number of random access response messages transmitted from the O-DU to the UE. |
| Received RRC Setup Request | F2 | | The number of RRC setup requests received by the O-DU coming from the UE. |
| Transmitted RRC Setup Complete | F4 | | The number of RRC setup complete messages transmitted from the O-DU to the O-CU. |
| Transmitted Authentication Request | F5 | | The number of authentication request messages transmitted from the O-DU to the UE. |
| Transmitted Context Release Request | F12 | | The number of context release requests transmitted by the O-DU to the O-CU. |
| Transmitted RRC Release | F15 | | The number of RRC release messages transmitted by the O-DU to the UE. |
| Transmitted RRC Setup | F3 | O-CU | The number of RRC setup messages transmitted from the O-CU to the O-DU. |
| Transmitted Authentication Response | F6 | | The number of authentication response messages transmitted from the O-CU to the AMF. |
| Received Registration Reject | F7 | | The number of registration reject messages received by the O-CU coming from the AMF. |
| Received Security Mode Command | F8 | | The number of security mode commands received by the O-CU coming from the AMF. |
| Transmitted Security Mode Complete | F9 | | The number of security mode complete messages transmitted by the O-CU to the AMF. |
| Transmitted Registration Complete | F10 | | The number of registration complete messages transmitted by the O-CU to the AMF. |
| Transmitted Deregistration Request | F11 | | The number of deregistration requests transmitted by the O-CU to the AMF. |
| Received Context Release Command with cause = radio | F13 | | The number of context release commands with cause being ”radio failure” received by the O-CU coming from the AMF. |
| Received Context Release Command with cause = normal | F14 | | The number of context release commands with cause being ”normal” received by the O-CU coming from the AMF. |

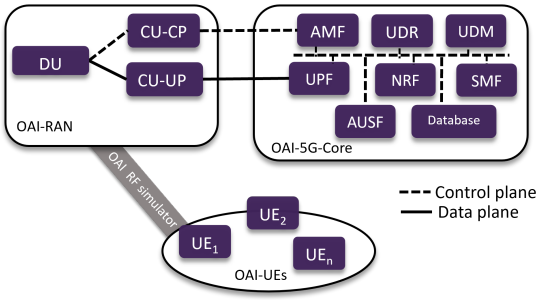



Fig. 4: Emulation testbed architecture

traffic patterns, commonly featuring peak loads around noon and lower ones during mornings and nights, we consider a Poisson [15] arrival of UEs under different loads such that no more than 16 UEs can coexist in the network as OAI's RAN is a femtocell. High and low loads are used in the emulation to introduce some abnormal non-malicious network conditions (i.e., connections rejection, or incomplete registrations).

### B. Attack Emulation

Assuming attackers have access to malicious UEs, we modify the UE code in OAI to induce malicious BTS-RD attack behavior (Section IV). More precisely, we emulate the RRC-Storm attack and three versions (representing different attack intensities) for each of the Invalid-SUCI, Valid-SUCI attacks and their mix. We refer to the mixed emulation later on by *Mix-SUCI*. The intensity level is defined as the number of RRC connection attempts per minute. We test three intensity levels: 4, 5, and 7 RRC connections per minute. For example, running the Invalid-SUCI attack at 7 RRC connections per minute is denoted as *Invalid-SUCI-7*. Our aim of keeping the intensity levels low (i.e., only 4, 5, or 7 connections per minute) is to identify the lowest needed level to disturb the network and the first one to cause a DoS. This is important for assessing BTS-BodyGuard's ability to detect stealthier attacks as it is already expected to easily identify much more apparent patterns in higher-intensity attacks. For every attack version, we begin by emulating benign network activity. The attack is then launched at the start of the second load of this emulation and continues for the duration of two consecutive loads. The final dataset and its documentation can be found here[1].

