Smart, Automated, and Reliable Security Service Platform for 6G

Deliverable D5.2
Report on the use of PLS in 6G

ROBUST-6G project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101139068.

Date of delivery:	23/12/2025	Version:	0.1
Project reference:	101139068	Call:	HORIZON-JU-SNS-2023
Start date of project:	01/01/2024	Duration:	30 months

Document properties:

Document Number:	D5.2
Document Title:	Report on the use of PLS in 6G
Editor(s):	Stefano Tomasin (UNIPD) Mattia Piana (UNIPD)
Authors:	Stefano Tomasin (UNIPD), Mattia Piana (UNIPD), Matteo Varotto (UNIPD), Ali Hossary (UNIPD), Francesco Ardizzon (UNIPD), Tommy Svensson (CHA), Azadeh Tabeshnezhad (CHA), Masoom Rabbani (CHA), Mehdi Sattari (CHA), Ramin Fuladi (EBY), Cem Ayyıldız (GOHM), Fatih Emre Yildiz (GOHM), Arsenia Chorti (ENSEA), Linda Senigagliesi (ENSEA), Laura Luzzi (ENSEA), Sara Berri (CYU/ENSEA), Eunjeong Jeong (LIU), Nikolaos Pappas (LIU)
Contractual Date of Delivery:	23/12/2026
Dissemination level:	PU[1]/SEN
Status:	Final
Version:	0.1
File Name:	ROBUST-6G D5.2_v1.0

Abstract

This deliverable comprises four sections presenting the results of Work Package 5. The first three parts describe the results of the work package's three tasks, while the fourth part contains further in-depth results in the form of appendices.

Part I covers the contributions of partners to Task 5.1, which focuses on classifying, identifying and mitigating attacks on physical layer security. In particular, we consider jamming, impersonation and eavesdropping attacks, and our aim is to design defence mechanisms based on signal processing at the physical layer. We also investigate the detection and characterisation of active attacks. Lastly, we demonstrate that image-based frequency-domain analysis of software-defined networking traffic can detect distributed denial-of-service attacks.

Part II proposes new schemes for physical-layer security based on 6G key enablers, providing measurable security guarantees, and developing new solutions for authentication and key agreement. In particular, we examine various aspects of authentication and secret key generation based on the physical layer. Firstly, we propose new techniques for physical-layer authentication, including reconciliation schemes that increase accuracy by mitigating channel variations and novel authentication schemes based on challenge-response at the physical layer.

Part III focuses on ensuring the trustworthiness of the 6G physical layer and building trust for autonomous agents through privacy by design; robust sensing and cross-layer anomaly detection are also covered. Specifically, we address the challenge of securing location-based services in 6G networks by introducing a joint optimisation framework that preserves location privacy through differential perturbation techniques, all the while meeting service-level latency and throughput constraints. We also demonstrate that joint sensing of angles of arrival and time of flight offers valuable tools for verifying the physical integrity and location of autonomous agents. Finally, we advance generalised

[1] SEN = Sensitive, only members of the consortium (including the Commission Services). Limited under the conditions of the Grant Agreement

PU = Public

cross-layer anomaly detection to support the identification of both known and previously unseen threats.

All contributions are related to the ROBUST-6G security architecture and the project demonstration components.

Keywords

Physical Layer Security, Physical layer Authentication, Radio Fingerprint, Integrated Sensing and Communications, 6G

Executive Summary

This deliverable includes four parts that present the results of work package 5. The first three parts describe the results of the three tasks of the work package, while the fourth part collects appendices with further in-depth results.

Part I includes the contributions of partners to Task 5.1, which focuses on the classification, identification, and mitigation of attacks at the physical layer security. First, we study how to protect the legitimate users against eavesdropping and spoofing using radar signals by designing a robust beamforming. Complementarily, we provide a comprehensive study of physical authentication based on the angles of arrival, from analytical proofs that this feature in digital arrays is difficult to spoof, to a security analysis of a system using reflective intelligent surfaces and under realistic multipath conditions. Next, we address device identification at scale using radio-frequency fingerprints first as an alternative to cryptographic authentication, especially for large IoT deployments based on hardware imperfections, and channel-based authentication that uses the radio channel as a location-dependent *signature*. We introduce a receiver-invariant radio frequency fingerprinting identification system based on domain adaptation that aligns the feature spaces of different receivers, achieving a large fraction of the accuracy.

We also investigate the detection and characterization of active attacks. We propose several complementary jamming-detection schemes, such as detecting signal jammers by using spectrograms with supervised and unsupervised learning, developing a dedicated jamming detector that analyzes radio signals with a convolutional neural network trained to implement a generalized likelihood ratio test, and a dynamic-graph framework for cell-free massive multiple-input-multiple-output, in which graph neural networks monitor the evolving access-point/user connectivity to reveal jammer activity.

We show that image-based frequency-domain analysis of software-defined networking traffic can detect distributed denial of service attacks by turning magnitude–phase spectra into *spectral fingerprints*, a technique that naturally extends to physical-layer anomaly detection. Finally, we quantify how residual hardware impairments, in particular carrier frequency offset and symbol timing offset, shape discriminative RF fingerprints, and show that machine learning models trained on these features can distinguish legitimate devices from adversarial transmitters even when an attacker uses the same protocol, modulation, and transmission pattern.

Part II details the partners' contributions to Task 5.2. The objective of this task is to propose new physical-layer security-based schemes leveraging 6G key enablers with measurable security guarantees and to develop new solutions for authentication and key agreement. In the project, we investigate several aspects of physical-based authentication and secret key generation.

First, we propose several new techniques for *physical-layer authentication*, from *reconciliation schemes* that allow increasing their accuracy by mitigating channel variations, to novel authentication schemes based on challenge-response at the physical layer schemes that leverage the use of *reflective intelligent surfaces and drones* in 6G systems. We also address the effect of hardware imperfections on authentication with reflective intelligent surfaces.

Next, we develop new *wireless secret key generation protocols* that are fast, lightweight, and robust against

eavesdroppers. First, we investigate the feasibility of secret key generation in different line-of-sight multipath channels, validated by theoretical modelling and experimental campaigns; subsequently, we optimize design parameters to maximize key rates, and finally, we test our solutions on a demonstrator using software-defined radio. We also investigate secret key generation in unmanned aerial vehicle contexts, which is a challenging scenario as the line-of-sight component might be dominant.
We provide *rigorous leakage guarantees for secret key generation* via conditional mutual information and conditional mean-entropy estimators, as well as bounds for the information leakage of *wiretap codes* for short packet and low latency constraints. We also address *attestation* requirements, and propose a new cryptographic attested secret key generation protocol that integrates physical layer features and ensures that session keys can only be generated by devices at a prescribed location, providing identity-binding, integrity, and resistance to replay attacks. Finally, we assess the security of new key enablers in 6G, such as *integrated sensing and communications systems*, against adversarial machine learning attacks.

Part III covers the contributions of partners to Task 5.3, which focuses on ensuring the trustworthiness of the 6G physical layer and supporting trust building for autonomous agents through privacy-by-design, robust sensing, and cross-layer anomaly detection.
In the area of privacy by design, we address the challenge of securing location-based services in 6G networks. We introduce a joint optimization framework that preserves location privacy through differential perturbation techniques while meeting service-level latency and throughput constraints. This approach ensures that user privacy is embedded into the resource allocation process itself, rather than being an afterthought.
To enable trustworthy and robust sensing, we identify physical-layer trust anchors that support cyber-physical systems. We show that joint sensing of angles of arrival and time of flight provides useful primitives for verifying the physical integrity and location of autonomous agents. Furthermore, we address the long-term reliability of hardware-intrinsic radio-frequency fingerprints by modeling the *aging drift* of radio-frequency fingerprints. By creating models that predict how hardware changes over time, we help security systems stay reliable while reducing the need for frequent retraining.
Finally, we advance generalized cross-layer anomaly detection to support the identification of both known and previously unseen threats. We propose unsupervised learning frameworks, including a solution based on generative adversarial networks for Cloud radio access networks, to detect unseen contention anomalies by analyzing cross-layer key performance indicators. We enhance this with semantic metrics, such as the Age of Consecutive Errors, to prioritize errors based on their semantic significance and persistence. Extending these concepts to distributed scenarios, we introduce a federated learning framework for collaborative authentication that handles non-IID data through local fine-tuning, and a position-based intrusion detection system that fuses physical channel estimates with upper-layer traffic patterns to detect impersonation attacks.

Table of Contents

Acronyms

6G sixth-generation.

A-RPCA adaptive robust principal component analysis.

A-SKG Attested Secret-Key Generation.

AD *advantage distillation*.

ADDA adversarial discriminative domain adaptation.

AE autoencoder.

AES advanced encryption standard.

AI artificial intelligence.

AKA authentication and key agreement.

AoA angle of arrival.

AoA-PLA angle of arrival based physical layer authentication.

AoCE Age of Consecutive Errors.

AoD angle-of-departure.

AoI Age of Information.

AP access-point.

AWGN additive white Gaussian noise.

BEC binary erasure channel.

BI Bellman iterative solution.

BI-AWGN binary input additive white Gaussian noise.

BPSK Binary Phase Shift Keying.

BS base station.

BSC binary symmetric channel.

BW bandwidth.

CA Certificate Authority.

CAE convolutional autoencoder.

CFO carrier frequency offset.

CFR channel frequency response.

CIR channel impulse response.

CLD cross-layer-detector.

CNN convolutional neural network.

CPS cyber-physical system.

CPU central processing unit.

CPWT Correct-Physical-Wrong-Traffic.

CR challenge-response.

CR-PLA challenge-response physical layer authentication.

CRB Cramér Rao bound.

CSI channel state information.

DDoS distributed denial of service.

DENM decentralized environmental notification message.

DET detection error tradeoff.

DFT discrete Fourier transform.

DL deep learning.

DMC discrete memoryless channel.

DoS denial of service.

DS delay spread.

eMBB enhanced mobile broadband.

Eve eavesdropper.

FA false alarm.

FL federated-learning.

FPR false positive rate.

FR1 frequency range 1.

GAN Generative Adversarial Networks.

GCN graph convolutional network.

GLRT generalized likelihood ratio test.

GNN graph-neural-network.

ICT information and communication.

IoD Internet-of-Drones.

IoT Internet of Things.

IQ in-phase quadrature.

ISAC Integrated sensing and communication.

KL Kullback-Leibler.

KPIs key performance indicators.

LeakyReLU leaky rectified linear unit.

LLM large language model.

LoS line of sight.

LSTM long-short term memory.

LT likelihood test.

MAP maximum a posteriori probability.

MBRLLC mobile broadband reliable low latency communications.

MCRB misspecified Cramér Rao bound.

MD misdetection.

MDP Markov decision process.

MI mutual information.

MIAE mutual information-driven autoencoder.

MILP mixed-integer linear programming.

MIMO multiple input multiple output.

MINE mutual information neural estimator.

MITM man-in-the-middle.

ML machine learning.

ML maximum likelihood.

MLP multilayer perceptron.

mm-Wave millimeter-wave.

mMIMO massive multiple input multiple output.

MPCs multi-path components.

MSE mean square error.

NE Nash equilibrium.

NLoS non line of sight.

NN neural network.

NOMA non-orthogonal multiple access.

NTN non-terrestrial networks.

OFDM orthogonal frequency-division multiplexing.

PD probability of detection.

PDCP Packet Data Convergence Protocol.

PDF probability density function.

PG purely greedy solution.

PGD projected gradient descent.

PKG physical layer key generation.

PLA physical-layer authentication.

PLCL physical layer closed loop.

PLS physical-layer security.

PoC proof of concept.

PSD power spectral density.

QoS quality of service.

QoSec quality of security.

RAN radio access networks.

RBF radial basis kernel.

RF radio frequency.

RFE recursive feature elimination.

RFFI radio frequency fingerprinting identification.

RHI residual hardware impairments.

RIC RAN intelligent controller.

RIS reconfigurable intelligent surface.

RMS root mean square.

RMSE root-mean square error.

ROC AUC Area Under the Receiver Operating Characteristic Curve.

RPCA robust principal component analysis.

RSS received signal strength.

SDN software defined networking.

SDR software-defined radio.

SIC successive interference cancellation.

SIMO single-input multiple-output.

SINR signal-to-interference-plus-noise ratio.

SKA secret key agreement.

SKC secret-key capacity.

SKG secret key generation.

SKR secret key rate.

SKR secret key rate.

SNR signal-to-noise ratio.

STD standard deviation.

STO symbol timing offset.

SVM support vector machine.

SVR support vector regressor.

TCP transmission control protocol.

TDD time division duplex.

TDD time division duplex.

TEE Trusted Execution Environment.

ToF time of flight.

TPR true positive rate.

TRL technological readiness level.

TRM global trust and reputation management.

TVD total variation distance.

UAV unmanned-aerial-device.

UDA unsupervised domain adaptation.

UDP User Datagram Protocol.

UE user equipment.

ULA uniform linear array.

URLLC ultra relable low latency communications.

UWAC underwater acoustic channel.

VAE variational autoencoder.

VANET vehicular ad hoc network.

WIPS wireless intrusion prevention systems.

WPCT Wrong-Physical-Correct-Traffic.

WPWT Wrong-Physical-Wrong-Traffic.

Chapter 1

Deliverable Overview and Contribution to The Architecture

1.1 Deliverable Overview

In this deliverable, we summarize the contribution of the project provided by Working Package 5 on "Artificial intelligence (AI) / machine learning (ML) Enabled physical-layer security (PLS)".
This chapter provides an overview of the contributions and their relation to the ROBUST-6G architecture. We first recall the physical-layer closed-loop module, and then we summarize the contribution of each of the following chapters.
The rest of the deliverable comprises three parts and the appendices. Each part corresponds to a specific task of the work package and presents, in short chapters, the scientific results obtained. Extended versions of chapters are available either in published works (conference and journals) or are in progress; the intermediate results are in the appendices of this deliverable.

1.2 Physical-layer Closed Loop

PLS solutions are mainly introduced to the ROBUST-6G architecture through the Physical Layer Closed Loop (PLCL), a high-level architectural mapping has already been underlined in deliverable D6.2. Fig. 1.1 illustrates a first stable version of the PLCL, which involves three key stages: i) monitoring, ii) analysis, and iii) actuation. It also shows a high-level presentation of the interconnection among the utilized components and the interaction of the physical layer closed loop to the components of the ROBUST-6G architecture, i.e., the zero-touch security management and the network layer.
In brief, PLCL receives: i) physical (PHY) layer inputs from the radio access network (radio access networks (RAN)) specifications, i.e., RAN network functions (RAN NFs), such as channel state information (channel state information (CSI)) and radar-based sensing metrics, representing radio frequency (RF) signals and sensing observations from the radio environment, and ii) Upper layer context inputs from the Zero-touch Security Management Layer, e.g., security configuration parameters and contextual data such as GNSS-based localization information and orchestration alerts. The monitoring stage receives the PHY and Upper layer inputs and exploits the following components: i) PHY monitoring (CENS01), utilized to estimate core channel metrics, e.g., SNR and determination of line of sight (LoS)/non line of sight (NLoS) conditions, and ii) Data sets generation and fingerprinting for Physical Layer Security (CCHA02), employed to generate RF datasets for fingerprinting-based research. The former outputs are fed to the analysis stage to perform: i) PHY attack identification, e.g., using the Jamming Detection (CUPD03), Signal/Attack Identification solution to Classify different types of EM Signals (CEBY04), RF-Predict (CGHM02) and Cross-layer Holistic Anomaly

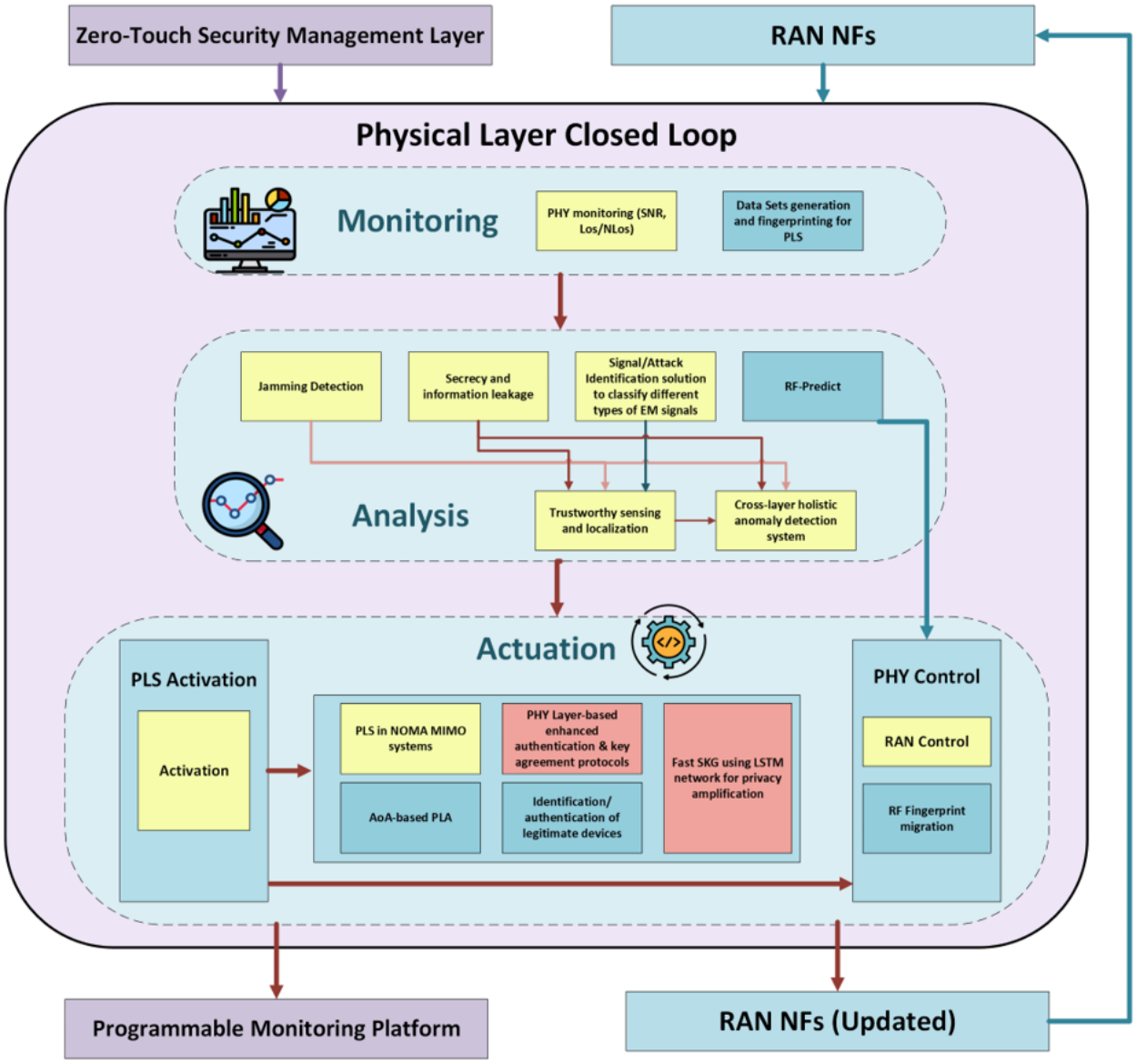


Figure 1.1: Physical layer closed loop.

Detection System (CUPD05) components, and, ii) overall trustworthiness evaluation of the physical layer, e.g., through the Secrecy and Information Leakage (CENS02) and the Trustworthy Sensing and Localization (CENS03) components. Finally, the actuation stage implements decisions based on the outputs of the first two stages to support PHY resource control and provisioning. These decisions include: i) the utilization of the appropriate PLS scheme, i.e., enable/disable security features depending on the required trustworthiness level; involved components are PLS in non-orthogonal multiple access (NOMA)-multiple input multiple output (MIMO) Systems (CCHA01), angle of arrival (AoA)-based physical layer authentication (PLA) (CENS04), Identification/Authentication of Legitimate Devices (CEBY05), PHY-layer based enhanced authentication and key agreement (AKA) Protocols (CUPD04) and fast secret key generation (SKG) using long-short term memory (LSTM) networks for Privacy Amplification (CENS05), and, ii) the activation of PHY control operations to adjust parameters regarding power/resource allocation, modulation schemes, synchronization sequences etc, through RAN Control and RF Fingerprinting Migration (CGHM01) component. This stage closes the PLCL by feeding back to: i) the PHY Layer the updated RAN specifications (RAN NFs) for continuous adaptation, and, ii) the orchestrator through the programmable monitoring platform, raising potential alerts from the infrastructure layer to the upper layers.

Table 1.1: Detailed Contribution to the Architecture

Chapter Number	Architecture Component	Dem. Comp.
2	Attack Identification (Analysis)	–
3, 19	Monitoring of the PHY (Monitoring)	CENS01
3, 19	Trustworthiness Evaluation of Localization and Sensing (Analysis)	CENS03
4	Monitoring and Analysis (Authentication)	CGHM01
5	PHY-Attack Identification (Analysis)	–
6	PHY-Attack Identification (Analysis)	–
7	PHY-Attack Identification (Analysis)	–
8	PHY-Attack Identification (Analysis)	–
9	Attack identification solution	CEBY04
10	Attack identification solution	CEBY04
11	PHY trustworthiness evaluation (Analysis)	–
12	Fast and Robust SKG (Actuation)	CENS05
13	Enhancing Performance of CSI-based PLA through Reconciliation (Actuation)	–
14, 12	Secrecy and Information Leakage Estimation	CENS02
15, 16	PHY-Attack Identification (Analysis)	CUPD04
17	PHY-Trustworthiness Evaluation	–
18	Identification/authentication of legitimate devices	CEBY05
19, 3	AoA and ToF based PLA (Actuation)	CENS04
20	Monitoring and Analysis (Authentication)	CGHM02
21	PHY-Trustworthiness Evaluation	CLIU02
22	PHY-Attack Identification (Analysis)	CLIU02
23	PHY-Attack Identification (Analysis)	–
24	PHY-Attack Identification (Analysis)	CUPD05

1.3 Detailed Contributions - Part I

In **Chapter 2**, CHA designs a dedicated sensing signal to detect the presence and estimate the location of a potential eavesdropper (Eve) with uncertain positioning. This is achieved by estimating key radar parameters such as the angle of arrival (AoA) and the range of the target. This solution is part of the PHY-Attack Identification module of the proposed ROBUST-6G architecture.
Chapter 3 presents a high-level discussion of four recent works from ENSEA and CYU AoA-based physical-layer authentication (PLA) and localization disseminated with ROBUST-6G acknowledgment in [1], [2], [3], [4] (joint work with UNIPD). These works map to CENS01, CENS03 and CENS04 in the ROBUST-6G architecture in all three stages of the physical layer closed loop (PLCL). They span multiple technological readiness level (TRL)s, including both i) theoretical results on the pertinence of the AoA as an unforgeable physical feature that can enable spoofing-resilient PLA, ii) analysis of AoA-PLA under advanced spoofing attacks using RIS (joint work with UNIPD) and finally iii) a PoC demonstration of its feasibility on a real, outdoor, massive multiple input multiple output (mMIMO) dataset at frequency range 1 (FR1).
In **Chapter 4**, GOHM addresses the scalability challenge of radio frequency fingerprinting identification (RFFI) caused by receiver variability, where models trained on one receiver fail when deployed on another. We propose a methodology based on unsupervised domain adaptation (UDA), specifically utilizing an adversarial discriminative domain adaptation (ADDA) framework to align feature distributions between source and target receivers. This approach eliminates the need for extensive labeled retraining data when migrating security models between different receivers or base stations. This solution defines the CGHM01 component within the analysis module of the ROBUST-6G architecture, enabling the orchestrator to maintain trust assessment of IoT devices as they move between different access points.
In **Chapter 5**, UNIPD presents an extensive survey on physical layer-based device fingerprinting, which is an emerging device authentication for wireless security. We focus on hardware impairment-based identity authentication and channel features-based authentication, which are passive techniques that are readily applicable to legacy IoT devices.
In **Chapter 6**, UNIPD proposes a new technique to detect jamming attacks based on the analysis of the spectrogram by a jamming detection device which is external to the network. The detection is based on a ML model that implements a one-class classifier by a convolutional autoencoder (CAE). This solution is part of the PHY-Attack Identification module of the proposed ROBUST-6G architecture, as it is aimed at analyzing measured signals (in this case, the spectrogram of the wireless spectrum used by the 6G cell) to identify threats (in this case, jamming attacks).
In **Chapter 7**, UNIPD offers another contribution on this topic by optimizing the classifier to ensure that it implements the likelihood test, thus connecting it to the statistical hypothesis testing theory. No specific demonstration component will be provided for this solution, but the solution has been extensively validated with experiments conducted with software-defined radios on a private 5G network set up for this purpose.
In **Chapter 8**, UNIPD resorts to dynamic graphs and graph convolution neural networks to detect jamming signals while capturing evolving communication links. No specific demonstration component will be provided for this solution, but the solution has been extensively validated with simulations.
Although in **Chapter 9** EBY focuses on detecting distributed denial of service (DDoS) attacks at the software defined networking (SDN) controller, its frequency domain imaging approach is directly relevant to the CEBY04 component for physical layer attack detection because both rely on identifying spectral and phase-based irregularities rather than simple time domain variations. By producing image-based spectral fingerprints that reflect full magnitude and phase relationships, the method closely resembles how CEBY04 analyzes radio frequency (RF) waveform structures to expose jamming, spoofing, and replay attempts. This alignment shows that the proposed technique can naturally extend into 6G physical layer level protection, where subtle spectral deviations are essential for detecting intelligent and adversarial signal behavior.
Chapter 10 is directly relevant to the CEBY04 component for physical layer attack detection because RF

fingerprinting exploits inherent hardware imperfections that cannot be replicated by attackers, allowing the system to distinguish legitimate devices from spoofed or cloned transmitters. By learning spectral and temporal features shaped by residual hardware impairments (RHI), carrier frequency offset (CFO), symbol timing offset (STO), and other device-specific distortions, the method developed by EBY aligns with how CEBY04 identifies subtle waveform inconsistencies that reveal impersonation and rogue device activity. These hardware-induced signatures remain stable even under noise and fading, making RF fingerprinting a strong and practical foundation for next-generation 6G physical-layer-level attack detection.

1.4 Detailed Contributions - Part II

In **Chapter 11**, CHA proposes a wireless setting where a resource-constrained node (N) establishes a secure session with an access point (AP), with both parties holding long-term certificate authority (CA)- issued credentials and no pre-shared secrets. By combining identity, integrity, and physical-layer information, the proposed system ensures that the resulting session key can only be derived by devices at the same physical location. Therefore, the design system model aims to integrate the node, the AP, and the CA within well-defined trust boundaries to achieve robust and location-bound secure communication. This solution is part of the Analysis module of the proposed ROBUST-6G architecture.
Chapter 12 discusses fast and robust secret key generation (SKG). Across our works [5–7] at ENSEA, we generated research outputs moving from low TRL works on the communication-theoretic modeling of LoS multipath channels, to a comprehensive study of SKG design parameters under worst-case eavesdropping attacks (on-the-shoulder), and finally to a context-aware, real-time SKG demonstrator on software-defined radios (SDRs). In these works, we delivered fast and lightweight, quantum-resilient SKG, with applications to 6G and Internet of Things (IoT) settings, well suited for low-end devices. We placed a strong emphasis on ensuring rigorous security guarantees (via conditional mutual information and conditional min-entropy estimators) and on the practical feasibility and real-time operation employing experimental measurement campaigns and demonstrators.
Chapter 13 discusses PLA enhanced with reconciliation [8, 9] based on recent ENSEA results. We demonstrate how the use of Slepian Wolf decoding can be applied in the case of CSI-based PLA and showcase that a careful fine-tuning of coding parameters, i.e., code-length and code-rate, can allow for arbitrarily low reconciliation error rates. Furthermore, in order to address evolving statistics in the time-domain for the CSI, we propose an adaptive robust principal component analysis pre-processing approach, which explicitly accounts for cross-corelation of channel realizations following a Markov chain model [10, 11]. These approaches can be incorporated into other PLA approaches, e.g., using RF fingerprints.
In **Chapter 14**, ENSEA discusses information leakage estimation and the possibility of using keyless transmissions when using finite blocklength wiretap coding. Wiretap coding allows for counter passive eavesdropping, provided that the legitimate receiver has a SNR advantage compared to the eavesdropper. In order to select a suitable wiretap coding scheme, it is necessary to adapt the channel coding rate to the channel conditions. For applications requiring short packets or low latency, it is not possible to guarantee a vanishing information leakage, and the back-off from secrecy capacity must be taken into account. Our goal is to obtain lower bounds on the achievable secrecy rate for a given block length, leakage, and error probability for general eavesdroppers' channels.
In **Chapter 15**, UNIPD provides an extensive analysis of attacks and their mitigation against PLA mechanisms implemented in a 6G network using reconfigurable intelligent surface (RIS). This technology is a key point for current and future communication networks in extending coverage and improving connectivity. Here, we also propose to use it to support PLA, using an innovative scheme based on a challenge-response mechanism. The resulting authentication mechanisms are part of the PHY-Attack Identification block, which performs the analysis of the received signal to establish, in this case, the authenticity of the message. Such solutions

will be in part included in CUPD04 as an enhanced authentication solution, even with respect to traditional PLA.
In **Chapter 16**, UNIPD proposes an innovative solution to generate secret keys using drones and to exploit the fading characteristics of the wireless communication channel. This pertains to the block PHY Components of the ROBUST-6G architecture, as it provides mechanisms of actuation by which drones change their route to implement the key generation mechanisms. It will be implemented in a second version of the CUPD04 component.
In **Chapter 17**, UNIPD proposes novel attacks based on adversarial ML to evaluate the robustness of Integrated sensing and communication (ISAC) systems, considering different scenarios and attacker capabilities.

1.5 Detailed Contributions - Part III

Chapter 18 is directly relevant to the CEBY05 component for authentication and identification because it demonstrates how residual hardware impairments, combined with RIS induced channel diversity, create distinctive physical layer characteristics that reliably separate legitimate transmitters from spoofers. By exploiting the temporal consistency of the Alice to Bob channel and the unpredictable CSI variations produced by Eve, the method proposed by EBY aligns with how CEBY05 verifies identity through sequential channel similarity checks. This RIS enriched channel structure strengthens discrimination, reduces miss detection and false alarms, and provides a robust basis for confirming that the received signals originate from the legitimate device.
In **Chapter 19**, we synthesized five contributions from ENSEA and CYU [12–16] on the role of the physical layer in 6G trust and trustworthiness. Across all works, we observed a common thread: future cyber–physical systems (CPS) and multi-agent networks will require *objective, quantifiable measures of trust*, deeply embedded in the wireless substrate itself. We made the case that joint AoA and time of flight (ToF) sensing emerge as crucial primitives that enhance trust, integrity, and accountability of autonomous devices.
In **Chapter 20**, GOHM introduces RF-PREDICT, a study focused on the temporal evolution of RF fingerprints to address "aging drift," where RFFI accuracy decreases over time due to factors such as hardware aging, temperature, and battery level. To analyze these factors, we have been collecting long-term data from custom-made, identical IoT sensors. A key objective of this study is to identify the optimum transmission interval required to keep RFFI accuracy stable. This work defines the CGHM02 component, which is dedicated to ensuring high identification accuracy is maintained over long operational periods without necessitating frequent device re-enrollment or model retraining.
In **Chapter 21**, LIU addresses physical-layer attack identification within the PLCL analysis stage by providing unsupervised anomaly detection capabilities for 6G cloud RANs. This work contributes to generalized cross-layer anomaly detection by integrating Generative Adversarial Networks (GAN) with transformer architectures to capture complex temporal dependencies in RAN performance data. The framework monitors key performance indicators spanning fronthaul traffic, thread scheduling, and precision time protocols: metrics essential for identifying network contention that may indicate physical-layer attacks or system degradation. By employing sliding window techniques and attention mechanisms, RANGAN enables the analysis stage to identify abnormal behaviors across physical and MAC layers, supporting trustworthiness evaluation and informing actuation decisions.
In **Chapter 22**, LIU contributes to physical layer trustworthiness evaluation within the PLCL analysis stage through semantics-aware remote estimation of Markov sources under resource constraints. This work addresses the challenge of determining information trustworthiness by integrating Age of Consecutive Errors (AoCE) to quantify the significance of estimation errors and Age of Information (AoI) to assess the usefulness of aged information. The framework formulates optimal transmission policies as constrained Markov

decision processes, demonstrating that switching policies achieve superior estimation quality. Together with Chapter 21, these works define the CLIU02 component, delivering LIU's contribution to Task 5.3 on cross-layer anomaly detection involving semantic attributes of information and learning attack phenomena. This semantic-aware approach enables distinguishing between critical errors requiring immediate action and benign variations.

In **Chapter 23**, UNIPD introduces a solution for authentication based on federated learning, where multiple base stations of a 6G network use local models to determine if the transmitter is transmitting from an authorized area or not, using the estimated CSI. The training is performed in a federated fashion while ensuring that each base station obtains a local, specific model that takes into account the specific propagation characteristics between the transmitter and the base station. This contribution is related to the PHY-Attack Identification block of the ROBUST-6G architecture, since it is related to PLA and thus to the analysis of received signals for security purposes. No component is dedicated to the demonstration of this solution, although extensive simulation results have been obtained to validate the effectiveness of the solution.

In **Chapter 24**, UNIPD introduces a new solution based on ML for the detection of threats using information from multiple layers of the network. In particular, we merge information coming from the physical and the network layers to establish the authentication of a message in an industrial automation context. This solution can be considered part of a PHY-Attack Identification module in the ROBUST-6G architecture, as it performs an analysis of signals to detect threats in a set of signals and will be tested in the CUPD05 component.

Part I

T5.1 - Classification, Identification and Mitigation of Attacks at PHY

Chapter 2

Physical Layer Security in NOMA MIMO Systems

ISAC is an important technology for sixth-generation (6G) mobile networks, enabling the joint use of communication and radar sensing within a unified system. While offering significant benefits in terms of spectral efficiency, ISAC introduces new security challenges. In particular, the joint use of resources for sensing and communication can increase vulnerability to eavesdropping and information leakage. In this chapter, CHA studies an uplink NOMA system where the base station (BS) simultaneously receives user data and senses a potential eavesdropper (Eve) with uncertainty in location. To enhance the PLS, a robust sensing signal is designed to both sense and jam the Eve. Specifically, we formulate a joint optimization problem that aims to maximize the sum rate of users and sensing performance while maintaining security against Eve. Since the optimization problem is challenging and non-convex, we propose an iterative algorithm that divides the problem into two subproblems, alternately optimizing precoding vectors and sensing power via quadratic optimization approaches. Simulation results demonstrate the effectiveness of our solution in terms of fast convergence and resource allocation.

2.1 Background and Motivation

The next generation of wireless networks is expected to support a diverse range of mission-critical applications, including autonomous systems, industrial automation, and smart defense infrastructure. These emerging services demand not only ultra-reliable and low-latency communications but also high-precision environmental awareness and stringent security guarantees. ISAC has emerged as a promising technology for 6G networks by enabling the joint operation of radar sensing and data communication through the use of shared spectrum and hardware resources [17], [18]. ISAC reduces hardware cost, improves spectral and energy efficiency, and facilitates real-time situational awareness via unified waveform design [19]. Despite this, ISAC introduces new design challenges, particularly in managing the interference between communication and sensing signals, and safeguarding such systems against malicious targets.
Concurrently, the power-domain NOMA has emerged as a compelling solution for multi-user access in ISAC systems. By enabling multiple users to simultaneously occupy the same time, frequency, and spatial resources-distinguished by their transmit power levels-NOMA improves spectral efficiency through superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver [20]. Thus, NOMA opens new opportunities for interference-resilient and resource-constrained ISAC system design, while also introducing new degrees of freedom for designing robust and secure waveforms against Eve in shared-spectrum environments. Therefore, this work is motivated by ISAC-NOMA beamforming to meet the quality of services, sensing accuracy, while protecting the users against Eve.

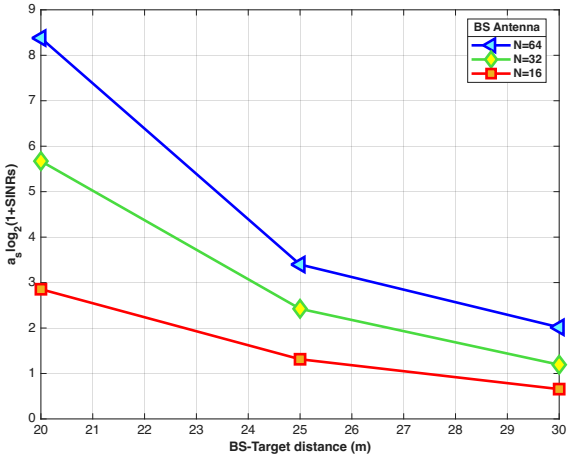


Figure 2.1: Sensing ratio as a function of the BS–target distance ($D = 20, 25, 30$ meters) for different BS antennas $N = 16, 32, 64$.

2.2 Proposed Methodology

In this work, we consider an uplink ISAC system empowered by power-domain NOMA, where a full-duplex radar BS equipped with two sets of well-separated antennas with $N_t = N_r = N$ to mitigate self-interference simultaneously receives from the uplink transmissions from two multi-antenna user equipment (UE)'s and the sensing echo, while transmitting radar waveforms to sense or jam a potential passive Eve. Each UE is equipped with M antennas, while Eve is assumed to have a single antenna. Eve acts as a malicious passive target who tries to overhear the uplink transmissions. While receiving the uplink signals from the UEs, the ISAC BS concurrently transmits radar signals to not only estimate Eve's location but also degrade her wiretapping performance. Such a scenario is typically considered in battlefield applications.

2.3 Numerical Results and Analysis

The simulation results demonstrate the impact of distance and system parameters on the overall performance. Fig. 2.1 demonstrates the impact of distance and system parameters on the overall performance. In the first figure, the capacity expressed as is plotted against the BS–target distance for different numbers of base station antennas $N = 16, 32, 64$. The curves show a consistent decrease in capacity as the distance increases from 20 m to 30 m, which is mainly due to the increase in path loss and the resulting reduction in received signal power. Among the three curves, the system with 64 antennas always achieves the highest capacity, followed by 32 and then 16 antennas. This agrees with the theoretical expectation that increasing the number of antennas enhances the beamforming gain and spatial diversity, thereby improving the effective SINR. At shorter distances, the performance gap between different antenna configurations is more pronounced, highlighting the significant advantage of using a larger antenna array in near- and mid-range scenarios, while this advantage slightly diminishes at longer distances due to dominant path attenuation effects.
Fig. 2.2 illustrates the convergence behavior of the proposed optimization algorithm in terms of the cost function versus the number of iterations for different distances. For all distances considered $D = 10, 15, 20, 25, 40$, the cost function decreases rapidly during the first few iterations and then gradually stabilizes after approximately 6 to 8 iterations. This confirms the fast convergence property of the algorithm and its suitability for practical implementation. It can also be observed that larger distances start with a higher initial cost and converge to a higher final value than smaller distances, which is a direct consequence of the lower SINR and more challenging channel conditions at longer ranges.

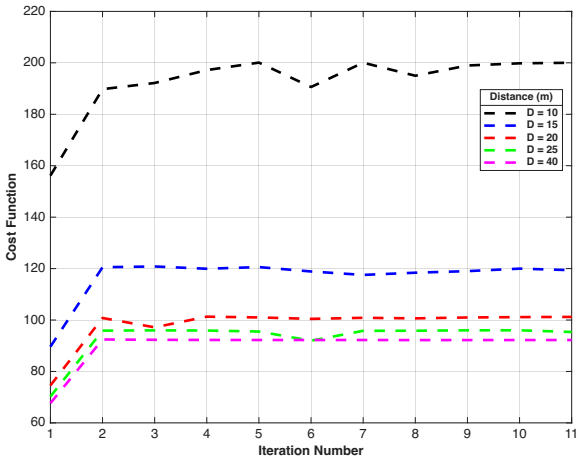


Figure 2.2: Cost function as a function of distance ($D = 10, 15, 20, 25, 40$ meters), under $N = 64$.

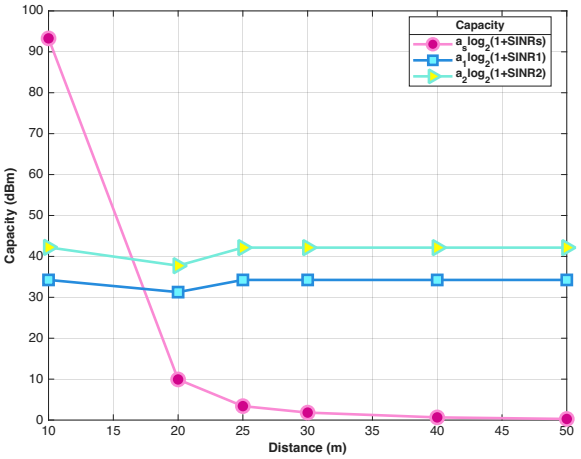


Figure 2.3: Capacity for both users and sensing as a function of Distance ($D = 10, 20, 25, 30, 40, 50$ meters) for $N = 32$ BS antenna.

In Fig. 2.3 we show the variation of capacity as a function of distance. As seen, due to the high attenuation, when the target is moving away from the BS, the sensing performance rapidly decreases as the distance increases. In contracts, the SINR for both users remains relatively stable over the entire distance range from 10 m to 50 m, with only minor fluctuations.

2.4 Integration with the Architecture

We have introduced a secure uplink ISAC systems that employ power-domain NOMA in the presence of a malicious target with unknown location, which has not been rigorously investigated in the literature. Existing works do not incorporate estimation-theoretic uncertainty into beamforming design, considering perfect CSI or sensing parameter estimation. Our work addresses this gap by introducing a CRB-informed design framework to capture the impact of parameter estimation errors on communication, sensing, and secrecy performance. The results of our work belong to the PHY-Attack Identification block, which secures the information from users against Eve.

Chapter 3

Unforgeable, Angle of Arrival based Physical Layer Authentication

3.1 Background and Motivation

This chapter presents an integrated discussion of four recent works from ENSEA and CYU on AoA based PLA and localization disseminated in [1], [2], [3], [4] (joint work with UNIPD). These works span multiple TRLs, including both i) theoretical results on the pertinence of the AoA as an unforgeable physical feature that can enable spoofing resilient PLA, ii) analysis of angle of arrival based physical layer authentication (AoA-PLA) under advanced spoofing attacks using RIS (joint work with UNIPD) and finally iii) a PoC demonstration of its feasibility on a real, outdoor, mMIMO dataset at FR1.
With respect to low TRL fundamental contributions, we have proven analytically that i) AoA in digital array MIMO systems is an unforgeable feature. Our proof is twofold and is based on: a) evaluating the mean square error (MSE) on AoA estimation in digital arrays under impersonation attacks [1] and b) deriving the misspecified Cramér Rao bound (MCRB) in the AoA estimation in a uniform linear array (ULA) under spoofing attacks [2]. These results jointly show that spoofing attacks can be identified in digital array systems using AoA-based PLA due to an attack-induced irreducible error floor that is independent of the SNR.
We note that in earlier joint works within the HEXA-X-II project (in which A. Chorti participated with her Barkhausen Institut affiliation), we have also shown that, on the contrary, AoA-PLA is not robust under impersonation attacks in the case of analog array MIMO systems [21]. Our analysis demonstrated that the loss of spatial degrees of freedom in analog array MIMO and the fact that rely solely on beam search patterns for source localization, renders the spoofing attack equivalent to a precoding optimization problem that can be solved precisely if sufficient information is available at the attacker side (legitimate node beamforming weights, locations of all nodes) as well as adequate power. Within ROBUST-6G we further provided a security analysis of RIS attacks on AoA-PLA under multipath propagation [21]. It was shown that attacks are only possible in the (unlikely) scenario that no multipath is present on the RIS-receiver side.
With respect to the suitability of AoA-PLA, we provided PoC results on a real outdoor dataset provided by Nokia (FR1 with carrier at 2.18 GHz, 64 mMIMO antenna array, 50 orthogonal frequency-division multiplexing (OFDM) subcarriers). It was shown that high-accuracy AoA-based outdoor localization is possible using hierarchical machine learning (ML) classifiers (first stage distinguishing between LoS and NLoS regions and a second stage classifier identifying the exact user track). **With respect to the WP5 stated key performance indicators (KPIs) we report reaching AoA-PLA accuracy of 100% at the first stage and more than 99.6% at the second stage**, surpassing the stated target of 90%.
Finally, we provided a first comparison in terms of computational complexity, between a) AoA-PLA and b) authentication using post-quantum cryptography. To this end, we estimated the number of central processing

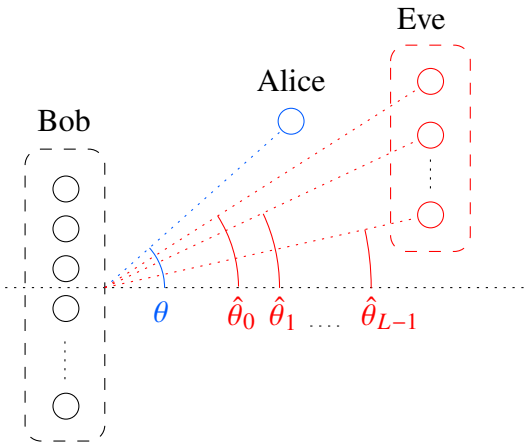


Figure 3.1: System model for AoA spoofing attack, in which Alice is equipped with a single antenna while both Bob and Eve are equipped with ULA.

unit (CPU) cycles of the proposed AoA-PLA against the state-of-the-art post-quantum signing algorithm Dilithium 5 (used in authentication handshakes).It was thus established that PLA is superior in terms of complexity, even when compared to only part of a post-quantum crypto-based authentication scheme. **With respect to the ROBUST-6G KPIs, we report run-times for AoA-PLA of less than 2 msec**. Further improvement in run-times and localization granularity is expected with the inclusion of time of flight (ToF) as an authentication feature, which is ongoing work, presented in Part II of this deliverable.
In the following, for each related publication, we outline key analytical or algorithmic contributions and the main numerical results. Finally, we conclude with a joint discussion, highlighting how these works collectively study the robustness and vulnerabilities of AoA as a robust feature for PLA, across different array architectures and propagation conditions.

3.2 Digital-array AoA-PLA Under Spoofing Attacks

3.2.1 Proposed Methodology

We considered a single-antenna legitimate user (Alice) communicating with a verifier (Bob) equipped with a digital ULA of M receive antennas, inter-element spacing d, and wavelength λ. Bob operated under far-field and narrowband assumptions, with carrier frequency f_c and bandwidth $B \ll f_c$.
The PLA protocol consisted of two phases:

- *Enrollment phase:* Bob collected AoA estimates from known legitimate transmitters (e.g., Alice), mapping estimated AoAs to node identities. During this offline phase, it was possible to employ higher-layer authentication mechanisms.
- *Verification phase:* A node declared their identity to Bob as Alice. Upon receiving this message, Bob estimated the AoA from the signal source and runs a hypothesis test (classifier) and declares whether the node is verified as Alice or not. An adversarial node, referred to as Eve, was also considered, equipped with an arbitrary number of antennas L.

We investigated spoofing (*impersonation attack*) by an adversary that attempts to choose its transmit strategy (location and precoding) so that Bob estimates an AoA indistinguishable from that of Alice. The adversary was allowed multiple antennas and can apply a complex precoding vector $\boldsymbol{q}$. The core question addressed

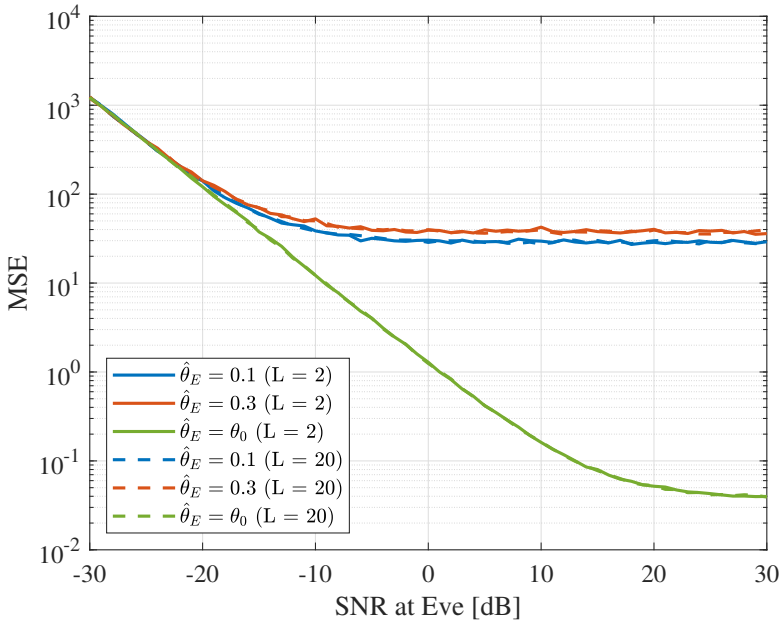


Figure 3.2: MSE vs SNR at Eve, with SNR at Alice equal to 15 dB, number of Bob's antennas $M = 16$, Alice's AoA $\theta = 0.4$ rad.

was: *under what conditions can Eve manipulate her transmitted signal such that the AoA at Bob is estimated to be the same as Alice's AoA, thus defeating AoA-PLA*?

3.2.2 Numerical Results and Analysis

In [1] the attack feasibility was characterized via the MSE between the observations when transmissions are received from the legitimate and adversarial nodes. In particular, closed form expressions for the MSE we derived, showing an irreducible with the SNR discrepancy between the received signals under Alice and Eve, even she employs optimal precoding to minimize the MSE, as exemplified in Fig. 3.2 for a particular numerical setting. In further detail, our theoretical analyses demonstrated that the spoofing attack results in an irreducible gap in AoAs. When Eve is not in the same direction as Alice , the AoA estimation error for Eve's signal cannot be made arbitrarily small by increasing the SNR or by optimal precoding. This leads to a non-vanishing gap between legitimate and adversarial AoAs. This gap is determined predominantly by geometric factors (AoA separation, array size), indicating that increasing the transmit power does not help the attacker unless the geometric conditions are already favorable. In future works we will look into the design of authentication beacons (pilots) to esnure a high AoA-PLA accuracy by indetifying the minimum number of antennas and SNR.
Furthermore, in the ongoing work [2] we extended the analysis in [1] to derive fundamental limits in AoA estimation under spoofing by emplying the machinery of MCRB. Our motivation lied in the fact that under spoofing attacks, the true observation model may deviate from the assumed model. Considering the same system model as in [1], we derived a closed form expression for the MCRB, as shown in AppendixMCRB. From this analysis we conclude that:

1. *MCRB versus Cramér Rao bound (CRB)*. The MCRB can be expressed as the sum of the CRB and an error term. As a result, it is always greater than or equal to the CRB. When there is no mismatch in the AoAs, the MCRB reduces to the CRB.

2. *Irreducible error floor.* The mismatch term does not depend on the noise variance. Consequently, increasing SNR reduces only the CRB term, but the second term can create a non-vanishing error floor at high SNR.

3. *Dependence on spoofer location.* The error term depends on the precoding, the adversarial position and on array geometry.

4. *Perfect alignment exemption.* If the adversary aligns exactly with the assumed steering direction, no mismatch exists.

5. *Impact of array size.* The CRB terms scales with M^{-3}, while the error term scales with M^{-2}; therefore the relative importance of the mismatch term depends on M in addition to the angular separations.

Furthermore, in a joint work with HEXA-X-II we considered the case where Bob is equipped with an *analog* antenna array: N antennas share a single RF chain and beamforming is implemented via analog combiners [21]. In contrast to digital arrays, AoA estimation in analog arrays relies on multiple transmissions with different beamforming vectors probing different directions.
Unlike in the case of digital array systems, analog arrays were shown to be vulnerable to AoA-based spoofing. Successful impersonation required knowledge of the locations of Alice and Bob and of the combiners at Bob, as well as sufficient transmit power at Eve. Nonetheless, when these conditions were met, both location-based and code-based attacks were effective in falsifying the AoA and compromising the security of the AoA-PLA. This contrasts with the robustness results for digital arrays MIMO systems.

3.3 Security Analysis of RIS-Assisted AoA-PLA Over Multipath Channels

3.3.1 Proposed Methodology

We studied AoA-PLA in a single-input multiple-output (SIMO) uplink scenario where the direct Alice–Bob link was blocked and communication occurred via a reconfigurable intelligent surface (RIS) controlled by Bob [4]. The system model included Alice as a legitimate transmitter with a single antenna, Bob as a (verifier) base station (BS) with a ULA of M antennas, and Trudy an single antenna adversary, using precoding to impersonate Alice. Furthermore, we assumed Bob controls an N-element reflecting surface applying a diagonal phase-shift matrix $\Omega = \mathrm{diag}(e^{j\phi_0}, \ldots, e^{j\phi_{N-1}})$. The system model is shown in Fig. 3.3.

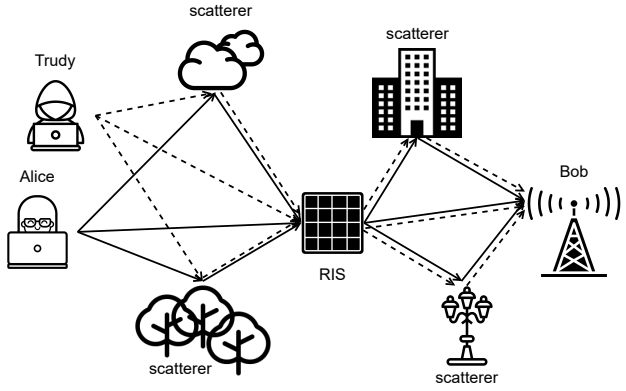


Figure 3.3: RIS-assisted PLA system: Alice and Trudy communicate with Bob via a RIS; the direct Alice–Bob link is blocked.

Operation	Number of CPU cycles
gen	819,475
sign	2,856,803
verify	871,609
Total	4,547,887

Table 3.1: Number of PU cycles for Dilithium 5 operations (source: IBM) for pk lenght= 2592, sign length=4595 (Intel Core-i7 6600U (Skylake) CPU, implementation in C)

The RIS configuration Ω was optimized for communication (e.g., spectral efficiency), using methods from prior work, and was assumed fixed during the PLA procedures. Bob performed AoA-PLA by estimating the cascaded channel and using it as a feature, assuming an enrollment phase and a verification phase during each new transmission. The adversary Trudy is assumed to know all channel matrices and is allowed to choose a precoding vector $\boldsymbol{q}$ and transmit power to maximize her impersonation success probability.

3.3.2 Numerical Results and Analysis

We derived mathematically an indistinguishability condition for this attack based on the AoA at the RIS. In more detail, we have shown that:

- For single-path RIS–Bob channels, impersonation was feasible even with mismatched channel parameters: Trudy could always choose the transmission parameters such that the cascaded channel matched Alice's in distribution.
- For multipath RIS–Bob channels, indistinguishability required very stringent conditions: matching AoAs at the RIS and proportional path gains across all paths. When these were not met, even the optimal transmission parameters result in an irreducible error floor.

Monte Carlo simulations confirmed these conclusions: increasing the number of RIS–Bob paths significantly enhances authentication robustness by limiting the attacker's ability to mimic the legitimate user [4].

3.4 ML-enabled AoA-PLA PoC on a Real Dataset

3.4.1 Proposed Methodology

We experimented with AoA-based localization in an outdoor mMIMO OFDM system, focusing on its robustness to impersonation attacks and its applicability to PLA [3]. We used a mMIMO digital array outdoor dataset collected at the Nokia campus in Stuttgart, Germany, depicted in Fig. 3.4. CSI vectors were available across multiple subcarriers and antennas, and AoA features were extracted using high-resolution algorithms. The goal was to identify user trajectories (tracks), in LoS and NLoS and test the efficiency of AoA-PLA, motivated by our work showing its robustness against impersonation in digital arrays [1, 2].
We performed extensive experiments comparing MUSIC and ESPRIT AoA estimators on the real outdoor CSI dataset, evaluating both accuracy and computational efficiency. The dataset was processed via a sliding-window technique; AoA features were then estimated per segment using MUSIC and ESPRIT, forming the feature vectors used for ML. To address the heterogeneity between LoS and NLoS regions, we proposed a hierarchical two-stage classifier as shown in Fig. 3.5

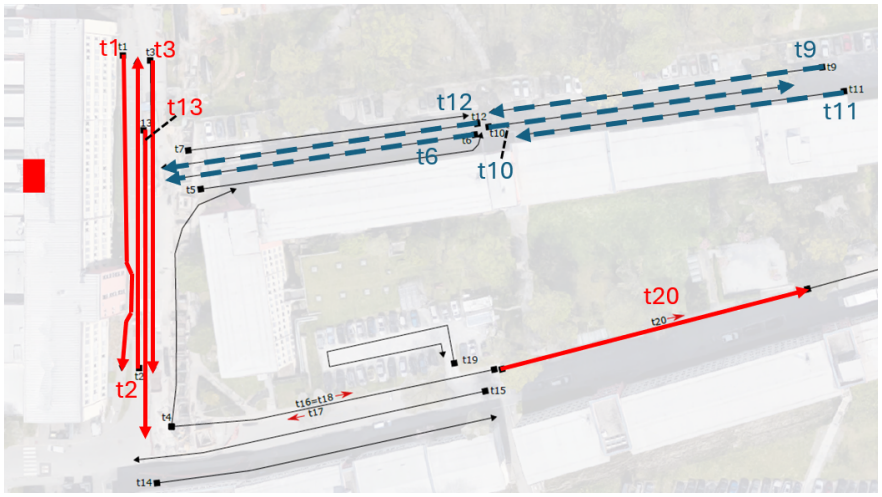


Figure 3.4: Nokia campus in Stuttgart, Germany. The red rectangle denotes the mMIMO antenna array mounted on top of a building, while the lines with arrows represent the trajectories (tracks) and their respective directions. Red solid lines indicate NLoS tracks, while blue dashed lines represent LoS tracks.

Operation	Number of CPU cycles
AoA estimation (window of 2000 samples)	2,423,998
LoS / NLoS	541,767
Specific track	553,541
Total	3,519,306

Table 3.2: Number of CPU cycles for AoA-based PLA including MUSIC and ESPRIT

3.4.2 Numerical Results and Analysis

We evaluated several base ML models, including logistic regression (LR), k-nearest neighbors (KNN), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and a stacking ensemble [3]. Since the publication of this work we have tuned SVM models and reached classification accuracy of 100 for stage 1 and 99.9% for LoS and 99.6% for NLoS for stage 2, with inference times of less than 2 msec.
Furthermore, we compared the AoA-based PLA with Dilithium 5 in terms of number of operations, and verified that the number of CPU cycles for PLA remains competitively low as shown in Tables 3.1 and 3.2. We also note that the AoA estimation is an inherent part of the link set-up to a particular user (and therefore a necessary step for any crypto algorithm to take place), therefore allowing the fusing of the link set-up with user authentication, resulting in reduced overall complexity and robustness.
This work serves as PoC for the feasibility of AoA-based PLA in a real outdoor environment. Online versions are also under development using meta-learning and online learning. In our preliminary results, we have reached accuracies of $> 90\%$ using a small initial fraction of the dataset and then using online learning. Furthermore, we are working on CNN-based deep learning AoA estimation to reduce the computational complexity demonstrated in 3.2 for the AoA estimation.

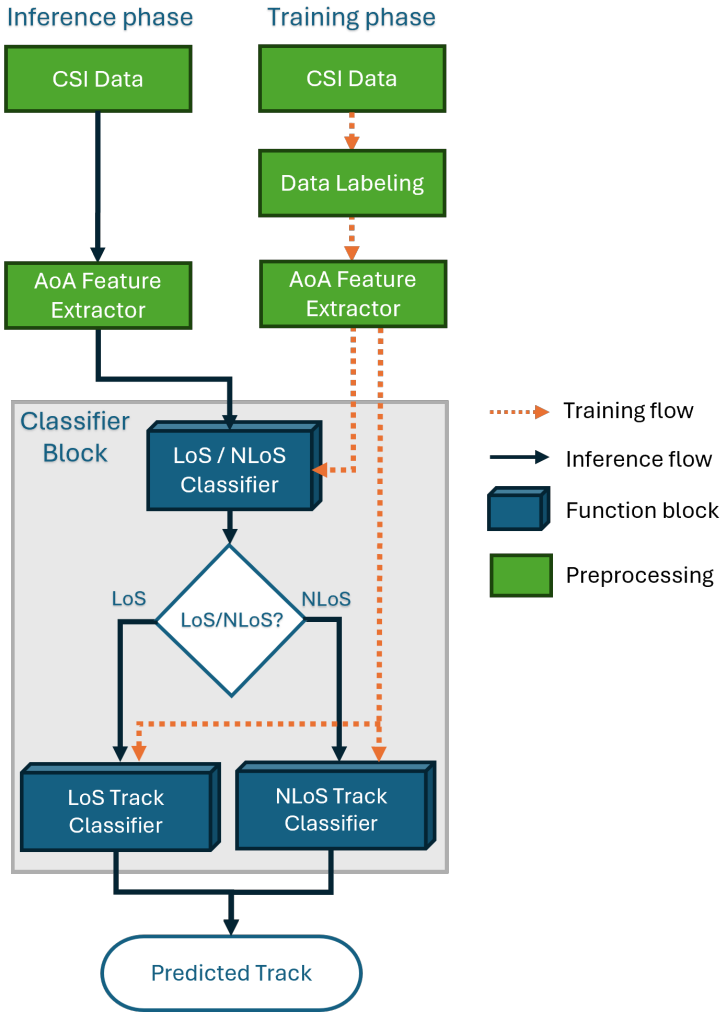


Figure 3.5: The proposed hierarchical two-stage classifier for user identification.

3.5 Integration with the Architecture

These works collectively provide a detailed picture of AoA-PLA under adversarial conditions and realistic propagation. We show that in digital array systems, impersonation attacks are feasible only under stringent conditions: the attacker's AoA must match the legitimate user's AoA, and corresponding precoding and phase alignment must be satisfied [1]. The work in [2] complements this by demonstrating that when the estimator is misspecified, an irreducible error floor arises from spoofing, dependent on the adversary's geometry and precoding but not on SNR, providing another proof of AoA's robustness as a PLA feature.
Furthermore, in [4] it is shown that multipath between a RIS and the verifier increases the rank of the effective channel and limits an attacker's ability to mimic the legitimate cascaded channel. Overall, spatial diversity provided by multi-antenna digital arrays and multi-path harden AoA against spoofing.
Finally, experimental results on a real outdoor dataset in [3] demonstrate that AoA features, when combined with hierarchical ML, can achieve high accuracy and robust LoS / NLoS discrimination. The fact that AoA features support 100% LoS / NLoS discrimination and high track classification accuracy suggests that AoA can serve as a jointly useful feature for both secure localization and authentication. It is further confirmed that AoA remains robust against impersonation in digital arrays, thus offering a stable basis for ML-driven

systems. We were able to achieve KPIs related to authentication accuracy (> 90%) and speed (< 2 msec). The above research was integrated in the following components of the PLS closed loop in the monitoring, analysis and actuation stages: **CENS01, CENS03, CENS04**.

Chapter 4

RF Fingerprint Migration

Future 6G networks are expected to support a massive number of connected IoT devices. Securing billions of connections presents a logistical challenge for standard protocols. While cryptographic authentication is secure, it requires significant energy and processing power, which can create bottlenecks for low-cost sensors with limited battery life. RFFI offers a non-cryptographic solution [22] to this problem. RFFI authenticates devices by detecting unique hardware impairments created during the manufacturing process [23, 24]. These analog imperfections, such as clock jitter, digital-to-analog converter non-linearities, and power amplifier distortions, act as a unique physical signature for the device. However, a primary obstacle for RFFI is receiver variability [25]. Models trained on one receiver often fail to identify the same devices when deployed on different hardware. In this chapter, GOHM presents a method to solve this scalability issue using UDA.

4.1 Background and Motivation

In future 6G deployments, IoT sensors will communicate with a heterogeneous infrastructure composed of various access points and base stations. One of the major barriers to the wide-scale adoption of RFFI is receiver variability [26]. This is a general challenge in RFFI, where models trained on one receiver often fail to recognize the same devices when deployed on a different receiver. This degradation occurs because the model inadvertently learns the unique analog characteristics of the receiver.
To maintain RFFI model stability in a large-scale network, packets from each new receiver need to be collected, labeled, and used to retrain the models. This must be repeated for every new receiver deployed. However, this process is prohibitive due to the high cost of data labeling and the operational downtime required. The goal of this work is to eliminate the need for extensive retraining by developing a receiver-invariant system capable of migrating a security model from a source receiver to a target receiver using minimal unlabeled data for adaptation, and a small amount of labeled data to select the best performing model. By enabling adaptation with minimal data from each new receiver, we aim to remove the logistical and cost barriers associated with continuous model retraining.

4.2 Proposed Methodology

Each receiver hardware introduces unique analog distortions, including phase noise, I/Q imbalance, and frequency-dependent gain variations [27]. When a model is trained on one receiver, it inadvertently learns these receiver-specific characteristics along with the device fingerprints, causing a domain shift that prevents the model from generalizing to new receivers. To address the domain shift caused by receiver hardware, we propose a methodology based on UDA. Specifically, we utilize an ADDA-based framework [28]. The core

objective is to align the feature distributions of the source and target domains so that the classifier remains effective across different receivers.
The methodology utilizes a source encoder, a target encoder, and a domain discriminator. The process operates in three distinct stages:

- **Supervised Pre-training:** First, a source encoder and a classifier are trained using labeled data from the initial receiver (source domain). This establishes a baseline capability to extract discriminative fingerprints from the source receiver.
- **Adversarial Adaptation:** We then initialize a target encoder for the new receiver. A domain discriminator is trained to distinguish between feature representations produced by the source encoder and those produced by the target encoder. Simultaneously, the target encoder is trained adversarially to mislead the discriminator. This adversarial game forces the target encoder to map its input signals into a feature space that is statistically indistinguishable from the source feature space, effectively removing receiver-specific distortions without using target labels for adaptation, relying on a small labeled set only for the final model selection.
- **Inference:** Finally, the adapted target encoder is combined with the original source classifier. This allows the system to correctly identify devices on the new receiver by mapping their signals into the shared, invariant feature space.

4.3 Numerical Results and Analysis

The performance of the proposed methodology was evaluated using the **RF Fingerprinting Migration Dataset**[1], created specifically for this project. Intermediate results demonstrate that the ADDA framework significantly mitigates the performance degradation caused by receiver variability. Prior to adaptation, transferring a model between different receivers resulted in a severe drop in classification accuracy. After applying the proposed unsupervised adaptation, the system restored a substantial portion of the lost performance. While the adaptation does not always fully recover the source-level accuracy, it improves the results enough to make the system operationally viable without requiring large-scale labeled data collection and retraining from scratch.
Furthermore, detailed numerical results, including F1-score comparisons and confusion matrices, are provided in Appendix A.

4.4 Contribution to 6G Physical Layer Security

This work contributes to the design of resilient 6G PHY by enabling scalable security. By minimizing the requirement for labeled retraining data, we reduce the operational overhead of PLS deployment. This aligns with the 6G goal of automated network management where security mechanisms can self-adapt to hardware changes and infrastructure upgrades with minimal human intervention.

4.5 Integration with the Architecture

The RF Fingerprint Migration solution defines the core logic of the **CGHM01** component within the "Monitoring" and "Analysis" block of the PLCL architecture. By ensuring that device identification remains robust across different receiver nodes, this component maintains a continuous trust assessment of IoT devices as they move between different access points.

[1] https://doi.org/10.5281/zenodo.14801935

Chapter 5

Physical Layer-Based Device Fingerprinting for Wireless Security: From Theory to Practice

The transmitter identification is part of the security mechanisms to ensure authentication in wireless communications. Conventional authentication approaches are cryptography-based, which, however, are usually computationally expensive and not adequate in the Internet of Things (IoT), while devices tend to be low-cost and with limited resources. In this study, UNIPD provides a comprehensive survey of physical layer-based device fingerprinting, which is an emerging device authentication for wireless security. In particular, this study focuses on hardware impairment-based identity authentication and channel features-based authentication. They are passive techniques that are readily applicable to legacy IoT devices. Their intrinsic hardware and channel features, algorithm design methodologies, application scenarios, and key research questions are extensively reviewed here. The remaining research challenges are discussed, and future work is suggested that can further enhance the physical layer-based device fingerprinting.

5.1 Background and Motivation

The IoT is expected to significantly impact our lifestyles. According to IoT Analytics, the number of connected devices reached 18.8 billion in 2024, an increase of 13% from 2023 [29]. These massively connected IoT devices have transformed our everyday lives with exciting applications such as smart homes, smart cities, connected healthcare, industry 4.0, etc. Wireless communications are preferred to connect these devices seamlessly. There have been many techniques for IoT, including WiFi (IEEE 802.11), ZigBee (IEEE 802.15.4), LoRa, Bluetooth-Low-Energy, and narrowband IoT (NB-IoT).

This revolution requires security at all levels. Security is quite a broad topic, involving confidentiality, integrity, availability, authentication, etc. This article will focus on device authentication, which is the first important step for network security. The receiver verifies the legitimacy of the received signal by checking specific features in the same signal. Our current computer and communications networks are protected by cryptography-based approaches, including both symmetric encryption, such as advanced encryption standard (AES), and public-key cryptography (PKC) such as Rivest-Shamir-Adleman (RSA). In particular, authentication is performed using a cryptographic challenge-response protocol based on symmetric encryption or PKC. However, cryptographic solutions may not be applicable to IoT devices. Symmetric encryption requires a pre-shared key, whose refresh turns to be challenging for IoT. PKC requires computationally expensive algorithms, which often have severe power and computational limitations; hence, they are unsuitable for IoT devices. In addition, at the dawn of quantum computing, PKC may be compromised due to the exponential

increase in the computational power of attackers. Due to the above limitations, there is a lack of competent IoT security solutions, and there have been many notorious security threats to IoT devices. This background is driving the development of lightweight, yet secure technologies for the IoT. Regarding device authentication, the two most promising non-cryptographic approaches are physical layer-based device fingerprinting, which includes hardware impairment-based RFFI and channel-based authentication. In detail,

- RFFI uses unique hardware impairments as the device identifier. Due to the imperfect manufacturing process, the nominal values of hardware components slightly deviate from their specification. These hardware impairments are unique and stable, which can be exploited as device fingerprints.
- Channel-based authentication exploits the channel characteristics through which the signal propagates to identify the source (or, better, its location) at the receiver, taking advantage of the fact that signals transmitted by devices at different locations travel through different channels (i.e., different delays and attenuations for each path). Thus, the propagation environment, rather than the transmitting device characteristics, and the relative position between transmitter and receiver, guarantee the authenticity of the transmitter.

5.2 Survey Aims

This survey, whose full version is in [30], complements and extends the published surveys with a comprehensive review of the physical layer-based fingerprinting for wireless security. We review the design principles of both RFFI and channel-based authentication. We also compare these two approaches and discuss their integration for more secure authentication mechanisms. Among the most promising and recent advances in these areas, we mention the availability of new technologies (such as RIS), the use of new transmission bands that fostered related technologies such as ISAC, the experimentation (thus with higher technology readiness level) of physical-layer security mechanisms, and the use of maximum likelihood (ML) techniques to secure transmissions by merging information coming from different communication layers. As unique features of our survey paper, we cover topics from theoretical development to practical implementation and share our experiences and insights on the design considerations of practical implementation. Thus, while looking at a specific domain, it still provides a general framework to discuss solutions across different domains.

5.3 Future Directions and Research Gaps

5.3.1 Generative AI Approaches

Generative AI represents transformative AI technologies to create new content, such as GAN and large language model (LLM). Generative AI has been widely used in securing communication from the physical layer, but its application in device fingerprinting is relatively limited.
Generative strategies/architectures include autoencoders (AEs), variational autoencoder (VAE), diffusion model, etc. They are used to design detectors in the anomaly detection context. They can also be used to generate the training dataset, e.g, VAE is used to generate satellite data. This may allow a binary classification-based detector to have an initial offline training with artificial but realistic data, later refined online. In the context of anomaly detection, generative models may be used to generate an artificial dataset. Regarding diffusion models, it is used for denoising in RFFI.
Recently, LLM have proven their effectiveness in multiple fields, even in the communication context. Still, no solution that exploits LLM has been proposed in the device fingerprinting context. Due to their generalization capabilities and if trained in a multimodal manner, thus taking as input also information concerning, for

instance, the environment, LLM may be used to generate high-fidelity artificial datasets, thus leading to even more robust detectors.
On the other hand, generative models may be used by the attacker to design effective attacks. In particular, an attacker provided with the legitimate detector (or dataset used for training it) may exploit a generative architecture to generate the attack samples that are most likely to fool the verifiers. Thus, future research directions should also include these attacks into account.

5.3.2 Emerging Communication Technologies

While the use of device fingerprinting for securing communication technologies such as WiFi is consolidated, for newer communication technologies, especially in the optical domain, only a few or even no work at all considers device fingerprinting for securing communication. This is the case for underwater optical communications, where, to the best of the authors' knowledge, very little research has been done. Thus, a research direction may involve the translation of the more consolidated solutions and algorithms into these new technologies.

5.3.3 Interplay between RFFI and Channel-based Authentication

RFFI and channel-based authentication represent two distinct but complementary approaches to wireless device authentication. RFFI relies on the unique hardware impairments inherent to individual devices, which are introduced during the component manufacturing process. The RFFI system is implemented at the receiver side, which is well-suited for scenarios where low-cost, infrastructure-independent security solutions are required. In contrast, channel-based authentication exploits the unique properties of the wireless channel, which are influenced by the surroundings; thus, it is effective in rich scattering environments, where it is hard for the attacker to predict, replicate, and compensate for the attack signal to effectively mimic the legitimate channel features.
The combination of RFFI and channel-based authentication offers a promising solution to enhance wireless security. Hybrid authentication protocols can be designed: RFFI ensures device-level identification based on unique hardware characteristics, while channel-based authentication validates location or monitors channel characteristics within a communication session. In this case, attackers would need to simultaneously replicate both the hardware impairments and the exact channel conditions to bypass the dual-layer protection, significantly increasing the difficulty of attacks.

Chapter 6

Detecting Signal Jammers Using Spectrograms with Supervised and Unsupervised Learning

Cellular networks are potential targets of jamming attacks to disrupt wireless communications. Since the fifth generation (5G) of cellular networks enables mission-critical applications, such as autonomous driving or smart manufacturing, the resulting malfunctions can cause serious damage. In [31], UNIPD proposes to detect broadband jammers by an online classification of spectrograms. These spectrograms are computed from a stream of in-phase and quadrature (IQ) samples of 5G radio signals. We obtain these signals experimentally and describe how to design a suitable dataset for training. Based on this data, we compare two classification methods: a supervised learning model built on a basic convolutional neural network (CNN) and an unsupervised learning model based on a convolutional autoencoder (CAE). After comparing the structure of these models, their performance is assessed in terms of accuracy and computational complexity.

6.1 Background and Motivation

The advent of fifth-generation (5G) technology promises very high data rates, low latency, and the support of mission-critical applications. However, 5G networks are vulnerable to jamming attacks, which may cause a denial of service (DoS) of critical applications, with potentially serious consequences on persons and things. One approach to cope with the threat of jamming is the use of wireless intrusion prevention systems (WIPSs) that monitor communication by analyzing features such as packet error rate (PER), bit error rate (BER), and signal-to-interference-plus-noise ratio (SINR). Using such features at a relatively high abstraction level (i) may be misleading since their high variation is typical in wireless channels and can, thus, only hardly be attributed to a single cause, and (ii) has been shown to fail at detecting jammers that target essential 5G signaling channels, such as the signal synchronization block (SSB). At a lower abstraction level, we find approaches that manipulate the 5G radio signals, e.g., by nulling some subcarriers and comparing the received power on such subcarriers with a threshold. This not only lowers the data rate of the system but also requires changes in current cellular network standards and systems. It is also inefficient since a simple threshold can be easily evaded by an intermittent jammer.
From a methodological perspective, some early machine learning (ML) and deep learning (DL) models have shown promising results through the direct analysis of received radio signals. Effective features were the number of transmissions, the clear channel assessments, or the aggregate measurements on the link layer.

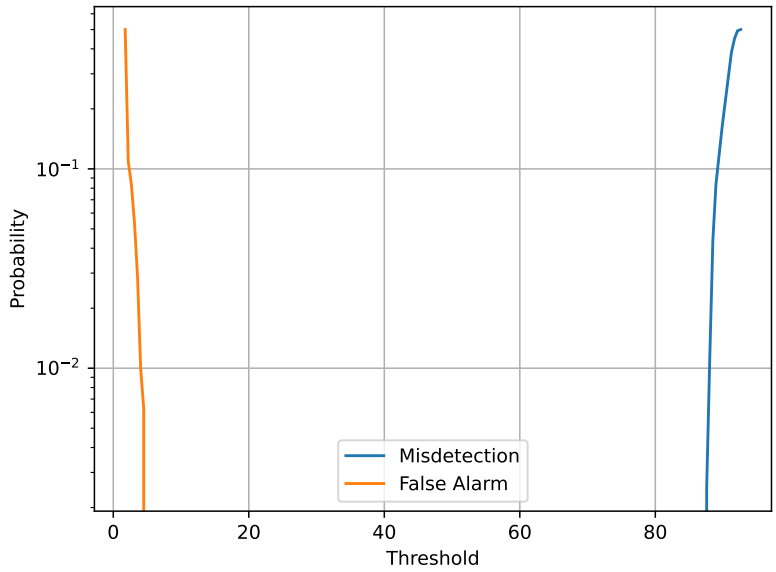


Figure 6.1: FA and MD probabilities as a function of the threshold τ for the uniform noise generator with the unsupervised learning approach.

6.2 Proposed Methodology

In [32], we propose a WIPS that obtains features directly from the radio signal at the physical baseband using ML. Based on received in-phase and quadrature (IQ) samples, a stream of spectrograms is computed, which is then used by a machine learning model to detect jammed signals. This process can be performed on a separate system (called watchdog) and requires neither changes to the 5G architecture nor to its signals. The watchdog can be functionally simple, as measuring received power requires only static parameterization, without further processing the radio signals, e.g., for equalization or decoding. A spectrogram, or more precisely, a power spectral density (PSD), can be still obtained from power measurements even when the received signal power is too low for communication. This allows to detect jamming attacks even at very low SINR – an important benefit compared to the mentioned approaches based on specific OFDM signals and to approaches using link-layer measurements.

6.3 Numerical Results and Analysis

6.3.1 Unsupervised Learning

The training set is composed of samples taken from trusted situations, with cases divided equally between an empty channel and an ongoing transmission.
From Fig. 6.1, we can see how the model distinguishes perfectly between the jammed and not-jammed cases. This perfect detection is possible because the reconstruction error of the jammed case is approximately 50 times higher than the reconstruction error of the case without jamming.

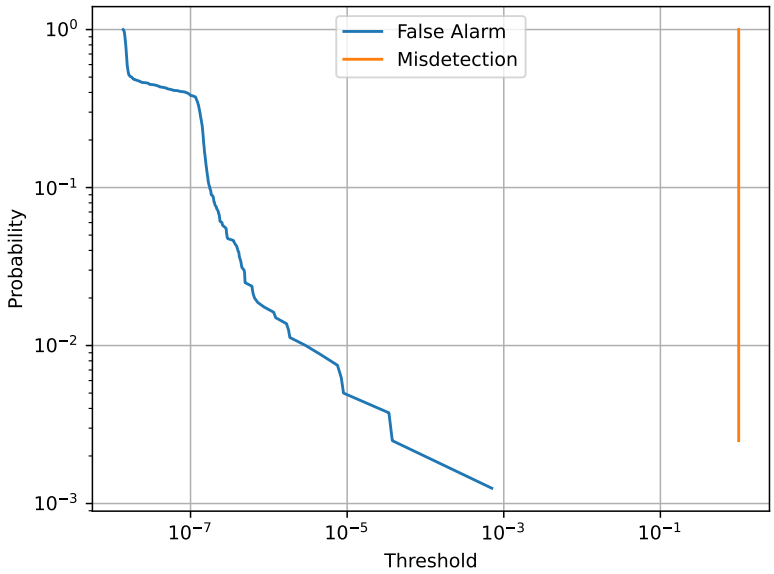


Figure 6.2: FA and MD probabilities as a function of the threshold $\tau \in [0, 1]$ (with y axis values normalized to 1) for the uniform noise generator with the supervised learning approach.

6.3.2 Supervised Learning

The training set was composed of samples, equally distributed between the three cases: jammed, not-jammed, and empty channel, not-jammed and ongoing transmission. The test set was composed of samples, distributed in the same way as the training and validation sets.
Comparing the detection rates in Fig. 6.2 to the results obtained in the unsupervised scenario shows that supervised learning reaches the highest accuracy. This becomes apparent by the absence of misdetection events and by the large threshold interval without false classification. This benefit of supervised learning, however, comes at a significant drawback that training is based on the signals of specific jamming attacks. Even slightly changing these signals may allow an attacker to evade detection. Albeit showing slightly worse performance, the unsupervised learning model is not based on specific attacks but models not-jammed signals. A jamming attack is then detected as a significant deviation from this trusted state.

6.4 Integration with the Architecture

We have introduced a novel method to detect jamming attacks by examining the spectrogram with an external jamming-detection device. The detection relies on a machine-learning model that functions as a one-class classifier implemented via a convolutional auto-encoder. This approach forms part of the PHY-Attack Identification module within the proposed ROBUST-6G architecture, as it analyzes measured signals—specifically the spectrogram of the wireless spectrum used by a 6G cell—to pinpoint threats such as jamming.

Chapter 7

One-Class Classification as GLRT for Jamming Detection in Private 6G Networks

Mobile networks are vulnerable to jamming attacks that may jeopardize valuable applications such as industry automation. In [33], UNIPD proposes to analyze radio signals with a dedicated device to detect jamming attacks. We pursue a learning approach, with the detector being a convolutional neural network (CNN) implementing a generalized likelihood ratio test (GLRT). To this end, the CNN is trained as a two-class classifier using two datasets: one of real legitimate signals and another generated artificially so that the resulting classifier implements the GLRT. The artificial dataset is generated mimicking different types of jamming signals. We evaluate the performance of this detector using experimental data obtained from a private 5G network and several jamming signals, showing the technique's effectiveness in detecting the attacks.

7.1 Background and Motivation

6G networks are expected to continue to be pervasive in everyday life scenarios, even more than the previous generation. Since they also support mission-critical applications such as smart manufacturing or autonomous driving, they should be adequately protected against security attacks.
Nowadays, wireless intrusion prevention systems (WIPS) monitor the security status of the transmission channel from the link layer up, aggregating measurements from the different communication layers. Several attackers, however, have learned to hide their malicious behaviors at layer 2 and above. Thus, a recent trend is to exploit the physical layer to provide security services. Here we leverage the recent work [32] that introduced deep learning (DL) to detect jamming attacks. Any device that injects noise into the band used for communication is considered a jammer, aiming at making the service unavailable to cellular devices. In this context, jamming and anti-jamming strategies have been recently surveyed.
We consider the WIPS as a one-class classification problem, also called *anomaly detection*. Note that the classifier also needs to detect jamming signals that have never been seen before, and on which it may not have been trained. Indeed, assuming a specific attack pattern may even lead to vulnerabilities in the learned detection model that an informed attacker may exploit. However, this constraint makes the design of anti-jamming techniques more challenging. A typical solution of such a one-class classification problem is the GLRT, which is used in various contexts. Still, this solution requires the knowledge of the statistics of received signals in legitimate conditions, which may be problematic to obtain, due to the different characteristics of the radio propagation environments where the private networks are deployed.

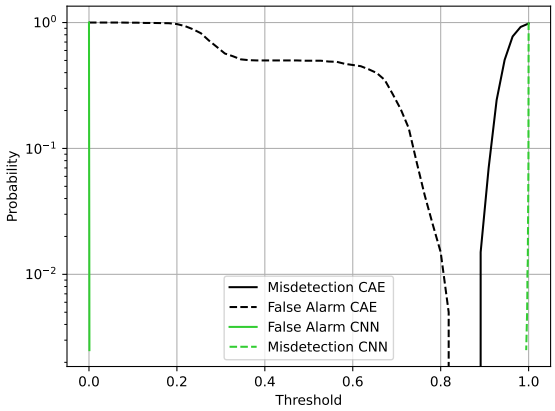


Figure 7.1: FA (continuous) and MD (dashed) rates achieved by CAE (black) and CNN (green) for $n = 256$ and uniform noise.

7.2 Proposed Methodology

We frame the WIPS as a one-class classification problem and tackle it by a GLRT implemented via supervised learning, in particular a CNN. As proven in [33], under suitable hypotheses, a DL model trained with supervised learning can indeed learn the GLRT, and thus can be used for one-class classification. Thus, drawing inspiration from [33], the detector builds an artificial dataset for the jammer with uniform distribution in the in-phase quadrature (IQ) sample domain and uses it during the training phase of the DL model. The accuracy of the trained model is evaluated using samples taken from a real-world jammer, thus modeling the discrepancy between the detector's prior knowledge and the actual attack statistics. The trained model performance is compared to the solution of [32] that uses CAE, a DL model that implements a full one-class classification problem. The comparison is based on experimental data, where the detector, jammer, and 5G base station are implemented as software-defined radios (SDRs).

7.3 Numerical Results and Analysis

We measure performance in terms of false alarm (FA) and misdetection (MD) rates for a variable threshold of the machine learning output between 0 and 1. To simplify comparison, the thresholds providing MD and FA rates of 10^{-2} are determined for each model. Then, for each pair of MD-FA curves, the distance between the respective MD and FA thresholds will be measured. The resulting distance value per model can then be compared among the models for a quick overview.
Fig. 7.1, 7.2, and 7.3 compare FA and MD rates achieved with both models for the three different jamming cases, i.e., uniform, uniform over a frame, and Gaussian jamming.
With uniform noise, the CNN clearly shows a better performance than the CAE.
With frame-like noise the CNN still outperforms CAE. This is indicated by the separation between the FA and MD curves for the CNN, which is 0.5 substantially wider than the separation of 0.35 with CAE. This is a performance gain of 43% over the baseline.
With Gaussian noise, the CNN model reaches an even higher performance gain. CAE produces a separation between the two curves of approximately 0.35, while the CNN produces a separation of approximately 0.75. Thus, the CNN with artificial data outperforms the baseline by 114%.
In addition to $n = 256$ IQ samples per bitmap, models created and tested with larger time windows were also

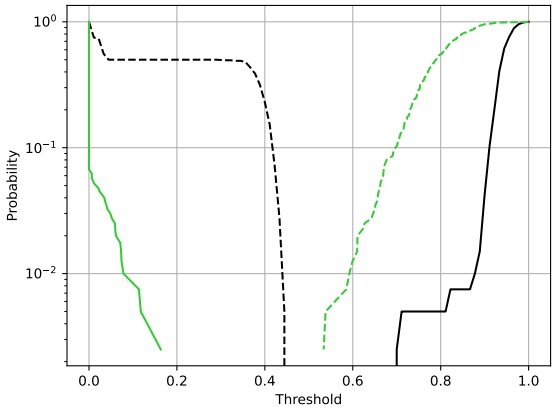


Figure 7.2: FA (continuous) and MD (dashed) rates achieved by CAE (black) and CNN (green) for $n = 256$ and uniform noise over a frame.

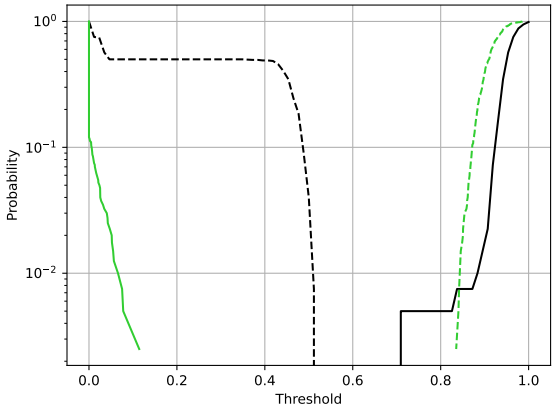


Figure 7.3: FA (continuous) and MD (dashed) rates achieved by CAE (black) and CNN (green) for $n = 256$ and Gaussian noise.

studied. With a window size of $n = 1024$ samples, the separation of the curves improves, compared to CAE, but the gain is smaller than with $n = 256$. A time window of 2048 samples, on the other hand, significantly improved separation and gain for the case of uniform noise.

7.4 Integration with the Architecture

We have presented a further contribution with respect to our previous contribution of Chapter 6, refining the classifier so that it performs a likelihood test, thereby linking it to statistical hypothesis-testing theory. Although no dedicated demonstration component accompanies this solution, it has been thoroughly validated through experiments with software-defined radios on a private 5G testbed created for this purpose.

Chapter 8

Jamming Detection in Cell-Free MIMO with Dynamic Graphs

Jamming attacks pose a critical threat to wireless networks, particularly in cell-free massive MIMO systems, where distributed access points and user equipment (UE) create complex, time-varying topologies. In this chapter (reported in its full version in Appendix I), UNIPD proposes a novel jamming detection framework leveraging dynamic graphs and graph convolution neural networks (GCN) to address this challenge. By modeling the network as a dynamic graph, we capture evolving communication links and detect jamming attacks as anomalies in the graph evolution. A GCN-Transformers-based model, trained with supervised learning, learns graph embeddings to identify malicious interference. Performance evaluation in simulated scenarios with moving UEs, varying jamming conditions, and channel fadings demonstrates the method's effectiveness, which is assessed through accuracy and F1 score metrics, achieving promising results for effective jamming detection.

8.1 Background and Motivation

Wireless communication increasingly adopts cell-free architectures to enhance connectivity and spectral efficiency. Cell-free MIMO relies on access-points (APs) that jointly serve UEs without predefined cell boundaries. This paradigm shift introduces new challenges related to network dynamics and security.
As reliance on wireless services continues to grow, security threats have become a major concern. Wireless networks, due to the shared nature of the radio spectrum, are particularly vulnerable to jamming. In MIMO wireless networks, traditional jamming detection methods rely on statistical models, which struggle to adapt to the complexities of dynamic wireless environments. In contrast, DL techniques can be applied using a data-driven approach. DL approaches, such as CNNs, have been employed to analyze spectrogram images for jamming detection, outperforming conventional feature-based methods. Furthermore, federated learning has been investigated for distributed jamming detection in flying ad-hoc networks. However, all these solutions are agnostic of the network structures and are not suited for cell-free communications where synchronization is looser. When users are mobile and channel conditions vary, modeling network behavior is crucial. *Dynamic graphs* offer a powerful representation for the evolving topology of wireless networks, where nodes correspond to APs and UEs, and edges represent communication links based on signal strength and interference levels. To process and analyze dynamic graphs data, *graph-neural-networks (GNNs)* provides a powerful framework. Inspired by CNNs, GNNs are designed to operate on graph structures, enabling tasks such as node classification, link prediction, and other graph-related learning problems.

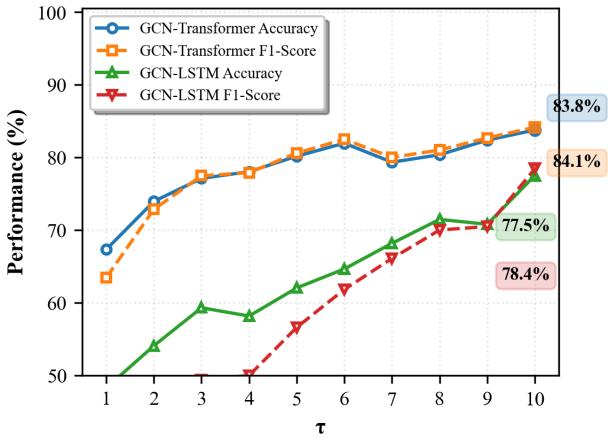


Figure 8.1: Accuracy and F1 score vs τ, for the fading scenario. Training performed with a dataset having $\tau = 10$.

8.2 Proposed Methodology

In this work, whose extended version is in Appendix I, we propose a novel framework to model cell-free massive MIMO communication, exploiting dynamic graphs to capture the time variability of the communication scenario. Then, we present a novel approach for jamming detection, leveraging dynamic graphs and GNN-based architectures. Our approach identifies jamming attacks by learning latent representations of network states and monitoring deviations from expected patterns. We evaluate the proposed method using simulations that model mobility, connectivity, and interference scenarios, demonstrating its effectiveness.

8.3 Numerical Results

Parameter τ represents the jammer activation frequency within each temporal sequence, where $\tau = 1$ indicates sporadic jamming (active for only 1 out of 10 timesteps), $\tau = 5$ represents moderate persistence (active for 5 out of 10 timesteps), and $\tau = 10$ denotes continuous jamming (active throughout the entire sequence).
The fading scenario, shown in Fig. 8.1, reveals the specialist's true generalization limitations and more pronounced performance variations. While maintaining strong overall performance (accuracy range: 67.2%-83.8%), the model shows increased sensitivity to jammer persistence patterns, however, a comparison has been done on the same dataset using the known Long Short Term Memory GCN (GCN-LSTM) [34] which combines the capabilities of LSTMs to extract temporal dependencies with the feature learning power of the GCN, and as the figure shows, our model performed better in all projected jamming behaviours. The performance progression from $\tau = 1$ (67.2% accuracy) to $\tau = 8$ (83.8% accuracy) demonstrates the model's adaptation to different temporal structures, with optimal detection occurring in the rhythmic jamming domain ($\tau = 6 - 8$).
More results are reported in Appendix I.

8.4 Integration with the Architecture

We have introduced a novel method to detect jamming attacks using graph neural networks in cell-free massive MIMO networks. We detect the jammer activity by monitoring the connectivity towards the base stations, represented as nodes of the graph. This approach is part of the PHY-Attack Identification module

within the proposed ROBUST-6G architecture, as it monitors the connectivity towards the 6G base-stations to detect potential jammers.

Chapter 9

Image-Based Frequency-Domain Analysis for Robust DDoS Detection

In the paper [35], EBY gives an overview of a frequency-domain imaging framework that is used for detecting DDoS attacks inside SDN networks. The main idea is that Packet-In traffic has many hidden patterns when we look at it in the spectral domain, so the authors transform the temporal signal into two-dimensional images by using full frequency information, including both magnitude and phase. This imaging approach allows deep models, especially CNN-based classifiers, to learn traffic behaviors in a more clear way. The study also reports numerical results, showing that the spectral images help the model separate normal traffic from different DDoS attack types with high accuracy, and that adding the phase information brings noticeable improvement compared to earlier works.

9.1 Background and Motivation

SDN centralizes the control plane, which concentrates traffic monitoring and decision making in a single controller. This architectural feature increases vulnerability, since adversaries can overwhelm the controller with spoofed Packet-In messages. Traditional anomaly detection either relies on simple thresholds or handcrafted statistical features that often fail when attackers vary traffic shapes or spread attacks across multiple spoofed addresses. Earlier work attempted to use frequency domain representations, but only used Fourier magnitude, which removed structural information related to phase alignment. This study is motivated by the need for a richer spectral representation that can capture small but consistent deviations in malicious traffic and provide a robust signature for learning based classification.

9.2 Proposed Methodology

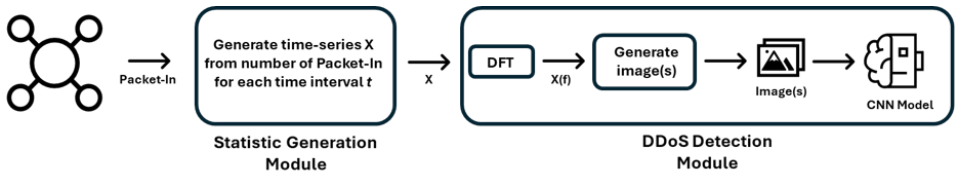
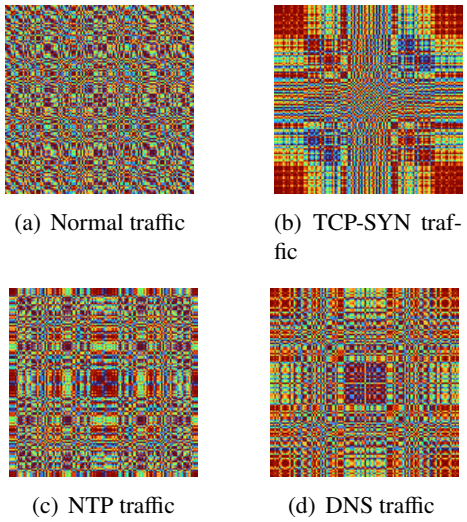


Figure 9.1: Proposed System.

The proposed detection framework begins by monitoring Packet-In messages received by the SDN controller



(a) Normal traffic (b) TCP-SYN traffic

(c) NTP traffic (d) DNS traffic

Figure 9.2: Images representing four traffic types: (a) Normal, (b) TCP-SYN, (c) NTP, (d) DNS.

and generating a time series of message counts. As illustrated in Figure 9.1, the system consists of two primary modules: (i) the *Statistic Generation Module*, which records Packet-In behavior and maintains a hash table of destination IP occurrences; and (ii) the *DDoS Detection Module*, which processes the time-series data using frequency-domain transformations. Discrete Fourier transform (DFT) is applied to sliding windows of the time series to obtain both magnitude and phase representations of frequency components. Unlike earlier work that used only magnitude values, the proposed method incorporates phase information to preserve the full spectral structure of the traffic.

To maintain semantic relationships between the complex-valued frequency components, a similarity matrix is constructed using the Hermitian product, forming a square image whose dimensions correspond to the window size. After normalization and logarithmic scaling, either the cosine or sine plane of this matrix serves as the final image representation. Multiple window sizes can be incorporated to capture both short- and long-range frequency patterns effectively. These images encode rich temporal–spectral characteristics, making them suitable for deep learning classification.

9.3 Results

Figure 9.2 demonstrates that each traffic type, including normal traffic, TCP-SYN floods, NTP amplification, and DNS reflection attacks, produces distinctive visual patterns. These patterns enable robust discrimination by a CNN trained on the generated images. The CNN architecture consists of convolutional, pooling, and fully connected layers, totaling more than 10 million parameters. Using real-world MAWI traces and emulated DDoS traffic, the proposed system achieves a true positive rate (TPR) of 99.96%, a false positive rate (FPR) of 0%, and an overall accuracy of 99.98%, significantly outperforming earlier magnitude-only frequency-domain approaches.

The authors acknowledge that the evaluation is performed within an emulated environment and that future work should consider scaling to broader attack types and real deployment conditions. However, the results demonstrate strong potential for integrating frequency-phase–based image representations into SDN security monitoring.

Contribution to 6G Physical-Layer Security

Although the study is positioned within the SDN domain, its methodology aligns closely with emerging requirements of 6G PLS. The use of magnitude–phase spectral analysis resembles physical layer anomaly detection techniques used for identifying jamming, spoofing, and replay behaviors. Converting spectral patterns into images enables RF fingerprinting, device authentication, and waveform-level threat classification, key components of 6G security architectures. The multi-window frequency representation mirrors how 6G systems will analyze non-stationary spectral behavior, while the modular structure of Figure 9.1 can be mapped directly into O-RAN near-RT RAN intelligent controller (RIC) security agents. Overall, the approach contributes a reusable methodological foundation for frequency-domain and image-based security mechanisms in 6G networks.

9.4 Integration with the Architecture

Although the proposed solution focuses on detecting DDoS attacks at the SDN controller, its core methodology, which converts time-series behavior into frequency domain images and captures both magnitude and phase characteristics, is highly relevant to physical layer attack detection (CEBY04). Many physical layer threats, such as jamming, spoofing, replay, and abnormal waveform injections, produce distinctive spectral signatures, irregular phase transitions, and non-stationary frequency patterns that cannot be reliably detected with simple time domain features. By constructing similarity matrices from the full spectral representation and training deep models to recognize image based spectral fingerprints, the method resembles how physical layer security systems analyze RF waveforms to identify malicious transmitters or unusual signal behavior. This spectral imaging approach naturally extends to 6G physical layer protection since future radios will rely heavily on detecting subtle changes in frequency domain structure to identify intelligent jammers, RIS enabled manipulation attempts, and adversarial waveform patterns, thereby making the proposed methodology a strong foundation for next generation physical layer attack detection.