## VII. Attack Impact

### A. Impact On The DU CPU Utilization

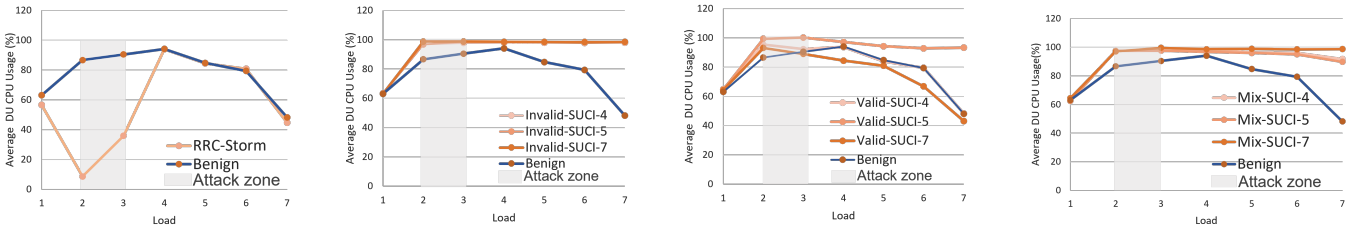
To evaluate the impact of the BTS-RD attack in its different variations, we collect the CPU consumption of the DU during the attack emulations every 2 seconds. We average the DU CPU utilization per network load for each attack version and present the results in Figure 5.

*1)* ***RRC-Storm****:* The DU CPU usage drops significantly during the RRC-Storm attack (Figure 5a) which can be attributed to the continuous streams of malicious RRC setup requests that get allocated SRBs by the DU without requiring additional computing resources. This is because malicious UEs do not respond to the RRC setup message; hence, the DU does not need to perform any processing for additional signaling.

*2)* ***Invalid-SUCI****:* All attack versions of the Invalid-SUCI attack variation heavily overloaded the DU, even after the attack ended (Figure 5b). Interestingly, during the *Invalid-SUCI-7* version, the extracted network logs showed that the DU went down midway through the attack, resulting in a complete loss of context for all established and queued connections. Once the DU reconnected, the attack resumed and quickly overwhelmed the DU again, which kept the CPU usage extremely high.

*3)* ***Valid-SUCI****:* The impact of Valid-SUCI varied between its versions. In *Valid-SUCI-4*, the attack caused a moderate increase in the CPU usage (Figure 5c) but the network recovered shortly after the attack ended. In contrast, in *Valid-SUCI-5* the CPU usage remained very high throughout the emulation. As for *Valid-SUCI-7*, again, the network logs show

[1]https://github.com/assrar/BTS_RD_Attack_Dataset




(a) RRC-Storm attack (b) Invalid-SUCI attack (c) Valid-SUCI attack (d) Mix-SUCI attack

Fig. 5: Average DU CPU usage per load

that the DU went down twice, once midway through the attack and once toward the end of the attack, losing context of all established and queued RRC connections, malicious and benign. This queue loss relieved the DU from handling malicious connections, which explains the low average CPU usage across the loads in *Valid-SUCI-7* (Figure 5c).

*4)* ***Mix-SUCI****:* As expected, all versions of this attack variation overwhelmed the DU (Figure 5d). Similarly, the network logs showed that the DU went down in *Mix-SUCI-5* and *Mix-SUCI-7* and exhibited high CPU usage after its reconnection.

### *B. Impact On Benign Users*

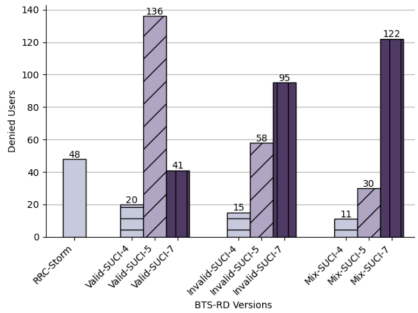



Fig. 6: Number of denied benign users per attack version

We also examine in Figure 6 the impact of the BTS-RD attack versions on benign users trying to access the network.

*1)* ***RRC-Storm****:* The RRC-Storm effectively denied 48 benign users access to the network nearly matching the total number of scheduled benign registration during the attack.

*2)* ***Invalid-SUCI****:* With the increase of this attack intensity, more benign users were denied access, proving network availability disruption. Accordingly, 15, 58, and 95 benign users were denied access to the network for intensity levels 4, 5, and 7 respectively.

*3)* ***Valid-SUCI****:* The number of denied users, along with the CPU usage, show network sensitivity to the different intensity levels of this variation. For instance, at (*Valid-SUCI-4*), the network denied access to 20 benign users. However, it eventually recovered after the attack ended. Conversely, at (*Valid-SUCI-5*), the DU remained persistently overwhelmed, denying network access to 136 benign users, (i.e., the highest number recorded across all tested attack versions). Yet, a *Valid-SUCI-7* which brought the DU down twice resulted in a total of 41 denied benign users.