Chapter 10

Radio Frequency Fingerprint-Based Classification Performance Analysis with ML Models in the Presence of Hardware Impairments

In the paper [36], EBY looks at how well RF fingerprinting can classify different devices when realistic hardware impairments are present, especially in the context of PLS. The idea behind RF fingerprinting is that every wireless device creates small and unique distortions because of manufacturing differences, such as RHI, CFO, and STO. These imperfections act like natural identifiers and can be used to check if a transmitter is genuine or possibly malicious. In the study, two OFDM-based devices are modeled with different impairment settings, and their received signals are processed to extract many physical-layer features, for example spectral flatness, skewness, kurtosis, autocorrelation, power spectral density (PSD), L-moments, mean, and variance. These features are then used to train several ML models that try to separate one device from the other.

10.1 Background and Motivation

The proliferation of IoT devices in next generation networks demands lightweight authentication methods that operate without relying entirely on cryptographic exchanges. RF fingerprinting offers such an approach, because each device has unique distortions that originate from manufacturing tolerances. These imperfections remain visible even when attackers transmit identical modulation formats. The study is motivated by the need to understand how impairment levels, spectral distortions, and temporal irregularities influence classification accuracy. Prior work did not systematically test multiple models under controlled impairment variations, nor did it analyze the impact of feature reduction on classification performance.

10.2 Proposed Methodology

This study simulates two OFDM based devices with different levels of residual transmit and receive impairments, including carrier frequency offset, symbol timing offset, and nonlinear transmitter distortions. Signals pass through Rayleigh fading channels with additive noise. Therefore, the signal model incorporates fading, additive Gaussian noise, CFO, STO, and RHI effects, enabling a realistic analysis of how impairments shape RF fingerprints. From the received signals, the authors extract multiple feature groups: PSD features

describing power distribution across frequency; spectral descriptors such as spectral flatness and bandwidth; statistical moments of the I/Q components; L-moments capturing distributional shape; autocorrelation-based parameters; amplitude–phase statistics; and a decision-boundary feature representing variance differences. These diverse feature sets are used to train ML models including logistic regression, naive Bayes, support vector machine (SVM), XGBoost, lightGBM, catBoost, and random forest. The dataset is split into an 80% training set and a 20% test set, and performance is evaluated in terms of accuracy, precision, recall, and F1-score.

10.3 Results

The results show that CatBoost consistently achieves the best classification performance, giving the highest accuracy and F1-score among all tested models. Logistic Regression also performs well in terms of precision, while naive Bayes remains weaker across several metrics. These differences can be clearly seen in Figure 10.1, where CatBoost provides the most stable and highest overall scores when all features are used.

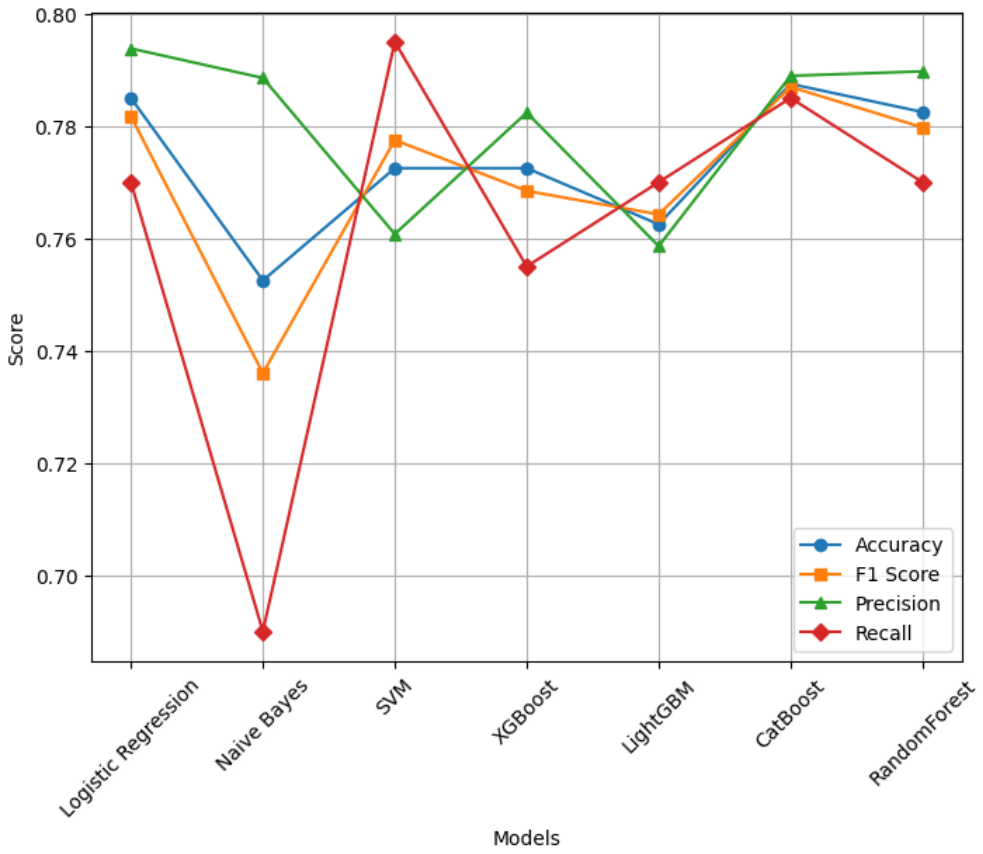


Figure 10.1: Performance score analysis of classification with all features.

To reduce computational load, the study applies recursive feature elimination (RFE) so that the models are trained with fewer but more important features. As shown in Figure 10.2, the performance after RFE remains almost the same, and CatBoost again provides the strongest accuracy and F1-score. This shows that the system can operate in a more lightweight way without losing classification quality.
Overall, the study demonstrates that hardware impairments generate measurable and discriminative RF fingerprints, and that ML-based classification stays robust even under noise, fading, and impairment variations.

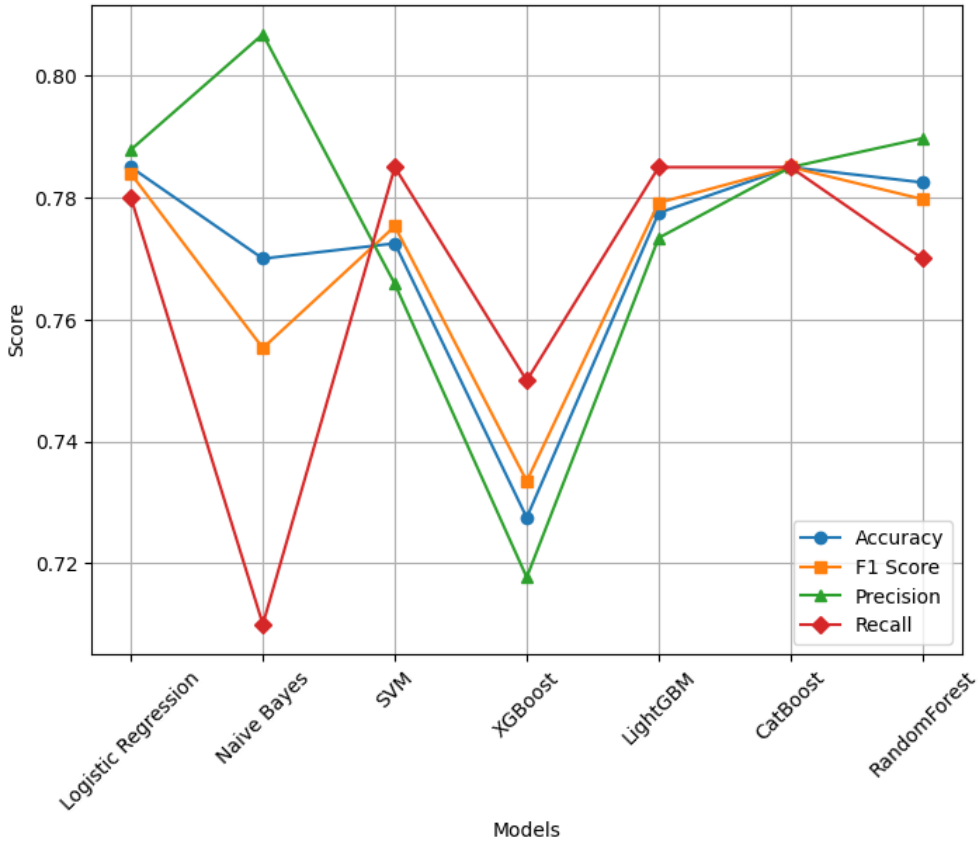


Figure 10.2: Performance score analysis of classification using selected features based on RFE.

The results also show that CatBoost is especially suitable for this task, both with full feature sets and with reduced ones after RFE.

Contribution to 6G Physical-Layer Security

The methodology directly contributes to future 6G physical-layer security by demonstrating how impairment-induced spectral and statistical signatures can be used for device authentication, rogue transmitter detection, and waveform-level trust mechanisms. As 6G moves toward zero-trust radio access, O-RAN security agents, and large-scale IoT deployments, the ability to classify devices using inherent RF fingerprints becomes essential. The findings show that even under fading and hardware distortions, RF fingerprint features remain stable and discriminative, enabling lightweight, real-time physical layer defenses. The feature groups used in this study, such as PSD variations, L-moments, skewness/kurtosis, and amplitude–phase statistics, closely align with signal-level intelligence envisioned for 6G networks, making the approach a strong candidate for scalable and secure 6G PLS frameworks.

10.4 Integration with the Architecture

This study is directly relevant to attack detection (CEBY04) because RF fingerprinting provides a powerful way to identify unauthorized or spoofed transmitters by exploiting natural hardware imperfections that cannot be easily imitated or cloned by an attacker. By analyzing how residual impairments, carrier frequency offset,

timing offset, and statistical irregularities appear in spectral and temporal features, the method allows ML models to distinguish legitimate devices from impersonators even when an adversary uses the same protocol, modulation, or transmission pattern. Since many physical layer attacks rely on transmitting forged signals that appear legitimate in the time domain, device specific hardware impairments become a reliable discriminator that reveals subtle inconsistencies in the attacker's waveform. The study shows that even under noise, fading, and varying impairment levels, these hardware induced signatures remain stable enough to support accurate classification, which directly translates to detecting rogue IoT nodes, cloned device identities, and sophisticated physical layer level spoofing attempts. This makes the proposed RF fingerprinting approach a highly relevant and practical foundation for next generation attack detection in future wireless and 6G environments.

Part II

T5.2 - Design of Resilient 6G PHY, Incorporating Physical Layer Security

Chapter 11

Secret Key Generation with Attestation and Physical Layer Fingerprinting

SKG is essential in wireless systems where links change quickly, and devices operate without fixed infrastructure. Existing methods either rely purely on cryptographic exchange (e.g., ECDH) or use physical-layer reciprocity, but neither verifies whether the participating devices are in a trusted software state. This leaves key establishment exposed to impersonation, software compromise, and channel manipulation. CHA in this work defines an Attested Secret-Key Generation (A-SKG) protocol that addresses these gaps. The protocol combines authenticated ECDH, certificate-based identity checks, and signed integrity measurements with physical-layer features (such as AoA, angle-of-departure (AoD), path loss). The final key is derived from both the DH secret and reconciled locally measured channel features, ensuring that only devices in a verified state and physically present on the link can derive it. The result is a minimal protocol that provides: (i) identity binding, (ii) integrity-verified participation, (iii) resistance to replay attacks, and (iv) channel-tied key derivation without transmitting raw measurements. This work specifies the system model, adversary model, and protocol functions used in the final A-SKG design.

11.1 Background and Motivation

SKG from physical-layer measurements is increasingly required in wireless systems where keys must reflect both device identity and channel conditions. Classical SKG mechanisms provide channel reciprocity but lack two properties needed in security-critical IoT deployments: (i) assurance that devices are running verified software, and (ii) cryptographic binding between physical-layer entropy and the authenticated key-exchange state.

Remote attestation protocols such as PROVE [37], SHeLA [38], SARA [39], and PADS [40] provide software-integrity evidence but operate independently of SKG and do not incorporate channel-specific entropy. Conversely, SKG protocols treat devices as inherently trustworthy and ignore adversaries capable of software compromise, key substitution, or replay. This separation leaves a gap: keys may be derived from valid channel measurements, but by an untrusted or compromised device. This work closes that gap by combining attestation with SKG in a single protocol run. We extend a standard attested key-exchange workflow with physical-layer features (AoA, AoD, and path loss) produced in section 2.2. Both parties derive the session key as:

$$K_{\text{A-SKG}} = \text{HKDF}\left(S_{\text{DH}} \parallel \Phi_{\text{PL}}, \text{salt} = N_N \parallel N_{AP} \parallel \text{transcript\_hash}\right)$$

The designed protocol ensures that no physical-layer data is exchanged; each device derives its own measurements locally and incorporates them into its key-generation state. The protocol provides (i) verified device

identity through certificate checks, (ii) integrity evidence via signed configuration hashes, (iii) freshness through nonces and transcript hashing, and (iv) a session key that is tied to link-specific physical-layer characteristics. Key confirmation is enforced through AEAD. By coupling attestation with SKG, the protocol reduces the time-of-check-to-time-of-use gap [41]: the device that proves its software integrity is the same device that derives the session key, and the derivation is bound to both its attested state and the handshake transcript. Only devices that are authenticated, verified, and physically present on the link can compute the final key.

11.2 State of the Art

11.2.1 Remote Attestation for IoT

Remote attestation verifies the integrity of software running on untrusted devices. Traditional attestation follows a challenge-response pattern where a trusted Verifier sends a nonce to an untrusted Prover, which computes a measurement (hash) of its firmware and returns a signed or MAC'd response.
Software-based attestation relies on timing constraints to detect compromised code, but requires strict network assumptions unsuitable for wireless IoT. Hardware-based attestation uses Trusted Execution Environment (TEE)s like ARM TrustZone [42] or TPM [43], providing stronger security at a higher cost. Hybrid attestation schemes like SMART [44], TrustLite [45], and TyTAN [46] use minimal hardware (e.g., Read-Only Memory (ROM) and Memory Protection Units (MPU)) to protect attestation code while remaining cost-effective.

11.2.2 Collective Attestation Protocols

SEDA [47] performs attestation over spanning tree topologies, with each device attesting its children and aggregating results upward. SEDA requires network stability and full connectivity during attestation. LISA [48] improves SEDA by supporting asynchronous attestation propagation without strict parent-child relationships. PADS [40] addresses highly dynamic topologies by using consensus algorithms rather than spanning trees. Specifically, devices self-attest at random intervals and broadcast results to neighbors, eventually converging to network-wide knowledge. SARA [39] attests distributed IoT services using publish-subscribe communication with vector clocks to track causality among asynchronously interacting services.

11.2.3 Physical Layer Secret Key Generation

Physical-layer SKG leverages the inherent randomness and reciprocity of wireless communication channels to enable two legitimate devices to extract shared cryptographic keys directly from the radio environment. To analyze the achievable secret key rate (SKR) under practical impairments, such as channel estimation and quantization errors, a mutual information neural estimator (MINE) [49] is employed in [50]. In [51], the authors propose a mutual information-driven autoencoder (MIAE) that learns reciprocal channel features while directly optimizing the SKR through end-to-end training. Alternatively, [52] introduces an approach that uses a predefined key and exploits obfuscation matrices with CSI as the carrier to aid secure key transmission.

11.2.4 Research Gap

Existing attestation protocols assume pre-established keys or computationally expensive per-device public-key operations [53, 54]. SKG protocols focus on key generation without addressing device integrity. To the best of our knowledge, no prior work integrates remote attestation with physical-layer key generation in a single protocol that jointly verifies device identity, software state, and physical proximity. Recent data-driven

methods for extracting channel features have improved SKG robustness, but they operate independently of attestation and do not provide end-to-end trust guarantees.

11.3 Proposed Methodology

11.3.1 System Model

We consider a wireless setting where a resource-constrained device (i.e., node N) establishes a secure session with an AP. Both entities hold long-term credentials issued by a trusted Certificate Authority (CA), and no pre-shared secrets are assumed. Trust is established through certificate verification and a cryptographic handshake.
A key requirement of this setting is that session establishment must include integrity evidence. The node and the access point must each supply a verifiable measurement of their current software state before a session key is derived. In addition, the final key must depend on both the exchanged cryptographic material and the wireless channel, ensuring that only devices at the same physical location can compute it.
The system model, therefore, consists of three components: the node and access point as communicating devices, and the certificate authority as the identity root. Each entity operates within a defined trust boundary that governs how identity, integrity, and channel information contribute to the resulting secure session.

Entities **Node (N).** A resource-constrained wireless device requiring authenticated network access. Nodes are provisioned with long-term cryptographic credentials, hardware-specific identifiers, and trusted execution capabilities for attestation and key derivation operations.
AP. Infrastructure device providing wireless connectivity and authentication services. APs maintain long-term credentials, hardware identifiers, and trusted execution environments for protocol operations. APs also maintain connectivity to backend services for policy enforcement and audit logging.
CA. Trusted entity responsible for issuing and managing device certificates that bind public keys.

11.3.2 Communication Model

The model consists of two legitimate entities, Alice and Bob (for convenience, we refer to the node and AP as Alice and Bob in the SKG communication model), operating in time division duplex (TDD) mode and aiming to establish secure communication. The objective is to generate shared secret keys by leveraging the reciprocity of the wireless channel. We assume the presence of a passive eavesdropper, Eve, who remains silent and is located more than half a wavelength away from both legitimate nodes. Under this assumption, the wireless channels from Alice and Bob to Eve are considered uncorrelated with the legitimate reciprocal channel, ensuring that Eve gains negligible information from the exchanged signals. The overall SKG pipeline includes channel probing, feature extraction, quantization, information reconciliation, and privacy amplification.

11.3.3 Trusted Execution Environment

Both Node and AP include a TEE providing isolation and attestation capabilities. The TEE maintains:

1. Isolated memory and code execution context protected from application software and device drivers.
2. Secure key storage with hardware-enforced access controls preventing unauthorized extraction.
3. Monotonic counter or secure clock for freshness verification and replay prevention.

4. Attestation code stored in read-only memory, protected from firmware modification attacks.

5. Key zeroization routines that securely erase ephemeral material immediately after use.

11.3.4 Protocol Overview

The protocol establishes a session key between a Node and an Access Point through three concrete operations: authenticated key agreement, integrity verification, and physical-layer binding. These operations occur in two phases.

Phase 1 (Functions 1–3) performs mutual authentication and integrity assessment. Node and AP exchange certificates and validate identities against the trusted CA. Both parties run ECDH to derive a shared secret S_{DH} and bind the ephemeral key exchange to long-term identities through digital signatures. Nonces N_N and N_{AP} provide freshness. Each device computes and signs a golden hash of its current firmware and configuration state. Signature verification confirms that both parties hold valid certificates and that no device has been compromised. This phase establishes an authenticated AEAD channel K_E for subsequent communication.

Phase 2 (Functions 4–5) derives the final session key using physical-layer features. Each device obtains its locally measured channel characteristics, such as angles, path loss, from the section 2.2 system. These measurements are quantized and locally processed without transmission. Both parties derive the session key:

$$K_{\text{A-SKG}} = \text{HKDF}\left(S_{DH} \parallel \Phi_{\text{PL}}, \text{salt} = N_N \parallel N_{AP} \parallel \text{transcript\_hash}\right)$$

11.3.5 Adversary Model

We assume a powerful adversary controlling the entire wireless network and capable of interacting arbitrarily with all protocol messages. The attacker can delay, drop, reorder, inject, replay, and modify packets. The attacker may compromise device software but cannot break trusted hardware, extract keys from secure storage, or modify ROM or TEE code. In line with the state-of-the-art swarm attestation protocols [53, 54] we assume the following adversarial capabilities:

Software Compromise ($\mathbf{Adv_{SW}}$). The adversary gains full control over non-TEE software on a device. This includes modifying application code, altering configuration, and accessing all non-protected memory. However, Adv_{SW} cannot:

1. Access private signing keys stored in secure storage

2. Modify ROM-resident reference values

3. Tamper with TEE-executed attestation routines

4. Forge signatures without the corresponding private key

Ephemeral-Covering Malware ($\mathbf{Adv_{MSW}}$). The adversary may restore the system to a clean state before attestation executes to hide prior modifications. Adv_{MSW} cannot forge nonce-bound attestation outputs or reconstruct valid measurement signatures. This adversary is standard in RA literature to capture self-erasing malware behavior.

Passive Physical-Layer Observer ($\mathbf{Adv_{PNI}}$). The adversary can observe all wireless transmissions but cannot infer local channel measurements that are not transmitted (AoA, AoD, path loss). AdvPNI cannot force two honest devices to observe matching physical-layer features without physically occupying the link. AEAD prevents the extraction of any attestation or key-agreement material from observed ciphertext.

Physical Invasive Attacker ($\mathbf{Adv_{PI}}$). The attacker obtains direct physical access to hardware, including bus probing or TEE key extraction. Such attacks are out of scope unless devices deploy tamper-resistant hardware.

11.3.6 Out-of-Scope Threats

We exclude: (i) compromise of the Certificate Authority or measurement subsystem, (ii) full network-wide denial-of-service attacks, (iii) attacks requiring quantum computational capabilities, and (iv) undetectable supply-chain compromise.

11.4 Protocol Description

We denote protocol participants, cryptographic values, physical-layer measurements, and protocol state using the following notation, organized by functional category.

11.4.1 Notation

Entities	
N	Node
AP	Access Point
Long-Term Credentials	
$(Priv_X, Pub_X)$	Long-term key pair for entity X
$Cert_X$	Certificate containing Pub_X signed by CA
Pub_{CA}	CA public key (trusted, provisioned at manufacturing)
Ephemeral Keys	
$(Priv_{DH_X}, Pub_{DH_X})$	Ephemeral key pair for entity X
S_{DH}	Shared ECDH secret
K_E	Handshake AEAD key derived from S_{DH}
$K_{\text{A-SKG}}$	Final session key
Freshness Values	
N_N	Node nonce (256 bits, fresh per session)
N_{AP}	AP nonce (256 bits, fresh per session)
Physical-Layer Measurements	
H_{N_A}	Node's channel state information from T5.1
H_{AP}	AP's channel state information from T5.1
V_N	Node's extracted physical-layer feature vector
V_{AP}	AP's extracted physical-layer feature vector
C_N	Node's commitment to V_N: $C_N = H(V_N \parallel N_N \parallel N_{AP})$
C_{AP}	AP's commitment to V_{AP}: $C_{AP} = H(V_{AP} \parallel N_N \parallel N_{AP})$
Φ_{PL}	Quantized physical-layer feature vector $= \{AoA, AoD, PL\}$
Protocol State	
transcript_hash	Cumulative hash of all protocol messages to current point
GH_N	Golden hash: $GH_N = H(HWID_N \parallel FW_N \parallel \text{config}_N)$
GH_{AP}	Golden hash: $GH_{AP} = H(HWID_{AP} \parallel FW_{AP} \parallel \text{config}_{AP})$
Timestamp	Monotonic counter or secure clock value for temporal ordering

11.4.2 Function 1: Initial Connection Request

Function 1 establishes the initial communication context by exchanging identities and fresh nonces. This step provides unilateral AP authentication through certificate disclosure and ensures freshness for all subsequent protocol messages. The nonces and initial transcript hash serve as binding values for later DH parameters, attestation evidence, and physical-layer measurements, anchoring them to a single protocol instance.

Protocol Steps

1. **Node → AP: Initial Nonce and Identity**

$$N_N \leftarrow \text{random}(256 \text{ bits})$$
$$M_1 = \{ID_N, N_N\}$$

   Node sends M_1 to AP.

2. **AP → Node: AP Identity, Certificate, and Fresh Nonce**

$$N_{AP} \leftarrow \text{random}(256 \text{ bits})$$
$$M_2 = \{ID_{AP}, N_{AP}, Cert_{AP}\}$$

   AP sends M_2 to Node.

3. **Both Parties: Transcript Initialization**

$$\text{transcript\_hash} \leftarrow H(M_1 \parallel M_2)$$

Security Properties

- **Freshness.** Random nonces N_N and N_{AP} ensure each protocol instance is unique and resistant to replay attacks across sessions.
- **Early AP Authentication.** The AP provides $Cert_{AP}$ in M_2, enabling the Node to validate AP identity before committing computational resources to key agreement operations.
- **Transcript Binding.** The initial transcript hash provides a cryptographic commitment to protocol messages. All subsequent messages extend this hash, preventing message interleaving or replay across different sessions.

11.4.3 Function 2: Certificate Exchange and Diffie–Hellman Establishment

Function 2 performs authenticated key agreement: each device verifies the peer's certificate, exchanges ephemeral DH public keys, and derives a handshake encryption key. Identity, nonces, and DH parameters are cryptographically bound into the transcript.

Protocol Steps

1. **Node → AP: Certificate and Ephemeral DH Public Key**
   (a) Verify $Cert_{AP}$ from M_2 against Pub_{CA}. Abort on verification failure.
   (b) Generate ephemeral key pair: $(Priv_{DH,N}, Pub_{DH,N}) \leftarrow \text{KeyGen}_{\text{ECDH}}()$
   (c) Construct: $M_3 = \{Cert_N, Pub_{DH,N}\}$

(d) Send M_3 to AP.

2. **AP → Node: Ephemeral DH Public Key**

   (a) Verify $Cert_N$ from M_3 against Pub_{CA}. Abort on verification failure.

   (b) Generate ephemeral key pair: $(Priv_{DH,AP}, Pub_{DH,AP}) \leftarrow \text{KeyGen}_{\text{ECDH}}()$

   (c) Construct: $M_4 = \{Pub_{DH,AP}\}$

   (d) Send M_4 to Node.

3. **Both Parties: ECDH and Handshake Key Derivation**

   (a) Compute shared secret:

$$S_{DH} = \text{ECDH}(Priv_{DH,\text{local}}, Pub_{DH,\text{remote}})$$

   (b) Construct DH-binding payload and sign:

$$\text{tosign} = (Pub_{DH,N} \parallel Pub_{DH,AP} \parallel ID_N \parallel ID_{AP} \parallel N_N \parallel N_{AP})$$

$$Sig_{DH,X} = \text{Sign}_{Priv_X}(H(\text{tosign}))$$

   (c) Derive handshake AEAD key:

$$K_E = \text{HKDF}\,(S_{DH},\ \text{salt} = N_N \parallel N_{AP},\ \text{info} = H(M_1 \parallel M_2 \parallel M_3 \parallel M_4))$$

   (d) Exchange AEAD-encrypted DH signatures:

$$M_5 = \text{AEAD}_{K_E}\left(\text{AAD} : H(M_1 \parallel \cdots \parallel M_4),\ \text{plaintext} : Sig_{DH,N}\right)$$

$$M_6 = \text{AEAD}_{K_E}\left(\text{AAD} : H(M_1 \parallel \cdots \parallel M_4),\ \text{plaintext} : Sig_{DH,AP}\right)$$

   (e) Verify received signatures. Abort on verification failure.

   (f) Update transcript: $\text{transcript\_hash} \leftarrow H(\text{transcript\_hash} \parallel M_3 \parallel M_4 \parallel M_5 \parallel M_6)$

Security Properties

- **DH Binding to Identity.** DH public keys are signed together with long-term identities and fresh nonces, preventing key substitution.
- **Handshake Confidentiality.** Signatures are encrypted under K_E derived from the shared secret, preventing eavesdroppers from observing DH bindings.
- **Early Verification.** Signature verification failures trigger immediate abort before further computation, preventing DH mismatch attacks.
- **Transcript Integrity.** All messages are incorporated into the transcript hash, ensuring that modification of any message invalidates all downstream cryptographic operations.

11.4.4 Function 3: Integrity Evidence Exchange

Function 3 provides mutual software-integrity validation. Each device measures its firmware and configuration, signs the result under nonce and transcript context, and transmits the attestation evidence under the handshake AEAD key.

Protocol Steps

1. **Node → AP: Configuration Hash and Attestation Signature**
   (a) Compute configuration hash:
   $$H'_N = H(FW_N \parallel \text{config}_N \parallel HWID_N)$$
   (b) Construct attestation signature binding to nonce and transcript:
   $$Sig_{H_N} = \text{Sign}_{Priv_N}\left(H(H'_N \parallel N_N \parallel N_{AP} \parallel \text{transcript\_hash})\right)$$
   (c) Send:
   $$M_7 = \text{AEAD}_{K_E}\left(\text{AAD} : \text{transcript\_hash},\ \text{plaintext} : (H'_N, Sig_{H_N})\right)$$
2. **AP → Node: Configuration Hash and Attestation Signature**
   (a) Decrypt and verify M_7: $(H'_N, Sig_{H_N}) \leftarrow \text{AEAD}^{-1}_{K_E}(M_7)$. Abort on verification failure.
   (b) Compare H'_N against stored golden hash GH_N. Abort on mismatch (device not in expected state).
   (c) Compute own configuration hash:
   $$H'_{AP} = H(FW_{AP} \parallel \text{config}_{AP} \parallel HWID_{AP})$$
   (d) Construct attestation signature:
   $$Sig_{H_{AP}} = \text{Sign}_{Priv_{AP}}\left(H(H'_{AP} \parallel N_N \parallel N_{AP} \parallel \text{transcript\_hash})\right)$$
   (e) Send:
   $$M_8 = \text{AEAD}_{K_E}\left(\text{AAD} : \text{transcript\_hash},\ \text{plaintext} : (H'_{AP}, Sig_{H_{AP}})\right)$$
3. **Node: Verification**
   (a) Decrypt and verify M_8. Abort on failure.
   (b) Update transcript: transcript_hash $\leftarrow H(\text{transcript\_hash} \parallel M_7 \parallel M_8)$

Security Properties

- **Nonce-Bound Attestation.** Attestation signatures include fresh nonces from the current session, preventing replay of measurements from prior sessions.
- **TOCTTOU Prevention.** Adv_{MSW} cannot restore firmware after attestation because the signature includes a transcript hash computed at attestation time; retroactively forging a valid signature requires the private key.
- **Authenticated Evidence.** Digital signatures prevent forgery of integrity measurements. Only the device holding the private key can produce a valid attestation.
- **Policy Integration.** Golden-hash comparison is performed locally; protocol does not mandate policy. Backend services enforce trust decisions based on comparison results.

11.4.5 Function 4: Attestation Token Generation

Function 4 generates signed attestation tokens binding device identity, integrity state, and physical-layer commitments. Tokens provide non-repudiable evidence for post-protocol audit.

Protocol Steps

1. **Node → AP**
   (a) Fetch physical-layer features: $H_{N_A} = \{AoA, AoD, PL\}_{T5.1}$
   (b) Extract and quantize: $V_N = \text{ExtractFeatures}(H_{N_A})$
   (c) Commit to features: $C_N = H(V_N \parallel N_N \parallel N_{AP})$
   (d) Construct token: $\tau_N = (ID_N, H(Cert_N), H'_N, C_N, \text{Timestamp})$
   (e) Sign: $Sig_{\tau_N} = \text{Sign}_{Priv_N}(H(\tau_N))$
   (f) Send: $M_9 = \text{AEAD}_{K_E}(\tau_N, Sig_{\tau_N})$
2. **AP → Node**
   (a) Verify M_9. Abort on failure.
   (b) Fetch and process own features: $H_{AP} = \{AoA, AoD, PL\}_{T5.1}$
   (c) Extract, quantize, and commit: V_{AP}, C_{AP}
   (d) Construct token: $\tau_{AP} = (ID_{AP}, H(Cert_{AP}), H'_{AP}, C_{AP}, \text{Timestamp})$
   (e) Sign: $Sig_{\tau_{AP}} = \text{Sign}_{Priv_{AP}}(H(\tau_{AP}))$
   (f) Send: $M_{10} = \text{AEAD}_{K_E}(\tau_{AP}, Sig_{\tau_{AP}})$
3. **Node**
   (a) Verify M_{10}. Abort on failure.
   (b) Update transcript: $\text{transcript\_hash} \leftarrow H(\text{transcript\_hash} \parallel M_9 \parallel M_{10})$

Security Properties

- **Feature Privacy.** Measurements never transmitted; only commitments exchanged.
- **Measurement Binding.** Commitments bind devices to local features, preventing adaptive attacks.
- **Non-Repudiation.** Signed tokens provide audit evidence tied to device keys.
- **Session Isolation.** Tokens include timestamp and transcript hash, unique per session.

11.4.6 Function 5: Secret Key Derivation and Confirmation

Function 5 derives the final session key from the DH shared secret and physical-layer measurements. Information reconciliation ensures identical key derivation. Bidirectional key confirmation verifies successful synchronization.

Protocol Steps

1. **Node: Prepare Reconciliation**
   (a) Quantize measurements: $\text{bits}_N = \text{Quantize}(H_{N_A})$

(b) Generate helper data: $(\text{sketch}, \text{parity}) = \text{Cascade.Gen}(\text{bits}_N)$

(c) Send: $M_{11} = \text{AEAD}_{K_E}(\text{sketch}, \text{parity})$

2. **AP: Reconcile and Derive Key**

(a) Quantize own measurements: $\text{bits}_{AP} = \text{Quantize}(H_{AP})$

(b) Reconcile: $\text{bits}' = \text{Cascade.Rec}(\text{bits}_{AP}, \text{sketch}, \text{parity})$

(c) Abort if reconciliation fails (spatial mismatch).

(d) Derive key: $K_{\text{A-SKG}} = \text{HKDF}(S_{DH} \parallel \text{bits}',\ \text{salt} = N_N \parallel N_{AP} \parallel \text{transcript\_hash})$

(e) Send confirmation: $M_{12} = \text{AEAD}_{K_{\text{A-SKG}}}(\text{“AP-confirm”})$

3. **Node: Derive and Confirm**

(a) Derive key: $K_{\text{A-SKG}} = \text{HKDF}(S_{DH} \parallel \text{bits}_N,\ \text{salt} = N_N \parallel N_{AP} \parallel \text{transcript\_hash})$

(b) Verify M_{12}. Abort on failure.

(c) Send confirmation: $M_{13} = \text{AEAD}_{K_{\text{A-SKG}}}(\text{“Node-confirm”})$

4. **AP: Verify and Zeroize**

(a) Verify M_{13}. Abort on failure.

(b) Zeroize: $\text{Zeroize}(Priv_{DH}, S_{DH}, K_E, \text{bits}_N, \text{bits}_{AP}, \text{sketch}, \text{parity})$

(c) Update transcript: $\text{transcript\_hash} \leftarrow H(\text{transcript\_hash} \parallel M_{11} \parallel M_{12} \parallel M_{13})$

Security Properties

- **Mutual Confirmation.** Bidirectional key confirmation ensures both parties derive identical keys.
- **Forward Secrecy.** Keys depend on ephemeral DH, not long-term credentials.
- **Ephemeral Secrecy.** All temporary material is zeroized immediately after key derivation.

11.5 Numerical Results and Analysis

11.5.1 Proof-of-Concept Implementation

The goal of the hardware PoC is to implement the Attested A-SKG protocol described in this chapter on real hardware and measure its execution time on both ends of the link. The PoC focuses on:

1. Implementing the full attested key-exchange path (Functions 1–3 and key derivation without real physical-layer entropy).
2. Measuring per-step and end-to-end latency on a constrained device and a less constrained host.

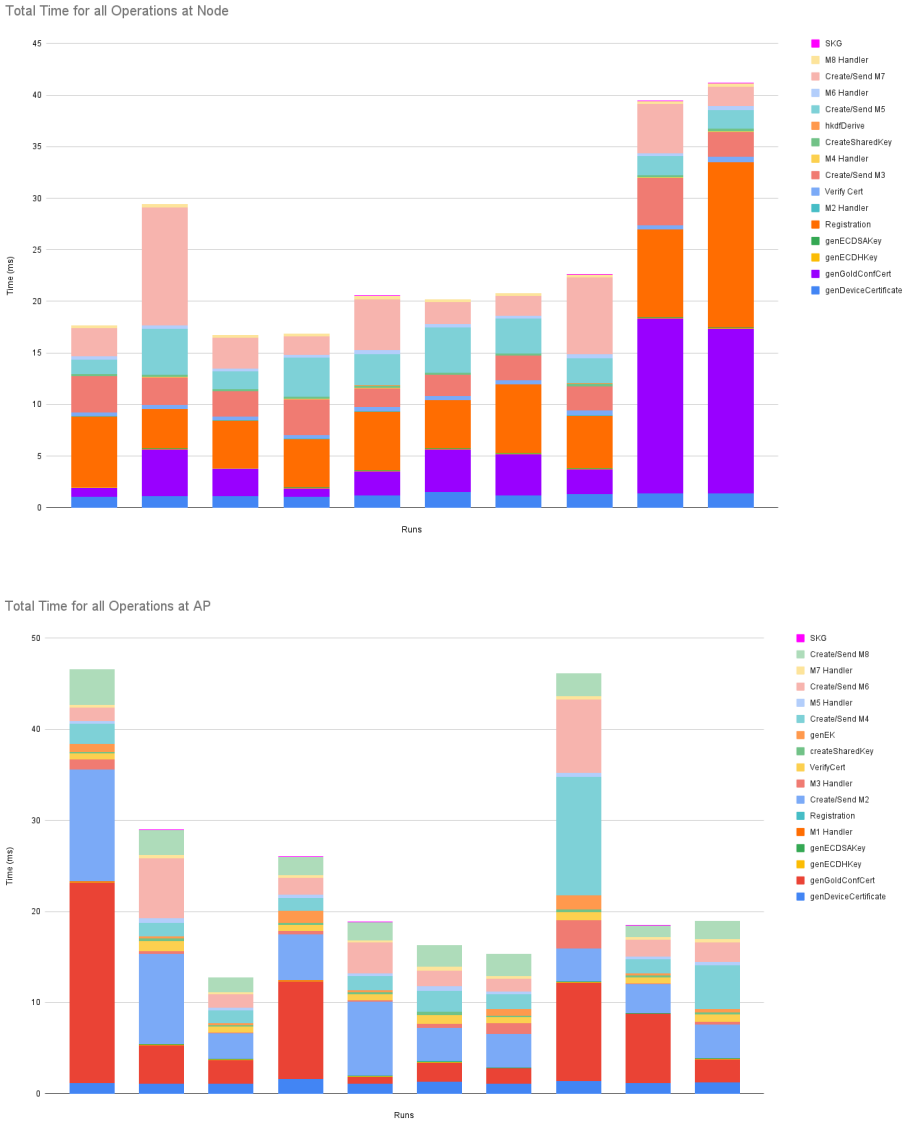


Figure 11.1: Runtime for A-SKG generation for the Node and the Access Point.

Access Point AP. The AP role is implemented on a Raspberry Pi 5 with 8 GB RAM, running a standard Linux distribution. All timing data reported for the AP side is collected on this device.
Node N. The node role is implemented as a separate process running on a local workstation (non-measured side). For this PoC, end-to-end timing on the node is measured separately from the AP runs, but both devices execute the same protocol logic for their respective roles.
Results. As shown in Figure 11.1, without network delays, the time from the initial handshake to completion of the A-SKG procedure is, on average, 24.56 ms for the Node and 24.86 ms for the AP. The median execution times across 10 runs are 20.70 ms and 18.91 ms, respectively. Within this total duration, the SKG generation phase itself contributes only a very small portion of the cost: on average, 21.80 µs for the Node and 31.51 µs for the AP, with corresponding median values of 21.89 µs and 23.92 µs. These results indicate that the cryptographic and attestation operations dominate runtime, while the SKG component adds negligible latency.

11.5.2 Security Analysis

Our protocol is designed to provide mutual authentication, software-integrity verification, and channel-dependent key generation in the presence of a strong network adversary. The analysis below follows the adversary model defined in 11.3.5.

11.5.3 Authentication and MITM Resistance

Mutual authentication is achieved through the certificate validation and signed DH bindings in Function 2. Each party signs a hash of $(Pub_{DH,N} \parallel Pub_{DH,AP} \parallel ID_N \parallel ID_{AP} \parallel N_N \parallel N_{AP})$ with its long-term key. An active adversary attempting MITM must either (i) forge a valid signature under Pub_N or Pub_{AP}, or (ii) obtain a CA-signed certificate for a dishonest key. Both contradict the EUF-CMA security of the signature scheme and the trust in the CA. Since the DH public keys are included in the signed data, key-substitution attacks are also prevented.

11.5.4 Integrity Verification and TOCTTOU

Integrity evidence in Function 3 is computed inside the TEE as $H'_X = H(HWID_X \parallel FW_X \parallel config_X)$ and compared against a golden hash. Attestation signatures include H'_X together with $(N_N, N_{AP}, \text{transcript\_hash})$. A software adversary changing firmware must either produce a collision in the hash (against collision resistance) or forge a valid signature on an inconsistent tuple. An ephemeral-covering adversary (Adv_{MSW}) cannot roll back to a clean state after using malicious code and still generate a nonce- and transcript-bound attestation matching a previous configuration. This reduces TOCTTOU attacks where the measured state differs from the state used during key establishment.

11.5.5 Key Secrecy and Forward Secrecy

The session key is derived as

$$K_{\text{A-SKG}} = \text{HKDF}\big(S_{DH} \parallel \Phi_{\text{PL}},\ \text{salt} = N_N \parallel N_{AP} \parallel \text{transcript\_hash}\big).$$

Under the Diffie–Hellman assumption, S_{DH} is pseudorandom given the public DH values; Φ_{PL} is never transmitted and remains unknown to the network adversary. Modeling HKDF as a PRF, $K_{\text{A-SKG}}$ is computationally indistinguishable from random for an adversary who does not know S_{DH} or Φ_{PL}. Since S_{DH}

is computed from ephemeral keys and these keys are zeroized after use, compromise of long-term signing keys does not reveal past session keys (forward secrecy). AEAD protects all encrypted messages; observing ciphertext does not leak information about $K_{\text{A-SKG}}$ beyond what is implied by the security of the primitives.

11.5.6 Replay, Transcript Manipulation, and Channel Binding

Fresh nonces (N_N, N_{AP}) are generated per session and incorporated, together with all messages, in the cumulative transcript_hash. Any replay or reordering of M_1–M_8 changes the transcript context and causes verification failure of subsequent signatures or AEAD tags. This prevents replay of old runs and interleaving of transcripts across sessions.
Channel binding is achieved by including Φ_{PL} in the key derivation while never transmitting raw physical-layer measurements. A passive observer (Adv_{PNI}) learns neither Φ_{PL} nor S_{DH} and therefore cannot reconstruct $K_{\text{A-SKG}}$. An off-path attacker at a different location cannot force two honest devices to obtain consistent AoA/AoD/PL values with non-negligible probability; any mismatch leads to divergent keys and failure in the AEAD-based key-confirmation step. Thus, only devices that are mutually authenticated, in a verified software state, and physically present on the same link can successfully derive and confirm the session key.
Unlike conventional authenticated key-establishment protocols, our protocol incorporates software-integrity evidence into the key-derivation phase. This ensures that only devices in a verified state participate in the exchange of ephemeral Diffie–Hellman values and subsequent key computation. Incorporation of physical-layer features provides an additional binding to the wireless link, giving the resulting key a dependence on channel conditions rather than solely on algorithmic secrets. Ephemeral DH keys and the erasure of temporary material support forward secrecy in line with standard attestation practices.
Limitations. The protocol relies on elliptic-curve Diffie–Hellman and therefore does not offer post-quantum security; replacing this step with a lattice-based KEM would address this limitation. The approach assumes trustworthy and stable physical-layer measurements from the 2.2, which may not hold under strong radio-frequency interference or adversarial manipulation of the environment. Finally, high mobility or severe channel variability may reduce measurement correlation and impact key reconciliation reliability.

11.6 Conclusion

This work presented a point-to-point A-SKG protocol that combines authenticated key exchange, software-integrity verification, and channel-dependent key derivation for wireless IoT settings. Physical-layer features 2.2 are incorporated directly into the key-derivation phase, removing the need for separate pilot-exchange steps and keeping the protocol compact. The system model and adversary assumptions follow established remote-attestation practice, and the protocol functions reflect the standard structure used in prior work: freshness and identity establishment, authenticated Diffie–Hellman exchange, integrity measurement, and session-key derivation. The final key is derived from both the ephemeral DH secret and locally observed physical-layer features, ensuring that only devices in a verified state and sharing the same wireless link complete the protocol. The security analysis shows that the protocol provides authentication, integrity validation, replay protection, and channel binding under the stated assumptions.

11.7 Integration with the Architecture.

We developed an A-SKG mechanism that binds physical-layer properties of the devices to the established session key. This approach ensures that only authenticated and integrity-verified devices that share the same physical channel can derive the key, thereby simultaneously establishing trust and enabling secure

communication. Within the ROBUST-6G architecture, this functionality maps to the Physical-Layer Trustworthiness and Analysis module, as it leverages channel characteristics to strengthen device authentication and session-key derivation.

Chapter 12

Fast and Robust Secret Key Generation

This chapter presents our contributions to the design of fast key agreement protocols, suitable for delay-constrained and real-time applications, and providing measurable security guarantees.

12.1 Background and Motivation

Across the three papers [5–7], we generated research outputs on wireless SKG, moving from low TRL works on the communication-theoretic modeling of LoS multipath channels, to a comprehensive study of SKG design parameters under worst-case eavesdropping attacks, and, finally to a context-aware, real-time SKG demonstrator on software-defined radios (SDRs). In these works, we delivered fast and lightweight, quantum-resilient SKG, with applications to 6G and IoT settings, well suited for low-end devices. We placed a strong emphasis on ensuring rigorous security guarantees (via conditional mutual information and conditional min-entropy estimators) and on the practical feasibility and real-time operation employing experimental measurement campaigns and demonstrators.
The SKG protocol followed the standard phases of i) randomness distillation and quantization, ii) reconciliation, and iii) privacy amplification. As shown in Fig. 12.1 we considered two legitimate users, referred to as Alice and Bob, exchanging complex pilots in a time division duplex (TDD) mode, while a passive eavesdropper, referred to as Eve, intercepted all exchanged messages. We explicitly accounted for dependencies in the legitimate and adversarial wireless links, moving beyond the idealized assumption of spatial decorrelation at distances of half-wavelength (using Jakes model). The key length $|\mathbf{k}|$ was upper bounded by the conditional min-entropy[1]

$$|\mathbf{k}| \leq H_\infty(\mathbf{r}_A|\mathbf{r}_E, \mathbf{s}_A), \tag{12.1}$$

where $\mathbf{r}_m, m \in \{A, B, C\}$ denoted the reconciliation input vectors at Alice, Bob and Eve, respectively. Key contributions of these works included:

- In [5] it was shown that LoS multipath channels can support non-trivial SKG rates when bandwidth (BW), delay spread (DS) and multipath resolvability are properly accounted for.
- In [6], a comprehensive design analysis was provided in realistic LoS / NLoS and dynamic / static settings, under "on-the-shoulder eavesdropping attacks"[2] (eavesdropper in very close proximity to one of the legitimate nodes). We analyzed how choices in sampling, quantization and code rate traded off

[1] While in pseudorandom number generators the min-entropy is used to evaluate randomness of generated sequences, in SKG we need to account for Eve's observations over the wireless links through the conditional min entropy.

[2] The term "on-the-shoulder-attack" was first coined in [55] to describe eavesdropping attacks at distances of a few wavelengths from a legitimate node, as a worst case scenario.

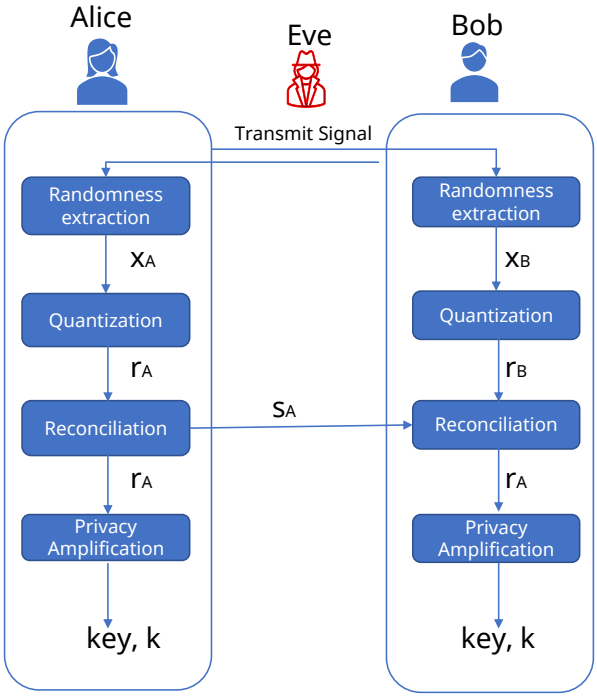


Figure 12.1: SKG protocol

with leakage and impacted the achievable key rates. **We reached reconciliation rates above** 99%**, which exceeds the WP5 stated KPI of** 90%. Furthermore, **we developed a ML based estimator of conditional min entropy with inference times less than** 0.2 **msec**.

- In [7], we participated with algorithmic contributions in a **real-time (run-time** $<$ 1 **msec), context-aware SKG demonstrator, aligned with the WP5 stated KPI of authentication and key agreement (AKA) of 5 msec**. the demonstrator won the best demonstrator award in IEEE CNCS 2025, further showcasing the impact of ROBUST-6G outputs.

Across all works, we kept an information-theoretic perspective on security while addressing real-world constraints such as limited complexity, static channels, and on-the-shoulder eavesdroppers. The demonstrator served as PoC for SKG under rigorous leakage guarantees and context-awareness; context-aware fast and robust SKG is not just a theoretical possibility, but a feasible, alternative for key generation and distribution in future 6G and IoT deployments.

12.2 Secret Key Generation Rates in LoS Multipath Channels

12.2.1 Proposed Methodology

In [5], we studied the feasibility of SKG when the received signal strength (RSS) is used as the source of shared randomness over a frequency-selective LoS multipath channel in the presence of a passive eavesdropper (this paper extends earlier results for the NLoS case). To this end, we derived the distribution of the received power under LoS Rician fading as a mixture of χ^2 and Γ distributions, distinguishing the case where multi-path components (MPCs) fell into the same delay bin versus different bins. In particular, we showed that when all MPCs collapsed into the same delay bin (narrowband channel), the RSS variance increased, leading to an increase in mutual information (MI); alternatively, when MPCs were resolved into distinct bins (wideband channel), the RSS variance and MI decreased.

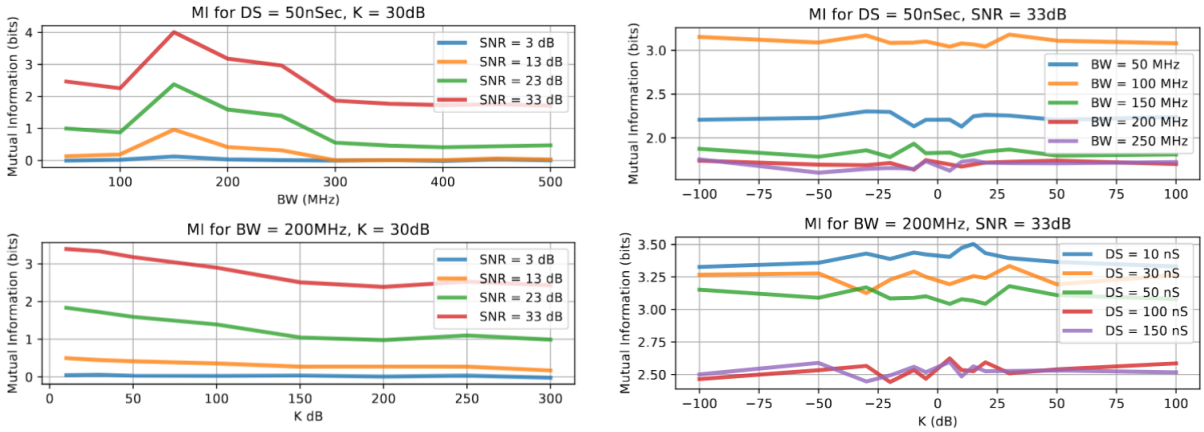


Figure 12.2: MI vs SNR and and vs Rician factor K.

12.2.2 Numerical Results and Analysis

Using 3GPP TDL-E channel models, we numerically estimated MI via a machine-learning-based estimator and studied its dependence on bandwidth (BW), delay spread (DS), Rician K-factor, and SNR. Numerical results are shown in Fig. 12.2. For fixed K and DS, increasing BW generally reduced MI, as MPCs become more resolvable and RSS fluctuations per bin decrease. On one hand, for fixed BW, increasing DS similarly reduced MI by spreading power over more resolvable taps. On the other hand, the Rician K-factor itself had limited direct impact on MI compared to BW and DS; its effect mostly manifested through how LoS power dominated or not the diffuse components. Finally, in higher SNR, MI saturated, but, remained sensitive to BW and DS. Our numerical results showed decreasing MI as BW increased from 50 MHz to 500 MHz at DS = 50 ns.

Our analysis revealed that LoS environments do not necessarily preclude SKG: the interplay of BW, DS and MPC resolvability can still yield significant MI and thus non-zero SKG rates. However, the susceptibility of SKG to on-the-shoulder eavesdropping is governed by the exact interference patterns of the MPCs at Eve, not merely by geometric distance. The results motivated channel-aware adaptation of SKG parameters in subsequent contributions.

12.3 Comprehensive Analysis of Achievable SKG Rates

12.3.1 Proposed Methodology

In [6], we bridged the gap between theory and practice by analyzing SKG rates as a function of both channel characteristics and SKG protocol design parameters, and by validating our analysis through an extensive measurement campaign with a realistic on-the-shoulder eavesdropper. In this work we addressed the following shortcoming of existing literature on SKG:

1. For ease of mathematical analysis, the vast majority of published works has in the past assumed spatial decorrelation when an eavesdropper is at distances more than half a wavelength away from the legitimate users, therefore seriously underestimating information leakage. In this work, no such assumption was made and we accounted for spatial dependencies even at very proximal distances (one wavelength) through the use of experimental datasets in LoS and NLoS conditions.

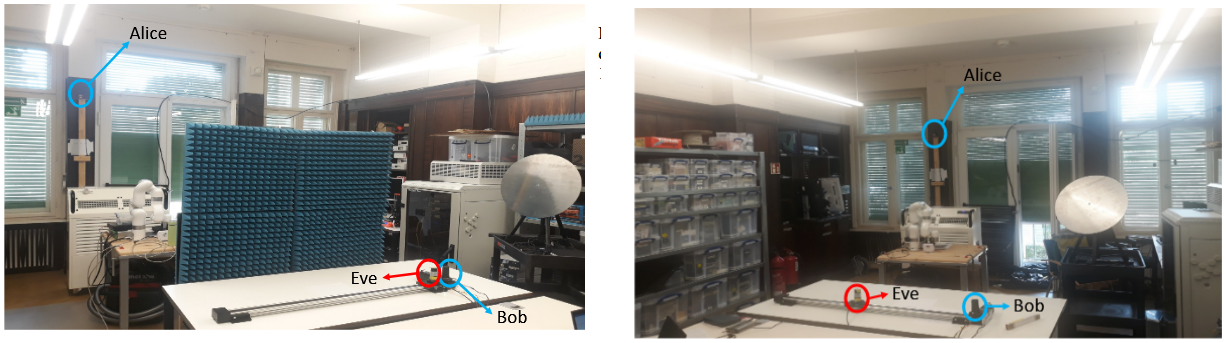


Figure 12.3: Set-up of experimental campaign for SKG.

2. Based on our experience, a key challenge in transitioning SKG from theory to practice was identified in the real-time implementation of privacy amplification, which was rarely addressed in earlier works.[3] In this work we employed the F-BLEU conditional min entropy estimator and further proposed a fast, real-time implementation using long-short term memory (LSTM) networks.

3. A large body of prior works focused on optimizing the quantization / non-reciprocity mitigation and the reconciliation rate. However, joint optimization of the SKG stages through careful parameter fine-tuning was entirely missing. In this work, we implemented and optimized jointly all stages. Interestingly, we have shown that increasing the quantization and reconciliation rates in silo, did not lead to the highest achievable key rates due to increased leakage towards Eve, captured through higher hashing rates at the privacy amplification.

12.3.2 Numerical Results and Analysis

We used the experimental campaign of our prior work [55] as shown in Fig. 12.3, in which three NI USRP-2974 SDRs, were configured as Alice, Bob and Eve. The center frequency was $f_c = 3.75\,\text{GHz}$ (wavelength $\lambda \approx 8\,\text{cm}$), with signal bandwidth $B = 70\,\text{MHz}$ and sampling rate $f_s = 140\,\text{MHz}$. Eve was placed on a linear positioner at distances 1λ to 10λ from Bob (8 cm to 80 cm), enabling a controlled on-the-shoulder attack. For each Eve position, 10^5 chirps are exchanged to guarantee convergence of conditional min-entropy and leakage estimates. We investigated four scenarios: LoS static, LoS dynamic, NLoS static and NLoS dynamic. In the static scenarios measurements were collected overnight, while in dynamic scenarios fluctuations were induced by moving a metal plate and by human or object motion in the room during daytime.
Randomness distillation was performed by convolving the received time-domain signals with a filterbank of K raised-cosine filters (e.g. $K = 16$ with roll-off 0.25), forming power measurements per subband. These observations were quantized using multi-level quantization with $Q \in \{4, 16\}$ uniform levels per subband and per channel realization. Information reconciliation was implemented via Slepian-Wolf Polar codes. Alice sent a syndrome $s_A \in \{0,1\}^{(1-r)K\log_2 Q}$, with code rate $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; both Bob and Eve used s_A to attempt error correction. The success of reconciliation depended on the number and positions of bit errors, and was shown to be negligible $< 1\%$ in low code-rates $0.1 - 0.2$.
For privacy amplification, we bench-marked our proposed solution against F-BLEAU [56], computing conditional min-entropy as the difference between min-entropy and leakage. We further introduced a conservative safety margin, compressing sequences by an extra 10% beyond the F-BLEAU estimate to

[3] Out of 43,305 papers published on IEEExplore, only 1600 papers even refer to privacy amplification, and only a handful investigate algorithmic implementations

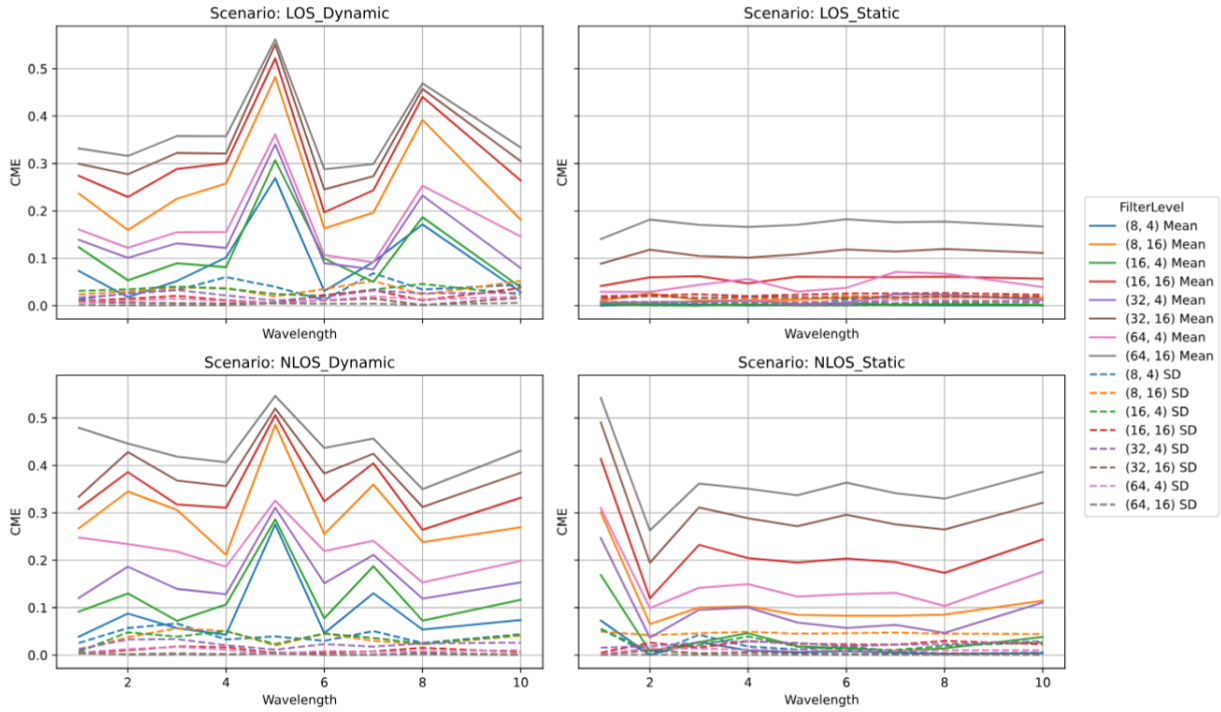


Figure 12.4: SKG rates over different combinations of design parameters as a function of the distance from Bob (measured in wavelengths).

mitigate underestimation risks. For real time operation, we proposed an LSTM that had a recurrent layer with 100 hidden units followed by a dense softmax output for multi-class conditional mutual information based rate selection. It was trained with the Adam optimizer and categorical cross-entropy.
Our measurements showed how SKG rates depended jointly on the filterbank size K, on the number of quantization levels Q, on the Polar code rate r over each scenario (LoS / NLoS, static / dynamic) and on the eavesdropper's positions (distance from Bob from 1λ–10λ). Dynamic scenarios generally yielded higher conditional min-entropy and SKG rates than static ones, as expected. However, even in static environments, careful tuning of (K, Q, r) yielded non-zero SKG rates. Our analysis highlighted that too high code rates reduced reconciliation success and thus effective key rates, while too aggressive quantization (large Q) could increase bit mismatch rate and reconciliation overhead, offsetting gains in raw entropy. Partial results are depicted in Fig. 12.4 and 12.5.

12.4 Context-Aware SKG Demonstrator with Real-Time Implementation

12.4.1 Proposed Methodology

In this work, we participated with our algorithms in the full, end-to-end, context-aware SKG demonstrator built by the Barkhausen Institut. The demonstrator integrated pilot exchanges, environment classification, parameter selection, SKG protocol execution, real-time privacy amplification and intuitive visualization via a physical LED key. The hardware setup used two USRP X410 devices acting as Alice and Bob, operating at a carrier frequency of 5.5 GHz with sampling rate 491.52 MS/s, providing a usable bandwidth of 400 MHz. The devices shared a common reference clock and time source to allow synchronized frame exchange. Each frame consisted of a chirp of length 2048 samples and bandwidth 400 MHz, and a burst consisted of 5 frames

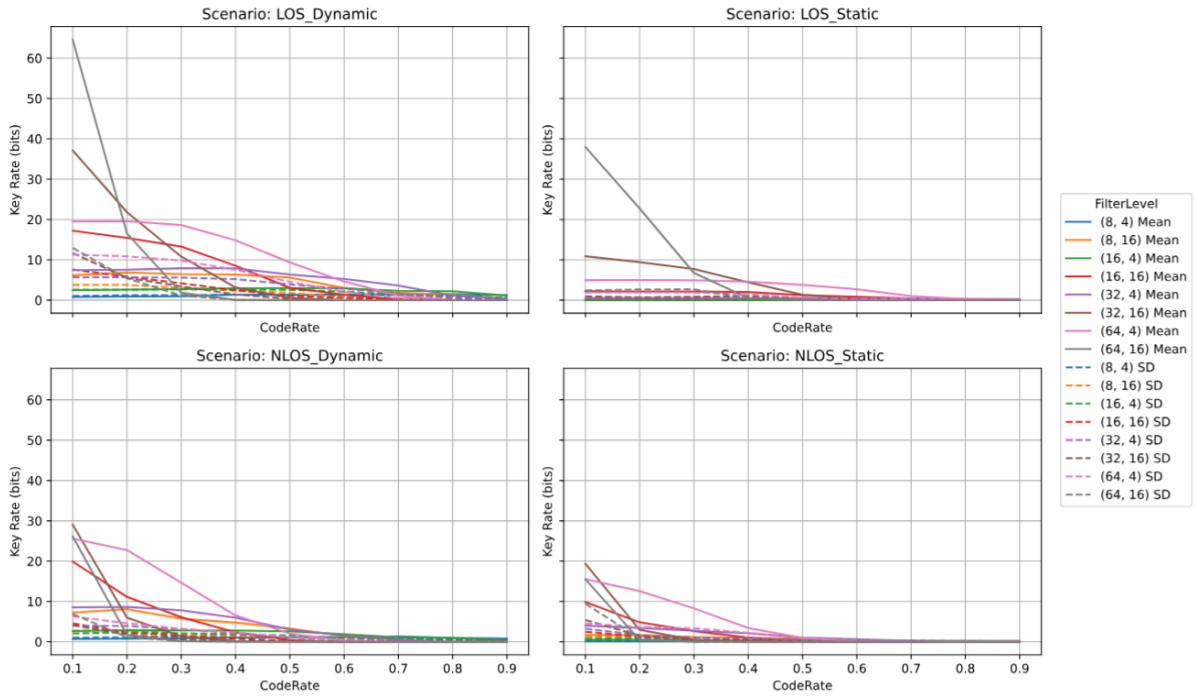


Figure 12.5: SKG rates over different combinations of design parameters as a function of the code rate.

separated by 0.1 s, with a transmission delay of 0.01 s between Alice and Bob.

12.4.2 Numerical Results and Analysis

We captured real-time channel frequency responses (CFRs) in three different physical contexts, namely, LoS static; NLoS static; and Dynamic (motion of antennas, people and objects). We extracted two families of features from CFRs: i) momentary features, i.e., absolute and phase differences between CFR samples, amplitude and phase of the dominant LoS component, and statistics (mean, variance, kurtosis, skewness). ii) temporal variation features, i.e., cross-correlation between consecutive CFRs, rates of change in LoS amplitude and phase, and PCA-based features summarizing temporal patterns. Based on the detected context, we selected SKG parameters (f, l, r), where f denoted the number of filters in the filterbank; l the number of quantization levels; and r the code-rate in reconciliation. The demonstrator implemented the full SKG chain in real time (less than 1 msec), hinting that the WP5 KPI of AKA is less than 5 msecs is within grasp. The final secure bits were displayed on a 3D-printed key equipped with LEDs controlled by an ESP32 microcontroller, providing an intuitive visualization of the abstract digital key, shown in Fig.12.6 This demonstrator was awarded the best demo award in IEEE CNCS 2025, shown in Fig. 12.7

12.5 Integration with the Architecture

Taken together, our works traced our path from foundational analysis of LoS multipath channels for SKG to a fully operational, context-aware, real-time demonstrator. Our development of a conditional min-entropy estimator, combined with realistic leakage estimation and context-aware parameter selection, ensures that we can quantify and control the information advantage of legitimate users over an eavesdropper, even under stringent attacks. We have reported reconciliation rates of $> 99\%$ while for privacy amplification the

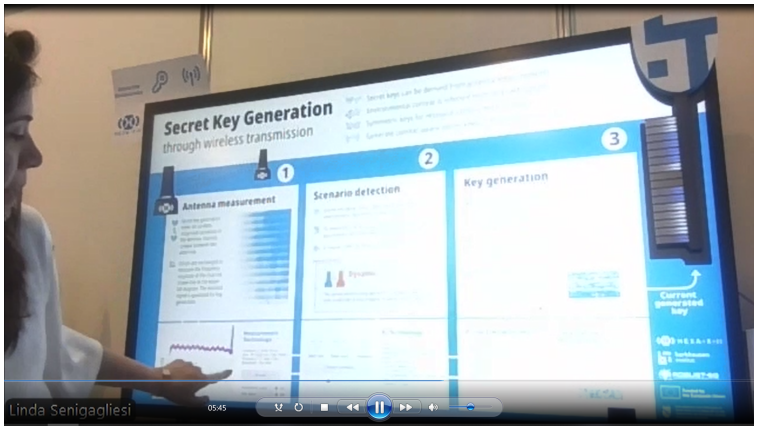


Figure 12.6: Context-aware, fast and robust SKG.

Figure 12.7: Best demonstrator award at IEEE CNCS 2025

inference times were less than 0.2 msecs, while overall run-time for key generation in the live demonstrator was inferior to 1 msec. As a result, these works are aligned with the stated WP5 KPIs of reconciliation more than 90% and AKA of 5 msec (in conjuction with AoA-PLA with run-time of 2 msec as discussed in Chapter 3). The integration of these results is contained in our component **CENS05** on fast SKG in the actuation stage of the physical layer closed loop. Furthermore, the conditional min-entropy leakage estimator is part of the secrecy and leakage analysis of the component **CENS02**.

Chapter 13

Enhancing the Performance of CSI-Based PLA Through Reconciliation and Preprocessing

In this chapter we provide a synthetic view of our works [8–11] on the use of information reconciliation and signal pre-processing to enhance the accuracy of PLA, addressing the issue of changing authentication physical channel features (such as the CSI) due to time, frequency or spatially evolving statistics. These works were shown to consistently enhance the accuracy of PLA when compared to state of the art approaches.

13.1 Background and Motivation

The computational complexity of classical cryptographic approaches based on public key distribution can be a limiting factor for user authentication. PLA can be an interesting solution to complement existing traditional approaches, e.g., in multi-factor authentication protocols. However, the precision and consistency of PLA is impacted by random variations of wireless channel realizations between different time slots, which can impair authentication performance. In [8] and [9], we focused on enhancing the accuracy of CSI-based PLA by introducing the use of forward error-correcting codes in the form of reconciliation, as depicted in Fig. 13.1. We note that this approach could be also applied in RF fingerprinting-based and AoA-PLA. The proposed method employed a Slepian-Wolf coding scheme with Polar codes that allowed the reconciliation of discrepancies between channel measurements over different time instances, in order to authenticate legitimate users. The authentication decision was therefore based on the comparison between the reconciled vectors followed by hypothesis testing. In addition, we derived closed-form expressions of the probability distribution of the hypothesis test statistical variable and of the probability of false alarm and detection. The method was first considered in a single user communication system using 1-bit quantization [8] and then extended in a multiuser system that employed Lloy-Max quantizers with an arbitrary number of bits [9].
Furthermore, in [10, 11], we proposed an adaptive preprocessing technique to enhance the accuracy of CSI-based physical layer authentication (CSI-PLA) alleviating CSI variations and inconsistencies in the time domain. To this end, we developped an adaptive robust principal component analysis (A-RPCA) preprocessing method based on unsupervised machine learning. The performance evaluation was conducted using a PLA framework based on information reconciliation, in which the Gaussian approximation (GA) for Polar codes was leveraged for the design of short codelength Slepian Wolf decoders. Furthermore, an analysis of the proposed A-RPCA methods was carried out. Simulation results showed that compared to a baseline scheme without preprocessing and without reconciliation, the proposed A-RPCA substantially reduced the error probability after reconciliation and also substantially increased the detection probabilities

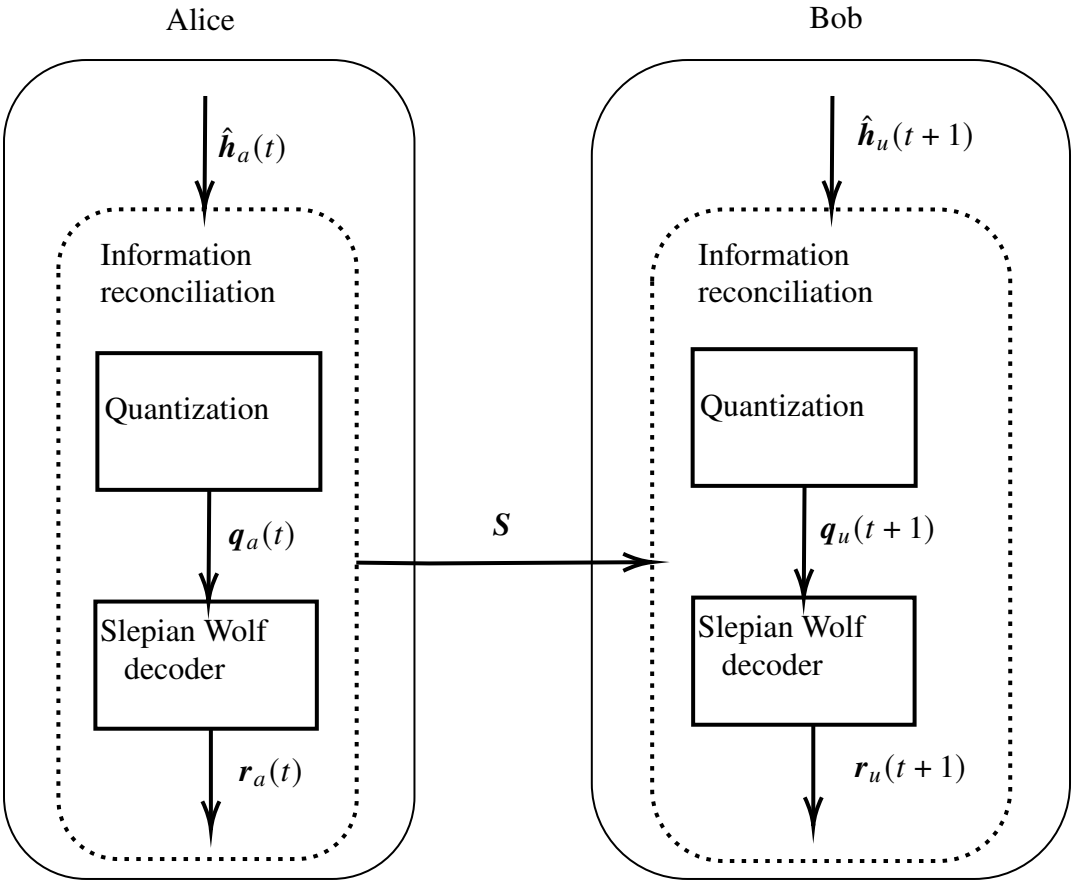


Figure 13.1: Proposed authentication scheme including an offline (enrollment) phase at time t and an online (verification) phase at time $t+1$, $u \in \{a, m\}$

that reached 100% in both LoS and NLoS scenarios on the real dataset from Nokia campus (introduced in Chapter 3). This work related to the KPI for more than 90% authentication accuracy, which was surpassed with the proposed methods.

13.2 Physical Layer Authentication Using Information Reconciliation

13.2.1 Proposed Methodology

In this work we presented a reconciliation scheme to mitigate the impact of disparities in the CSI observed in subsequent time slots by employing the principle of Slepian-Wolf decoding. This approach aimed to reconcile the channel measurements during the enrollment phase to the ones during the verification phase. The motivation for employing reconciliation lied in the observation that it is trivially used in physical unclonable function based authentication (typically is referred to with the term fuzzy extractors) and secret key generation from channel measurements as discussed in Chapter 12. It was proposed in this work, for the fist time, to be used in CSI-PLA.

Each phase involved a quantization of the CSI, with the output vectors at time t and $t+1$, both treated as dithered codewords at the input of the reconciliation decoder, estimated with the use of helper data $\mathbf{S}$. The basic assumption was that if the measurements came from the same legitimate user, then the codewords at time t and time $t+1$ would coincide (i.e., the reconciliation would be successful), which would not be the case

otherwise. The reconciliation decoder was assumed to output one reconciled vector at each time instance. Note that to this end, the helper data **S** generated in the first phase was used by Bob in the authentication phase to reconcile the newly obtained CSI at time $t + 1$ to the previous one at time t. Then, to make a decision during the online authentication phase, a hypothesis test was performed by Bob in order to identify the legitimate user versus a potential impersonator.

13.2.2 Numerical Results and Analysis

In Fig. 13.2 and 13.3 we provided comparisons with state of the art methods without reconciliation.

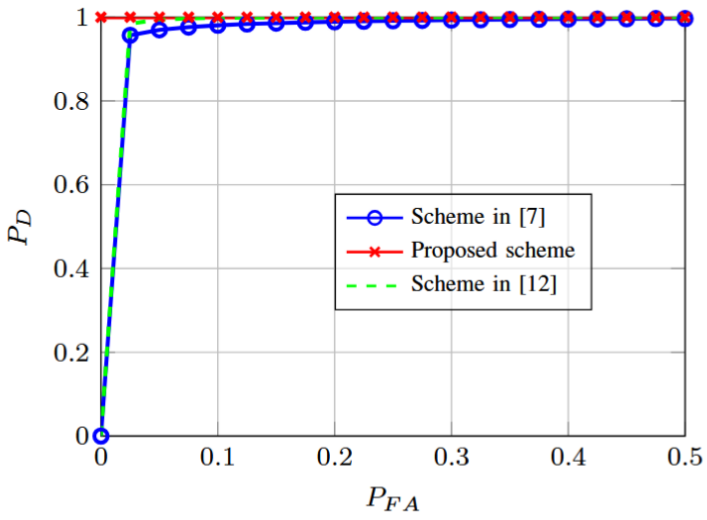


Figure 13.2: Probability of detections vs Probability of false alarm compared to state of the art.

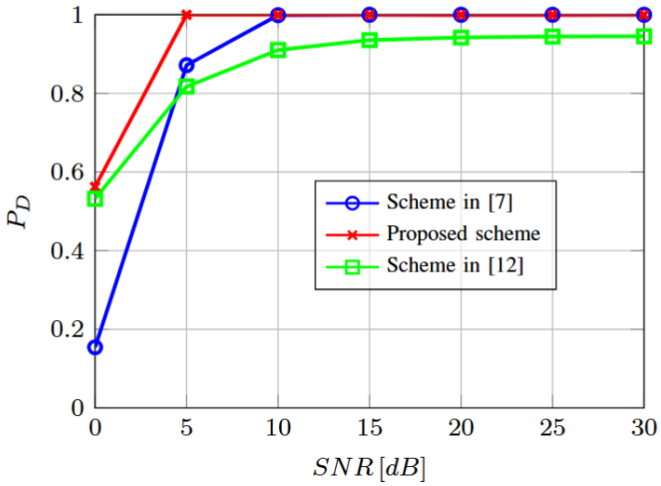


Figure 13.3: Probability of detection vs the SNR, compared to state of the art.

We provided closed-form expressions for the false alarm probability and the detection probability. In a scenario of SNR = 15 dB and a false alarm rate of 10^{-3}, our results showed that that the probability of detection was close to one for code rates less than or equal to 0.2 and very small for code rates greater

that 0.2. Simulation results confirmed also that our reconciliation-based method had better performance than prior schemes. The fact that we used a small value of 10^{-3} for the false alarm probability allowed to confirm the performance of our work in practical systems that need very low false alarm probabilities (e.g., in vehicular ad hoc network (VANET)).

13.3 Enhanced Multiuser CSI-based Physical Layer Authentication Based on Information Reconciliation

13.3.1 Proposed Methodology

Our previous work was made more general by accounting multiuser interference. The wireless communication network considered is depicted in Fig. 13.4 and included legitimate nodes, referred to as Alice and Bob (base station) as well as U benign interfering legitimate users, In this network, Bob wanted to authenticate Alice in the presence of the other legitimate users and an adversary Mallory that attempted to impersonate Alice through a simple attack without the use of any precoder or other pre-processing techniques. The objective was to design a scheme based on CSI to distinguish Alice from Mallory in the presence of interfering users. Each user was equipped with a single antenna and Bob was equipped with N_b antennas.
We assumed that the communication occurred in a rich scattering environment and the distance among users exceeded half of a wavelength. To simplify the system model, the channel attributes between different transmitter-receiver pairs were assumed to be spatially uncorrelated. The channel between the same transmitter-receiver pair was described by a first order Gauss-Markov process and the dependence between samples in the time domain is captured through the correlation coefficient β.

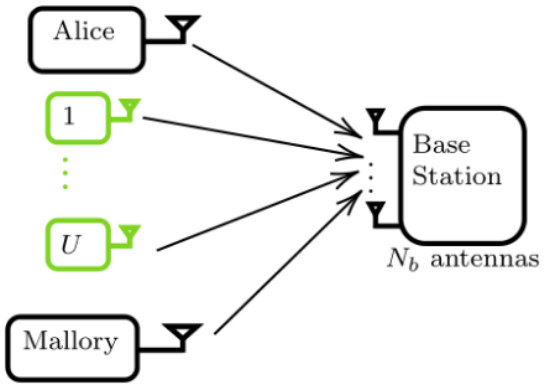


Figure 13.4: Multiuser system with interference

13.3.2 Numerical Results and Analysis

In Fig. 13.5 the impact of the SNR on the detection probability was studied. As the SNR increased, probability of detection (PD) increased as subsequent CSI measurements were more correlated due to decreased noise. The proposed scheme performed very well with a PD greater than 99.86%, while the performance of prior methods was poor for low SNRs less than 10 dB.
Fig. 13.5 showed the detection probability as a function of β. We had a PD very close to 1 for $0.4 \leq \beta \leq 1$. We therefore had excellent performance even for challenging scenarios of medium correlation coefficients and our proposed scheme performed better than state of the art approaches. An assessment of the time and memory complexities is given below.

- Time complexity
  The time complexity of the enrollment phase was evaluated as $O(N)+O(L)+O(L_s nN\log(nN))$. The time complexity of the authentication phase and the decision using hypothesis testing were respectively given by $O(UN)+O(L)+O(L_s nN\log(nN))$ and $O(K)$. L_s is the list size of the cyclic redundancy check successive cancellation list decoding of the polar decoder and K is the length of the reconciled vectors.

- Memory complexity
  The overall memory complexity required to perform the authentication process taking into account all steps of the scheme was given by $O(U)+O(L)+O(L_s nN)$.

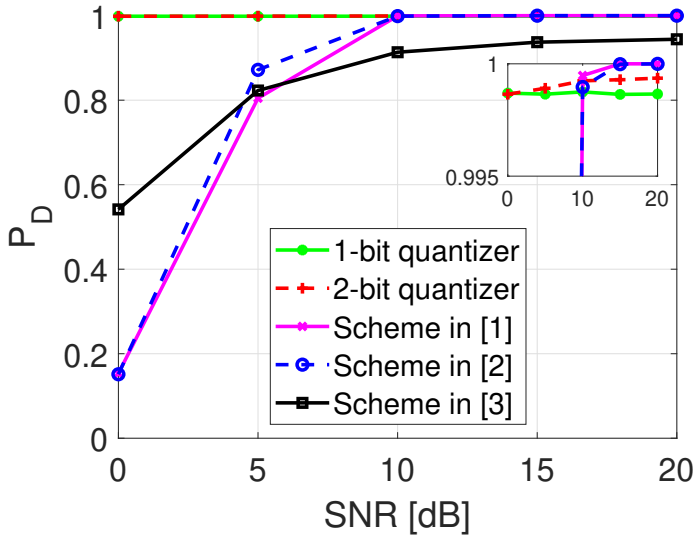


Figure 13.5: PD vs SNR, $P_{FA} = 10^{-3}$

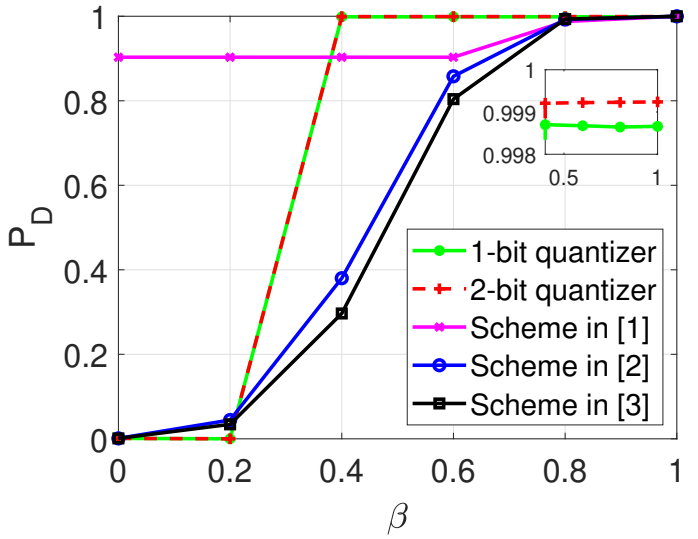


Figure 13.6: PD vs β for SNR=10 dB, $P_{FA} = 10^{-3}$

13.4 Channel State Information Preprocessing for CSI-based Physical-Layer Authentication Using Reconciliation

13.4.1 Proposed Methodology

In these works we studied how preprocessing could further enhance CSI-PLA accuracy. We observed that the main premise of CSI-PLA depended on the correlation between the channel measurements across the offline and online phases and the underlying assumption of the hypothesis test that CSI observations would be more correlated in the normal case than in the alternative. Additionally, even if two signals would be highly correlated in some features, applying a preprocessing schemes independently across subsequent observations could destroy that correlation, especially if the correlation was higher in the noisiest features dimension. Consequently, adaptive schemes that took into account channel measurements from previous authentication phases were needed.
In our recent works we presented a PLA framework based on an adaptive preprocessing techniques and information reconciliation using polar codes to enhance the accuracy of CSI-based PLA and alleviate CSI estimation variations and inconsistencies in the time domain. In particular, an adaptive robust principal component analysis (RPCA) (A-RPCA) preprocessing was proposed. An analysis of the proposed A- RPCA methods was carried out along with a study of its computational cost. Unlike the in our previous works [8] and [9], in this work we employed a Gaussian approximation (GA) for designing polar codes instead of the code construction that relies on the binary erasure channel (BEC). The main contributions of these works were summarized as follows:

- A PLA framework based on an adaptive preprocessing technique denoted by A-RPCA was proposed to enhance the accuracy of CSI-PLA and alleviate CSI estimation uncertainties and time-varying nature.
- The proposed adaptive preprocessing took into account both the CSI from the enrollment and the authentication phases instead of applying the preprocessing indepen- dently across the phases. This was achieved through an (offline) estimation of the correlation coefficient between the CSI from the enrollment and authentication phases and it allowed to discriminate the the adversarial CSI from the legitimate user's CSI.
- A convergence analysis of the time-regularized principal component pursuit (TR-PCP) optimization problem that was used in the proposed A-RPCA algorithm was carried out.
- We assessed the preprocessing and the PLA authentication performance with synthetic data and in real-world scenarios, where a dataset from Nokia was used to distinguish between different users (partial results presented below).

13.4.2 Numerical Results and Analysis

Figure 13.7 showed the track segmentation for both tracks we want to differentiate. One segment represented the track of interest (Alice 1) to be detected, where the green squares represent the user position at time t and the orange squares are the user location at time $t + 1$. The red squares represented the user location at $t + 1$ on the track (Alice 2) from which we wanted to distinguish the targeted user. Different communication scenarios were considered where (Alice 1) and (Alice 2) were on LoS or NLOS tracks. We then had the pairs (Alice 1, Alice 2): (LoS, LoS) = (track 6, track 11), (LoS, NLoS) = (track 6, track 1), (NLoS, LoS) = (track 1, track 6) and (NLoS, NLoS) = (track 1, track 13). In the following, unless otherwise specified, the parameters were defined as follows: code rate = 0.1, codelength $N = 128$, SNR= 10 dB, number of subtracks is $N_s = 46$, $n = 1$ bit quantizer.

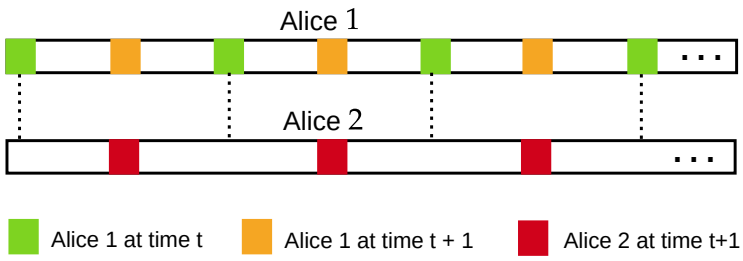


Figure 13.7: Tracks of Alice 1 and Alice 2 to be differentiated. Each Alice is on a single track because the data corresponds to a walking user. The green and orange colors represent respectively Alice 1 locations at time t and $t + 1$. The red color is Alice 2 location at time $t + 1$.

First, we showed the ROC curves. For a code rate equal to 0.1 in Fig. 13.8 it was demonstrated the effectiveness of the reconciliation scheme in real scenarios. In all cases, including that of CSI without preprocessing, the RPCA algorithm and the proposed A-RPCA algorithm performed very well, while A-RPCA had detection probabilities equal to 1 even for false alarm rates almost 0. The difference was more obvious when the code rate was increased to 0.2 in Fig. 13.9. We could see that the performance of A-RPCA did not change much even for a code rate = 0.2. The difference was clear for the cases of no preprocessing and RPCA, where the performance declined.

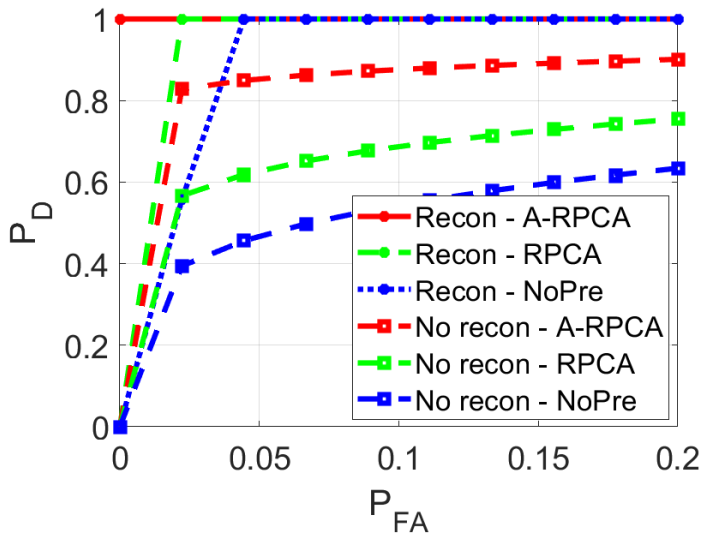


Figure 13.8: Code rate = 0.1, P_D vs P_{FA}: (LoS, NLoS) = (6, 1), SNR = 10 dB, N = 128.

The miss-detection and false alarm probabilities were evaluated as a function of the hypothesis testing threshold in Fig. 13.10 where the code rate was, respectively, equal to 0.3. The figure confirmed our previous results with ROC curves. The equal error rate (EER) that is the point where the false alarm and miss-detection probability cross is lower for A-RPCA as compared to the case without preprocessing or RPCA. This confirmed the higher accuracy obtained by A-RPCA preprocessing.
In conclusion, we have developed a framework for PLA that employs a novel A-RPCA preprocessing algorithm along with a recon- ciliation technique based on polar codes to further enhance the performance of existing CSI-based PLA schemes. The proposed A-RPCA preprocessing technique was obtained by solving a TR-PCP optimization problem. Numerical results showed that A-RPCA substantially improved the error

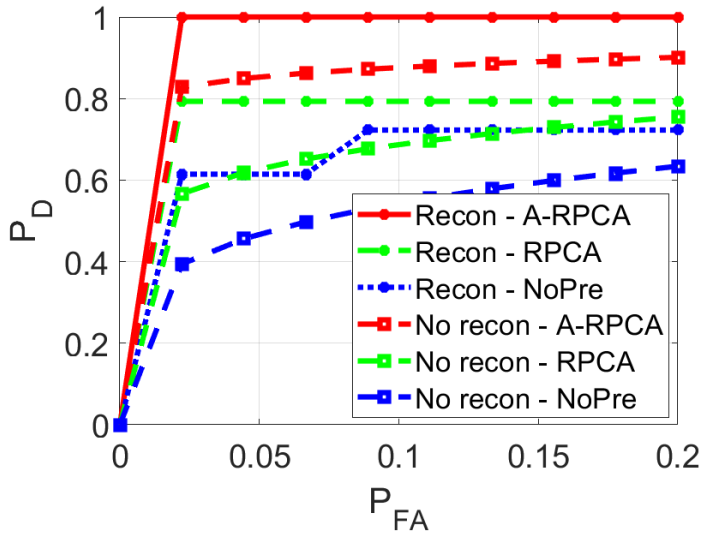


Figure 13.9: Code rate =0.2, P_D vs P_{FA}: (LoS, NLoS) = (6, 1), SNR = 10 dB, $N = 128$.

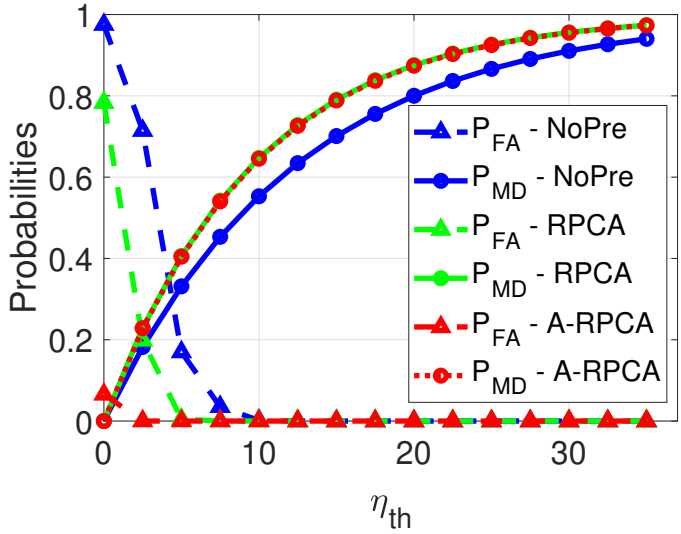


Figure 13.10: P_D and P_{FA} vs η_{th}: (LoS, NLoS) = (6, 1), SNR = 10 dB, Code rate = 0.3, $N = 128$, $n = 1$

probability after reconciliation.

13.5 Integration with the Architecture

We have not proposed any CSI-PLA components for the ROBUST-6G architecture as we did not perform an analysis of their vulnerability to spoofing and man-in-the-middle (MITM) attacks. However, these techniques could be easily integrated into the components of other partners, e.g., for RF fingerprinting.

Chapter 14

Bounds on Information Leakage of Short Packet Wiretap Codes

This chapter provides a characterization of the available secrecy rates of practical, low complexity and delay-constrained wiretap code constructions as a function of the codelength and of the quality of security (QoSec) requirements. Further technical details about this work[1] can be found in Appendix E.

14.1 Background and Motivation

Wiretap coding allows to counter passive eavesdropping provided that the legitimate receiver has an SNR advantage compared to the eavesdropper. In order to select a suitable wiretap coding scheme, it is necessary to adapt the channel coding rate to the channel conditions. For applications requiring short packets or low latency, it is not possible to guarantee a vanishing information leakage; for a target leakage constraint δ, the back-off of the finite-blocklength secrecy rate from the secrecy capacity must be taken into account. In our preliminary work [57], we have investigated the performance of practical wiretap schemes based on polar codes for a simplified channel model where the main channel is noiseless and the eavesdropper's channel is a binary erasure channel (BEC). Our goal is to extend these results in order to obtain lower bounds on the achievable secrecy rate $R(n, \delta, \epsilon)$ for a given blocklength n, leakage δ, and error probability ϵ for more general channels.

14.2 Proposed Methodology

For simplicity, in this project we focus on the case of binary inputs and assume that the wiretap channel is degraded. Following recent theoretical results on the second-order approximation for the secrecy rate of wiretap codes [58], we measure the information leakage in terms of the total variation distance (TVD) between the joint distribution of the confidential message M and the eavesdropper's observation Z^n, and the ideal distribution where M is uniform and independent of Z^n. We consider practical wiretap coding schemes based on polar codes, which are low-complexity, asymptotically secrecy capacity-achieving, and are already integrated in modern communication standards such as 5G New Radio.
In our previous work [57], we have shown that the information leakage in TVD for polar codes is upper bounded by the sum of the TVDs of the bit-channels corresponding to the confidential bits. However, for general channels, the exact computation of the TVDs of the bit-channels is prohibitively complex since the

[1] V. Bioglio, L. Luzzi, paper in preparation, to be submitted to *ISIT 2026*.

cardinality of the output alphabet grows exponentially. To solve this problem, we use the upgrading merge algorithm proposed in [59] in order to obtain an upper bound of these TVDs with arbitrary precision.

14.3 Numerical Results and Analysis

Numerical results are presented in Appendix E.3. Our simulations show that although they asymptotically achieve the secrecy capacity, wiretap schemes based on polar codes incur a significant loss in terms of secrecy rate in finite blocklength, which depends on the target parameters ϵ and δ and the channel gains, and must be taken into account for practical implementation. For instance, when the main channel and eavesdropper's channel are additive white Gaussian noise (AWGN) channels with Binary Phase Shift Keying (BPSK) inputs and noise variance $\sigma_b^2 = 0.2$ and $\sigma_e^2 = 2$ respectively, at blocklength $n = 512$, polar codes are guaranteed to achieve approximately 43% of the optimal secrecy rate.
We note that the proposed lower bound on the secrecy rate of polar codes is not tight for general channels; a tighter bound was derived in our previous work [57], but it seems difficult to evaluate numerically except for the simple case where the eavesdropper's channel is a binary erasure channel (BEC) (see Appendix E.3).

14.4 Integration with the Architecture

Our code outputs a lower bound for the achievable secrecy rate with polar codes as a function of the blocklength n, the instantaneous SNR and the target information leakage δ (or, alternatively, it outputs an upper bound for the information leakage for a given SNR, blocklength and secrecy rate) and can be used in the Secrecy and Information Leakage component **CENS02** to generate secrecy maps.

Chapter 15

Challenge-Response Authentication At The Physical Layer

15.1 Analysis of Challenge-Response Authentication With Reconfigurable Intelligent Surfaces

PLA mechanisms exploit signals exchanged at the physical layer of communication systems to confirm the sender of a received message. In [60], UNIPD proposes a novel challenge-response PLA (CR-PLA) mechanism for a cellular system that leverages the reconfigurability property of a RIS (under the control of the verifier) in an authentication mechanism. In CR-PLA, the verifier BS sets a random RIS configuration, which remains secret to the intruder, and then checks that the resulting estimated channel is modified correspondingly. In fact, for a message sent by an attacker in a different location than the legitimate UE, the BS will estimate a different channel, and the message will be rejected as fake. Such a solution reduces the communication and computational overhead with respect to higher-layer cryptographic authentication. We derive the maximum a posteriori probability attack when the attacker observes a correlated channel and the RIS has many elements, and the attacker transmits to Bob either directly or through the RIS. Using a generalized likelihood ratio test to test the authenticity at the BS, we derive approximate expressions of the false alarm and misdetection probabilities when both the BS and the UE have a single antenna each, while the RIS has a large number of elements. We also evaluate the trade-off between security and communication performance, since choosing a random RIS configuration reduces the data rate. Moreover, we investigate the impact of various parameters (e.g., the RIS randomness, the number of RIS elements, and the operating signal-to-noise ratio) on security and communication performance.

15.1.1 Background and Motivation

Determining if a received message is coming from its claimed sender, i.e., establishing its *authenticity*, is a key security problem in communication systems. The current and future networks will include several interconnected devices with diversified energy and computational constraints, and PLA is an attractive security solution since it requires simple signal processing capabilities and exploits existing signals without introducing communication overhead. In tag-based PLA, the channel operates as a tag: the receiver authenticates newly received messages that appear to have traveled through the same channel as those (authentic) received in the past. When an attacker transmits from another location, the resulting channel is different from that of the legitimate transmitter and is detected as fraudulent.
Since PLA is based on channel characteristics, devices that enable the manipulation of propagation properties should be considered to enhance security. To this end, a RIS is an interesting component, as it includes

several reflective elements, each introducing a controllable phase shift in the equivalent baseband reflected signal. RISs have also been considered for PLA to increase the SNR and improve the authentication process. However, the possibility of reconfiguring RISs also paves the way for a new PLA procedure, called challenge-response (CR) PLA. In CR-PLA, the receiver first randomly modifies the propagation environment (which represents the challenge) and then estimates the channel through which the received signal has passed (which represents the response) to verify that it matches the modified environment. For an attacker that does not know the current challenge (i.e., the current RIS configuration) it will be harder to perform an effective authentication attack than in the PLA setting.
While RIS-based CR-PLA is a promising security solution, its performance has not been investigated in the literature. This study should include not only the security performance – in terms of false alarm (FA) and misdetection (MD) probabilities – but also communication performance, since the choice of the RIS configuration has an impact on the achieved data rates. Moreover, the ability to withstand advanced attacks that exploit the partial channel knowledge of the attacker is still to be investigated. By a thorough analysis of the RIS-based CR-PLA, it will be possible not only to assess the merits of this security solution but also to design it properly, i.e., to select the RIS size and the randomness.