*4)* ***Mix-SUCI****:* The number of denied benign users increased with the increase of the attack's intensity, resulting in 11, 30, and 122 denied users at levels 4, 5, and 7 respectively. The network logs showed that the DU went down then reconnected in both *Mix-SUCI-5*, and *Mix-SUCI-7*. however, the number of denied users in *Mix-SUCI-7* was much higher than in *Invalid-SUCI-7*.

## VIII. Experimental Results

### *A. Selecting BTS-BodyGuard's Architecture*

After evaluating many architectural configurations of BTS-BodyGuard, we selected a mirroring structure comprising two BiLSTM layers with 230 units each, encapsulating a latent representation of 60 neurons. This code bottleneck filtered key patterns while discarding noise. We also applied early stopping as a regularization technique to prevent overfitting and preserve the model's optimal state.

### *B. Selecting The Threshold*

We utilize the Receiver Operating Characteristic (ROC) curve a widely used metric that illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different threshold settings [16]. We thoroughly evaluate each threshold, computing the average F1-score (i.e., the harmonic mean between the correctly predicted positives and the TPR) across all BTS-RD attack variations and versions (i.e., 10 test datasets) to select the threshold $\alpha$ as the one that yields the highest average F1-score.

### *C. BTS-BodyGuard Performance*

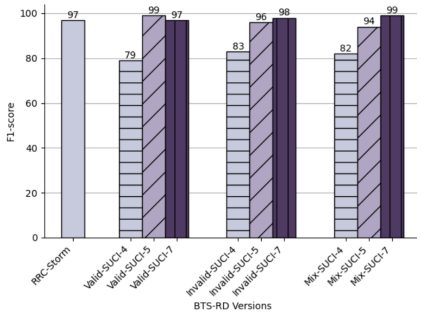
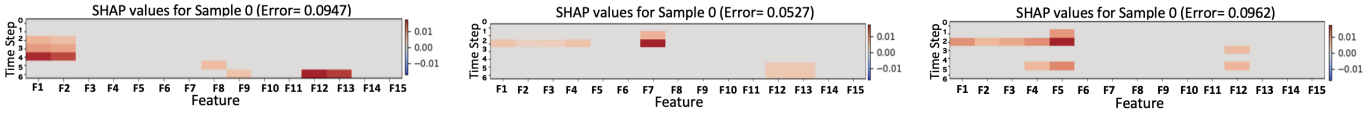



Fig. 7: F1-score per attack version




(a) RRC-Storm (b) Invalid-SUCI (c) Valid-SUCI

Fig. 8: Heatmap of SHAP values per feature per timestep

After selecting the threshold, we evaluate BTS-BodyGuard performance against all 10 attack versions and depict the F1-scores per attack version in Figure 7. We observe that the higher the intensity levels are, the better the BTS-BodyGuard is at detecting the attack, scoring as high as 99% for both intensity levels 5 and 7. These intensity levels are much lower than those reported in the literature (i.e., 20 RRC connections per second [3]), suggesting BTS-BodyGuard may perform even better under higher attack intensities. At the lowest level of 4 RRC connections per minute, the model yielded lower F1-scores, as the high stealthiness increases the chance of misclassifying attack data as benign.

### *D. BTS-BodyGuard Explainability*

We set up the SHAP explainer using 200 training samples to calculate the baseline model output. Then we select representative input sequences from each identified attack variation and pass them to the explainability unit to compute the corresponding SHAP values. This unit produces SHAP-based heatmaps that visually capture the features' contribution to the model decision per timesteps (Figure 8). The x-axis represents the features' tags and the y-axis depicts the timestep from the latest to the earliest where the red tones highlight features that significantly increase the prediction error, thereby pushing the model toward identifying anomalous behavior.