15.1.2 Proposed Methodology

To address these issues, in this work, we consider an RIS-supported CR-PLA mechanism for cellular networks, where a BS verifies the authenticity of messages received from a UE. The CR-PLA procedure includes two stages. In a preliminary stage, the UE transmits a sequence of pilot samples (properly authenticated by a higher-layer procedure) to the BS via the RIS with several configurations, and the BS estimates the corresponding UE-RIS-BS cascaded channels. This will enable the BS to predict the cascaded channel under any other RIS configuration. In the second stage, aiming at authenticating a message potentially coming from the UE, the BS randomly chooses a RIS configuration while a new message is transmitted. The BS then compares the channel estimated from the received signal with that predicted for the selected RIS configuration, using the information obtained in the preliminary stage. From this comparison, the BS decides on the message authenticity.
We consider a GLRT to decide about the authenticity of the message and analyze the performance of the CR-PLA scheme in terms of both FA and MD probabilities. Note that the random RIS configuration also affects the data rate of the communication link between the UE and the BS. To limit the rate loss, we restrict the random selection of each phase shift of the RIS to an angular sector centered around the phase shift that maximizes the data rate, and we investigate the security performance as a function of the size of the angular sector. We also derive the maximum a posteriori probability (MAP) attack to be used when the attacker knows his channel to the legitimate UE, and this channel is partially correlated with that from the legitimate UE to the RIS. The attacker can transmit the attack signal either directly to the BS or through the RIS. In the latter case, the attacker signal will also be determined by an instantaneous random RIS configuration. Approximate expressions for the FA and MD probabilities and the data rate are obtained when both the UE and the BS are equipped with a single antenna, and the RIS has a large number of elements.

15.1.3 Numerical Results and Analysis

Here we report the main result, for more details please refer to [60].
We examine in detail the effects of the correlation between the Alice-RIS and Eve-RIS channels. Fig. 15.1 shows the analytical mean MD probability as a function of ρ for an FA probability $P_{\text{FA}} = 10^{-3}$, $N = 50$, and for values of ρ that either give the minimum MD probability, i.e. $\gamma = \pi$ (independent of spectral efficiency), or give a spectral efficiency loss $\eta = 2\%$, 10%, or 50%. First, we consider the scenario in which Eve transmits directly to Bob. We also include the performance of the tag-based PLA for comparison. First, we observe

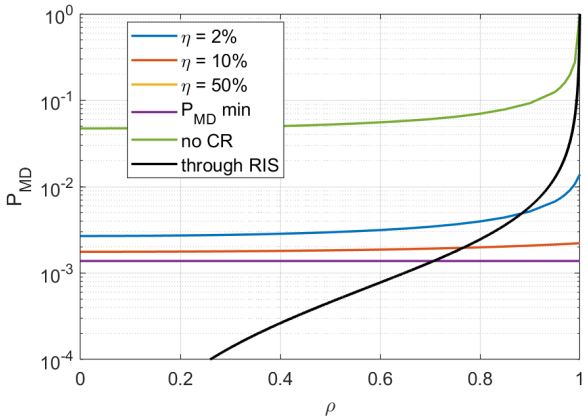


Figure 15.1: Analytical average MD probability as a function of ρ for FA probability $P_{FA} = 10^{-3}$, for values of γ providing the minimum MD probability (irrespective of the spectral efficiency), or giving values of $\eta = 2\%$, 10% or 50%, and in the absence of CR PLA mechanism.

that using the CR-PLA mechanism significantly reduces the MD probability compared to the case without CR. Regarding the behavior as a function of ρ, we observe that even for $\rho = 0$ (uncorrelated channels), the MD probability is high for the scheme without CR, since choosing the RIS configuration that maximizes the spectral efficiency makes the resulting channel non-zero mean, and this bias can be exploited by the attacker. Using the CR approach instead reduces the bias (which becomes zero for $\gamma = \pi$) and results in a lower MD probability. Indeed, for a spectral efficiency loss of only 2%, the MD probability drops to about 10^{-2} with $\rho = 1$, and even lower values are obtained for lower ρ or larger η. Also note that in the absence of CR, the MD mostly increases for $\rho > 0.8$, while the increase is smoother with CR PLA.
For the RIS attack, we can see from Fig. 15.1 that the correlation factor has a higher impact on the MD probability. As observed before, a lower correlation makes the RIS attack less effective than the direct attack, while at high correlations the direct attack is more effective.

15.2 Divergence-Minimizing Attack Against Challenge-Response Authentication with IRSs

In [61] UNIPD proposes a new attack against challenge response physical layer authentication (CR-PLA) with RISs. Drawing from prior work, we establish bounds on performance metrics, such as probabilities of false alarm and missed detection, using Kullback-Leibler (KL) divergence. Leveraging prior results in [62], we extend the analysis to the CR-PLA scenario with RIS. We derive the optimal attack strategy to minimize the divergence between authentic and forged signals when the attacker has either partial or no knowledge of the legitimate cascade channel. We evaluate the attack performance under different conditions, by varying the correlation between Eve's observation and the legitimate cascade channel, or the SNR at the legitimate receiver.

15.2.1 Background and Motivation

Source authentication is the problem of establishing if a received message truly comes from the declared sender or has been forged by an impersonating attacker. Risks in accepting unauthenticated messages go from denial of service to privacy to the loss of control of devices, e.g., in Internet of Things (IoT) contexts.

Several authentication mechanisms have been explored, beyond those operating at the application level and using cryptographic approaches. Here we focus on PLA, which leverages the propagation characteristics of the physical channel as signatures of the transmitting devices or the communication links. The basic approach, introduced by Simmons [63] includes two phases, the identification acquisition, and the identification verification. In the former, the receiver Bob (verifier) estimates the channel using signals transmitted by Alice (the authentic source) that are authenticated at higher layers (e.g., by cryptographic mechanisms). In the latter, whenever Bob receives a new message, he also estimates the channel over which the signal traveled and compares this estimate with that obtained in the first phase. If the two estimates are consistent (note that they are still affected by noise), then the received message is deemed authentic, otherwise it is considered fake. PLA has been studied for several technologies, including OFDM and MIMO, underwater acoustic communications [64], and using several techniques for the test, from Neyman-Pearson tests to machine learning approaches.
Recently, a further evolution of PLA has been introduced by exploiting the controllable nature of wireless channels provided by new communication technologies. In particular, RISs are controllable devices that can change the propagation of wireless signals by changing the phase shift introduced by their elements. When the RIS is under Bob's control, he can set a random configuration of the RIS which remains secret to the attacker, and verify that the estimated channel on a received message corresponds to the set configuration, [65]. This approach provides a *challenge-response physical layer authentication (CR-PLA)* mechanism and can be applied also when other *controllable channels* are available (e.g., Bob is a drone that can change its position, [66]).

15.2.2 Proposed Methodology

We investigate novel attacks to be performed when the CR-PLA uses an RIS to perform the challenge-response approach. In particular, we leverage the results of [62, Th. 2] that has established bounds on the performance for a conventional (non-interactive) PLA mechanism, in terms of probabilities of FA and MD. The bounds exploit the KL divergence of observed channels at Alice, Bob, and Eve, and it turns out that when the observations in the legitimate case are jointly Gaussian distributed, the optimal attack strategy is also Gaussian distributed. Here we derive the bounds for the considered CR-PLA scenario with RIS and, under the assumption of a large number of RIS elements, we ensure that the assumption in [62, Th. 2] are satisfied and derive the optimal attack by Eve. We assess the performance of the obtained attack by considering different correlations between Eve's observation and the legitimate cascade channel, and different SNR values for the legitimate channel.

15.2.3 Numerical Results

We consider correlated legitimate and attack channels, with $\rho \in [0, 1]$ representing the correlation factor (see [61] for more details), and Alice, Bob, and Eve are equipped with 5 antennas each. The channel is AWGN and Bob's noise power is $\sigma_B^2 = N/10$, where N is the number of elements of the RIS.
Fig. 15.2 shows the detection error tradeoff (DET) curves for different value of ρ. As ρ increases, the DET curves move towards the edge of the shaded area, which represents the trivial limit case when the decision is taken randomly, without using the signal. Thus, the FA probability for a given MD probability increases when the correlation between Eve's observation and the actual cascade Alice-IRS-Bob channel is higher.

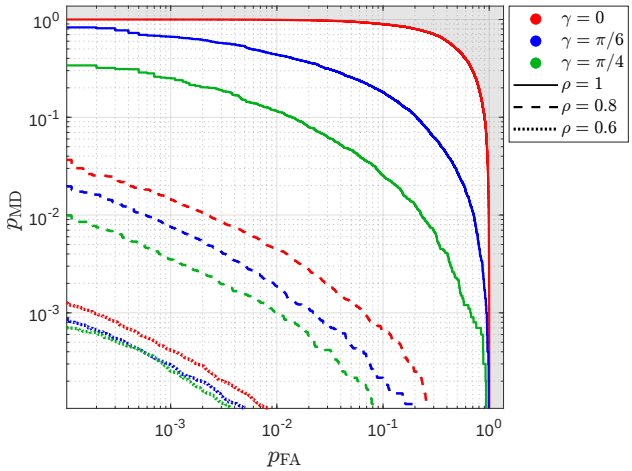


Figure 15.2: DET curves for the GLRT for different values of ρ. The area corresponding to $p_{\mathrm{MD}} \geq 1 - p_{\mathrm{FA}}$ is shaded, and its edge represents the trivial limit case in which the decision is made tossing a biased coin.

15.3 Physical-Layer Challenge-Response Authentication with IRS and Single-Antenna Devices

In [67], UNIPD has focused on a novel challenge-response physical layer authentication (CR-PLA) mechanism for wireless communications. It integrates an RIS under the control of the receiver, which operates as a verifier for the identity of the transmitter. In CR-PLA, the verifier randomly configures the RIS and then checks that the resulting estimated channel is correspondingly modified. We address the trade-off between communication and security performance, in terms of average SNR and MD probability of an impersonation attack, respectively. In particular, we design the probability distribution of the random RIS configuration that maximizes the average receiver SNR under an upper bound constraint on the MD and FA probabilities, for the special case where both the transmitter and the receiver are equipped with a single antenna. Numerical results demonstrate effective balancing of communication metrics and security requirements, suggesting that CR-PLA is a promising solution for future secure wireless communication.

15.3.1 Background and Motivation

Establishing whether a received message truly comes from the legitimate sender or has been forged by an impersonating attacker describes the user authentication problem. If unauthenticated messages are accepted, several risks might occur that go from denial of service to privacy or the loss of control of devices, e.g., in Internet of Things (IoT) contexts.
In the literature, several authentication mechanisms have been proposed, mostly operating at the application layer and using cryptographic approaches. Here, we exploit the propagation characteristics of the physical channel as a signature of the communication link or the transmitting device, in what is known as PLA. In [63], the basic approach is introduced: it consists of two phases, i.e., the identification acquisition and the identification verification phases. In the first phase, the receiver Bob (verifier) estimates the channel from signals transmitted by Alice (the authentic source). Higher-layer mechanisms, e.g., based on cryptography, are used to authenticate the signals. In the second phase, whenever Bob receives a new message, he also estimates the channel over which the transmitted signal has propagated and compares this estimate with that in the first phase. If the two are consistently similar (considering that they are both affected by noise), the received message is stated as authentic; otherwise, it is assumed to be fake. Several technologies, including

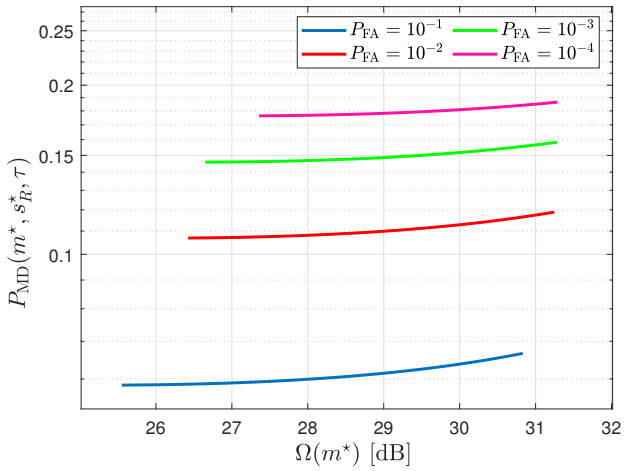


Figure 15.3: $\overline{P}_{\rm MD}(m^\star, s_R^\star, \tau)$ as a function of $\Omega(m^\star)$ for $\overline{P}_{\rm FA} \in \{10^{-4}, 10^{-3}, 10^{-2}10^{-1}\}$, number of RIS elements $N = 100$, and noise power at Bob $\sigma_B^2 = 0.6$

OFDM, MIMO, and underwater acoustic communications, employed PLA, using different testing techniques, from Neyman-Pearson tests to machine learning approaches.
Recently, the controllable nature of wireless channels provided by new communication technologies has been exploited for further improvement of PLA. Specifically, the propagation of wireless signals can be modified using RISs, i.e., controllable devices, where the phase shift introduced by each element can be changed. Indeed, when the verifier controls the RIS, he can set a random configuration of the RIS which remains secret to the attacker, and verify that the channel estimated from a received message corresponds to the predicted channel according to the set configuration. Such an approach provides a *CR-PLA* mechanism, where the random configuration is the challenge and the predicted channel is the expected response. Such an approach can also be applied when other *controllable channels* are available, e.g., when Bob is a drone that changes its position to pose a challenge.

15.3.2 Proposed Methodology

In this paper, we aim to design the random RIS configuration of the CR-PLA mechanism. We focus on the simple scenario where both the legitimate transmitter and the verifier are equipped with a single antenna, and the number of elements in the RIS is large. First, we observe that the random RIS configuration affects the data rate of the communication link between the UE and BS. In particular, increasing its randomness yields in general a lower MD probability while also lowering the communication performance. To measure the communication performance we consider the SNR averaged over the random RIS configuration. Then, we consider a GLRT at the verifier to make the decision about the authenticity of the message and analyze the performance of the CR-PLA scheme in terms of both FA and MD probabilities. Lastly, we design the probability distribution of the randomly selected phase shifts that maximize the average SNR under an upper bound constraint on the MD probability for a desired FA probability. In particular, we identify two statistical properties (represented by two real numbers) that capture the effects of the probability density function (PDF) on both the communication and the security metrics, so that the PDF design problem boils down to the optimization of these two parameters, under other constraints. In the design, we consider the worst-case scenario for the defense, by assuming that the attacker has complete channel knowledge, which is a challenging condition in practice.

15.3.3 Numerical Results

As an example of the obtained results in [67], we report here the results obtained to find a tradeoff between the signal-to-noise-ratio (SNR) Ω and the average misdetection probability $\overline{P}_{\mathrm{MD}}$. Fig. 15.3 shows the $\overline{P}_{\mathrm{MD}}(m^\star, s_R^\star, \tau)$ as a function of $\Omega(m^\star)$. In particular, we consider different values of the false alarm probability $\overline{P}_{\mathrm{FA}}$ in the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, $N = 100$, and $\sigma_B^2 = 0.6$. It can be seen as for a desired $\overline{P}_{\mathrm{FA}}$, a higher $\Omega(m^\star)$ comes at the expense of a higher $\overline{P}_{\mathrm{MD}}(m^\star, s_R^\star, \tau)$. Furthermore, the smaller the desired $\overline{P}_{\mathrm{FA}}$ is, the smaller the reduction of $\Omega(m^\star)$ in dB would be, if the minimum possible $\overline{P}_{\mathrm{MD}}(m^\star, s_R^\star, \tau)$ is assured. However, the smaller the $\overline{P}_{\mathrm{FA}}$ is, the smaller the minimum $\overline{P}_{\mathrm{MD}}(m^\star, s_R^\star, \tau)$ that can be ensured.

15.4 Energy-Based Optimization of Physical-Layer Challenge-Response Authentication with Drones

Drones are expected to be used for many tasks in the future, requiring protocols to ensure that they operate in a secure state. In [68], UNIPD proposes a novel PLA-based CR protocol in which a drone Bob authenticates the sender Alice with the presence of a malicious agent Eve by exploiting his prior knowledge of the wireless channel statistic (fading, path loss, and shadowing). In particular, Bob will move to a set of positions on the map, and by estimating the attenuations of the received signals he will authenticate the sender. Considering the energy consumption in the design, we provide three solutions of our protocol that differ in performance and computational complexity: we propose a purely greedy solution (PG), an optimal Bellman iterative solution (BI), and a heuristic solution based on the evaluation of the standard deviation (STD) on the map. Finally, we demonstrate the effectiveness of our approach through numerical simulations.

15.4.1 Background and Motivation

Today, drones are being used for various tasks such as precision agriculture and disaster management. Moreover, the integration of drones in machine-to-machine communication is helpful in a variety of contexts, such as the Internet-of-Drones (IoD) and non-terrestrial networks (NTN). However, these capabilities also make drones possible targets of attacks, including, for instance, spoofing to disrupt the drone's navigation system, and jamming as a denial of service.
We address the problem of drone authentication, where a drone, Bob, communicates with a transmitter, e.g., on the ground, which we call Alice, while a third-party agent, Eve, aims to impersonate Alice and send fake malicious messages to Bob, e.g., to convince him to land in a certain area. The goal of Bob is to verify the authenticity of the sender and distinguish Alice from Eve.
A classical authentication strategy is the (cryptography-based) CR protocol, which is based on a secret key shared between Alice and Bob. In this protocol, the verifier sends a request, called a *challenge*, which only a legitimate user with a valid key can correctly answer. Here, we consider PLA-based CR solutions. The key component of CR-PLA is the availability of a *partially controllable channel*: the verifier Bob authenticates Alice by manipulating the channel and verifying that the received signal is consistent with the expected change. The change induced by Bob corresponds to the *challenge* of crypto-based CR. On the other hand, since the change in the channel is decided on the fly by the verifier, it is difficult for the adversary to predict Bob's manipulations and guess the expected channel.

15.4.2 Proposed Methodology

We propose a CR-PLA protocol for drone authentication, where the controllable parameter used as the challenge is the drone's position since moving the drone implies changing the path loss of the wireless link.

Even a sophisticated Eve (which can freely tune its transmission power and pre-compensate the Eve-Bob channel) that does not know the current drone's position has high uncertainty on how to manipulate the channel, thus the authentication procedure is secure. This means that the attacker's probability of success depends on the variability of shadowing over space since an attenuation that does not vary with Bob's position can be easily guessed by the attacker. Thus, we use the shadowing effect as a source of randomness. In particular, Bob checks whether the received response matches the distribution of attenuation due to shadowing given his position. Moreover, since more positions can be associated with the same path loss and shadowing (i.e., the challenge), we also propose an energy-saving CR-PLA strategy that minimizes the long-term energy expended by Bob without sacrificing the security of the protocol.

In [66], the authors proposed a CR-PLA protocol for drone communication. However, the channel was only partially modeled (e.g., no shadowing was considered), and the problem of energy minimization was not addressed.

We consider the scenario of Fig. 15.4, where a drone (agent Bob), is communicating with a ground device (agent Alice) while moving on a gridded region of points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, where $\boldsymbol{x}_i$ is the coordinate vector of position i. The drone receives several messages from the ground device, including those for navigation and the instructions required by the mission (e.g., taking pictures). We aim to provide an authentication mechanism to ensure that the drone processes only messages coming from the ground device, rather than from an attacker impersonating the ground device. In turn, an attacking agent, Eve, aims to impersonate Alice by sending messages to Bob purposely designed to be confused with those of Alice. Such messages aim for example to detour the drone from its designed route.

Transmissions are narrowband and model the channel gains are obtained as the result of (free space) path-loss, shadowing, and fading phenomena, which properly capture the main components of wireless propagation. Note that, as we are considering shadowing and path-loss as authentication features, our solution is independent of the number of used antennas and devices.

For CR-PLA, as described in the following, Bob needs to estimate the channel over which the receiver message went through. Several well-known strategies can be used for such a task, e.g., via least squares estimation. The complex gain of the baseband equivalent channels Alice-Bob and between Eve-Bob are denoted as h and g, respectively. However, Bob does not know whether the estimated channel γ he has just obtained, is h or g. From the estimate γ, Bob computes the attenuation a, and that includes the free space path-loss (modeled by the Friis formula), shadowing, and fading. We also assume the transmit power of Alice and antenna gains of the legitimate agents are publicly known.

Shadowing includes the effects of obstacles placed between the transmitter and receiver. The shadowing term depends on the location of the transmitter and the receiver, and channel gains of couples of transceivers in proximity have a high correlation. To model such a phenomenon, we resort to the well-known Gudmundson [69] model of the correlation of the shadowing components for two receivers at a given distance. Fading accounts for shorter channel variations and is affected by Doppler spread and multipath. As it is hardly

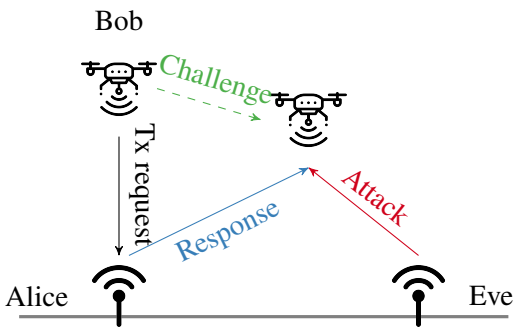


Figure 15.4: Example of CR protocol.

predictable, it cannot be exploited for our authentication scheme. Shadowing instead is shown to have a higher coherence distance. Moreover, it varies only slowly over time (i.e., slow fading). Thus, we will exploit the shadowing for the CR protocol while considering fading as estimation noise.
Concerning the attacker, we assume Eve can alter her transmission power to match any desired channel. On the other hand, we assume the drone position to be secret until signal reception, i.e., the attacker does not know Bob's position while designing the spoofing signal.
Since the protocol requires the drone to move, we model the energy required to move following the results in [70].

15.4.3 Numerical Results and Analysis

We modeled a scenario where a drone moves around a plane area of size $50\,\mathrm{m} \times 50\,\mathrm{m}$, at a height of $20\,\mathrm{m}$ from the center of the plane area where Alice is. We consider the drone to be always in LoS, and the free space path loss is modeled via the Friis formula. The area is sampled using a step size of $A = 1\,\mathrm{m}$ along both directions, ending up with a total of $N = 2500$ sampled positions. The considered carrier frequency is $f_0 = 1.8\,\mathrm{GHz}$. The shadowing has been simulated considering coherence distance $D_{\mathrm{coh}} = 10\lambda$ with standard deviation $\sigma_\xi = 6\,\mathrm{dB}$.
First, we assess the security performance of the protocol. We set the noise variance to $\sigma_w = 0.02, 0.05, 0.1$, and $0.2\,\mathrm{dB}$, $|\mathcal{R}| = 10\,\mathrm{dB}$, and, for each σ_w, we run $10,000$ simulations for legitimate and under attack case. Fig. 15.5 compares the DET curves obtained for different values of σ_w considering both simulation and the analytical derivation (B.13). We confirm the validity of our model, as the analytical model and simulation results almost perfectly match. As expected, when σ_w increases, it becomes harder for the defender to distinguish legitimate from attacker transmission, thus the p_{md} increases.
Next, we evaluate the results of the energy minimization policies. Concerning the parameters of such solutions, for the BI we fixed the discount factor $\gamma = 0.95$; for the STD-based solution, we consider instead windows $W(\boldsymbol{x})$ of size $L = 5$, normalization factor $\alpha = 100$, and decaying factor parameter $\beta = 20$.
Fig. 15.6 shows the statistics of the energy spent by the drone, computed over 1000 simulation runs, after random initialization, obtained using the PG, the BI, and the STD-based approach. As it can be seen from the magnification, on the first movements the greedy policy requires on average (slightly) less energy. However, after just a few steps, both BI, and the STD-based solutions start to outperform the PG solution, with a gap that increases over time. As expected, the best performance is achieved by the BI. Still, the gap between the optimal BI and the STD-based approach is limited. Thus, in scenarios where the Markov decision process (MDP)-based solution is not computationally feasible, it is reasonable to resort to the heuristic STD approach.

15.5 Challenge-Response to Authenticate Drone Communications: A Game Theoretic Approach

As drones are increasingly used in various civilian applications, the security of drone communications is a growing concern. In this context, in [71] we present strategies for CR-PLA of drone messages. The ground receiver (verifier) requests the drone to move to a defined position (challenge), and authenticity is verified by checking whether the corresponding measured channel gain (response) matches the expected statistic. In particular, the challenge is derived from a mixed strategy obtained by solving a zero-sum game against the intruder, which in turn decides its own positions. In addition, we derive the optimal strategy for multi-round authentication, where the CR-PLA procedure is iterated over several rounds. We also consider the energy minimization problem, where legitimate users want to minimize the energy consumption without compromising the security performance of the protocol. The performance of the proposed scheme is tested

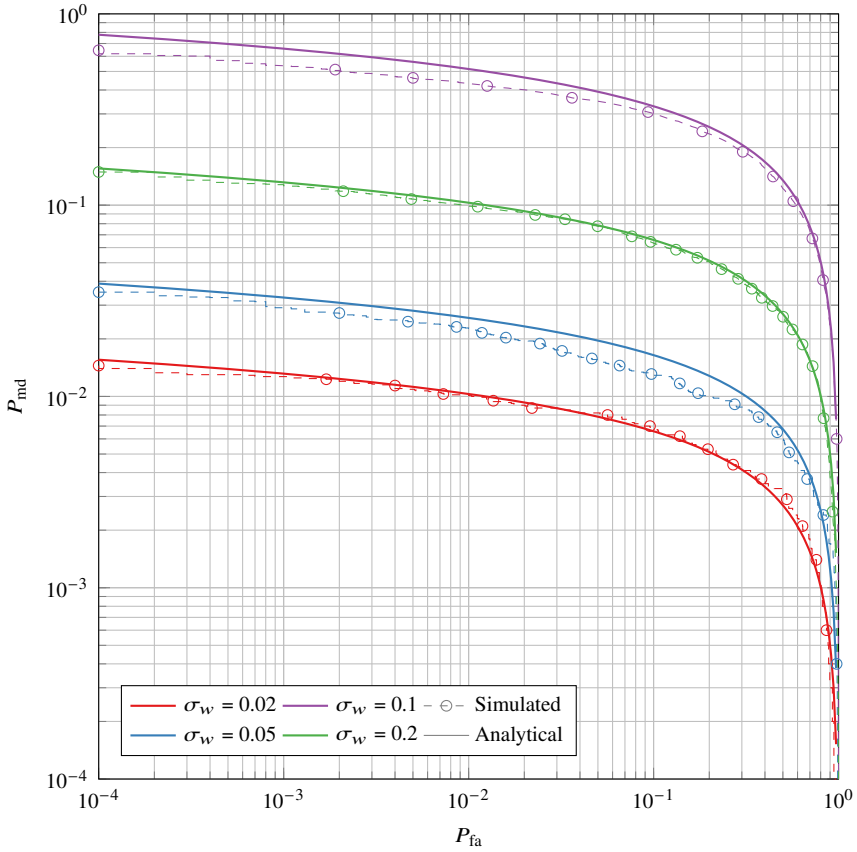


Figure 15.5: DET curves of the proposed authentication procedure for different σ_w values. Analytical (continuous) vs simulated (circles mark, dashed).

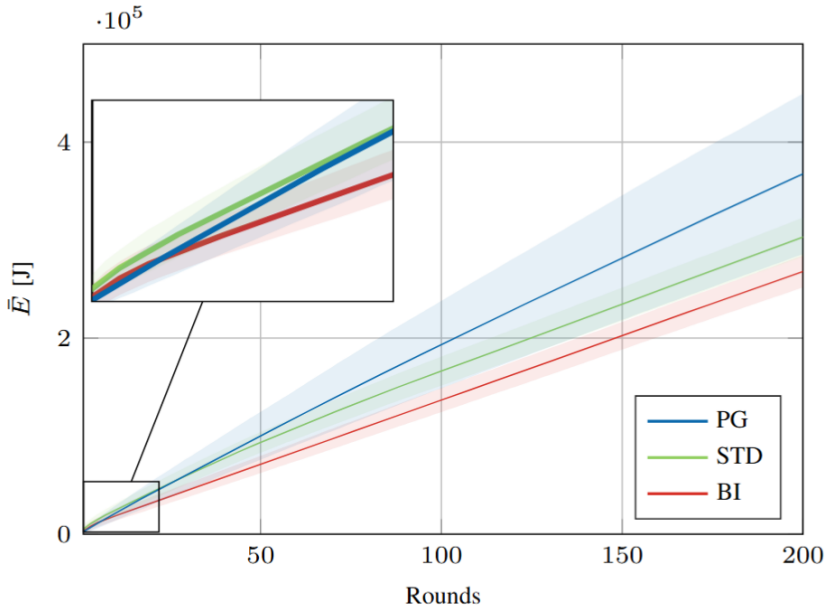


Figure 15.6: Energy spent over time, using the PG (blue), the BI (red), and the STD-based approach (green). Average (continuous line) plus/minus standard deviation (shaded area), computed over 1000 simulations.

in terms of both security and energy consumption through numerical simulations, considering different protocol parameters, different scenarios (urban and rural), different drone altitudes, and also in the context of drone swarms.

15.5.1 Background and Motivation

The use of drones has rapidly increased over the last few years. Starting as a military tool, they are now used in many civil applications such as precision agriculture, environmental monitoring, and disaster management and relief. As drones become integrated into more complex systems, security is a rising concern. Among the major threats, the transmission of jamming or spoofing signals is particularly dangerous as these can disrupt the navigation system and jeopardize the drone's mission.
This contribution addresses the problem of authenticating messages transmitted by drones. In particular, a drone or ground device (Bob) wants to authenticate messages potentially transmitted by a legitimate drone (Alice) and distinguish them from those transmitted by an *intruder* drone (Trudy) attempting to impersonate Alice. Cryptography-based authentication solutions have several drawbacks: a) they are typically computationally expensive, b) they often provide only computational security, which may be vulnerable to quantum computing-based algorithms, and c) they introduce significant communication overhead. Both communication overhead and computational complexity lead to high energy consumption, which is a limited resource for drones. Therefore, in this work, we focus on PLA mechanisms that provide security by exploiting the statistical properties of the channels. These techniques typically consume less energy than their crypto-based counterparts while providing information-theoretic security.
PLA has recently been studied, and mechanisms specifically targeting drone communication may rely on fingerprinting or on channel characteristics, eventually supported by a federated learning architecture. Here we focus on a recent evolution of PLA using a CR approach. In cryptographic CR authentication, Alice and Bob share a secret key. Then, Bob sends a message called a *challenge* over a public channel, and Alice computes and sends back to Bob a *response* obtained using a hash function of both the secret key and the challenge. Finally, Bob verifies the authenticity of the sender of the response by comparing the response to a local response obtained using the same secret key and challenge. Instead of relying on crypto-based solutions, we consider CR-PLA. CR-PLA is based on *partially controllable channels*, where the challenge is issued by the verifier Bob by manipulating the propagation environment, while the response is the received signal from Alice, which must be consistent with the expected change. Since the challenge, i.e., the propagation conditions, is randomly chosen by the verifier, it is difficult for Trudy to predict it and transmit a signal that is consistent with it.

15.5.2 Proposed Methodology

In this paper, we propose novel strategies for Bob and Trudy when using CR-PLA in drone communication. In particular, in a preliminary phase, Bob measures the channel gain when Alice is at a set of predefined positions. Then, the authenticated transmission protocol requires that Bob first randomly select a set of positions (with a suitable distribution) and secretly communicate them to Alice. Alice goes to the indicated positions and, for each of them, she transmits a pilot signal together with the message to be authenticated. Next, Bob assesses the authenticity of the received signal by checking that the measured channel gains correspond to the expected ones, estimated during the preliminary phase. In this context, the challenge is represented by the set of positions, and the corresponding response is the set of channel gains estimated by Bob. In turn, for her attack, Trudy randomly selects (with a suitable distribution) a set of positions from which to transmit the pilot signal and her message. The distributions used by Bob and Trudy to select the position sets are optimized to their advantage by finding the Nash equilibria (NEs) of a zero-sum game. In addition, we also consider the problem of minimizing the energy of Alice's movements by proposing both

optimal and heuristic strategies to minimize the (average) distance traveled by Alice without sacrificing the security of the protocol.
The contributions are as follows

- We model the channel between the drone and the receiver, including both the preliminary channel estimation phase and the security protocol.
- We design the statistical distribution of positions generated by Bob and Trudy by modeling the problem as a zero-sum game between legitimate users and Trudy, where the payoff is the MD probability for a target FA probability.
- We consider both optimal and low-complexity solutions for optimizing Bob's position selection statistics.
- We test the proposed technique by numerical simulation, based on a realistic model of both Alice-Bob and Trudy-Bob channels, including shadowing effects in urban and rural scenarios.

15.5.3 Numerical Results and Analysis

We now investigate the effects of several parameters of the proposed technique on its performance.

Shadowing variance $\sigma_{(s)_{\text{dB}}}$ First, Fig. 15.8 compares the security performance in environments characterized by different shadowing variances. The figure shows the DET curves, for $N = 1$, $\sigma_{(s)_{\text{dB}}} = 6$, 10, 13, and 16 dB. Lines are obtained with the closed-form formulas of the probability, while black markers show the results obtained with Monte Carlo simulations. We note a perfect agreement between the analytical and numerical results. When comparing the different variances of the shadowing, we note that the MD probability decreases with $\sigma_{(s)_{\text{dB}}}$ (for a fixed FA probability). This is because a high $\sigma_{(s)_{\text{dB}}}$ value increases the map diversity, i.e., positions at the same distance from the receiver have different gains, thus it is harder for Trudy to guess a position leading to the same gain as Alice to fool Bob.

Number of Pilot Symbols for Channel Estimation K About the impact of the number of pilot symbols used for channel estimation in the authentication phase of CR-PLA, Fig. 15.7 shows the DET curves of the proposed solution for $N = 1$ round and $K = 50$, 70, and 100 pilot symbols. We see that a higher K yields better channel estimates and thus lower MD probabilities. However, the benefits are limited, indicating that it may be possible to achieve a good security performance even with low K values, thus potentially saving drone energy and time.

Number of Rounds N We now consider the impact of the number of rounds N on the security performance. Fig. 15.9 shows the MD probability as a function of rounds N for FA probabilities 10^{-2}, 10^{-3}, and 10^{-4}. From the results, we see that the MD decreases exponentially with the number of rounds; thus, we can easily reduce the MD probability to desired values by adding a few more rounds (at the cost of increased consumed energy).

15.6 Integration with the Architecture

This Chapter has delivered a comprehensive examination of attacks targeting Physical Layer Authentication (PLA) mechanisms in a 6G network that leverages Reconfigurable Intelligent Surfaces (RIS), as well as the corresponding mitigation strategies. RIS technology is pivotal for both current and upcoming communication

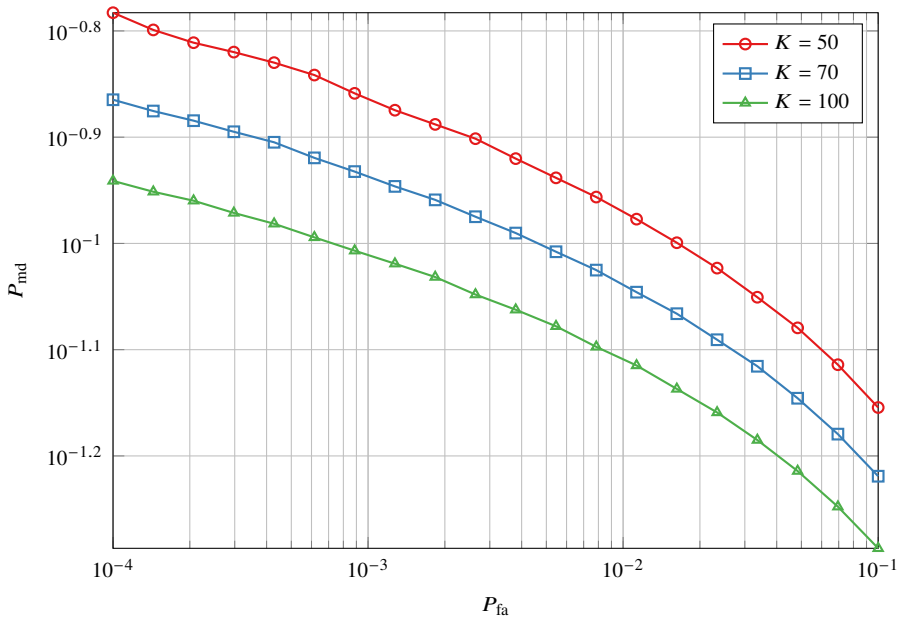


Figure 15.7: DET curves for different numbers of pilot symbols in the authentication phase, K.

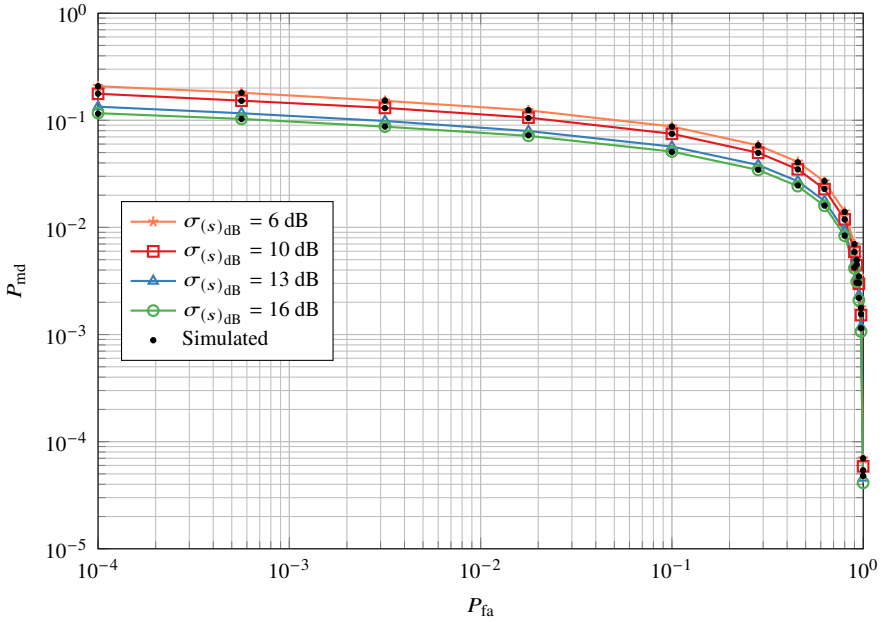


Figure 15.8: new pmd
DET curves for $\sigma_{(s)_{\text{dB}}}$ = 6, 10, 13, and 16 dB for simulated and theoretical analysis.

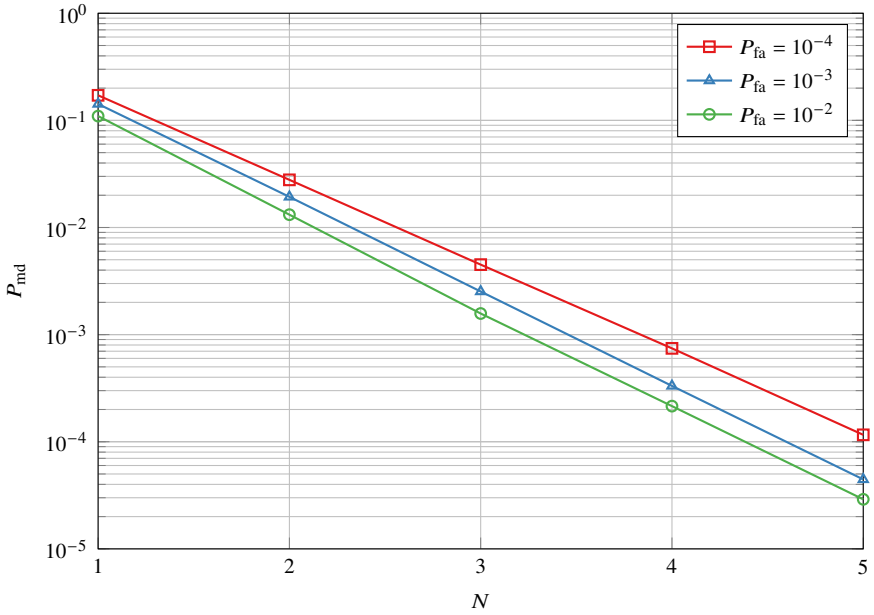


Figure 15.9: P_{md} as a function of the number of rounds, N, for $P_{\text{fa}} = 10^{-2}$, 10^{-3}, and 10^{-4}.

systems, enhancing coverage and connectivity. In this work, we also explore using RIS to bolster PLA through an innovative challenge-response scheme. The resulting authentication protocols belong to the PHY-Attack Identification block, which scrutinizes the received signal to verify the legitimacy of each message. Portions of these solutions will be incorporated into CUPD04 as an advanced authentication option, surpassing the capabilities of conventional PLA.

Chapter 16

Secret Key Generation On Aerial Rician Fading Channels Against A Curious Receiver

Secret key generation at the physical layer is expected to be a fundamental enabler for next-generation networks. In [72], UNIPD considers a network where the user equipment is a drone and proposes a novel secret key generation solution when the eavesdropper is another node belonging to the network (curious device). We exploit drone mobility over realistic Rician fading channels. In our protocol, after a prior training phase, drone Alice chooses a trajectory of positions in space and transmits a message to Bob, on the ground, from each position. From the received messages, Bob estimates the channel gain from which a secret key is extracted. The choice of the positions is made to maximize a lower bound on the secret key capacity. Numerical simulations are used to prove the effectiveness of the proposed approach.

16.1 Background and Motivation

SKG at the physical layer is a mechanism that enables two devices to agree (or refresh) a stream of bits (key) that remains secret to other eavesdropping devices. Such a key can then be used to support cryptographic techniques, e.g., to achieve confidential transmissions, and provide authentication mechanisms. Two main approaches are available for SKG at the physical layer: the source-based SKG, where the randomness to generate the key is provided by the channel over which communication occurs; the channel-based SKG where instead one of the two parties transmits a random key that is kept secret from the eavesdropper as its channel does not allow to infer the key properly (e.g., the attacker channel is more noisy than the legitimate one). Here, we focus on the channel-model technique.

Conventional source-model SKG mechanisms at the physical layer typically exploit two characteristics of the wireless channel: reciprocity and fading. Usually, the former is guaranteed by both the reciprocity theory for antennas/electromagnetic propagation and the synchronization of the devices whose communication must occur within the coherence time of the channel. The latter, on the other hand, is often guaranteed by random reflections of the signal in the environment.

However, with poor scattering and/or slow fading, typical for example of unmanned-aerial-device (UAV) applications, SKG is challenging, as the LoS component might be dominant, thus the channel can be, in the worst case, deterministic. Few recent studies tackle SKG in UAV contexts; in particular, they exploit MIMO and three-dimensional (3D) spatial angles to extract keys in LoS environments or inject randomness into the transmitter, thus creating an artificial frequency-selective fading channel. Still in the context of single-antenna devices, these approaches cannot be applied as the angle of arrival cannot be estimated.

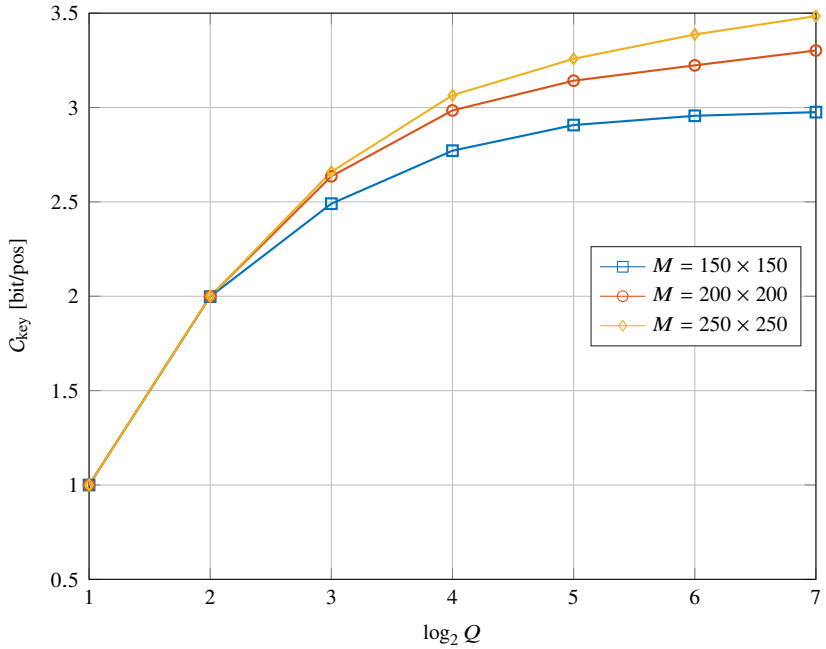


Figure 16.1: SKC vs the number of quantization levels Q (in log scale), with different map sampling spaces.

16.2 Proposed Methodology

In this paper, we consider a realistic Rician fading channel for UAV transmissions and propose a novel solution that exploits the mobility of the UAVs to perform SKG at the physical layer, allowing a robust yet effective key agreement in low-scattering environments in the presence of a curious device. In particular, the eavesdropping device Eve belongs to the same network and has a fixed position on the ground, known to the other devices. In detail, our protocol starts with drone Alice moving in space to gather information on the environment, more precisely, on the channel gain in a pre-defined grid of possible positions. In the next phase, Alice chooses a trajectory of positions, moves there, and transmits a message to Bob, on the ground, which will estimate the channel gain and then quantize it to a fixed number of levels Q. The choice of the positions is made to maximize the secret-key capacity (SKC).

16.3 Numerical Results and Analysis

16.3.1 Map Geometry and Number of Quantization Levels

We see from Fig. 16.1 that by increasing the number of quantization levels Q, the secret key capacity first increases and then saturates. This justifies our design decision to quantize the estimated gains at Bob and work with discrete levels. The same is true for the number of map positions M: a higher number of points on the map improves the performance, yet oversampling the space would result in sampling the same gains, leading to a saturation of the key entropy at Bob.

16.3.2 Shadowing Variance and Eve-Bob Distance

Fig. 16.2 shows that higher shadowing STD at Bob leads to higher capacity; this is because, given a trajectory, the resulting gains at Bob have higher entropy. Finally, the proximity of Eve to Bob worsens the performance because all the trajectories lead to similar gains to Bob: if Eve were exactly in Bob's position,

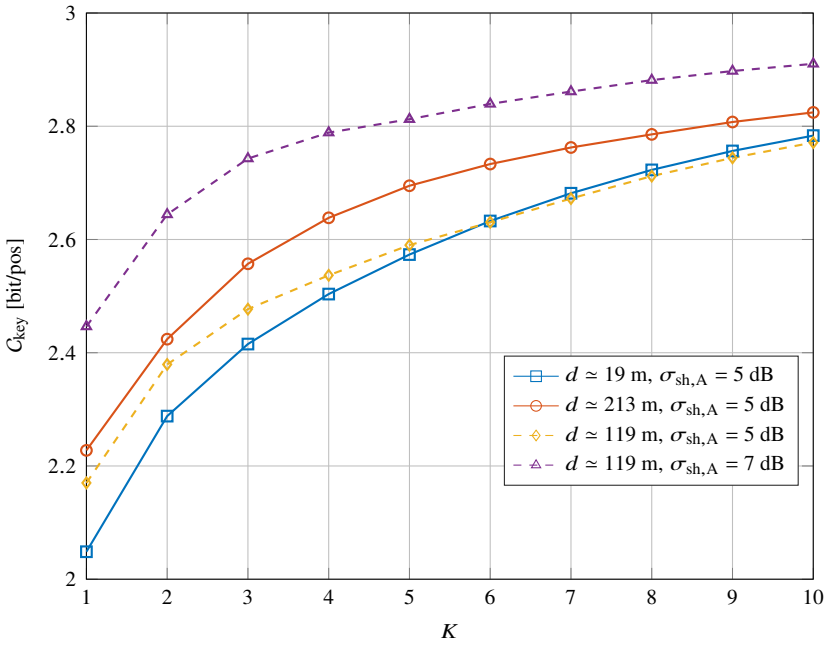


Figure 16.2: SKC as a function of the number of pilot symbols K, with different Alice-Bob channel shadowing STD $\sigma_{sh,A}$ and Eve-Bob distance d.

all the trajectories would lead to the same gain also to Bob, and the mutual information between him and Alice would be zero as well.

16.4 Integration with the Architecture

We introduce an innovative approach for generating secret keys by employing drones and leveraging the fading properties of the wireless channel. This solution belongs to the PHY Components block of the ROBUST-6G architecture, as it defines actuation mechanisms whereby drones alter their trajectories to facilitate key-generation processes. The method will be integrated into a forthcoming second version of the CUPD04 component.

Chapter 17

Adversarial Attacks on ISAC Systems

ISAC is expected to be a key enabler for next-generation networks, posing unprecedented security issues. In this paper, we evaluate the security of ISAC systems under adversarial ML attacks. In particular, at UNIPD, we have studied a scenario where Alice and Bob cooperate to perform bistatic sensing of the environment. As the scatterers are located in different regions, Alice and Bob can obtain a coarse estimation of the scatterers' locations by classifying to which area the received signal belongs. On the other hand, the attacker Trudy aims at disrupting such a procedure by properly designing her transmitting beamformer to fool Bob, and make him estimate a target region. We evaluate the effectiveness of the proposed attack via numerical simulations.

17.1 Background and Motivation

ISAC systems emerge as a cornerstone technology for the sixth-generation era, seamlessly incorporating sensing functionality into wireless networks as a native capability. The object's localization capabilities of such a technology are of crucial interest, as the ability to monitor physical factors is crucial for optimizing the network's performance, enhancing security, driving automation, and carrying out other vital tasks. Still, this technology poses unprecedented security and privacy issues. In the literature, spoofing attacks are of particular interest, as the attacker can perform beamforming and disrupt the sensing phase of ISAC systems, and often ML strategies are adopted for tackling such issues.

17.2 Proposed Methodology

In this contribution, whose extended version is in Appendix D, we have Alice and Bob cooperating to perform bistatic sensing of the environment, while Trudy aims at disrupting such a procedure by properly designing her transmitting beamformer to fool Bob. In particular, the contributions are as follows:

- We model a realistic ISAC channel, using a geometrical channel model that takes into consideration the scatterer's location.
- We train a standard CNN to classify the received signal into the area where the scatterers are located.
- We design a projected gradient descent (PGD) attack that Trudy can perform to induce her desired classification area, taking into account the required transmitting power.
- We numerically evaluate the attack, demonstrating its effectiveness

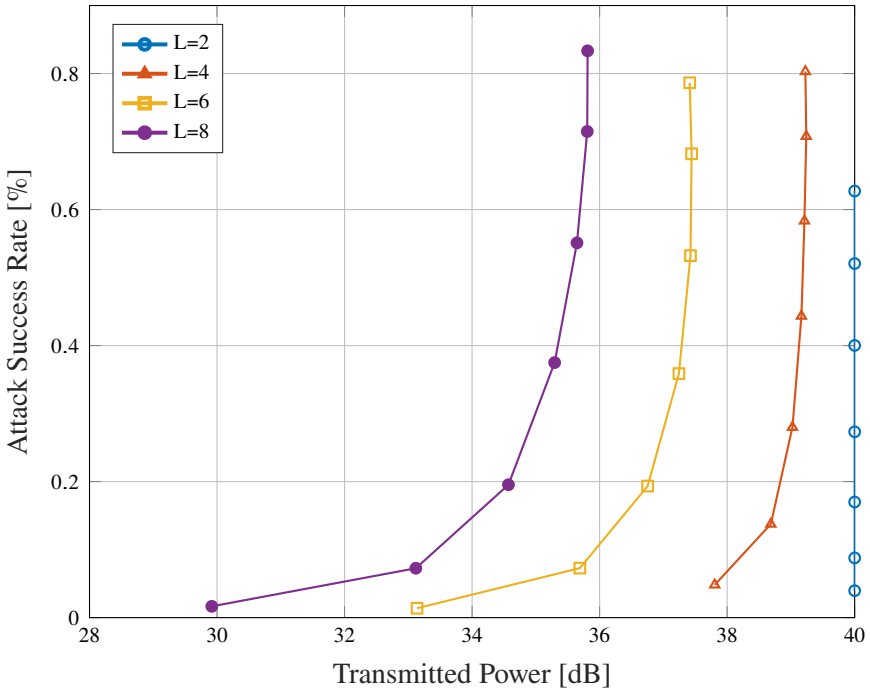


Figure 17.1: Attack Success Rate as a function of the transmitted power, for different numbers of scatterers L.

17.3 Numerical Results

The scenario we simulated has $N = 5$ squared areas with size $S = 30\,\text{m}$. We used a number of antennas of $N_T = N_R = 4$, a maximum transmitter power by Trudy of $P_{\text{max}} = 40\,\text{dB}$. The target area for Trudy is the region $\tilde{a}_n = 2$, when the true region is $a_n = 1$.

17.3.1 Attack Success Rate VS Number of Scatterers

We see from Fig. 17.1 that by increasing the transmitted power P_{T}, Trudy can perform more effective attacks, reaching an attack success rate, i.e., an accuracy on the target class, of $\geq 80\%$ when the number of scatterers $L \geq 4$. We also notice that the performance is extremely dependent on the number of scatterers L: in fact, the greater L, the easier it is for Trudy to find a beamforming matrix $\boldsymbol{W}$ to fool Bob. This effect relies on the rank of the cascaded channel $\boldsymbol{Z}^{(m)}$: in fact if $L \geq N_R, N_T$ then $\boldsymbol{Z}^{(m)}$ becomes invertible, thus it easier for the attacker to find the optimal beamformer. This effect is particularly evident with $L = 2$: in that case $\boldsymbol{Z}^{(m)}$ has at most two non-zero eigenvalues, thus the optimal beamformer saturates at $P_{\text{max}} = 40\,\text{dB}$. Note also that in that case, multiple solutions are available to the attacker: for each maximum perturbation ϵ in, the attacker can find the beamformer that respects the power constraint. Another solution would be to modify the PGD algorithm by directly taking into account the power constraint into the solution $\tilde{\boldsymbol{Z}}$, and this is left for future works.

17.3.2 Attack Success Rate VS SNR

In Fig. 17.2, we observe that as the SNR increases, the required power for Trudy decreases, yet the results remain very similar. This effect can be justified by the fact that when the channels in input to the PGD are less noisy, it is easier for the algorithm to find suitable channels to fool Bob's model.

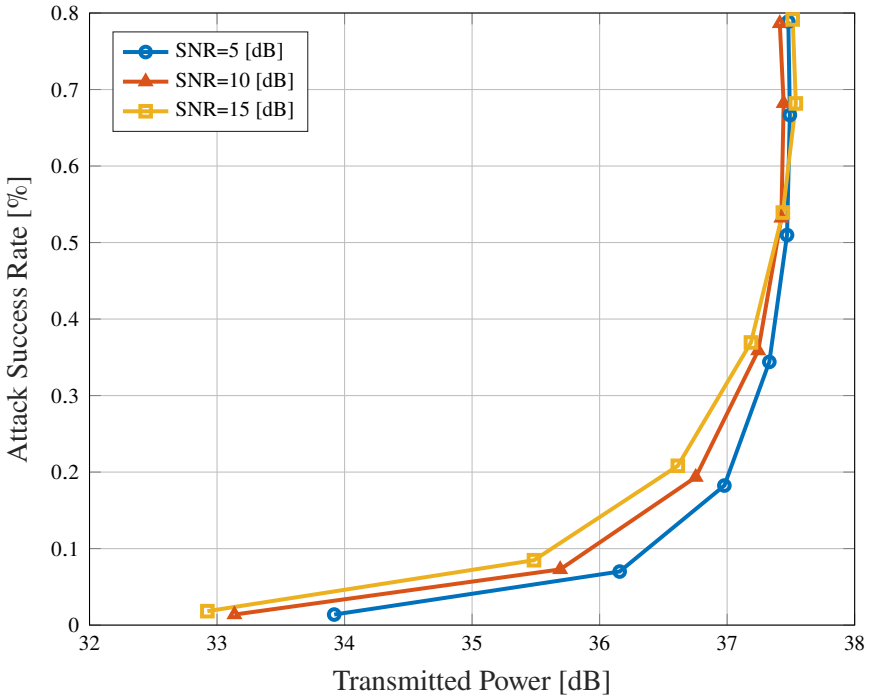


Figure 17.2: Attack Success Rate as a function of the transmitted power, for different numbers SNR levels.

17.4 Integration with the Architecture

This contribution fits with the Trustworthiness Evaluation part of the architecture, as it poses serious benchmark attacks based on adversarial ML to evaluate the robustness of ISAC algorithms.

Chapter 18

Impact of Residual Hardware Impairments on RIS-Aided Authentication

In this study [73], EBY examines how residual RHI influences PLA in RIS-aided communication systems in the presence of spoofing attacks. RIS technology, widely studied for 6G networks, enhances channel manipulation and provides additional spatial diversity; however, its interaction with hardware impairments in PLA mechanisms remains insufficiently explored. This study addresses this gap by analyzing how RHI at both transmitter and receiver nodes affect authentication performance when RIS is used to support legitimate users and suppress attacker signals.

18.1 Background and Motivation

PLA works by checking the stability of channel observations and how they change over time. Since the wireless channel carries unique patterns that are difficult for an attacker to copy, PLA compares consecutive CSI values to decide if the transmitter is legitimate or not. Because of this, the temporal behavior of the channel is very important for reliable authentication. RISs bring a new dimension to this idea. By controlling reflections in the environment, a RIS can change the propagation conditions and create extra diversity in the channel. In theory, this additional diversity can help PLA, because it makes the channel more distinctive and harder for a spoofing device to imitate. Therefore, using RIS together with PLA looks promising for improving system-level security. But in practical systems, the hardware is not ideal. Real devices include different imperfections that directly affect the received signal. When a RIS is added, these imperfections interact with both the natural channel and the RIS-generated paths, creating a compound channel with more complicated behavior. Although many studies focus on RIS-aided communication performance, the effect of hardware imperfections on RIS-based authentication has not been examined in detail. Because of this gap, it is not clear if the expected benefits of a RIS remain valid when the devices and reflections are influenced by non-ideal hardware.
For this reason, the study focuses on understanding how hardware impairments influence PLA when a RIS is part of the system. The goal is to quantify the impact on authentication reliability and to see if the extra diversity created by a RIS still improves detection performance in spoofing scenarios. This helps us understand the practical behavior of RIS-assisted PLA in realistic deployments and whether it can be trusted as a security enhancement in future systems.

18.2 Proposed Methodology

The aim of this study is to examine how RHI affects the performance of PLA in a RIS-aided communication scenario and to understand whether the additional channel diversity created by the RIS can still support reliable authentication under realistic hardware conditions. The study also aims to quantify how different RIS sizes, SNR values, and impairment levels influence the false alarm and miss detection behavior of the system. To achieve this aim, the proposed methodology constructs a detailed system model that includes a legitimate transmitter, an attacker, a receiver, and a RIS with adjustable phase elements. Since the direct path is blocked, both legitimate and malicious signals reach the receiver only through RIS reflections. The compound channel is modeled by considering fading, RIS phase adjustments, transmitter and receiver RHI, and noise. This provides a realistic basis for analyzing the impact of hardware non-idealities on PLA.
The methodology then applies a likelihood ratio test at the receiver, comparing the current estimated channel with the previous estimate. This exploits the autoregressive temporal stability of the legitimate RIS-aided channel, allowing the receiver to identify abnormal variations caused by spoofing attempts. Analytical expressions for the false alarm probability are derived, while the miss detection probability is evaluated numerically due to its non-tractable form. Finally, extensive simulations are performed for different RIS sizes, SNR conditions, and RHI levels to validate the analytical findings and to reveal how each parameter contributes to authentication performance. Through this methodology, the study aims to provide clear insights regarding when RIS improves PLA, when RHI becomes a limiting factor, and how system parameters should be selected for robust authentication.

18.3 Results

The numerical results clearly show that RIS assistance provides strong gains for PLA. As illustrated in Figure 18.1, increasing the number of reflecting elements leads to a noticeable reduction in the miss detection rate. For example, when the RIS has 32 elements, the miss detection rate decreases by more than 50 percent compared to a 4-element RIS at a fixed false alarm rate of 0.3. This shows that larger RIS surfaces create richer channel diversity, which improves the reliability of the decision mechanism.
The results also demonstrate that channel quality plays a major role in authentication performance. When SNR increases from -20 dB to -5 dB, the false alarm rate drops from 0.69 to 0.25 for a miss detection rate of 0.3, meaning that better channel conditions support more stable and reliable authentication. Impairment differences between Alice and Eve further contribute to distinguishing attackers. For instance, when Eve has impairment variance 2 while Alice has 0.25, the miss detection rate improves from 0.42 to 0.30 at a false alarm rate of 0.5.
Additionally, intelligent RIS operation yields a clear advantage. At a false alarm rate of 0.2, the miss detection rate drops from 0.5 under blind RIS operation to 0.33 with intelligent phase control. In contrast, systems without RIS assistance perform significantly worse, with miss detection levels more than double those of RIS-aided operation under identical SNR and impairment settings. These results confirm that RIS substantially strengthens authentication performance, and that impairment asymmetry provides additional discriminatory value against spoofing attempts.

Contribution to 6G Physical-Layer Security

The findings of this study directly contribute to emerging 6G PLS frameworks by demonstrating how RIS can be used not only for communication enhancement but also as an active authentication element that amplifies physical layer uniqueness. By analyzing hardware-impairment–driven channel variations and exploiting temporal CSI consistency, the proposed PLA mechanism enables secure and lightweight

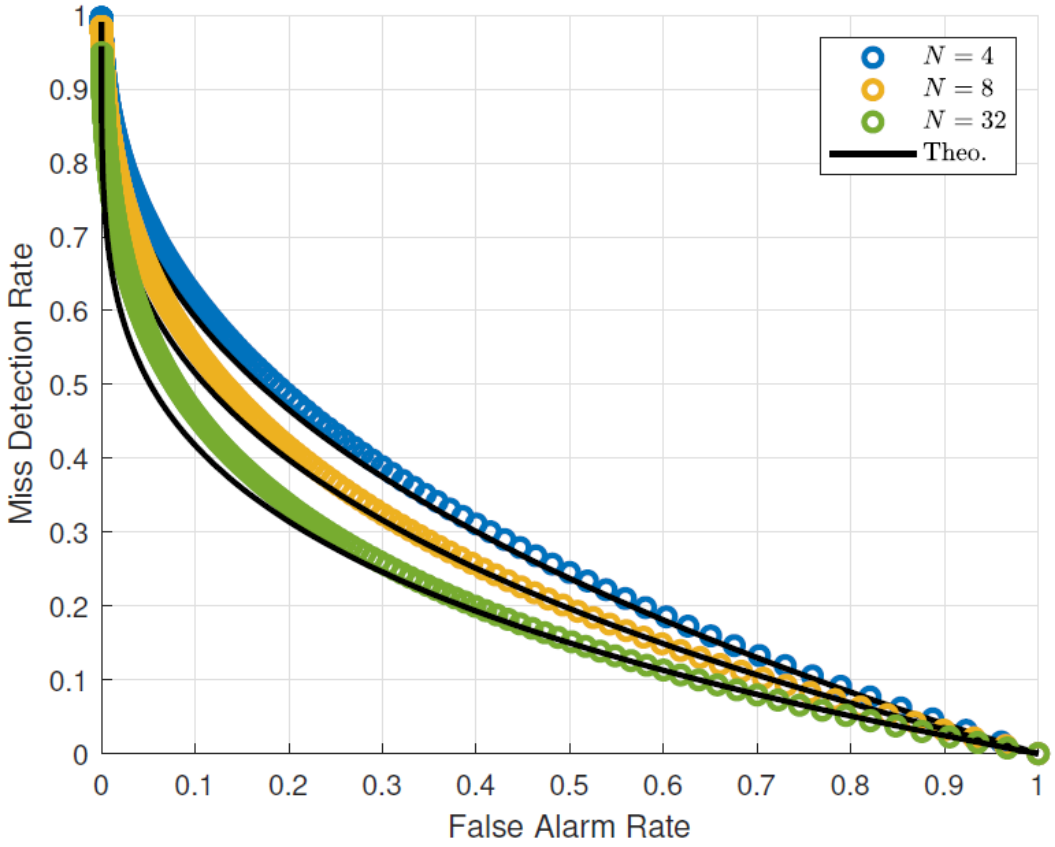


Figure 18.1: MDR vs. FAR analysis showing the impacts of different number of RIS reflecting elements where $\alpha = 0.9$, SNR = 0 dB, $\kappa_A = \kappa_B = 0.1$ and $\kappa_E = 0.4$.

authentication suitable for dense 6G IoT deployments. The demonstrated benefits of RIS, such as channel enrichment, improved discriminability, and robustness under varying RHI and SNR, align with 6G goals for endogenous security, zero-trust radio access, and real-time RIS-assisted physical layer threat detection. As 6G systems increasingly rely on RIS-enabled environments, the presented methodology offers a scalable and mathematically grounded approach for attacker detection and physical-layer integrity assurance.

18.4 Integration with the Architecture

This study is directly relevant to authentication and identification of legitimate devices (CEBY05) because it shows how RHI together with RIS-induced channel diversity creates distinctive physical-layer patterns that separate genuine transmitters from spoofers. The legitimate Alice–Bob link follows a stable autoregressive behavior, while Eve's channel varies unpredictably, allowing identity verification through consecutive CSI measurements. The results demonstrate that RIS strengthens this contrast and reduces miss detection and false alarms. Since device impairments create unique distortions that cannot be perfectly copied, especially after RIS reflections, the method becomes a reliable way to detect impersonation and confirm that the received signal comes from the legitimate user.

Part III

T5.3 - Trustworthiness of 6G PHY and Enabling Trust Building in 6G Autonomous Agents

Chapter 19

6G PHY Trustworthiness

19.1 Background and Motivation

In this chapter, we synthesized five contributions from ENSEA and CYU [12–16] on the role of the physical layer in 6G trust and trustworthiness. Across all works, we observed a common thread: future cyber–physical systems (CPS) and multi-agent networks will require *objective, quantifiable measures of trust*, deeply embedded in the wireless substrate itself. We made the case that joint AoA and ToF sensing emerge as crucial primitives that enhance trust, integrity, and accountability of autonomous devices.

In [12], we presented a panoramic view of the threats and opportunities emerging in 6G networks of cyber-physical system (CPS). We highlighted that future networks will not merely carry data; instead, they will mediate complex multi-agent tasks enabled by sensing and distributed computing, where malicious actors could disrupt functionality beyond classical attacks such as denial of service (DoS), eavesdropping, or link dropping. A key insight from this extensive review was the call for more *quantifiable measures of trust* that go beyond binary authentication and classical reputation models. We argued that trustworthy multi-agent systems would require a multi-dimensional trust quantification pipeline, grounded in sensing and physical measurements, and accounting jointly for cyber as well as physical behaviors.

In this sense, the traditional decoupling between application and communication layers was argued to be no longer viable for evaluating trust in CPS. Future 6G systems should instead support a tight interaction between the physical world and the decision-making processes acting upon it. We emphasized that the "physicality" of 6G networks—sensing, perception, and channel engineering—provided unique opportunities for trust: wireless channels inherently encoded physical constraints on mobility, geometry, and propagation. Certain of these physical features have been shown to be unforgeable, as discussed in Chapter 3, making physical-layer trust anchors essential for 6G CPS. Physical signals thus serve as *first-hand witnesses* of agent behavior. These concepts tie directly to the trust architectures we proposed in [13], in which different lower and upper layer data fusion models were discussed, along with a toy PoC demonstration on how positioning can allow identifying mis-information attacks in VANETs.

Based on the above, we made the case that AoA and ToF could enable trust via continuous physical integrity checks, anomaly detection through physical motion consistency, and authenticated locations. The next key question that needed addressing was the achievable accuracy of AoA and ToF estimation. In [14] we evaluated the accuracy in joint AoA and ToF based authentication on a real dataset, while in [15], aspects of sensing accuracy enhancement were revisited. Furthermore, in [16] we exemplified how physical context awareness could be incorporated in resource allocation for future, trustworthy networks, starting with the current fifth generation.

In [74], we proposed a joint optimization of privacy-preserving and resource allocation in location based services through differential privacy and mixed-integer linear programming (MILP) formulation techniques.

19.2 Physicality, Trust Anchors, and the Need for New Trust Models

19.2.1 Proposed Methodology

In [12], we provided a holistic view of how trust can be built in future networks by leveraging the physical layer and contextual information. Context was shown to serve as a base for autonomous controls and CPS that react accordingly to their physical and cyber realities, and, is conceived as a construct that goes beyond the pure description of the environment in terms of *where*, *when*, *what* and *who* (known as the four Ws). Therefore, context-awareness was closely related to situational awareness and could be distilled jointly from physical and cyber sources, i.e., from radio frequency (RF) inputs, hardware, and sensors, as well as network attributes.

Furthermore, as 6G is dubbed as the first AI-native[1] wireless generation, it was only natural to invest in AI to interpret semantics and context in future systems. Incorporating context awareness in trust building could allow handling more efficiently aspects related to identifying risk or threat level and required security level, particularly for applications such as autonomous driving, robots, e-Health, etc.

Albeit, we also identified that the vision of incorporating the enabling technologies of sensing, positioning and channel engineering to derive trust was not free of its own set of challenges. For example, in sensing, we showed that there exist key questions of privacy and reliability. Nevertheless, *the ability to infer context from the sensing, computing, and the channel engineering capabilities expected from future generations of wireless technologies, were deemed to be valuable for paving the way to trust-centered cyberphysical systems.* Context could be used to better assess, and perhaps quantify, the legitimacy (i.e., trustworthiness) of a link or of a user (agent), e.g., using AoA to identify abnormal physical behavior, see Fig. 19.1 on the potential use of AoA-fingerprints to identify Sybil attacks.

Furthermore, context was identified as an enabler for the deployment of PLS security controls in 6G, as demonstrated in Fig. 19.2. We made the case that PLS controls could not be deployed without having first ensured the physical context is favorable, or made favorable through feedback control and channel engineering. These operations were captured in the monitoring stage and in the PHY feedback loop of the proposed PLS closed loop in the ROBUST-6G architecture, and was conceptually depicted in Fig. 19.2.

19.2.2 Numerical Results and Analysis

In Fig. 19.1 the potential use of RF fingerprints was demonstrated to identify Sybil attacks. The separation between fingerprints from legitimate vs virtual Sybil agents was established. This distinction was captured quantitatively as a stochastic trust value, through signal processing of the wireless channels. Furthermore, in Fig. 19.2 examples of PHY context aware PLS were shown, with several possible inputs on the left, and several possible outputs on the right. This could include measurements of channel quality, agent mobility, distance between agents or between agents and infrastructure such as communication towers; and the processing of these quantities to arrive at physical unclonable functions, secret key generation, and radio fingerprinting, which in turn enhance trust and security of the system. This figure also indicates the possible role of AI and data-driven tools for processing raw inputs to arrive at outputs of interest for physical layer security. Below, key takeaways of our analysis on trustworthiness and resilience of future networks were summarized:

- **Key Point 1:** Future CPS would integrate communication, sensing, and control—creating new vulnerabilities that traditional cybersecurity could not fully address. Cryptography (e.g., PKI, certificates) was developed to verify identity, but not behavior, protect against outsiders but would fail against

[1] Edge and on-device intelligence will enable real-time operation, without a human-in-the-loop, of autonomous agents in 6G, thus rendering it an "AI-native" generation.

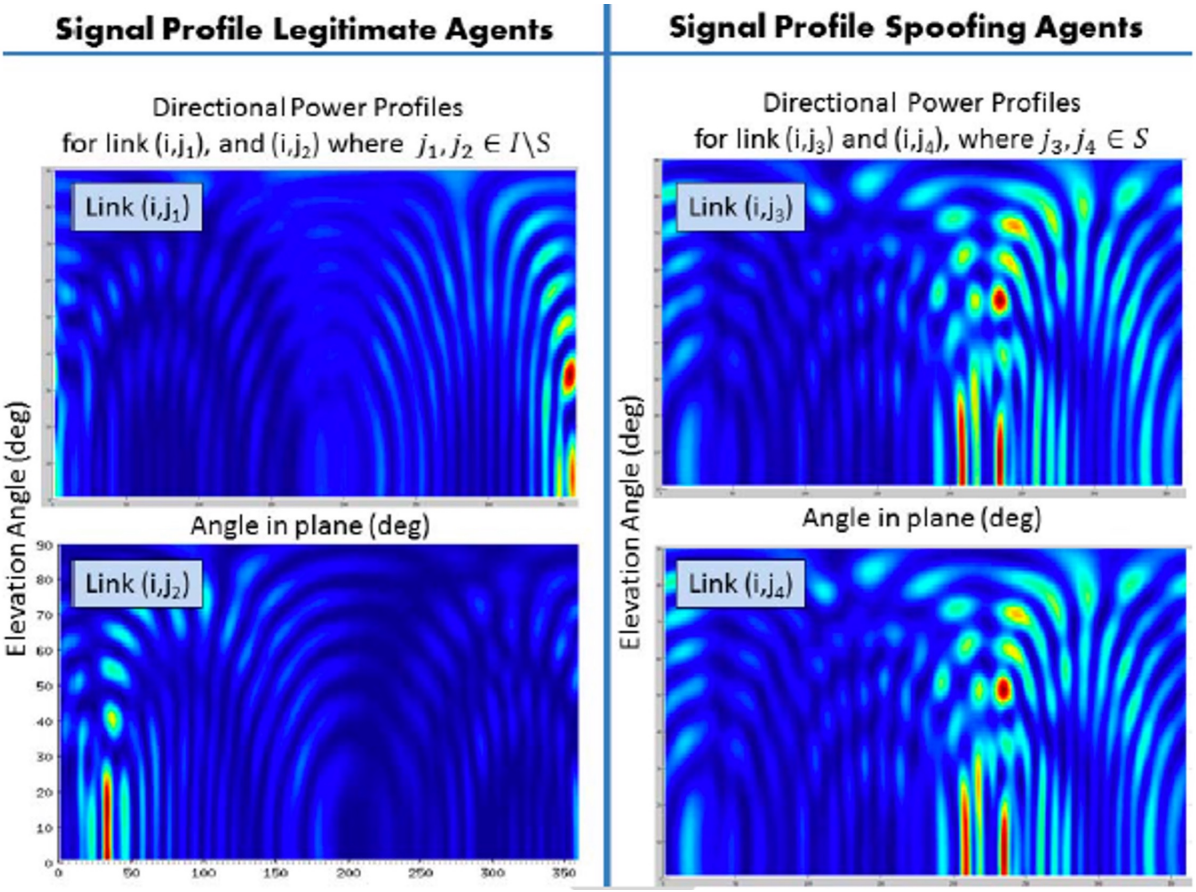


Figure 19.1: Example of an Angle-of-Arrival wireless "fingerprint" for two unique senders (left) and two Sybil agents (right).

malicious insiders and could be too slow or heavy for spontaneous, low-latency CPS interactions (e.g., between vehicles or drones).

- **Key Point 2:** The PHY could now aid in building trust using PLS and more broadly trustworthy localization and physical, context-aware, verification. In many practical scenarios, trust and reputation systems could be enhanced to detect behavioral attacks by cross-checking the location, speed, and movement patterns of autonomous CPS.
- **Key Point 3:** The path forward consisted in combining authentication, contextual sensing, and ML-driven behavioral trust into adaptive, multi-layered trust frameworks for resilient CPS.
- **Core insight:** Authentication was necessary but not enough — true trust demanded online, context-aware validation of both identity and behavior.

19.3 Trust and Reputation Management

19.3.1 Proposed Methodology

Motivated by the above framework, in [13], we proposed a global trust and reputation management (TRM) framework for 6G CPS networks. We highlighted that authentication alone was insufficient for autonomous

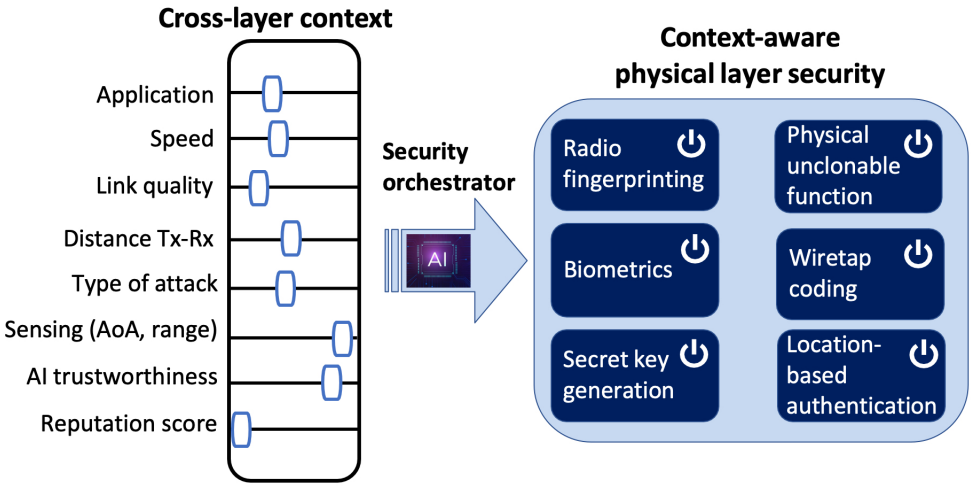


Figure 19.2: Conceptualization of PHY context-aware PLS.

agents that continuously exchange physical-state data such as position, speed, or sensor readings. These measurements could be falsified in propagation across multiple agents, raising questions about how to validate information integrity.

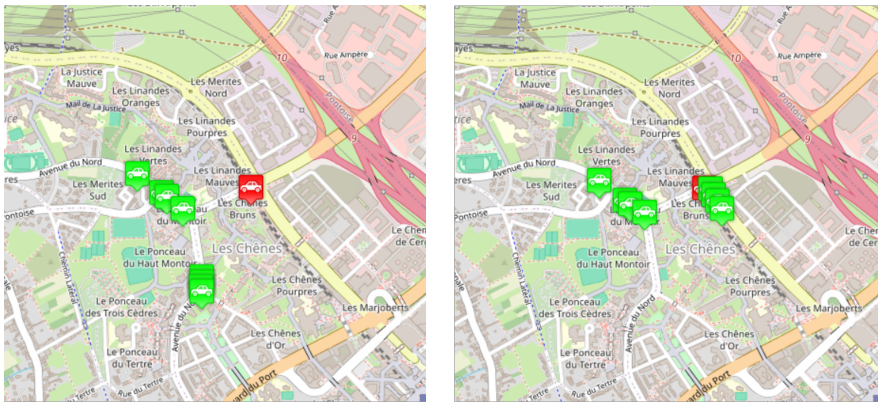


Figure 19.3: Map visualization for two test cases (a) successful bogus information attack successfully and (b) unsuccessful attack – detected and prevented. The red marker indicates the malicious vehicle while the green markers represent benign vehicles.

19.3.2 Numerical Results and Analysis

To support our arguments, we performed simulations, during which a VANET collected real-time data exchanged between vehicles and infrastructure, including information on location, speed, and acceleration; these were analyzed to identify mis-behavior, as shown in Fig. 19.4. We used Eclipse MOSAIC to develop a small-scale attack scenario. It took place in an urban setting around the area of the ETIS campus in Cergy, France, with a simple road topology, including a main route and an alternative route. One of the vehicles in the middle of a convoy played the role of an attacker, broadcasting fake decentralized environmental notification message (DENM) messages to notify about non-existing obstacles on the main route, causing vehicles within a certain radius to redirect to an alternative (sub-optimal) route, as shown in Fig. 19.3.

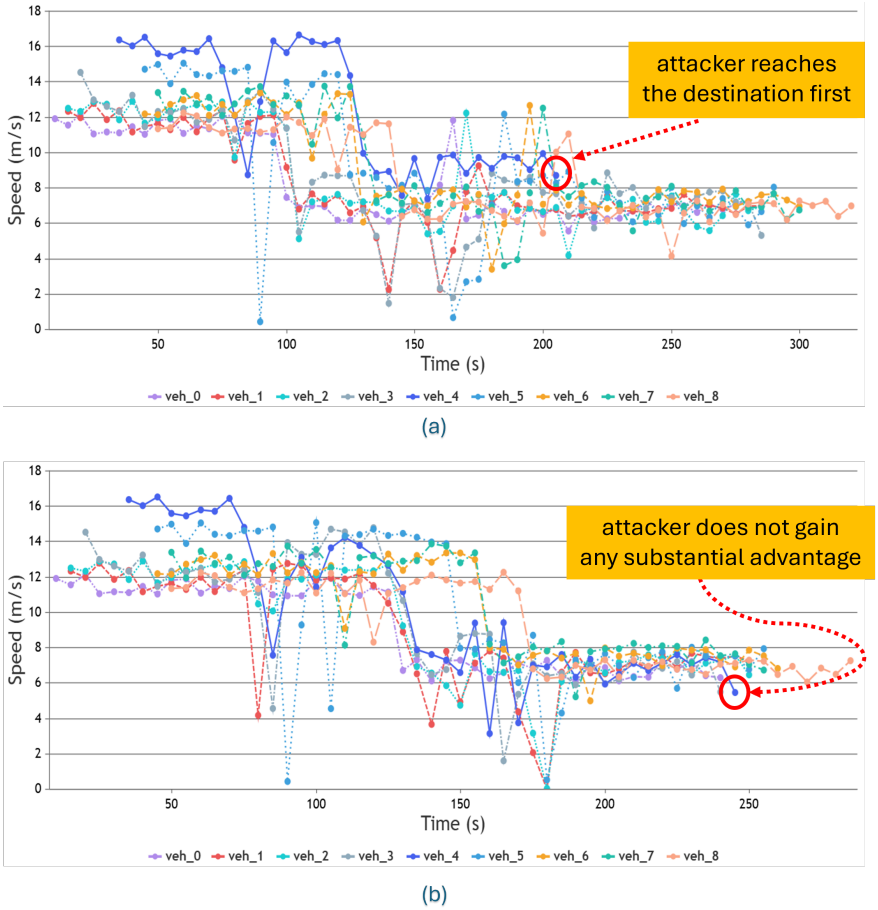


Figure 19.4: Vehicle speed graph for two test cases (a) bogus information attack successfully propagated and (b) attack detected and prevented.

Furthermore, we proposed two approaches for TRM. First, a framework was designed to maintain network security by combining multi-layer data from physical layer (PHY) signals and the information from the upper layers, leveraging ML to improve the capability of detecting attacks or misbehavior while enhancing the system's robustness. In Fig. 19.5, we presented this first possible architecture.
Finally, we proposed to incorporate additional planes including communication, sensing, and semantics to collect further information from the environment and correlate them with the data collected from the network. For example, the data collected in VANET could include GPS coordinates, velocity, acceleration, weather sensor data, camera images from the sensing plane; signal strength, packet transmission rates from the communication plane, and the meaning and context behind the data being transmitted from the semantic plane. This was captured in the second proposed TRM approach, shown in Fig. 19.6.

19.4 AoA–ToF-Based Impersonation Attack Detection

19.4.1 Proposed Methodology

In [14], we exemplified the use of physical-layer signal processing of CSI to enhance positioning accuracy, for trust enhancement as discussed in previous sections. We proposed three key techniques: (i) CSI sanitization, (ii) subarray-based AoA estimation, and (iii) extraction of other unforgeable sensing features such as ToF.

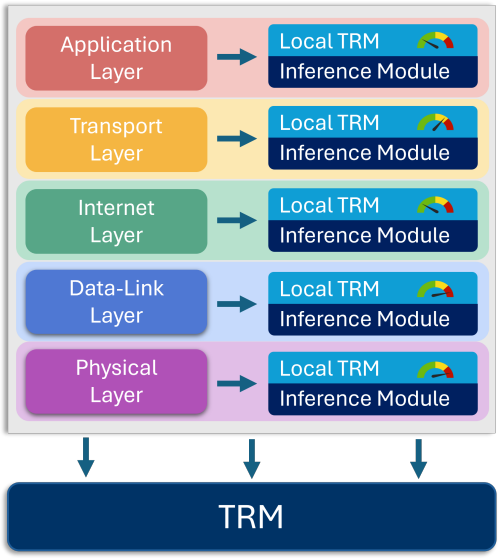


Figure 19.5: Multi-layer architecture for TRM framework.

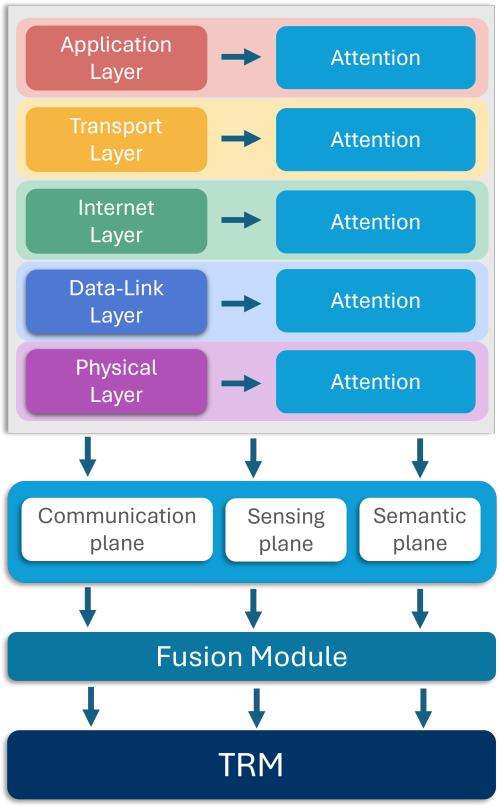


Figure 19.6: Multimodal attention fusion architecture for TRM framework.

These features allowed us to detect proximal impersonation attacks.

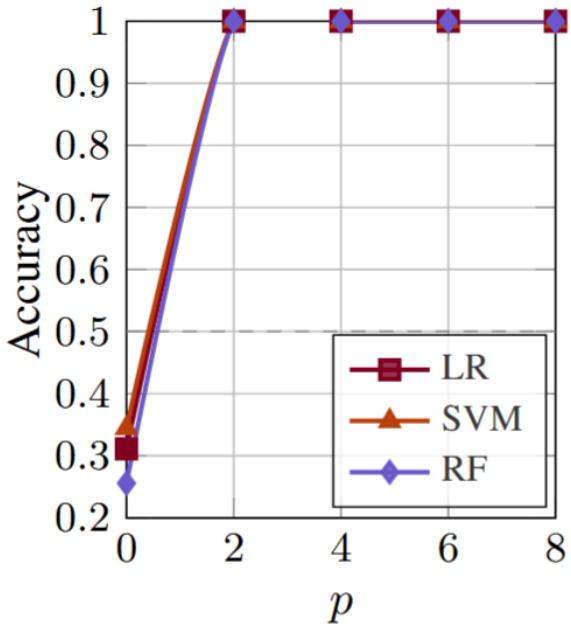


Figure 19.7: 3D localization accuracy using AoA and ToF (x axis in meters), suing simple classifiers (linear regression, support vector machine, random forest

19.4.2 Numerical Results and Analysis

It was demonstrated that AoA alone was limited when attackers were on similar directions as legitimate users. Albeit, jointly using AoA and ToF enabled very high positioning accuracy (up to 100% under certain shifts). This work showed how joint multiple physical unforgable features, such AoA and ToF, provided finer spatial discrimination than theoretical ToF resolution limits (based solely on frequency domain analysis), dramatically improving the trustworthiness of PHY-layer authentication.
It was concluded that combining AoA and ToF sensing significantly increased trustworthiness by enabling robust physical-layer impersonation detection even under challenging proximity conditions, as shown in Fig. 19.7, in which a positioning accuracy of 2 meters was achieved on the Nokia dataset discussed in Chapter 3. We are currently working on distance bounding protocols using AoA and ToF.

19.5 Joint Sensing–Communication Channel Estimation

19.5.1 Proposed Methodology

In [15], we flipped our focus on the inter-relation between CSI accuracy and sensing accuracy and introduced a novel channel estimation algorithm enhanced by radar sensing information to improve the accuracy of channel estimation. This algorithm combined environmental insights gathered from radar sensing with compressed sensing methods for channel estimation, enabling accurate assessment of channel states without relying on a large number of pilot signals. Key advantages of this approach included reduced pilot overhead in MIMO systems and improved estimation performance, both of which were crucial for optimizing spectral efficiency and minimizing system complexity.

19.5.2 Numerical Results and Analysis

To evaluate the effectiveness of the proposed algorithm, we conducted extensive simulations comparing its performance with traditional channel estimation techniques across various conditions. The results consistently demonstrated that our sensing-aided channel estimation algorithm achieved higher accuracy, regardless of the complexity of the propagation environment. Integrating radar sensing into channel estimation thus represented a promising direction for enhancing the efficiency and performance of wireless communication systems. The numerical results depicted in Fig. 19.8 demonstrated the enhancement in terms of normalized mean square error of the proposed approach against state of the art.

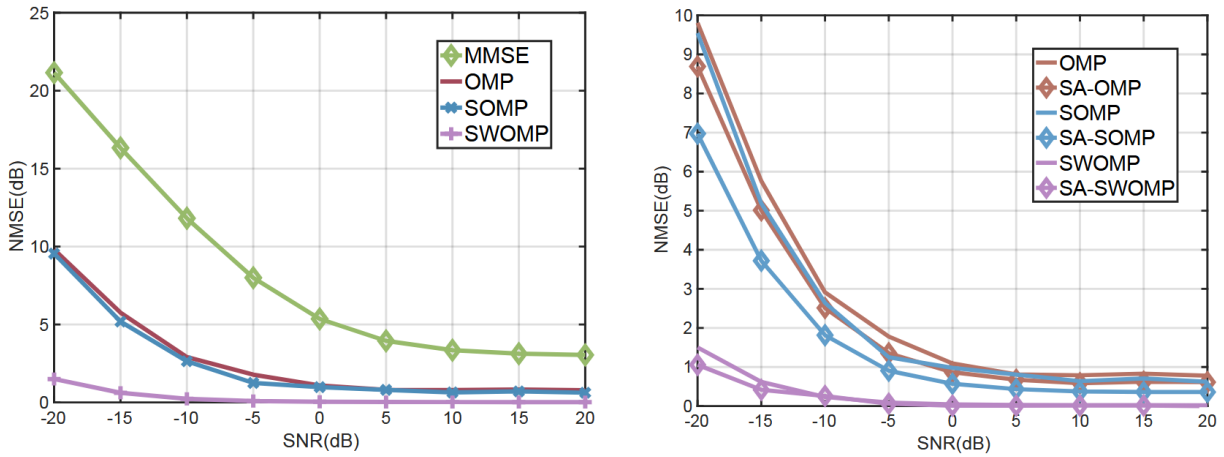


Figure 19.8: Normalized mean square error and error performance comparison of radar assisted channel estimation with the state of the art.

Future research could explore adaptive sensing strategies that adjust radar parameters in real time, extend the approach to multi-user and multi-cell environments to manage interference, and implement hardware tests to evaluate practical performance and scalability in real-world applications

19.6 Enhancing the Trustworthiness of Multi-Slice 6G Networks Through Hierarchical, Environment-Aware Resource Allocation

19.6.1 Proposed Methodology

In [16] we addressed trustworthiness in multi-slice 6G networks. We formulated network slicing and user scheduling as a multi-objective, constrained optimization problem and proposed a hierarchical reinforcement learning (RL) framework with an inter-slice agent that allocates resources fairly across slices to maximize the number of satisfied users, and intra-slice agents that independently schedule users within slices to enhance trustworthiness, quantified through metrics related to reliability, availability, and fairness. A major contribution was the integration of *environment-aware* knowledge such as LoS / NLoS conditions, which directly related to physical-layer trust metrics. Working in an indoor factory floor set-up, as clutter density increased, trustworthiness declined due to poor channel conditions, but our RL agents adapted effectively to maintain higher availability and fairness.

19.6.2 Numerical Results and Analysis

We evaluated the proposed method under two scenarios. In both scenarios, ultra relable low latency communications (URLLC) and mobile broadband reliable low latency communications (MBRLLC) slices demanded random latencies between 2.16 and 2.6 ms, while the enhanced mobile broadband (eMBB) and MBRLLC slices requested data rates of [8, 11.25] Mb/s in Scenario 1 and [11.25, 16] Mb/s in Scenario 2. Since no existing method directly optimized the defined trustworthiness metrics through resource allocation and network slicing, we evaluated the performance of the proposed method against a heuristic baseline that used the same inter-slice resource allocation from our framework but applied a heuristic intra-slice schedule. The simulation results demonstrated that the proposed method found a balance between these trustworthiness metrics in dense and cluttered 6G network environments. Future work will incorporate more trustworthiness aspects, such as security and privacy. A sample of numerical results is shown in Fig. 19.9 with respect to user satisfaction in the two scenarios.

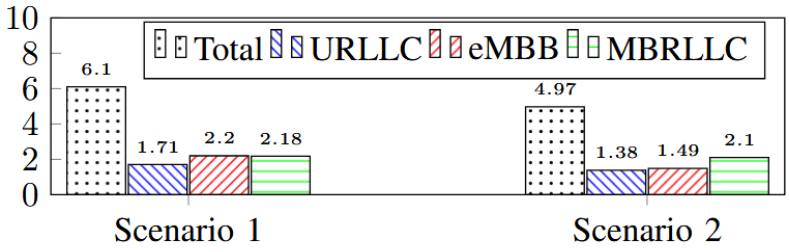


Figure 19.9: Satisfied user in scenarios one and two.

19.7 Privacy by Design

19.7.1 Proposed Methodology

In [74], we considered a location-based application that leverages edge servers to process users' data while meeting QoS requirements, such as minimum average throughput and maximum latency. We developed a scheme that maintains the desired quality of service (QoS) while preserving user location privacy. The proposed model assumes that the users share perturbed locations generated using differential privacy, which adds controlled noise to obscure their real positions. Based on the reported location, each user is perceived to be associated, at the radio level, with the access point closest to the perturbed location. In this solution, while the application remains unaware of the underlying routing details, a trusted party routes data from real locations to the perturbed ones before handing it off to the application servers. Although the application is aware of its reserved cloud computing resources and can scale them according to the perceived needs, it does not have access to internal network routing information. This design permits multiple network routes to enhance privacy, at the cost of increased link utilization. The illustration of the system model is given by Fig. 19.10.

19.7.2 Numerical Results and Analysis

We formulated the joint radio, link, routing, and processing resource allocation problem under privacy and latency constraints using a two-phase MILP framework. The first phase aimed to enhance the privacy guarantees of users by exploiting the system resources, while the second allocated the resources after perturbing the location in accordance with the assigned privacy guarantees. We evaluated multiple MILP

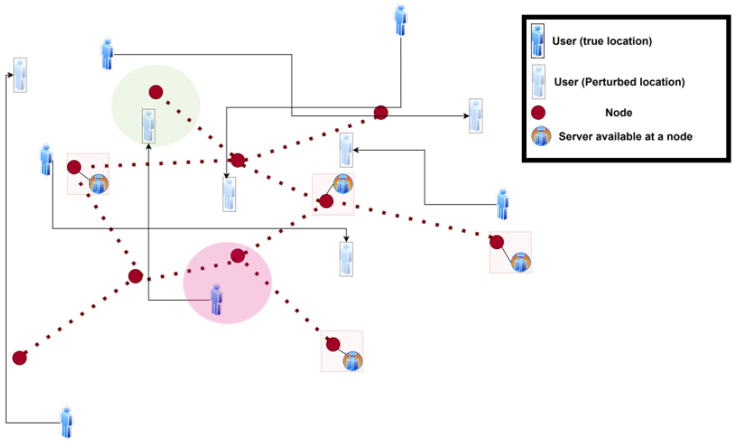


Figure 19.10: System model illustration.

variants : ***1)*** maximizes privacy first and then the number of served users under latency constraints, ***2)*** defines privacy to its required level and maximizes the served users under latency constraints, ***3)*** defines privacy to its required level fixes privacy maximizes the served users while minimizing total latency, ***4)*** a variant where privacy constraints were canceled, focusing solely on satisfying latency requirements, ***5)*** a version where privacy constraints were canceled and latency was handled on a best-effort basis in the objective function.
Key takeaways:

- **Uncontrolled privacy maximization harms QoS.** Increasing privacy at the cost of latency reduces user admission, necessitating a cap on additional privacy.
- **Latency relaxation can boosts performance.** Relaxing delay constraints in high-load scenarios to serve more users.
- **Privacy enhancement is resource-sensitive.** As load increases, increasing privacy becomes harder for variant **1**.
- **MILP suits static conditions.** It is usable when changes are infrequent; dynamic environments need heuristics or learning.
- **PS-LS offers the best tradeoff.** It meets privacy and latency requirements with the least execution time.

19.8 Integration with the Architecture

Across all the above works, we observed a unifying vision of 6G trust based on i) **Physical integrity** through AoA/ToF sensing; ii) **Behavioral trust** for multi-agent systems judged by actions, not only identities; and iii) **Context-awareness**, starting at the PHY with LoS / NLoS, clutter, and radar information strengthened trust estimation. These analyses and related algorithms were incorporated in the PLS closed loop within the following components **CENS01, CENS04**.

Chapter 20

Predictive Modeling for RF Fingerprint Evolution

In chapter 4, GOHM addressed the challenge of receiver invariance in RFFI, focusing on how to identify devices across different receivers. However, RFFI systems face another critical challenge: temporal drift. The accuracy of a RFFI model can degrade over time. For instance, a model that is trained and optimized on data collected during a single day typically performs well immediately. However, if that same model is tested on data collected from the same devices the following day or week, it often experiences a significant drop in classification accuracy [75]. This phenomenon suggests that the physical signature of a device is not entirely static but evolves over time. This chapter introduces **RF-PREDICT**, the work focused on understanding these temporal changes to maintain high identification accuracy over long operational periods.

20.1 Background and Motivation

Physical layer authentication works effectively if the device's signal remains consistent. However, hardware characteristics change slightly as time passes. The analog components inside a device are affected by heat, usage patterns, and environmental stress [76]. Changes in ambient temperature or battery voltage may cause the resulting radio signal to appear different to the receiver [77]. If this temporal drift is not accounted for, the RFFI model may increasingly misclassify legitimate devices over time.
Maintaining RFFI system accuracy over time can be challenging for practical large-scale IoT deployments [78]. Some existing approaches require periodic fingerprint maintenance and/or model updates, which can become increasingly costly and operationally difficult as device populations scale. The goal of this work is to characterize and predict temporal evolution patterns in RF fingerprints, supporting adaptive authentication systems that aim to maintain accuracy while reducing the need for frequent recalibration. By understanding how signatures change over time, we can develop predictive maintenance strategies rather than reactive approaches.

20.2 Proposed Methodology

To study these issues, we designed an experimental setup to collect long-term data. This setup utilizes the same controlled environment and equipment described in the RF Fingerprint Migration work (Chapter 4). The methodology focuses on identifying which factors cause the signal to change:

- **Device Diversity:** We use 30 custom IoT transmitters based on the TI CC13XX platform.

- **Power Profiles:** To compare how batteries affect the signal versus stable power, half of the devices (15) use DC power. The other half (15) run on batteries.
- **Interval Transmission:** Devices are programmed to send packets at different time intervals. These intervals range from 15 seconds to 24 hours.
- **Controlled Environment:** We record data continuously using synchronized SDR receivers (Fairwaves XTRX and Ettus B200).

Data Processing and Learning Framework: To analyze the long-term data, we have designed a processing scheme based on strict chronological partitioning. Unlike random splitting, which may introduce data leakage, the model will be trained exclusively on historical data and evaluated on subsequent future data. This validation procedure simulates a realistic deployment where the security system must authenticate devices based on their previously established profiles. The planned classification model will employ a 1-Dimensional CNN architecture optimized for time-series signal processing.