We observe in Figure 8a, explaining RRC-Storm detection, that early timesteps dominated by F1 and F2 contribute significantly to the prediction error. Also, features F12 and F13 emerge in dark red as key contributors toward the end of the sequence. Knowing that it was an RRC-Storm, the explainability unit accurately highlighted the features critical to its detection which aligns with the attack indicators identified in Section IV-1. For the Invalid-SUCI variation (Figure 8b), F1-F4 and F12-F13 contributed to the error. Yet, F7 had the highest impact, aligning with the discussion in Section IV-2. Similarly, in the Valid-SUCI variation (Figure 8c), the presence of features F1-F5, and particularly F5, drives the model's output, again consistent with our threat insights (Section IV-3).

## IX. BTS-BodyGuard deployment

**In An O-RAN Deployment** the features listed in Table I can be accessible through the O1 interface as vendor supplied measurements [17] and handed directly to the BTS-BodyGuard which can then be deployed as an rApp in the Non-Real Time Radio Intelligent Controller of the RAN as the data segmentation is performed every 3 seconds aligning with non-real-time control loops of O-RAN.

**In A Non-O-RAN Deployment** where the RAN does not have the O1 interface, the BTS-BodyGuard remains deployable where the BTS-Stalker module becomes indispensable in the setup to monitor the network and prepare the data for the BTS-BodyGuard, ensuring seamless integration in the system.

## X. Conclusion

We present BTS-Band, an XAI-based anomaly detection solution that identifies BTS-RD attack variations with an average F1-score of 92.4%. Using SHAP, it explains detection results across three emulated attack scenarios. Experiments on a 5G disaggregated testbed showed these attacks can cause DU DoS, blocking benign UEs. As a future work, we aim at extending the BTS-Band to also mitigate BTS-RD attacks.

## References

[1] O-RAN WG 1, "O-ran architecture description r004-v13.00," 2025.
[2] X. Hu, C. Liu, S. Liu, W. You, Y. Li, and Y. Zhao, "A systematic analysis method for 5g non-access stratum signalling security," IEEE Access, 2019.
[3] H. Kim, J. Lee, E. Lee, and Y. Kim, "Touching the untouchables: Dynamic security analysis of the lte control plane," in IEEE Symposium on Security and Privacy, 2019.
[4] H. Wen, P. Porras, V. Yegneswaran, A. Gehani, and Z. Lin, "5g-spector: An o-ran compliant layer-3 cellular attack detection service," in Proceedings of the 31st Annual Network and Distributed System Security Symposium, 2024.
[5] A. Scalingi, S. D'Oro, F. Restuccia, T. Melodia, and D. Giustiniano, "Det-ran: Data-driven cross-layer real-time attack detection in 5g open rans," in IEEE INFOCOM 2024-IEEE Conference on Computer Communications, 2024.
[6] T. Senevirathna, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, and S. Wang, "A survey on xai for 5g and beyond security: Technical aspects, challenges and research directions," IEEE Communications Surveys Tutorials, pp. 1–1, 2024.
[7] OpenAirInterface, "OpenAirInterface 5G," https://gitlab.eurecom.fr/oai/openairinterface5g, 2025, accessed: 2025-04-04.
[8] K. Baccar and A. Lahmadi, "An experimental study of denial of service attacks on a 5g cots hardware." in 2023 7th Cyber Security in Networking Conference, 2023.
[9] 3GPP, "NR; NR and NG-RAN Overall Description; Stage 2 ," 3GPP TS 38.300 v18.5.0, (2025-03).
[10] 3GPP, "5G; NG-RAN; Architecture Description ," 3GPP TS 38.401 V18.5.0 (2025-03).
[11] 3GPP, "Procedures for the 5G System (5GS); Stage 2," 3GPP TS 23.502 v19.3.0, (2025-03) .
[12] 3GPP, "Non-Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3," 3GPP TS 124 501 V19.2.0, (2025-03).
[13] I. You and H. Kwon, "Toward enhancing 6g security and resilience with blockchain: A case study on mitigating suci replay attacks," in 2025 IEEE International Conference on Consumer Electronics, 2025.
[14] G. L. Santos, P. T. Endo, D. Sadok, and J. Kelner, "When 5g meets deep learning: a systematic review," Algorithms, 2020.
[15] B. Chandrasekaran, "Survey of network traffic models," Waschington University in St. Louis CSE, 2009.
[16] T. Fawcett, "An introduction to roc analysis," Pattern recognition letters, vol. 27, no. 8, pp. 861–874, 2006.
[17] O-RAN WG 10, "O-RAN O1 Interface Specification R004-v15.00," 2025.