20.3 Numerical Results and Analysis

As of this deliverable, the RF-PREDICT testbed is active. We have completed the "Phase 1" baseline data collection. In this phase, we captured approximately 25,000 packets for each device transmitting continuously (every 6 seconds). This gives us a starting reference point for every sensor.
The ongoing "Phase 2" collection is currently recording data to capture the long-term changes. Figure 20.1 presents preliminary results from Phase 2, showing spectrograms from two devices (T29 and T30) at different time points during the monitoring period.
The packet structure used for this dataset includes fields such as internal temperature and power level to help us analyze the results later (details in Appendix G).

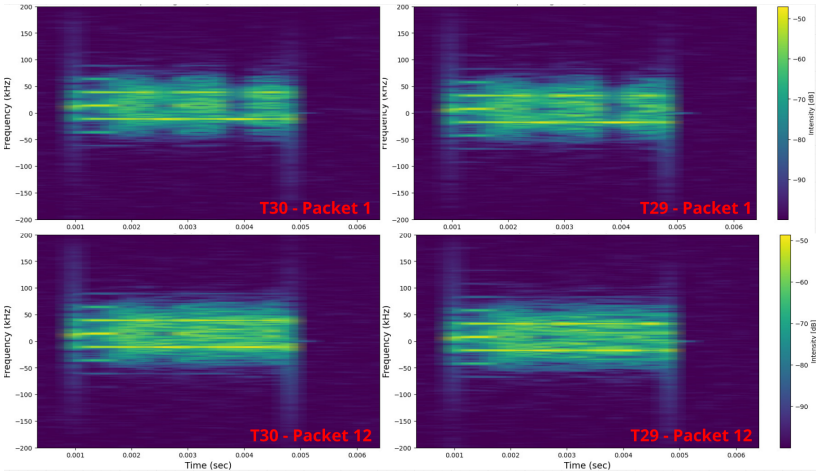


Figure 20.1: Spectrograms from Phase 2 long-term collection. Packet 1 (top row) and Packet 12 (bottom row) are shown for devices T29 and T30, demonstrating device-specific spectral characteristics over time.

20.4 Contribution to 6G Physical Layer Security

This work aims to improve the robustness of PLS. By understanding how RF fingerprints change over time, it becomes possible to build models that adapt to these changes. This could help maintain the accuracy of the RFFI models without needing to re-enroll devices frequently.

20.5 Integration with the Architecture

This predictive modeling work defines the **CGHM02** demonstration component. Its primary objective is to characterize the temporal evolution of RF signatures to address the challenge of signal drift. By integrating these predictive capabilities into the PLCL architecture, the system can enhance the **accuracy stability** of RFFI models. This ensures that device identification remains reliable over extended operational periods, effectively mitigating performance degradation caused by component aging, environmental factors, or irregular transmission intervals.

Chapter 21

GAN-based Unsupervised Anomaly Detection for 6G Cloud RANs

In this work, LIU addressed the generalized cross-layer anomaly detection component through unsupervised learning approaches, aligning with Task 5.3's objective. In particular, the framework operates at the intersection of physical and MAC layers by leveraging multiple KPIs spanning fronthaul traffic, thread scheduling, and precision time protocol logs, resonating with the proposed cross-layer methodology that integrates information from layers proximate to the physical layer. The research presented in this work [79] originated as a master's student project at Linköping University and has progressed to a peer-reviewed publication at Asilomar in 2025. The work remains at a proof-of-concept stage, having been tested exclusively on simulated data from controlled scenarios with five user equipments, and the authors acknowledge several limitations requiring further development, including the model's current inability to generalize across different anomaly types (such as Packet Data Convergence Protocol (PDCP) thread contention, radio interference, and MAC thread contention) and the absence of root cause analysis capabilities beyond detection.

21.1 GAN-based unsupervised anomaly detection for 6G cloud RANs

RAN systems exhibit inherent complexity [80], necessitating continuous monitoring to prevent performance degradation and maintain optimal user experience. These networks employ numerous key performance indicators to assess system performance, generating substantial data volumes every second. This extensive data production significantly complicates troubleshooting processes and accurate diagnosis of performance anomalies [81]. Additionally, the highly dynamic characteristics of RAN performance require adaptive methodologies capable of capturing temporal dependencies for reliable anomaly detection. Addressing these challenges, this work presents RANGAN, an anomaly detection framework integrating a GAN with a transformer architecture. To strengthen the capability of capturing temporal dependencies within data streams, RANGAN utilizes a sliding window approach during data preprocessing. The evaluation of RANGAN was conducted using the publicly available RAN performance dataset from the Spotlight project, which is based on 5G network infrastructure. Experimental findings demonstrate that RANGAN achieves promising detection accuracy, notably attaining an F1-score reaching 83% in identifying network contention issues.

21.1.1 Background and Motivation

RAN constitute a critical infrastructure component enabling mobile connectivity for voice communications, applications, and digital services. The increasing complexity of modern telecommunication systems, driven

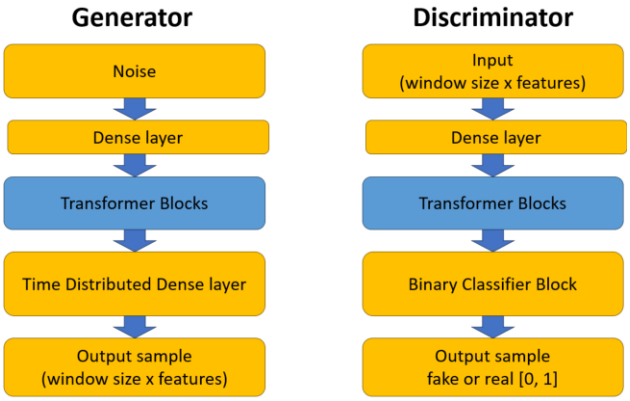


Figure 21.1: A schematic view of the architecture of RANGAN.

by exponential growth in mobile data consumption and continuous technological evolution, has created substantial challenges for network monitoring and maintenance [82]. Communication service providers face mounting pressure to simultaneously deliver enhanced performance, ensure reliability, manage energy efficiency, and maintain service quality in an increasingly competitive marketplace.

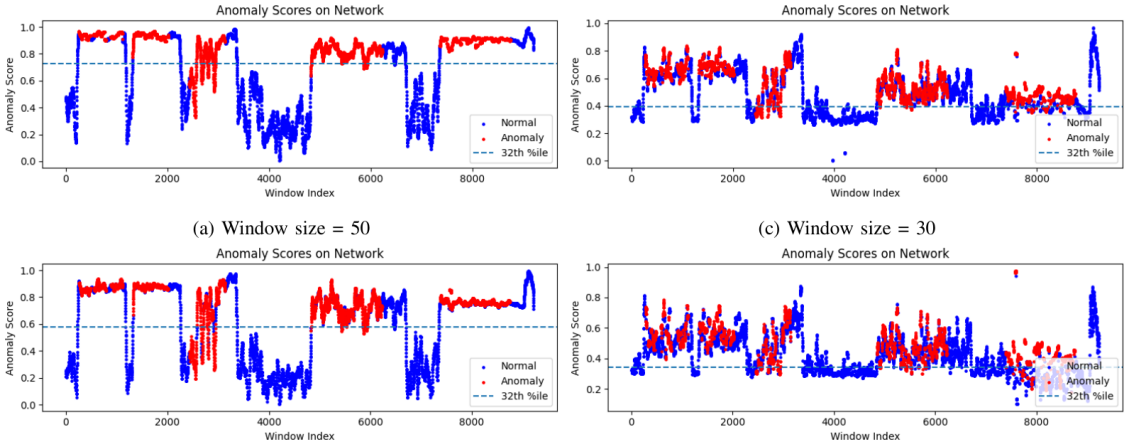
Anomaly detection in RAN environments encounters several fundamental challenges. The scarcity of labeled datasets for training and validation represents a primary obstacle, as obtaining reliable labels in operational networks is often impractical or prohibitively expensive [83]. Network environments exhibit continuous changes in traffic patterns and software configurations, demanding adaptive detection mechanisms that maintain effectiveness over time [84]. Furthermore, techniques successful in one domain frequently fail to generalize across different contexts [85], and telecommunication data inherently contains noise and missing values that complicate the distinction between normal variations and genuine anomalies. Traditional approaches based on rule-based systems and statistical techniques rely heavily on predefined thresholds and heuristics [86–89], demonstrating diminished effectiveness in dynamic environments. Recent research has shifted toward advanced machine learning techniques, including variational autoencoder (VAE) with long-short term memory (LSTM) networks [90], sparse autoencoders [91], and graph convolutional network (GCN) with transformers [92].

The key contribution of RANGAN is the integration of GAN with transformer-based architectures specifically designed for unsupervised cross-layer anomaly detection in RAN environments. Unlike existing reconstruction-based methods or density-based clustering techniques, RANGAN leverages adversarial training to learn complex data distributions while employing transformer attention mechanisms to capture long-range temporal dependencies in time-series KPIs data, enabling identification of contextual anomalies across multiple time steps and layers without requiring labeled training data.

21.1.2 Proposed Methodology

The RANGAN framework introduces an unsupervised anomaly detection approach specifically designed for identifying network contention in RAN. The architecture illustrated in Fig. 21.1 integrates a GAN with transformer-based components to effectively capture complex temporal dependencies inherent in time-series performance data.

The methodology employs the SpotLight dataset, which simulates network traffic generated by five user equipments operating under distinct scenarios encompassing various traffic types: transmission control protocol (TCP) and User Datagram Protocol (UDP) traffic in uplink and downlink directions, file downloads and uploads, video streaming, web browsing, and random ping patterns. The system monitors key performance indicators aligned with 3GPP standards, spanning three primary categories: fronthaul traffic metrics



(a) Window size = 50

(c) Window size = 30

(b) Window size = 40

(d) Window size = 20

Figure 21.2: Anomaly score assessed under different window sizes.

measuring uplink and downlink link usage, thread scheduling data capturing on-CPU and off-CPU runtimes, and precision time protocol logs recording frequency, root mean square, delay, and maximum offset values. Data preprocessing involves selecting informative KPIs and applying min-max normalization to scale values within the normalized range. A sliding window technique partitions the time series into overlapping fixed-length segments, enhancing the model's capacity to learn temporal patterns effectively. This windowing approach proves particularly crucial for identifying contextual anomalies where the significance of a data point depends on its temporal neighborhood.

The GAN architecture consists of two primary components: a generator receiving latent input vectors and producing outputs matching training sample dimensionality, and a discriminator trained to distinguish between authentic and synthetically generated samples. Transformer blocks integrated into both components employ attention mechanisms that assign dynamic weights to each time step based on relevance to others in the sequence. This enables selective focus on the most informative segments when generating or discriminating time-series data, substantially improving anomaly detection performance.

21.1.3 Experimental Results and Analysis

The evaluation employed standard anomaly detection metrics including precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC AUC). RANGAN achieved superior overall performance compared to established baseline methods, attaining an F1-score of 0.83, precision of 0.75, and recall of 0.93. The ROC AUC value of 0.78 indicates robust discriminative capability between normal and anomalous instances, although the model produced 1,585 false positives.

Among comparative methods, the autoencoder approach demonstrated competitive but inferior results to RANGAN, while traditional methods exhibited substantially weaker performance. The investigation of sliding window size impacts revealed that larger windows generally enhance performance, with optimal results achieved at window sizes of 50 and 60 time steps. Decreasing window sizes corresponded to declining precision and increasing false positive rates, with the smallest tested window size producing significantly degraded performance. Visualization of anomaly scores across different window sizes, shown in Fig. 21.2, demonstrates that larger windows yield more distinct separation between normal and anomalous segments. These findings highlight the importance of sufficient temporal context for effectively distinguishing anomalous from normal behavior in RAN environments.

21.2 Integration with the Architecture

The research demonstrates that deep learning models can achieve substantial performance in detecting network contention within RAN. The proposed RANGAN framework attained the highest F1-score among evaluated methods while maintaining a moderate false positive rate, with many false positives occurring temporally proximate to genuine anomalies. Although deep learning approaches exhibited slightly elevated false positive rates compared to traditional methods, this trade-off was offset by substantially higher recall and the capability to identify contextual anomalies that simpler techniques frequently overlooked. The experimental findings confirm that feature engineering through sliding window techniques significantly enhances detection performance for time-dependent datasets, emphasizing the critical role of appropriate temporal context selection in practical deployments.

The RANGAN framework is directly linked to the **CLIU02** component, advancing Task 5.3's objective of cross-layer anomaly detection through unsupervised learning. Within the PLCL architecture, RANGAN functions in the analysis stage, supporting physical-layer attack identification by detecting network contention patterns across physical and MAC layers. The framework processes time-evolving KPIs data (fronthaul traffic, thread scheduling, PTP logs) from the monitoring stage, distinguishing semantically significant anomalies from benign variations through temporal dependency analysis. Outputs inform both trustworthiness evaluation and actuation decisions regarding resource control and security feature activation.

Chapter 22

Convergence Analysis of Semantics-Aware Estimation Algorithm Enabling Cross-Layer Anomaly Detection

In this work, LIU developed cross-layer anomaly detection methodologies involving semantic attributes of information and learning attack phenomena, directly addressing LIU's committed scope within the project. This study [93] introduces a semantics-aware remote estimation framework for finite-state Markov chains that explicitly integrates the AoCE as a semantic metric to quantify the significance and lasting impact of estimation errors, aligning with the proposal's emphasis on continuous learning and unsupervised learning approaches for detecting zero-day physical-layer attacks. The framework operates across physical and MAC layers by leveraging key performance indicators and employing the maximum a posteriori (MAP) estimator, which utilizes AoI to assess the usefulness of aged information at the receiver, thereby capturing temporal dependencies critical for detecting evolving attack patterns. The research has been submitted for publication in a peer-reviewed journal (IEEE Transactions on Information Theory).

22.1 On the Role of Age and Semantics of Information in Remote Estimation of Markov Sources

This component investigates semantics-aware remote estimation of finite-state Markov chains employing the maximum a posteriori estimator, aiming to devise transmission policies that optimize estimation performance under transmission frequency constraints. Two metrics were utilized: AoCE to quantify the significance of estimation error at the transmitter, and AoI to measure the predictability of outdated information at the receiver. The optimal transmission problem is formulated as a constrained Markov decision process with unbounded costs. It is demonstrated that there exists an optimal simple mixture policy that randomly selects between two deterministic switching policies with fixed probability. Notably, each switching policy initiates transmission only when AoCE exceeds a threshold value depending on both AoI and instantaneous estimation error. Sufficient conditions are further derived under which the switching policy simplifies to a threshold policy admitting identical thresholds for all estimation errors. Leveraging these structural results, the team developed an efficient structure-aware algorithm, Insec-SPI, that computes the optimal policy with reduced computational overhead. These findings demonstrate that incorporating both AoI and AoCE yields significantly improved estimation quality compared to using either metric alone.

22.1.1 Background and Motivation

In emerging cyber-physical systems, the timeliness and contextual relevance of information frequently outweigh mere signal fidelity, motivating semantics-aware remote estimation where emphasis transitions to ensuring that conveyed information is fresh, significant, and aligned with system control objectives [94,95]. A critical challenge is the lasting impact of consecutive estimation errors; prolonged errors generate increasingly severe consequences, particularly in applications such as autonomous driving where failing to detect nearby obstacles incurs costs growing exponentially with error persistence [96–98]. While AoI has emerged as the predominant metric for quantifying information freshness, it has been considered inefficient for monitoring Markov chains with zero-order hold estimators [99, 100].
This work reveals that for maximum a posteriori estimators, AoI becomes relevant by measuring the usefulness of outdated information at receivers. By integrating AoI with the significance-aware AoCE metric that assigns content-aware nonlinear age functions to different estimation error types, the framework captures two complementary aspects: how long information has aged and how severe the consequences of current errors become over time. This dual-metric approach enables transmission policies that balance communication costs against escalating penalties of prolonged erroneous states, critical for detecting and responding to evolving physical-layer attacks where delayed corrections can lead to cascading system failures.

22.1.2 System Model, Problem Formulation, and Methodology

The research considers remote estimation systems where sensors monitoring finite-state Markov chains decide at each time step whether to transmit measurements based on current source states and delayed channel feedback. Channel states follow Bernoulli processes indicating successful or failed transmissions, with receivers employing MAP estimation rules that maximize conditional probabilities given all received measurements. The MAP estimator's AoI-based belief representation enables information usefulness to depend on both age and content. Performance incorporates the AoCE metric through cost functions with non-decreasing age functions imposing escalating penalties for prolonged errors. The optimal transmission problem minimizes average semantics-aware cost subject to transmission frequency constraints, formulated as a constrained Markov decision process.
The research establishes that λ-optimal policies exist and demonstrates that sensors can discard historical information by restricting to transmission rules depending only on current source states, latest updates, their ages, and AoCE values. A fundamental result shows that switching policies suffice, with transmissions triggered only when AoCE exceeds threshold values depending on source state, latest update, and AoI. The research characterizes optimal policies: when transmission frequency under some λ-optimal policy equals the maximum allowed frequency, that switching policy is optimal; otherwise, optimal policies become randomized mixtures selecting between two switching policies differing in one threshold value.

22.1.3 Algorithm Development and Computational Efficiency

The developed Insec-SPI algorithm comprises structured policy iteration modules computing switching λ-optimal policies and intersection search modules updating Lagrangian multipliers, offering substantial complexity advantages over classical unstructured approaches like relative value iteration that evaluate all state-action pairs. The structured policy iteration exploits the known switching structure by searching for threshold values in increasing order of AoCE, ensuring that if optimal actions for certain states involve transmission, all states with higher AoCE values also transmit without further computation. For numerical tractability, the algorithm operates over finite state spaces by truncating both AoI and AoCE using sufficiently large bounds, justified because belief vectors converge exponentially fast in AoI and age process truncation impacts diminish as maximum AoCE increases. The intersection search method far outperforms bisection search by exploiting piecewise-linear and concave properties of Lagrangian costs, computing intersection

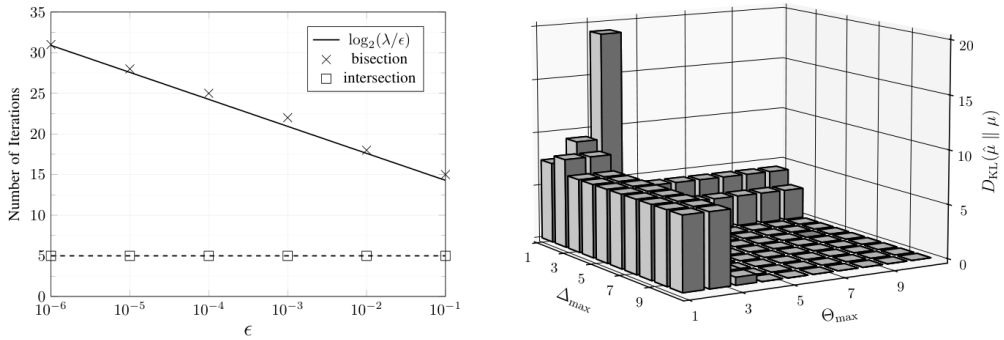


Figure 22.1: (Left) Time complexity of different multiplier update methods. (Right) The Kullback-Leibler divergence between the stationary distributions.

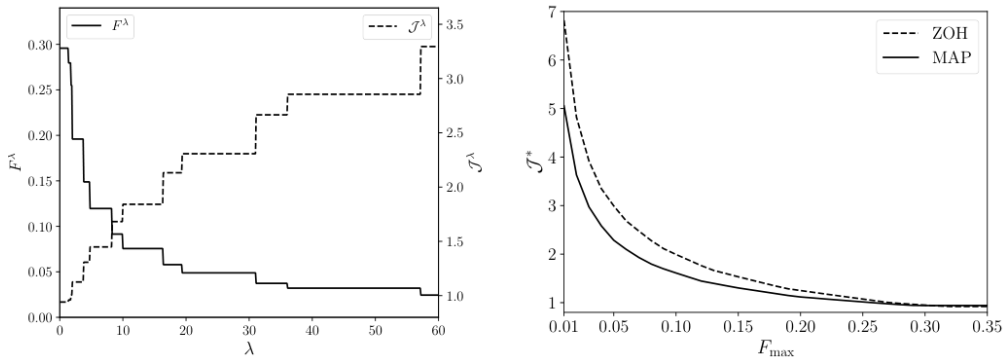


Figure 22.2: (Left) The system performance as a function of transmission cost λ; (Right) The minimum average cost achieved by different estimators.

points of tangents at search interval boundaries and terminating when intersection points locate on Lagrangian cost curves, finding optimal multipliers in only a few iterations with time complexity independent of accuracy tolerance.

22.1.4 Experimental Validation and Performance Analysis

Numerical results using Hamming distortion, exponential age functions, and three-state Markov sources with AoI and AoCE truncation sizes of twenty time steps demonstrate that intersection search exhibits significantly lower time complexity than bisection search, finding optimal multipliers in few iterations regardless of accuracy requirements (see Fig. 22.1). Asymptotic optimality analysis using Kullback-Leibler divergence shows that truncation errors decrease as truncation sizes increase, with increasing AoI truncation offering dual benefits of reducing approximation errors and alleviating AoCE truncation impacts; while zero-order hold estimators with minimal AoI truncation maintain significant distribution distances even for large AoCE truncation, MAP estimators with larger AoI truncation achieve close approximations with relatively small AoCE truncation. Policy structure analysis (Fig. 22.2) reveals that system performance exhibits monotonic and piecewise constant characteristics, with deterministic optimal policies existing only when maximum allowed transmission frequencies match value domains of transmission frequencies under λ-optimal policies, otherwise requiring simple mixture policies. Performance evaluation demonstrates that MAP estimators significantly outperform zero-order hold estimators especially when transmission frequencies are constrained, and incorporating both AoI and AoCE yields significantly improved estimation

quality compared to using either metric alone, validating the effectiveness of exploiting both age and semantics of information in remote estimation systems.

22.2 Integration with the Architecture

The Insec-SPI framework is directly affiliated to the **CLIU02** component, realizing remote state estimation under heterogeneous significance in semantic errors and fulfilling Task 5.3's objective. Within the PLCL architecture, Insec-SPI contributes to the analysis stage's physical layer trustworthiness evaluation by processing Markovian sources and outputting stochastic optimization policies that balance estimation accuracy against communication costs. The framework evaluates semantic significance through AoCE and provides trustworthiness assessments based on AoI to guide actuation decisions regarding resource allocation and security feature activation. The developed Insec-SPI algorithm efficiently computes optimal transmission policies exhibiting simple switching structures, directly addressing CLIU02's success criteria of reduced transmissions while maintaining estimation quality.

Chapter 23

Federated Authentication for 6G Networks

In this work, UNIPD studies a scenario where several authorized devices (collectively denoted as Alice) are transmitting from several *authorized areas*. An attacker device, Trudy, is instead impersonating Alice, i.e., transmitting messages and claiming to be Alice. However, Trudy does not have access to the areas where Alice is. Therefore, we propose that multiple BSs of a 6G network (collectively indicated as Bob) determine whether the received signal comes from the authorized areas or not, to authenticate the received messages, and determine whether they have been transmitted by Alice or not. The BSs collaborate to authenticate the transmitting device at the physical layer. This problem can also be seen as a problem of a distributed in-region location verification problem, [101].

23.1 Background and Motivation

In recent years, federated-learning (FL) has gained much interest as it allows different devices to collaborate on a common objective without explicitly sharing their data. Each device, in fact, uses local data for local training, then uploads the model to the server for aggregation, and finally, the server sends the global model back to the participants. Different aggregation strategies can be employed by the server. One is FedAvg, in which the server simply averages over the devices' updates. This is a very effective method when the data are i.i.d. across clients, but can perform poorly in the case of non-i.i.d. client data distributions. The problem of non-i.i.d. data distributions is still open, and several strategies are proposed in the literature. The application of federated learning in PLA is at its dawn, and to the best of our knowledge, only a few works have been done, especially in non-IID contexts. This work thus focuses on federated authentication at the physical layer in non-i.i.d. contexts, comparing state-of-the-art algorithms with a new, more lightweight proposed method.

23.2 Proposed Methodology

In this work, whose complete version is in Appendix H, we study a scenario where multiple BSs collaborate to authenticate the transmitting device at the physical layer. The legitimate transmitter, Alice, can be located in different areas; thus, each BS, Bob, needs to identify the transmitting area and, by knowing the legitimate one, can authenticate Alice. Trudy, on the other hand, aims at impersonating Alice by transmitting from another area with respect to Alice, thus fooling the BSs. The main contribution of this work is FedLoss, a novel FL framework that tackles the non-IID data heterogeneity present in common wireless environments via local fine-tuning. Each device, by implementing FedLoss, monitors the local training loss and switches dynamically to local and federated training on demand. In particular, the contributions are as follows:

1. We propose a realistic channel model, taking as baseline the 3GPP specifications.

2. We present FedLoss, the FL framework able to tackle the non-IID data distributions via local finetuning.

3. We numerically evaluate FedLoss, demonstrating its effectiveness even in the case of challenging channel conditions and scarcity of data.

23.3 Numerical Results

To validate the effectiveness of FedLoss, we compare it with three baselines, named Global, Single, and FedAvg, considering all the possible transmitting positions of both Alice and Trudy. In the Global case, there is a "virtual" single BS that has a dataset containing all the BS data. This is optimal if the datasets were formed by IID data, i.e., when the BS were located close enough to one another. In the Single case, on the other hand, all the BS train on their own dataset, even if small. This approach is supposed to work well when the BSs are far from each other; thus, sharing local information is damaging the overall performance. Finally, in FedAvg, the BS perform the FedAvg algorithm.

23.3.1 Accuracy VS Epochs

Fig. 23.1 shows the average accuracy across the BSs as a function of the number of training epochs when the average distance between BSs is $D_{\text{bs}} \simeq 1.1$km. We notice that both Global and FedAvg perform poorly, as expected for the very non-IID datasets across the BSs. The Single performs well despite the small training dataset, but it gets outperformed by the proposed FedLoss, which achieves the best accuracy across all the methods.

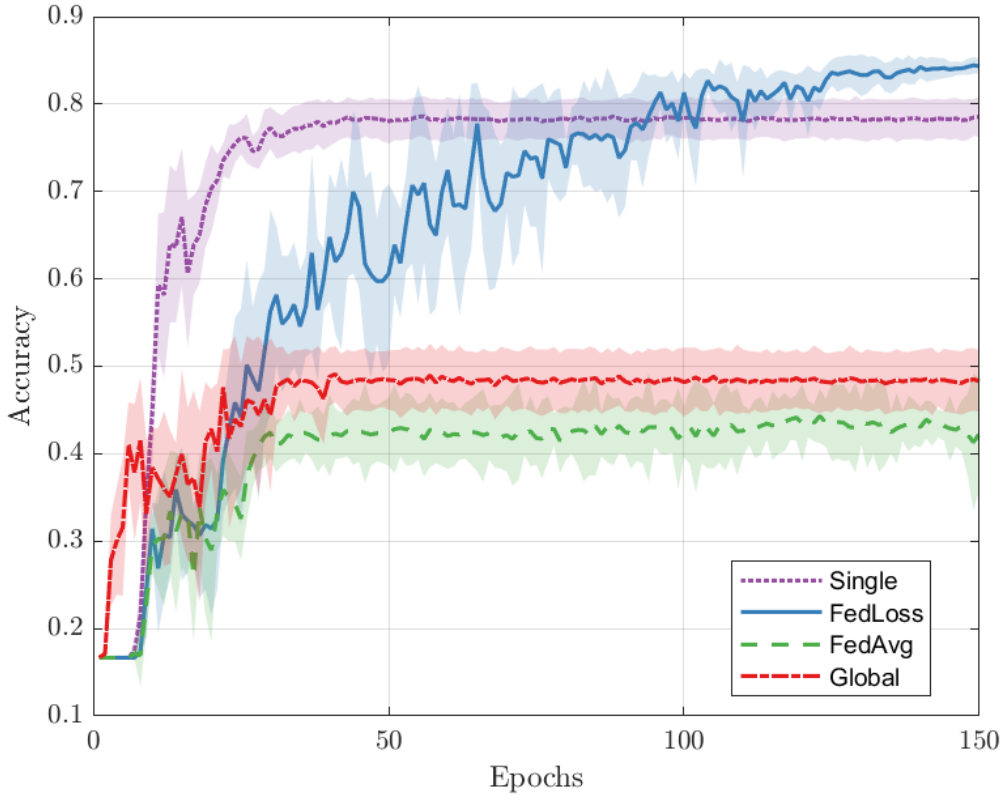


Figure 23.1: Accuracy VS Epochs for Single Client, Global, FedAvg, and the proposed FedLoss.

23.4 Integration with the Architecture

We have presented an authentication scheme built on federated learning. In this approach, several 6G base stations each train a local model—using their own estimated CSI – to decide whether a transmitter originates from an authorized region. The models are trained collaboratively in a federated manner, yet allowing each base station to capture the unique propagation conditions between itself and the transmitter by local finetuning. This work falls under the PHY-Attack Identification block of the ROBUST-6G architecture, as it pertains to PLA and the security-focused analysis of received signals. While no dedicated demonstration component is provided, extensive simulations have been carried out to confirm the solution's effectiveness.

Chapter 24

Position-Based Cross-Layer Authentication For Industrial Communications

In this study, UNIPD considers a robot (Alice) moving in an industrial environment while transmitting messages to nearby endpoints through fixed APs. An intruder robot (Trudy) aims at transmitting malicious messages to the endpoints, impersonating Alice. We aim at detecting Trudy transmissions by comparing the expected position of the transmitter with two estimates of it obtained from a) the CSI estimated on the signals received by the APs, and b) the traffic information in the network. Such estimates are obtained with CNN and support vector regressor (SVR) models along with Kalman filters to exploit the trajectory evolution. Numerical results obtained using the DICHASSUS dataset confirm the effectiveness of our proposed solution.

24.1 Background and Motivation

With the dawn of Industry 4.0, AI, the IoT, and robotics are gaining much interest to improve efficiency, productivity, and quality. In such a context, new information and communication (ICT) systems are used to support entire supply chains, increasing the attack surface to malicious devices aiming to disrupt the industrial infrastructure. Authenticating transmitters in such networks is a crucial task to ensure the integrity of the transmissions. To this end, different strategies can be adopted, from conventional cryptographic schemes to novel quantum cryptography, to lightweight physical-layer security mechanisms. Focusing on the latter, the literature offers different strategies to authenticate transmitters directly at the physical layer. Concerning cross-layer solutions, different strategies can be adopted. The first is to design hybrid protocols that combine physical-layer-based with conventional key-based authentication schemes. The second provides detectors that combine information from both layers. Still, when conventional PLS protocols cannot be applied (e.g., because the wireless channel conditions are not suitable), other strategies need to be adopted. This work focuses on hybrid strategies that combine the physical layer with upper-layer information, ensuring more robust security protocols.

24.2 Proposed Methodology

In this work, whose extended version is in Appendix F, we design an authentication protocol that fuses information coming from the physical and upper layers in an industrial network context. Alice is a robot, moving on a factory floor and communicating to neighboring endpoints via several APs, under the supervision of a MAP, that monitors the physical-layer along with the traffic information. Trudy, on the other hand,

aims at transmitting malicious messages to the endpoints, impersonating Alice. We aim at detecting Trudy transmissions by comparing the expected position of the transmitter with two estimates of it obtained by a) the CSI estimated on the signals received by the APs and b) the traffic information in the network. In particular, a CNN is trained to estimate the transmitter position from the CSI, while a SVR uses the traffic information in the network again to infer the robot position. The two predicted positions are then fed into Kalman filters to refine them with prior estimates of the trajectory, and the refined position estimation is lastly compared with the expected legitimate one. If the three positions are close enough, the message is considered authentic; otherwise is rejected as fake.
In comparison with existing literature, the proposed solution has several novel features. Notably, we fuse information from various layers to create position information, which is then compared using a statistical test. This provides a better understanding of the detector's behavior. Furthermore, we exploit the temporal correlation of the information using the well-established Kalman filter.
The contributions of this work are as follows:

1. The fusion of information coming from different layers passes through a common estimate of the device position rather than as a mixed input to an ML model.
2. The refinement of the estimated positions by Kalman filters to take into account the temporal evolution.
3. A cross-layer-detector (CLD), a lightweight ML framework constituted by a CNN and a SVR, that estimate the transmitter position using CSI and connectivity data, respectively.
4. A continuous learning approach is adopted to adapt the model to various conditions, keeping the model simple and without forgetting what it has learned in the past.
5. The performance assessment of the proposed solution on both synthetic and real-world data.

24.3 Numerical Results

In this section, we describe the data we used to validate our framework and evaluate the security performance in terms of FA probability, i.e., the probability that Alice is misled to Trudy

$$P_{\text{fa}} = \mathbb{P}(\hat{\mathcal{H}} = \mathcal{H}_1 | \mathcal{H} = \mathcal{H}_0) \,, \tag{24.1}$$

and the MD probability, i.e., the probability that Trudy is misled to Alice

$$P_{\text{md}} = \mathbb{P}(\hat{\mathcal{H}} = \mathcal{H}_0 | \mathcal{H} = \mathcal{H}_1) \,. \tag{24.2}$$

The FA/MD probabilities against the attacks, as well as the localization performance for different ML strategies and scenario parameters, are explained in the following subsections.

24.3.1 Performance With Static End-points

We begin our analysis by showing the behavior of our framework against the WPWT, WPCT, and CPWT attacks for various Trudy distances $D_{\max}$. Fig. 24.1 shows the DET of the proposed authentication mechanism, obtained by varying the decision threshold ϕ. Among the attacks, the least effective is WPWT, as it contains anomalies both at the physical and upper layers. Comparing the WPCT and CPWT attacks, we notice that their effectiveness depends on the localization accuracy, which in turn highly depends on the system conditions. Lastly, we observe that our framework is also effective against the CPWT attack, which would otherwise remain undetected in [102] and [103], thus we demonstrate the superiority of our protocol over state-of-the-art CSI-only based protocols.

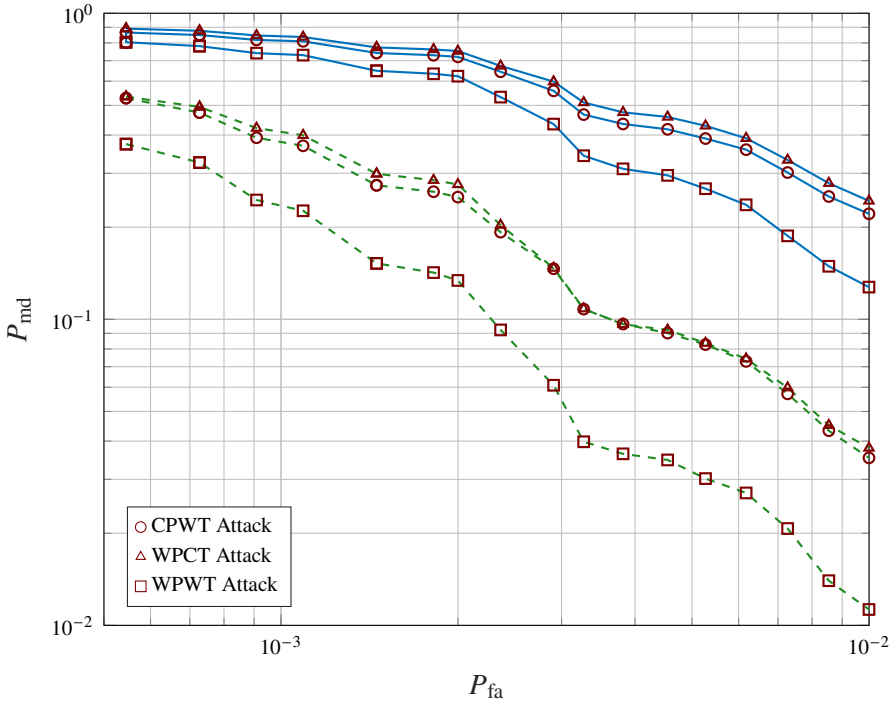


Figure 24.1: DET curve of CLD against the Wrong-Physical-Wrong-Traffic (WPWT) (squares), Wrong-Physical-Correct-Traffic (WPCT) (triangles) and Correct-Physical-Wrong-Traffic (CPWT) (circles) attacks and at different Trudy distances $D_{\max} = 1.5$ m (solid-blue line) and $D_{\max} = 2$ m (dashed-green line).

24.4 Integration with the Architecture

We presented a machine-learning–based approach for threat detection that fuses data from multiple network layers. Specifically, we combine insights from the physical layer and the network layer to authenticate messages within an industrial-automation scenario. This method fits within the PHY-Attack Identification module of the ROBUST-6G architecture, as it analyzes signal characteristics to uncover malicious activity, and it will be evaluated as part of the CUPD05 component.

Part IV

Appendices

Appendix A

RF Fingerprint Migration

A.1 Neural Network Architecture

The deep learning model employed for the domain adaptation task comprises two main modules: a feature extractor (Encoder) and a classifier. The Encoder utilizes a series of 1D convolutional layers to process the raw I/Q samples. The specific hyperparameters for the layers are chosen to balance computational complexity with feature extraction capability. The Classifier consists of three fully connected layers ending in a Softmax activation for the 30 device classes. The specific architecture used to generate the results in this deliverable is detailed in Table A.1.

Table A.1: Deep Learning Model Architecture (Encoder and Classifier)

Layer Type	**Filters/Units**	**Kernel Size**	**Stride**	**Activation**
Feature Extractor (Encoder)				
Conv1D	32	7	2	ReLU
Conv1D	64	3	2	ReLU
Conv1D	128	3	2	ReLU
Conv1D	256	3	2	ReLU
Global Avg Pool	-	-	-	-
Linear	256	-	-	ReLU
Classifier Head				
Linear	128	-	-	ReLU
Linear	64	-	-	ReLU
Linear	30 (Classes)	-	-	Softmax

A.2 Experimental Environment

The dataset was collected in a controlled indoor environment to minimize external interference while isolating hardware-specific effects. The following figures depict the physical deployment of the testbed used for the migration experiments.
Figure A.1 shows the synchronization of the three SDRs used as receivers (Two Fairwaves XTRX and one Ettus B200 Mini).

Figure A.1: The receiver array setup consisting of synchronized XTRX and B200 Mini SDRs.

Figure A.2 displays the array of the custom-built IoT transmitters used in the experiment. These devices are manufactured to identical specifications, serving as the source devices for the fingerprinting task. Since the devices are manufactured to identical specifications, the classification is driven not by obvious design differences across devices, but by hardware-induced characteristics in the RF signal.

Figure A.2: The set of 28 identical, custom-manufactured IoT transmitters used to evaluate the RF fingerprinting migration.

A.3 Intermediate Results

To assess the effectiveness of the proposed ADDA method, we conducted an evaluation across six different source-target receiver pairs. The results in Table A.2 represent a snapshot of the current development status. The accuracy results demonstrate that the adaptation process effectively recovers a significant portion of the classification accuracy lost due to receiver variability. This is particularly evident in transfers between receivers, where improvements often exceed 20 %.

Table A.2: Intermediate Accuracy Scores: Impact of Domain Adaptation

Source RX	Target RX	Source Acc.	Target Acc. (Pre)	Target Acc. (Post)	Improvement
R01 (XTRX)	R02 (XTRX)	93.08 %	85.08 %	90.39 %	+5.31 %
R01 (XTRX)	R03 (B200)	87.65 %	47.66 %	74.71 %	+27.05 %
R02 (XTRX)	R01 (XTRX)	90.92 %	66.01 %	85.58 %	+19.57 %
R02 (XTRX)	R03 (B200)	91.64 %	37.87 %	71.76 %	+33.89 %
R03 (B200)	R01 (XTRX)	95.13 %	42.73 %	70.33 %	+27.60 %
R03 (B200)	R02 (XTRX)	89.57 %	48.39 %	72.66 %	+24.27 %

A.4 Detailed Confusion Matrices

This section presents the comprehensive results of the domain adaptation experiments. The confusion matrices below illustrate the classification performance for each source-target receiver pair. For every scenario, three matrices are provided: the Source Baseline, the Target Before Adaptation (showing the impact of hardware impairments), and the Target After Adaptation (showing the recovery of performance).

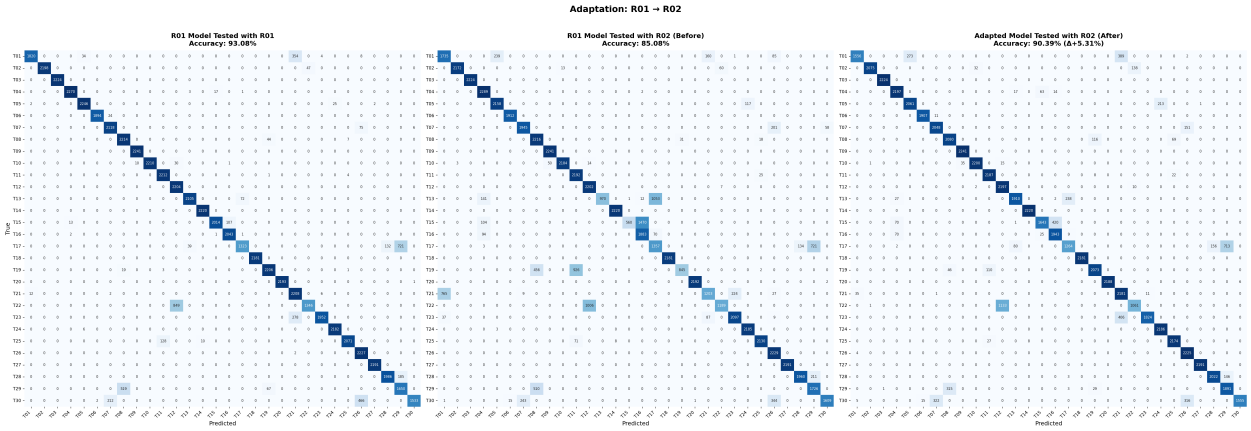


Figure A.3: Adaptation Scenario: R01 → R02.
Source Domain: XTRX (R01). **Target Domain:** XTRX (R02).
Performance: Accuracy improved from 85.08% to 90.39% (**+5.31%**).

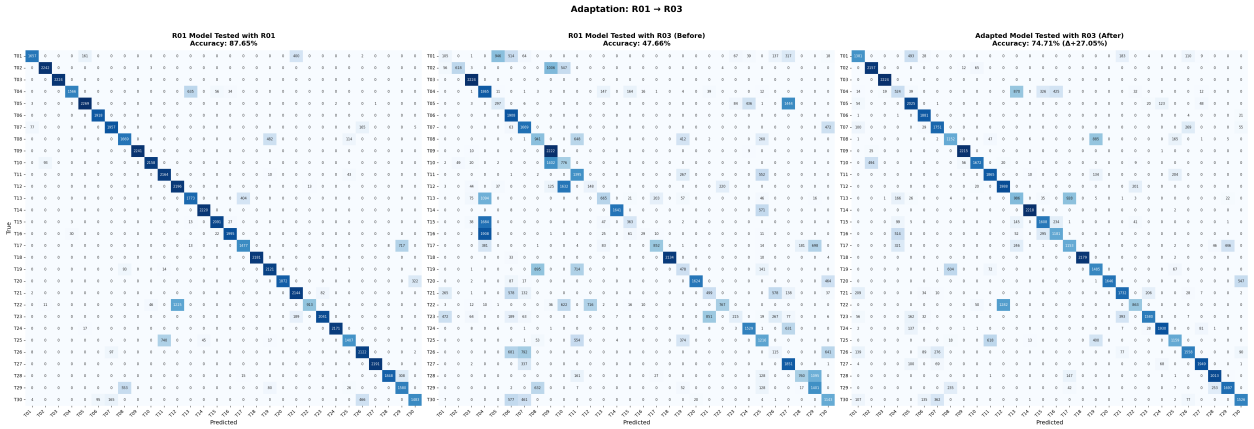


Figure A.4: Adaptation Scenario: R01 → R03.
Source Domain: XTRX (R01). **Target Domain:** B200 (R03).
Performance: Accuracy improved from 47.66% to 74.71% (**+27.05%**).

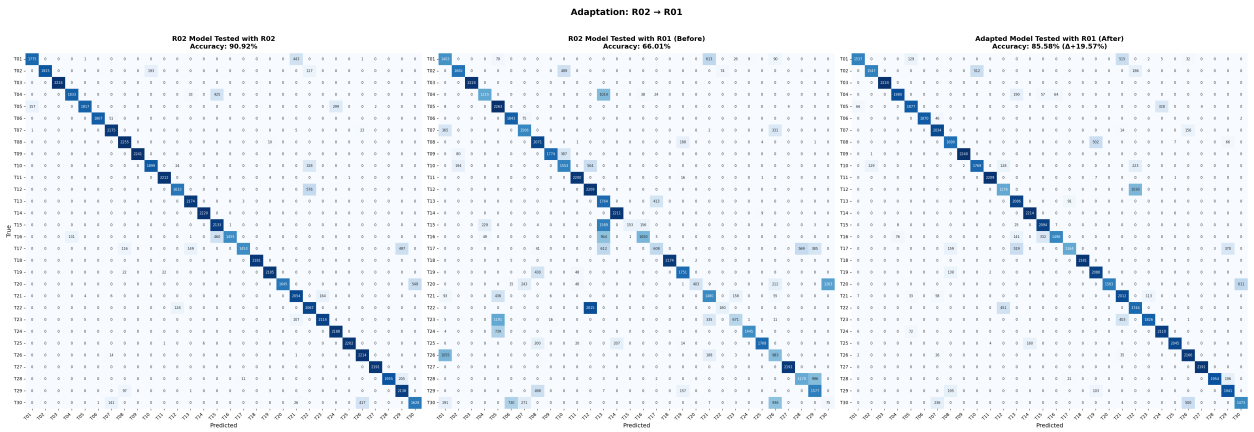


Figure A.5: Adaptation Scenario: R02 → R01.
Source Domain: XTRX (R02). **Target Domain:** XTRX (R01).
Performance: Accuracy improved from 66.01% to 85.58% (**+19.57%**).

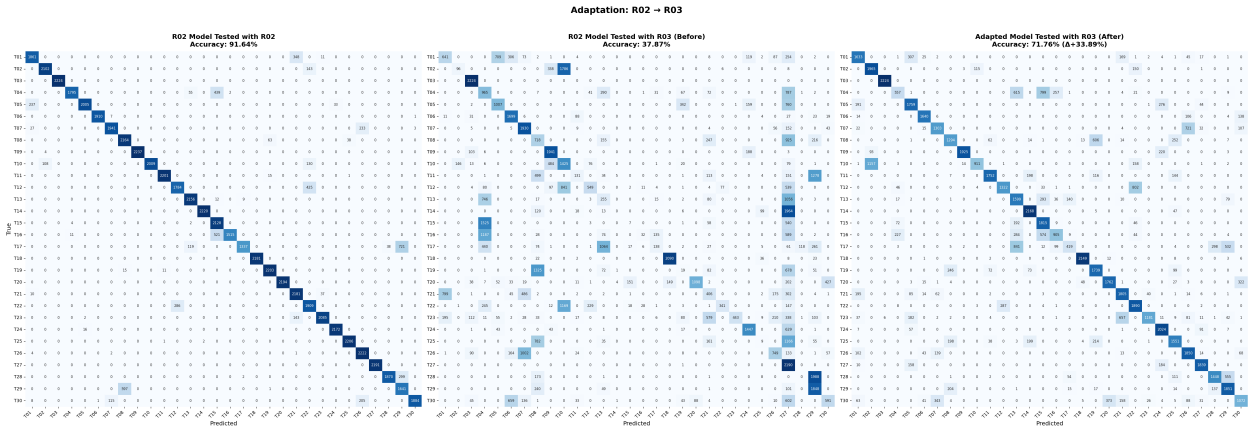


Figure A.6: Adaptation Scenario: R02 → R03.
Source Domain: XTRX (R02). **Target Domain:** B200 (R03).
Performance: Accuracy improved from 37.87% to 71.76% (**+33.89%**).

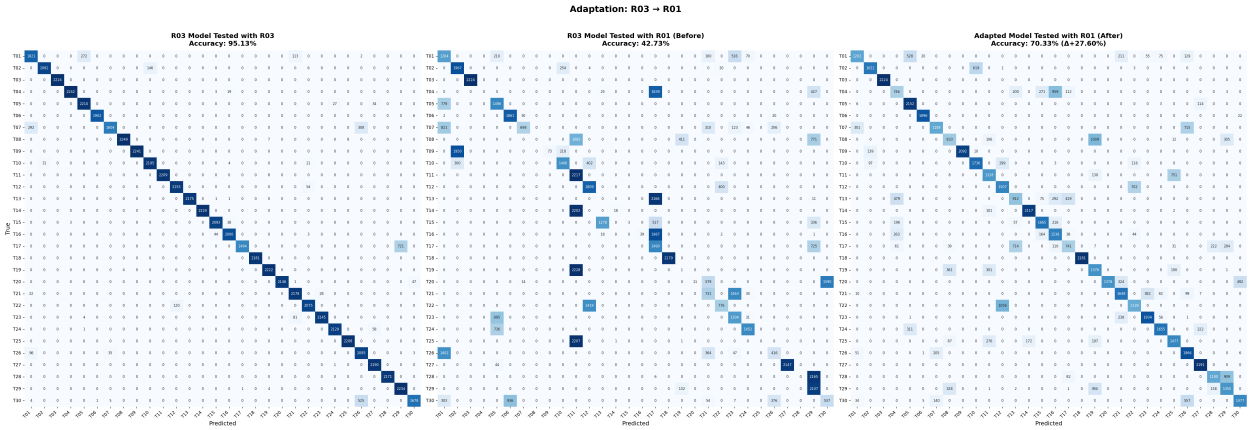


Figure A.7: Adaptation Scenario: R03 → R01.
Source Domain: B200 (R03). **Target Domain:** XTRX (R01).
Performance: Accuracy improved from 42.73% to 70.33% (**+27.60%**).

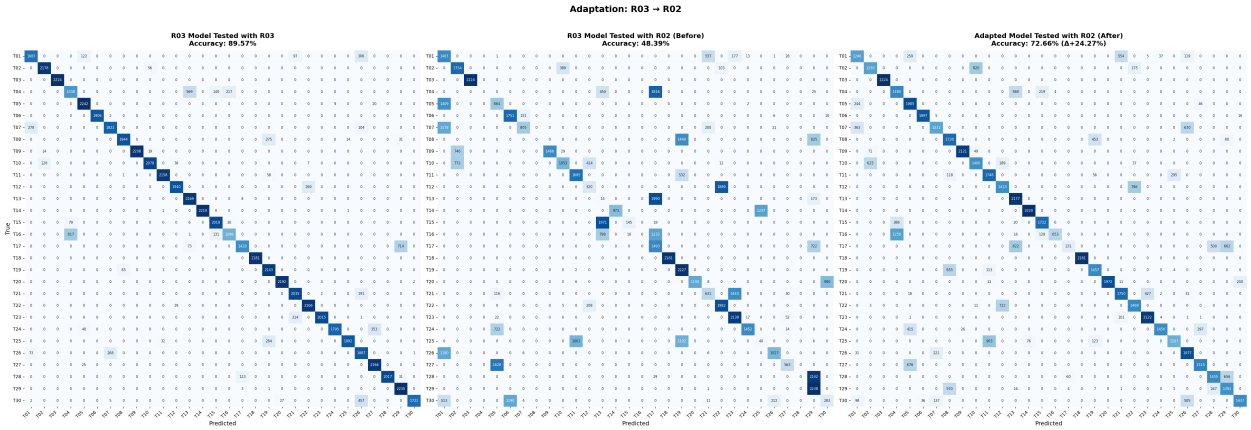


Figure A.8: Adaptation Scenario: R03 → R02.
Source Domain: B200 (R03). **Target Domain:** XTRX (R02).
Performance: Accuracy improved from 48.39% to 72.66% (**+24.27%**).

Appendix B

Security Analysis of RIS-Assisted Physical-Layer Authentication Over Multipath Channels

Accepted for IEEE WIFS 2025 conference.
by Linda Senigagliesi, Anna V. Guglielmi, Marco Baldi, and Stefano Tomasin

Abstract: In physical layer authentication, verification of a user's identity is based on the characteristics of the transmission channel through which signals are delivered to the authenticator (Bob). In this paper, we assume that the signals received by Bob pass through a RIS (controlled by Bob) and that the legitimate transmitter (Alice) is equipped with one antenna. Conversely, the attacker (Trudy) has multiple antennas and uses precoding to deceive Bob's verification. Assuming that Trudy knows all the channel matrices, we first derive her optimal attack strategy. Then, we analyse the conditions under which the channel estimated by Bob is indistinguishable when either Alice or Trudy is transmitting. When Trudy has a single antenna, we show that the indistinguishability condition cannot be met when the channels to the RIS are the result of propagation over multiple paths. For single-path line-of-sight (LOS) conditions, instead, Trudy can impersonate Alice, although transmitting from a different position. We verify these results numerically and assess the security of the considered scenario, even when the indistinguishability conditions cannot be met.

B.1 Introduction

Authentication is the process by which a receiver can verify the identity of a transmitter. Authentication mechanisms based on cryptographic algorithms remain secure provided that no computational breakthrough occurs, i.e., for new attack algorithms or the introduction of quantum computing. They typically entail high complexity, unsuitable in scenarios with limited power and computational resources, e.g., the Internet of Things. Alternative approaches are based on information-theoretic or physical-layer security, which are not affected by the computational capability of attackers. In PLA, transmitters are differentiated only based on the electromagnetic characteristics of their transmission channels.

PLA has been studied in the literature for quite some time, using various features of received signals, such as channel frequency response (CFR) and channel impulse response (CIR), to distinguish a legitimate user from a potential attacker, [30]. Recently, the AoA of the signal has been shown to be a robust feature for PLA, [21, 104]. In addition, user classification has been done using both classical statistical approaches and modern tools based on machine learning.

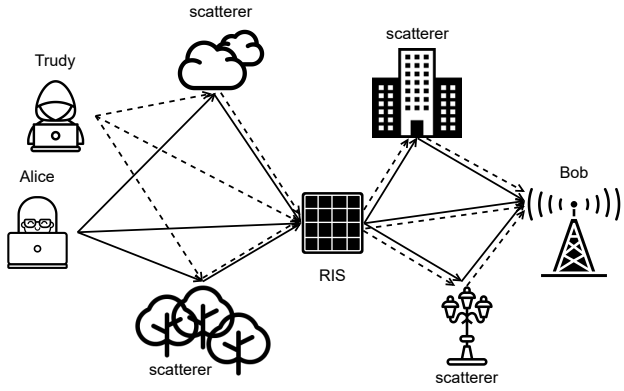


Figure B.1: System model.

In parallel, wireless communications have evolved through the introduction of RISs that, with their ability to shape the propagation environment, improve energy efficiency, reduce hardware complexity, and improve coverage. RISs have also been considered to improve PLA. Variable and random configurations can be set on the RIS to generate challenge-response pairs and propose a challenge-response PLA protocol based on the CSI, [60, 61, 65, 105]. In [106], the authors consider CFR-based PLA in the presence of a hybrid RIS, also capable of acting as a receiver and estimating the channels of impinging signals; thus, this estimate is exploited for authentication. Authentication in a scenario with an RIS is studied also in [107], however, also exploiting pre-shared keys used for asymmetric cryptography; thus, it cannot be considered working purely at the physical layer. In [108] PLA based on the CIR in a dynamic wireless communication environment, is studied, and convolutional neural networks are used to perform classification: this overcomes the limitations of the classical statistical approach based on hypothesis testing when the wireless channel is time-varying.
In this paper, we consider that signals received by Bob are reflected through a RIS that he controls, and the legitimate transmitter, Alice, is equipped with a single antenna. In contrast, the adversary, Trudy, possesses multiple antennas and employs precoding techniques to attempt to bypass the verification process. Assuming Trudy has full knowledge of all channel matrices, we first determine her optimal attack strategy. We then examine the conditions under which Bob's channel estimation is identical regardless of whether Alice or Trudy is transmitting. When Trudy is limited to a single antenna, we derive conditions based on the angle of arrival at the RIS. Our analysis shows that under multipath propagation conditions to the RIS, the indistinguishability requirement cannot be satisfied. However, in the case of a single-path line-of-sight (LOS) scenario, Trudy can successfully impersonate Alice by transmitting from a different location. These findings are supported by numerical simulations. We also evaluate the system's security in situations where indistinguishability cannot be achieved.
The rest of the paper is organized as follows. Section B.2 presents the system model. Section B.3 describes the PLA mechanism and, then, in Section B.4, a security analysis is performed, focusing on conditions that make the attack indistinguishable from a legitimate signal. Numerical results are discussed in Section B.5 and, finally, conclusions are drawn in Section B.6.

B.2 System Model

We consider the uplink scenario shown in Fig. B.1, where the base station (BS) (Bob) aims to authenticate a user equipment (UE) (Alice) in a single-input multiple-output (SIMO) communication system, with Alice equipped with a single antenna and Bob with a uniform linear array (ULA) of M antennas. The signal

transmitted by Alice reaches Bob through a reconfigurable intelligent surface (RIS), while a blockage obstructs the Alice-Bob direct link. An attacker device, Trudy, attempts to impersonate Alice by transmitting messages that Bob may mistake as originating from Alice. Trudy is equipped with a ULA of N_{T} antennas. We also assume that no direct communication is possible between Trudy and Bob, and that all of her messages are transmitted through the RIS.

Transmissions occur at millimeter-wave (mm-Wave) frequencies. ULA antennas are uniformly spaced by a distance $d = \lambda_c/2$, where λ_c is the carrier wavelength. Moreover, we assume that the field of view of Bob is 120°.

The RIS, controlled by Bob, has N reflecting elements spaced by the same distance d. The n-th element, $n = 0, 1, \ldots, N-1$, of the RIS introduces a phase shift $\omega_n = e^{j\varphi_n}$ on the equivalent baseband signal and has unitary gain. The RIS configuration matrix is defined as

$$\boldsymbol{\Omega} = \mathrm{diag}\{[e^{j\varphi_0}, \ldots, e^{j\varphi_{N-1}}]\}. \tag{B.1}$$

We denote the baseband equivalent vector for the channel from Alice to the RIS as $\boldsymbol{f} \in \mathbb{C}^{N\times 1}$, the channel matrix from the RIS to Bob as $\boldsymbol{G} \in \mathbb{C}^{M\times N}$. Thus, the resulting Alice-RIS-Bob cascaded channel is

$$\boldsymbol{h}_{\mathrm{ARB}} = \boldsymbol{G}\boldsymbol{\Omega}\boldsymbol{f}\,. \tag{B.2}$$

Alice transmits suitable pilot symbols to let Bob estimate the channel, which is used for authentication. The pilot signal is assumed to be known to Trudy.

We denote as $\boldsymbol{T}$ the matrix of the channel from Trudy to the RIS. To impersonate Alice, Trudy precodes the transmitted signal (including pilots) with vector $\boldsymbol{q}$ and the resulting Trudy-RIS-Bob channel is then

$$\boldsymbol{h}_{\mathrm{TRB}} = \boldsymbol{G}\boldsymbol{\Omega}\boldsymbol{T}\boldsymbol{q} \in \mathbb{C}^{M\times 1}. \tag{B.3}$$

All channels ($\boldsymbol{f}$, $\boldsymbol{G}$, and $\boldsymbol{T}$) are time-invariant.

B.2.1 Channel Model

In the presence of objects around the transmitter and the receiver, the transmitted signal reaches the receiver through multiple paths. At the mmWave band, channels typically have only a few relevant paths; thus, we use a geometric model for their description. We define the K-size array response column vector for angle of arrival (AoA) θ as

$$\boldsymbol{e}_K(\theta) = \frac{1}{\sqrt{K}}[1, e^{-j\frac{2\pi}{\lambda_c}d\sin\theta}, \ldots, e^{-j(K-1)\frac{2\pi}{\lambda_c}d\sin\theta}]^T. \tag{B.4}$$

For a generic channel with L paths, we define the L-paths array response matrix with AoA angles $\boldsymbol{\theta} = [\theta_1, ..., \theta_L]^T$ as

$$\boldsymbol{E}_N(\boldsymbol{\theta}) = [\boldsymbol{e}_N(\theta_1), ..., \boldsymbol{e}_N(\theta_L)]. \tag{B.5}$$

Let L_f be the number of paths between Alice and the RIS, and $\phi_{f,l}$, $\theta_{f,l}$, and $\gamma_{f,l}$ represent the angle-of-departure (AoD) at Alice, the AoA at the RIS, and the complex path gain for the l-th path i.e., $l = 1, ..., L_f$, respectively. Let us also define $\boldsymbol{\phi}_f = [\phi_{f,1}, ..., \phi_{f,L_f}]^T$ and $\boldsymbol{\theta}_f = [\theta_{f,1}, ..., \theta_{f,L_f}]^T$. Moreover, $\mathbf{1}_{L_f}$, $\boldsymbol{E}_N(\boldsymbol{\theta}_f)$, and $\boldsymbol{\Gamma}_f = \mathrm{diag}([\gamma_{f,1}, ..., \gamma_{f,L_f}]^T)$ denote the L-size column vector of ones corresponding to Alice's array response matrix, the RIS array response matrix, and diagonal path gain matrix, respectively. The baseband channel matrix between Alice and the RIS is modeled as [109]

$$\boldsymbol{f} = \sqrt{\frac{KN}{L_f}}\sum_{l=1}^{L_f}\gamma_{f,l}\boldsymbol{e}_N(\theta_{f,l})\boldsymbol{e}_1^H(\phi_{f,l}) = \boldsymbol{E}_N(\boldsymbol{\theta}_f)\boldsymbol{\Gamma}_f\mathbf{1}_{L_f}\,. \tag{B.6}$$

The RIS-Bob channel matrix is modeled as

$$\boldsymbol{G} = \boldsymbol{E}_M(\boldsymbol{\theta}_G)\boldsymbol{\Gamma}_G\boldsymbol{E}_N^H(\boldsymbol{\phi}_G) \in \mathbb{C}^{M\times N}\,, \tag{B.7}$$

where $\boldsymbol{\theta}_G$ and $\boldsymbol{\phi}_G$ are the vectors of AoAs to Bob and AoDs from the RIS, and $\boldsymbol{\Gamma}_G$ is the diagonal matrix of path gains.
Similarly, the Trudy–RIS channel is modeled as

$$\boldsymbol{T} = \boldsymbol{E}_N(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}_t\boldsymbol{E}_{N_t}^H(\boldsymbol{\phi}_t) \in \mathbb{C}^{N\times N_\mathrm{T}}, \tag{B.8}$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{\phi}_t$ are the vectors of AoAs to the RIS and AoDs from Trudy, and $\boldsymbol{\Gamma}_t$ is the diagonal $L_t \times L_t$ matrix of the L_t path gains.

B.2.2 Assumptions on Trudy

Trudy is assumed to perfectly know all the channels, including the Alice-RIS and RIS-Bob channel matrices $\boldsymbol{f}$ and $\boldsymbol{G}$. This assumption is very generous to Trudy, because she typically is neither co-located with Alice nor Bob. Moreover, the channels corresponding to $\boldsymbol{f}$ and $\boldsymbol{G}$ are only experienced in cascade through the RIS. Note that Alice and Bob can easily estimate the overall cascaded Alice-RIS-Bob channel, while it is harder for them, and even more so for Trudy, to estimate the individual channels represented by $\boldsymbol{f}$ and $\boldsymbol{G}$. Consequently, considering the attacker with complete channel knowledge will result in a conservative estimate of the security performance, corresponding to a worst-case condition for the legitimate receiver.
We also assume that Trudy chooses the transmit power without restrictions. Finally, we assume that neither Alice nor Bob knows the instantaneous channels with Trudy nor their statistics. In particular, Alice and Bob do not know where Trudy is located, so they cannot infer anything about the propagation of signals transmitted or received by Trudy.

B.2.3 Communication-Optimal RIS Configuration

Since the RIS is used for communication purposes between Alice and Bob, its configuration should be optimized accordingly by Bob. We indicate the *communication-optimal RIS configuration* maximizing the spectral efficiency as

$$\overline{\boldsymbol{\Omega}} = \mathrm{diag}(e^{j\bar{\varphi}_1}, \ldots, e^{j\bar{\varphi}_N}), \tag{B.9}$$

where $\bar{\varphi}_n$, $n = 0, \ldots, N-1$, represent the communication-optimal phase shifts of the N RIS elements. Various works in the literature have proposed methods for optimizing the RIS configuration. Here we consider the technique of [110].

B.3 Physical Layer Authentication Mechanism

We consider a physical-layer authentication (PLA) mechanism, where Bob aims at deciding between the two hypotheses

$$\mathcal{H}_0 : \text{the signal comes from Alice,}$$
$$\mathcal{H}_1 : \text{the signal comes from the attacker Trudy.}$$

To this end, the channel vector estimated by Bob operates as a distinguishing feature between the transmissions done by Alice and Trudy.

The PLA mechanism includes two phases, namely the association and verification phases. Since we assume that Bob does not know the cascade channel when Trudy is transmitting, we will not exploit this information for PLA.

In the association phase, Alice transmits some known pilot signal s_0 to Bob, who exploits its knowledge to obtain a noisy estimate of $\boldsymbol{h}_{\text{ARB}}$ that we denote $\bar{\boldsymbol{h}}$. We assume that such a phase is authenticated at a higher layer; thus, it provides a reliable estimate of the Alice-Bob channel. The association phase has to be repeated every time the Alice-Bob channel changes. In the subsequent verification phase, upon reception of a signal Bob estimates the channel over which such a signal traveled, assuming that s_0 was transmitted, and obtaining the estimate $\hat{\boldsymbol{h}}$. Then, Bob performs a test on the obtained estimate to decide whether the transmitter was Alice or not.

Let $\boldsymbol{r}$ denote the signal received by Bob when Alice is transmitting. Assuming that Bob knows s_0 and the communication-optimal RIS configuration $\overline{\boldsymbol{\Omega}}$, the received signal is $\boldsymbol{r} = \boldsymbol{h}_{\text{ARB}} s_0 + \boldsymbol{n}$, where $\boldsymbol{n}$ is a circularly-symmetric complex Gaussian vector with zero mean and variance σ_n^2 per entry. Bob obtains an estimate of the channel as

$$\hat{\boldsymbol{h}} = \frac{\hat{\boldsymbol{r}}}{s_0} = \boldsymbol{h}_{\text{ARB}} + \frac{\boldsymbol{n}}{s_0}. \tag{B.10}$$

Since we do not exploit any information on Trudy's channel for this test, we resort to the likelihood test (LT) on $\hat{\boldsymbol{h}}$, based on the norm-2 distance between the current channel estimate and that obtained in the association phase [111], i.e.,

$$\zeta = \|\hat{\boldsymbol{h}} - \bar{\boldsymbol{h}}\|^2. \tag{B.11}$$

The LT providing a decision $\hat{\mathcal{H}}$ between the two hypotheses is obtained by thresholding ζ as follows

$$\zeta < \tau : \ \hat{\mathcal{H}} = \mathcal{H}_0, \quad \zeta \geq \tau : \ \hat{\mathcal{H}} = \mathcal{H}_1, \tag{B.12a}$$

where τ is a suitably chosen threshold.

B.3.1 Security Metrics

Two possible error events might occur in the authentication mechanism: the false alarm (FA), when Bob discards a message as forged by Trudy while it is coming from Alice, and the misdetection (MD), when Bob accepts a message coming from Trudy as legitimate.

Specifically, an FA occurs when, under hypothesis $\mathcal{H}_0$, $\zeta \geq \tau$, whereas, an MD occurs when, under hypothesis $\mathcal{H}_1$, $\zeta < \tau$. As security metrics, we then consider the probabilities of FA and MD, i.e.

$$P_{\text{FA}} = \mathbb{P}[\zeta \geq \tau | \mathcal{H}_0], \quad P_{\text{MD}} = \mathbb{P}[\zeta < \tau | \mathcal{H}_1]. \tag{B.13}$$

B.4 Security Analysis

We now analyze the security of PLA for the considered scenario. The obtained results will highlight how the structure of the channel, due to the few reflection paths, has an impact on the error probabilities of PLA. First, we compute the optimal precoding vector for Trudy that maximizes the probability of her attack succeeding, i.e., maximizes the MD probability. Then, we discuss the impact of the number of paths on the security.

Let us define the cascade channels when Alice and Trudy are transmitting as

$$\boldsymbol{c}_{\text{A}} = \boldsymbol{E}_M(\boldsymbol{\theta}_G)\boldsymbol{\Gamma}_G \boldsymbol{E}_N^H(\boldsymbol{\phi}_G)\overline{\boldsymbol{\Omega}}\boldsymbol{E}_N(\boldsymbol{\theta}_f)\boldsymbol{\Gamma}_f \mathbf{1}_{L_f}, \tag{B.14}$$

$$\begin{aligned} \boldsymbol{c}_{\text{T}} &= \boldsymbol{E}_M(\boldsymbol{\theta}_G)\boldsymbol{\Gamma}_G \boldsymbol{E}_N^H(\boldsymbol{\phi}_G)\overline{\boldsymbol{\Omega}}\boldsymbol{E}_N(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}_t \boldsymbol{E}_{N_t}^H(\boldsymbol{\phi}_t)\boldsymbol{q} \\ &= \boldsymbol{c}'_{\text{T}}\boldsymbol{q}, \end{aligned} \tag{B.15}$$

where $\boldsymbol{q}$ is the precoding vector used by Trudy to try to falsify Alice's channel. Then, the channel estimated by Bob when Alice is transmitting can be written as $\hat{\boldsymbol{h}}_A = \boldsymbol{c}_\mathrm{A} + \boldsymbol{n}$, while the estimated channel when Trudy is transmitting with precoding vector $\boldsymbol{q}$ is $\hat{\boldsymbol{h}}_T = \boldsymbol{c}'_\mathrm{T}\boldsymbol{q} + \hat{\boldsymbol{n}}$.

B.4.1 Trudy Optimal Transmit Power

Trudy's goal is to maximize the probability that Bob accepts her message as legitimate, i.e., to maximize P_MD. Considering the likelihood (B.11) used in the LT, Trudy must choose $\boldsymbol{q}$ to minimize ζ, as Trudy knows the Alice-Bob cascade channel $\boldsymbol{c}_\mathrm{A}$. However, she does not know the noise of the estimate obtained by Bob in the association phase. Therefore, we obtain the following impersonation optimization problem

$$\boldsymbol{q}^\star = \arg\min_{\boldsymbol{q}} \|\boldsymbol{c}'_\mathrm{T}\boldsymbol{q} - \boldsymbol{c}_\mathrm{A}\|^2 . \tag{B.16}$$

Now, we have

$$\begin{aligned}\zeta &= ||\boldsymbol{c}'_\mathrm{T}\boldsymbol{q} - \boldsymbol{c}_\mathrm{A}||^2 \\ &= \boldsymbol{r}^H\boldsymbol{c}_\mathrm{A} - \boldsymbol{c}_\mathrm{A}^H\boldsymbol{c}'_\mathrm{T}\boldsymbol{q} - \boldsymbol{q}^H\boldsymbol{c}_\mathrm{T}^{'H}\boldsymbol{r} + \boldsymbol{q}^H\boldsymbol{c}_\mathrm{T}^{'H}\boldsymbol{c}'_\mathrm{T}\boldsymbol{q},\end{aligned} \tag{B.17}$$

and by nulling the derivative with respect to q, the solution of the minimization problem (B.16) is

$$\boldsymbol{q}^\star = \boldsymbol{c}_\mathrm{T}^{'H}(\boldsymbol{c}'_\mathrm{T}\boldsymbol{c}_\mathrm{T}^{'H})^{-1}\boldsymbol{c}_\mathrm{A}. \tag{B.18}$$

B.4.2 Indistinguishability Conditions

When $\zeta = 0$, the Alice-Bob channel is indistinguishable from the Trudy-Bob channel, and Bob cannot detect an attack. Let us investigate which are the conditions under which this may occur. Clearly, when Trudy is in the same position as Alice, they have the same channel to Bob. The interesting point here is to understand if there are other positions of Trudy that (together with some optimum precoding vector $\boldsymbol{q}$) provide the same indistinguishability condition. Such positions may exist, since Bob estimates only the *cascade channel* from Alice, and signals transmitted by Trudy pass through the same RIS used by Alice. From (B.16), we note that indistinguishability is achieved when the system of complex linear equations

$$\boldsymbol{c}'_\mathrm{T}\boldsymbol{q} = \boldsymbol{c}_\mathrm{A} \tag{B.19}$$

is solvable. However, determining the general conditions on the Trudy-RIS channel that ensure the solution is challenging. Therefore, in the following, we focus on the special case in which Trudy also has a single transmit antenna, for which a theoretical analysis is feasible.

B.4.3 Indistinguishability Conditions for $N_\mathrm{T} = 1$

Let us focus on the case in which Trudy has a single antenna and both Alice-RIS and Trudy-RIS channels have L paths. Thus (B.15) becomes

$$\boldsymbol{c}_\mathrm{T} = \boldsymbol{E}_M(\boldsymbol{\theta}_G)\boldsymbol{\Gamma}_G\boldsymbol{E}_N^H(\boldsymbol{\phi}_G)\overline{\boldsymbol{\Omega}}\boldsymbol{E}_N(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}_t\mathbf{1}_L q\,, \tag{B.20}$$

and the precoding vector boils down to the scalar q.
To understand the conditions for indistinguishability in this case, let us define $\boldsymbol{W} = \boldsymbol{E}_M^H(\boldsymbol{\theta}_G)\boldsymbol{E}_M(\boldsymbol{\theta}_G) \in \mathbb{C}^{L_G\times L_G}$ as the matrix with entry $[\boldsymbol{W}]_{ii} = M$ and

$$[\boldsymbol{W}]_{ij} = \sum_{m=1}^{M} e^{-j(m-1)\kappa(\sin\theta_{G,i} - \sin\theta_{G,j})}\,, \quad \text{for } i \neq j \tag{B.21}$$

z_A as a L_G-size vector with entry $l_1 = 1, \ldots, L_G$

$$[z_A]_{l_1} = \sum_{l_2=1}^{L_f} \gamma_{f,l_2} \sum_{n=1}^{N} e^{-j[\kappa(n-1)\mu_{A,l_1 l_2} + \bar{\varphi}_n]}, \tag{B.22}$$

for $\mu_{A,l_1 l_2} = (\sin\phi_{G,l_1} - \sin\theta_{f,l_2})$, and z_T as a L_G-size vector with entry

$$[z_T]_{l_1} = \sum_{l_2=1}^{L_t} \gamma_{t,l_2} \sum_{n=1}^{N} e^{-j[\kappa(n-1)\mu_{T,l_1 l_2} + \bar{\varphi}_n]}, \tag{B.23}$$

for $\mu_{T,l_1 l_2} = \sin\phi_{G,l_1} - \sin\theta_{t,l_2}$. We also have

$$\boldsymbol{c}_{\mathrm{A}}^H \boldsymbol{c}_{\mathrm{A}} = \boldsymbol{z}_A^H \boldsymbol{\Gamma}_G^H \boldsymbol{W} \boldsymbol{\Gamma}_G \boldsymbol{z}_A, \tag{B.24}$$

$$\boldsymbol{c}_{\mathrm{T}}^{'H} \boldsymbol{c}'_{\mathrm{T}} = \boldsymbol{z}_T^H \boldsymbol{\Gamma}_G^H \boldsymbol{W} \boldsymbol{\Gamma}_G \boldsymbol{z}_T, \tag{B.25}$$

$$\boldsymbol{c}_{\mathrm{A}}^H \boldsymbol{c}'_{\mathrm{T}} = \boldsymbol{z}_A^H \boldsymbol{\Gamma}_G^H \boldsymbol{W} \boldsymbol{\Gamma}_G \boldsymbol{z}_T, \tag{B.26}$$

$$\boldsymbol{c}_{\mathrm{T}}^{'H} \boldsymbol{c}_{\mathrm{A}} = \boldsymbol{z}_T^H \boldsymbol{\Gamma}_G^H \boldsymbol{W} \boldsymbol{\Gamma}_G \boldsymbol{z}_A = (\boldsymbol{c}_{\mathrm{A}}^H \boldsymbol{c}'_{\mathrm{T}})^H. \tag{B.27}$$

Now, substituting (B.24), (B.25), (B.26), and (B.27) into (B.17), and for $\tilde{\boldsymbol{W}} = \boldsymbol{\Gamma}_G^H \boldsymbol{W} \boldsymbol{\Gamma}_G$, we have

$$\zeta = \boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_A - q \boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T - q^* \boldsymbol{z}_T^H \tilde{\boldsymbol{W}} \boldsymbol{z}_A + q q^* \boldsymbol{z}_T^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T \,. \tag{B.28}$$

Defining $b{=}\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_A$, $c{=}\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T$, and $d{=}\boldsymbol{z}_T^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T$, (B.28) becomes

$$\zeta = d|q|^2 - cq - (cq)^* + b \,. \tag{B.29}$$

We are now ready to investigate the indistinguishability condition. Replacing $q = \beta e^{j\alpha}$ in (B.29), such condition can be written as

$$d\beta^2 - 2|c|\beta\cos(\alpha + \rho) + b = 0 \,, \tag{B.30}$$

with $c = |c|e^{j\rho}$. We firstly note that (by definition) $\zeta \geq 0$ and it is minimized for $\alpha^\star = -\rho$. Substituting $\alpha^\star$ in (B.30), we have $d\beta^2 - 2|c|\beta + b = 0$, which has solutions only if $|c|^2 - bd \geq 0$, or, equivalently, if

$$|\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T|^2 \geq (\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_A)(\boldsymbol{z}_T^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T). \tag{B.31}$$

However, by the Cauchy-Schwarz inequality

$$|\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T|^2 \leq (\boldsymbol{z}_A^H \tilde{\boldsymbol{W}} \boldsymbol{z}_A)(\boldsymbol{z}_T^H \tilde{\boldsymbol{W}} \boldsymbol{z}_T), \tag{B.32}$$

and thus (B.31) must hold with equality. However, this happens if and only if $\sqrt{\tilde{\boldsymbol{W}}}\boldsymbol{z}_A$ and $\sqrt{\tilde{\boldsymbol{W}}}\boldsymbol{z}_T$ are linearly dependent. Note that this does not generally imply $\boldsymbol{z}_A$ and $\boldsymbol{z}_T$ to be linearly dependent unless $\tilde{\boldsymbol{W}}$ is a full rank matrix. By definition, the rank of $\tilde{\boldsymbol{W}}$ is the same of $\boldsymbol{W}$ (due to $\boldsymbol{\Gamma}_G$ being diagonal), which is full rank if and only if the vectors $\{\boldsymbol{e}_M(\theta_{G,i})\}_{i=1}^{L_G}$ (i.e., the columns of $\boldsymbol{E}_M(\boldsymbol{\theta}_G)$) are linearly independent. This condition is satisfied when $L_G \leq M$ and the angles $\theta_{G,i}$ related to the different paths are distinct, i.e., $\sin\theta_{G,i} \neq \sin\theta_{G,j}$, $\forall i, j = 1, \ldots, L_G$, with $i \neq j$. Since each entry of $\boldsymbol{W}$ is given by the inner product of array response vectors (B.21), which depend only on $\sin(\cdot)$ and are periodic over π for ULAs with half-wavelength spacing, we must have

$$\theta_{G,i} \neq \theta_{G,j} + u\,\pi, \tag{B.33}$$

for any integer u. Since we assume Bob has a field of view of $\frac{2}{3}\pi$, we are also ensuring $\boldsymbol{W}$ to be full rank when $L_G \leq M$. In this case, it can be stated that (B.31) holds with equality if and only if $\boldsymbol{z}_A$ and $\boldsymbol{z}_T$ are

linearly dependent. From the definitions in (B.22) and (B.23), we conclude that the indistinguishability conditions require that Alice and Trudy have the same number of paths ($L_t = L_f$), the AoA angles at the RIS corresponding to Alice and Trudy match exactly, yielding

$$\sin\theta_{f,l} = \sin\theta_{t,l}, \quad l = 1, \ldots, L_t = L_f\,, \tag{B.34}$$

and their path gains are proportional, i.e.,

$$\gamma_{f,l} = \lambda\,\gamma_{t,l}\,, \quad l = 1, \ldots, L_t = L_f\,. \tag{B.35}$$

These are then the indistinguishability conditions for $N_{\mathrm{T}} = 1$.

B.4.4 Single-Path RIS-Bob Channel

When the RIS–Bob channel is single-path (L_G=1), z_A and z_T collapse to complex scalars. This dimensionality reduction significantly simplifies the attacker's task, as linear dependence now can be trivially achieved in $\mathbb{C}$, where any two non-zero scalars are always linearly dependent if one is a scaled version of the other. Hence, it becomes easier for the attacker to find values of α and β such that (B.30) is satisfied. Indeed, in this case, even when Trudy does not show the same angles and path gains of Alice ($z_T \neq z_A$), indistinguishability can still be achieved by appropriately tuning α and β so that (B.30) holds. In formulas, this happens for

$$\alpha = -\rho + u\pi, u \text{ even}, \alpha \in [-\pi, \pi], \text{ and } \beta = \frac{|z_A|}{|z_T|} \tag{B.36}$$

or

$$\alpha = -\rho + u\pi, u \text{ odd}, \alpha \in [-\pi, \pi], \text{ and } \beta = -\frac{|z_A|}{|z_T|}. \tag{B.37}$$

The case L_G=1 inherently poses a higher impersonation risk, as it offers fewer spatial degrees of freedom to differentiate between Alice and Trudy.
This result could also be directly inferred from the structure of the cascaded channels in (B.14) and (B.20). Since the common term $\boldsymbol{E}_M(\boldsymbol{\theta}_G)\boldsymbol{\Gamma}_G\boldsymbol{E}_N^H(\boldsymbol{\phi}_G)$ of the RIS-Bob channel has rank 1, the cascaded channels lie in the same one-dimensional subspace. Therefore, no matter how different Trudy's and Alice's angles and path gains are, once they pass through it, the result is always confined to a single spatial direction, limiting Bob's ability to distinguish between them. In fact, any differences in Alice and Trudy transmissions are effectively collapsed into a single direction by the rank-one projection of $\boldsymbol{G}$ and, then, Trudy can more easily mimic Alice's cascaded channel.

B.5 Numerical Results

In this section, we assess the performance of the considered authentication method investigating both single-path (i.e., $L_G = 1$) and multipath (i.e., $L_G = 3$) scenarios for the RIS-Bob channel. We consider $L_f = L_t = 3$ and path gains $\gamma_{f,l}$, $\gamma_{G,l}$, and $\gamma_{t,l}$ distributed as $\mathcal{CN}(0, 1)$. We assume that the angles at the RIS and the AoDs from the transmitters are uniformly distributed in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, while the AoAs at Bob are uniformly distributed in the range $\left[-\frac{\pi}{6}, \frac{\pi}{6}\right]$. Angles and gains are generated independently for Alice and Trudy. Bob is equipped with $M \in \{4, 8, 16, 32\}$ antennas, while Alice and Trudy are single-antenna devices. The number of RIS elements is $N = 64$.
Fig. B.2 shows a contour plot of the test function ζ under attack conditions for a single-path RIS-Bob channel (i.e., $L_G = 1$). Note that different angles and path gains for the Trudy-RIS and Alice-RIS channels are considered. The red cross marks the values of α and β that minimize ζ: when Trudy chooses the value of $q^\star$ corresponding to these optimal values of α and β, we have $\zeta = 0$.

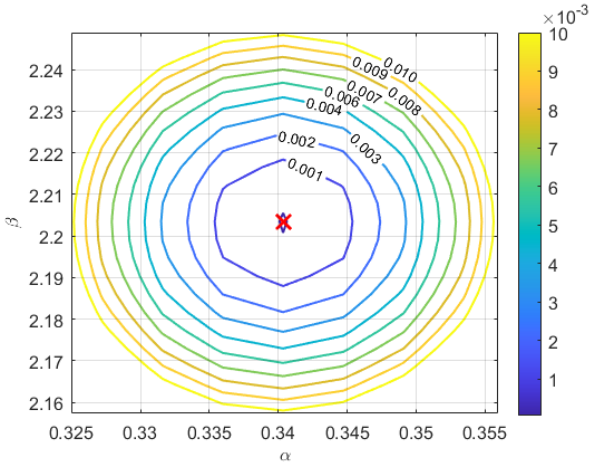


Figure B.2: Contour plot of ζ (under hypothesis $\mathcal{H}_1$) for L_G=1, L_f=L_t=3, M=16, N=64. The red cross marks the values of α and β that minimize ζ. We consider different angles and path gains for the Trudy-RIS and Alice-RIS channels.

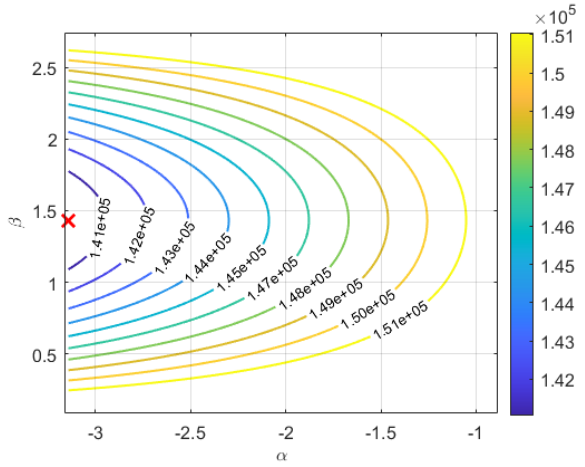


Figure B.3: Contour plot of ζ (under hypothesis $\mathcal{H}_1$) for L_G=L_f=L_t=3, M=16, N=64. The red cross marks the values of α and β that minimize ζ. We consider different angles and path gains for the Trudy-RIS and Alice-RIS channels.

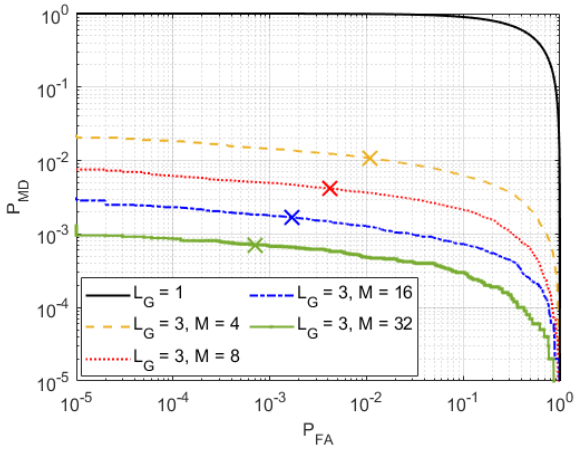


Figure B.4: DET curves for different value of M, $L_G \in \{1, 3\}$. The crosses mark the points for which $P_{\text{MD}} = P_{\text{FA}}$.

Similarly, Fig. B.3 shows a contour plot of the test function ζ under attack conditions for $L_G = 3$. Comparing Figs. B.3 and B.2, we observe that, for $L_G > 1$, even if Trudy uses the optimal $q^\star$, the resulting minimum of the test function ζ is strictly greater than zero. This confirms that, unlike the scenario with $L_G = 1$, perfect impersonation becomes impossible to achieve. Indeed, the presence of L_G paths increases the rank of the RIS–Bob channel matrix, thereby introducing additional spatial diversity that makes it harder for Trudy to align her cascade channel with that of Alice by setting the proper $q^\star$.
The result is also confirmed by Fig. B.4, which shows the detection error trade-off (DET) curves for different values of M and $L_G \in \{1, 3\}$. The crosses mark the points for which $P_{\text{MD}} = P_{\text{FA}}$. All the curves show that reducing P_{FA} results in an increase in P_{MD}, and vice versa. It can also be noticed that for $L_G = 1$, we have $P_{\text{MD}} = 1 - P_{\text{FA}}$, regardless of the number of Bob's antennas M. In fact, in this case, Trudy can always find an attack strategy that yields to indistinguishability with Alice; thus the probability that Bob decides for hypothesis $\mathcal{H}_1$ (i.e., attack condition) is the same irrespective of who is transmitting. For $L_G > 1$, instead, the optimal attack does not usually lead to indistinguishability (since the AoAs from Trudy and Alice are independent). Indeed, the DET curves do not start from the top-left corner as is typically the case. This is due to the statistical nature of the test and imperfections in Trudy's impersonation of Alice. In fact, when $L_G > 1$, the perfect alignment between Trudy's and Alice's cascaded channels is not achievable, even if Trudy uses $q^\star$. Hence, the minimum achievable P_{MD} is strictly less than 1, emphasizing a significant limit on the success of the impersonation attack. Hence, we can conclude that a higher L_G enhances authentication robustness by limiting the ability of Trudy to fully mimic Alice's cascaded channel. Moreover, we observe that, as M increases, the DET curves move towards smaller P_{MD} for a target P_{FA}. This shows that having more receive antennas allows for better distinction between Alice and Trudy.

B.6 Conclusions

We analyzed the security of a RIS-assisted PLA scheme in scenarios with no direct link between the transmitter and the receiver, and multipath propagation conditions of the channels to and from the RIS. Assuming the worst case scenario of an attacker Trudy having full channel knowledge, we determined her optimal attack strategy. Then, we examined the conditions under which Bob's channel estimation may have the same statistics regardless of whether Alice or Trudy is transmitting, deriving the conditions based on the AoAs at the RIS for single antenna attacker. Numerical results show that when the RIS–Bob channel is

single-path, impersonation is feasible even with mismatched channel parameters. Conversely, increasing the number of RIS–Bob paths significantly enhances authentication robustness by limiting the attacker's ability to mimic the legitimate user.

Appendix C

Adversarial ML for Channel-based Key Agreement for 6G Newtorks

To be submitted.
by Mattia Piana, Francesco Ardizzon, and Stefano Tomasin

Abstract: With the increasing use of 5G and B5G, there is a need for mechanisms that guarantee security features to such protocols, such as confidentiality, integrity, and availability. Such features typically require that legitimate devices be provided with secret keys. In this paper, we propose a physical layer key generation (PKG) scheme that takes advantage of the deployment of dense devices in cellular networks to extract keys from the observation of the near field channel. In particular, the legitimate transmitter and receiver process the respective near-field channel observation by using two neural network (NN)-based key extractors to extract a raw key. Drawing inspiration from the lower bound on the secret key capacity, we train the key extractors to maximize the reciprocity and uniformity, while minimizing the information leakage to the eavesdropper. Specifically, a custom architecture and loss function have been designed to accomplish these tasks. The results highlight the benefit of the proposed mechanism with respect to the current state of the art.

C.1 Introduction

With the advent of 6G, more devices are expected to be deployed in the user environment, providing a wide range of services that go far beyond the scope of providing communications.
However, while new technologies and protocols guarantee better performance in quality of service (QoS), there is an increased need to also provide more secure and robust services. Additionally, it is worth noting that most of the communication is typically wireless, thus exposed to any sufficiently close malicious user, who may access the channel to eavesdrop on the communication or to inject false signals. Examples include denial of service (DoS), where the attacker aims to disrupt the link between the nodes of the network, eavesdropping, where they try to disclose the private communication between the legitimate user, and even impersonation, where the attacker pretends to be a legitimate user to intrude on the network.
Many security mechanisms can be adopted to counter these threats, often based on cryptography, but a common requirement is that they require the legitimate party to have a common but secret key. This includes, for instance, symmetric encryption schemes such as advanced encryption standard (AES), where the key is used to encode and decode the secret message. Thus, all these mechanisms must be coupled with key management and key distribution systems for secret keys. However, to avoid storing a large dataset of keys on each device, it is advisable to resort to key generation schemes, where fresh keys are periodically generated and used for other services.

We consider a physical layer key generation (PKG) scheme that exploits the dense deployment of devices in a 5G and B5G scenario, to generate secure keys by processing the transmitter and receiver near-field observations by exploiting the channel reciprocity. With respect to their crypto-based counterpart, PKG schemes are information-theoretic secure, quantum-resistant, and require less computational power [112, Ch. 1]. In particular, legitimate users Alice and Bob use a PKG scheme to obtain a key that must stay secret from eavesdropper Eve. The protocol involves four steps. First, during *channel probing*, Alice and Bob exchange pilot signals over the public channel. Next, during the *advantage distillation* (AD), Alice and Bob independently process the observation to obtain a raw key. Next, during the *information reconciliation*, Alice and Bob communicate over the public channel to perform error correction between the raw keys. Finally, during *privacy amplification*, Alice and Bob reduce the information common to Eve, typically by using hash functions. More details about PKG can be found in [112, Ch. 4]. Here, we will focus on the first steps, in particular, the AD.
Recently, several strategies for PKG have been proposed. In the context of 5G, an attention-based convolutional autoencoder (AE) was proposed in [113], where the encoder processes the channel impulse response (CIR), where Eve's channel is statistically independent of Alice and Bob, and the Alice-Bob channel is reciprocal.
Specifically concerning the near field, in [114], the authors propose a design of the precoders to maximize the secret key rate (SKR). In particular, the precoders are optimized to inject randomness while adding artificial noise to the channel orthogonal to the legitimate transmission, thus limiting the information leakage to the eavesdropper.
In [115], an advantage distillation protocol for underwater acoustic channel (UWAC) was proposed, where the raw key extractor was an neural network (NN) trained to maximize randomness and Alice's and Bob's raw bit sequence reciprocity, while minimizing the leakage to Eve. However, this approach assumes the legitimate channels to be symmetrical, which is not true in general, e.g., due to hardware receiver non-idealities.
Specifically, regarding the near-field, the authors of [116] propose a mechanism that uses a random precoder to induce artificial randomness, thus increasing the entropy from which the user extracts the secret keys, therefore countering the possibly low-entropy static channel.
In this work, we propose a novel strategy for the design of the quantizers used by Alice and Bob in the advantage distillation step of the PKG procedure, used to extract raw keys from the measurements obtained from a near-field channel. This strategy is based on an adversarial learning procedure, where Alice's and Bob's quantizers are jointly trained to output bit sequences that achieve high reciprocity and uniformity.
The rest of the work is organized as follows. Section C.2 details the system model. Section C.3 describes the proposed protocol. Numerical results are presented and discussed in Section C.4. Section C.5 draws the conclusions.

C.2 System Model

User Alice is connected via an UWAC to user Bob. Alice and Bob wish to exploit the UWAC for PKG as a source of randomness to extract a common key, i.e., a binary sequence of length b so that Alice's and Bob's keys should coincide while being secret to any eavesdropper observing the channel, Eve. We assume Eve to be a passive device, thus not transmitting signals over the channel.
The first step of the secret key agreement (SKA) procedure, channel probing, involves Alice and Bob transmitting signals with public pilot symbols via UWAC. From such an exchange, each user measures via channel estimation a CIR, from which a set of channel features is extracted. In particular, we call $\boldsymbol{x}_{\mathrm{A}}$, $\boldsymbol{x}_{\mathrm{B}}$, and $\boldsymbol{x}_{\mathrm{E}}$ the feature observations obtained by Alice, Bob, and Eve, Drawing from previous works, e.g., [117], we consider as features: the number of channel taps, the average tap power, the relative root mean square (RMS) delay, and the smoothed received power. These features have been shown be strongly characterize the

transmitter and receiver relative position for the UWAC context [64], thus are well suited for the considered application. Still, the proposed strategy may be extended to consider a broader set of features, e.g., see [118], at the cost of a higher complexity. Thus, in our case, $\boldsymbol{x}_\mathrm{A} = [x_{\mathrm{A},1}, \ldots, x_{\mathrm{A},K}]$, $\boldsymbol{x}_\mathrm{B} = [x_{\mathrm{B},1}, \ldots, x_{\mathrm{B},K}]$, and $\boldsymbol{x}_\mathrm{E} = [x_{\mathrm{E},1}, \ldots, x_{\mathrm{E},K}] \in \mathbb{R}^K$.

We assume the presence of a publicly available dataset. This will be exploited by Alice and Bob to train the NNs used for the raw key extraction. Being this dataset public, it is also available to Eve. Thus, while it cannot be used to compute the actual sequences, it can be used for training purposes, assuming that, in turn, Eve will exploit the very same dataset to train her own extractor. In details, we consider a training dataset $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}\}$, where each entry is a triplet collecting Alice, Bob, and Eve observations, as $\boldsymbol{x} = (\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}, \boldsymbol{x}_\mathrm{E})$.

C.3 Proposed Strategy

Our aim is to design Alice's and Bob's raw key extractor functions, i.e., the function that extracts the bit sequences

$$\boldsymbol{y}_\mathrm{A} = f_\mathrm{A}(\boldsymbol{x}_\mathrm{A}), \quad \boldsymbol{y}_\mathrm{B} = f_\mathrm{B}(\boldsymbol{x}_\mathrm{B}), \tag{C.1}$$

where $\boldsymbol{x}_\mathrm{A}$ and $\boldsymbol{x}_\mathrm{B}$ are the feature vector extracted from the CIR obtained after channel probing. The extracted sequences $\boldsymbol{y}_\mathrm{A} \in \{0,1\}^b$ and $\boldsymbol{y}_\mathrm{B} \in \{0,1\}^b$ must have good agreement, be highly correlated, and must be random but secret to Eve. In turn the eavesdropper extracts the sequence as $\boldsymbol{y}_\mathrm{E} = f_\mathrm{E}(\boldsymbol{x}_\mathrm{E})$.

To evaluate the performance of the extracted raw key sequences via PKG, i.e., the amount of useful information for the PKG, we employ a lower bound on the secret key capacity,

$$C_\mathrm{sk}^\mathrm{low} = I(\boldsymbol{y}_\mathrm{A}; \boldsymbol{y}_\mathrm{B}) - \max\left\{I(\boldsymbol{y}_\mathrm{A}; \boldsymbol{y}_\mathrm{E}), I(\boldsymbol{y}_\mathrm{B}; \boldsymbol{y}_\mathrm{E})\right\}, \tag{C.2}$$

where $I(\boldsymbol{y}_1; \boldsymbol{y}_2)$ is the mutual information between sequences $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, i.e.,

$$I(\boldsymbol{y}_\mathrm{A}; \boldsymbol{y}_\mathrm{B}) = H(\boldsymbol{y}_\mathrm{A}) + H(\boldsymbol{y}_\mathrm{B}) - H(\boldsymbol{y}_\mathrm{A}, \boldsymbol{y}_\mathrm{B}), \tag{C.3}$$

while $H(\cdot)$ and $H(\cdot,\cdot)$ are the entropy and the joint entropy of the extracted sequences, respectively.

Interestingly, as previously pointed out also in [115], the definitions (C.2) and (C.3) provide several insights that can be used to design the key extractor. First, (C.3) hints that the entropies of each sequence, in particular $H(\boldsymbol{y}_\mathrm{A})$ and $H(\boldsymbol{y}_\mathrm{B})$, should be high. This is achieved when both sequences are random, with the maximum entropy achieved when $y \sim \mathcal{U}(\{0,1\}^b)$. We call this property *randomness*. Additionally, (C.3) also requires that the joint entropy $H(\boldsymbol{y}_\mathrm{A}, \boldsymbol{y}_\mathrm{B})$ to be low, which is achieved when the two sequences are one a deterministic function of the other. We call this *reciprocity*. Finally, the last term in (C.2) deals with the information obtained by Eve about Alice or Bob. We refer to this as *information leakage*.

These requirements will be taken into account to design both the architectures and the loss function of the proposed NN-based raw key extractors, described in detail in the next sections.

In particular, we consider a *training phase* during which Alice and Bob jointly train the NNs-based raw key extractors. Next, during the *exploitation* or *inference* phase and after channel probing, Alice and Bob use their own NN paired with a uniform quantizer with the desired number of levels to extract the raw key sequences $\boldsymbol{b}_A$ and $\boldsymbol{b}_B$. Such keys will be the input of the information reconciliation and then, the privacy amplification.

The loss function $L(\cdot)$ will have three components, which models randomness, the reciprocity, and information leakage. More in detail, given the training dataset $\boldsymbol{\mathcal{X}}$, where each entry is $\boldsymbol{x} = (\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}, \boldsymbol{x}_\mathrm{E})$, and the weights collection $\boldsymbol{\theta} = (\theta_\mathrm{A}, \theta_\mathrm{B}, \theta_\mathrm{E}, \theta_\mathrm{dec}, \theta_\mathrm{dis})$, the raw key extractor will be trained to minimize loss.

$$\mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}) = \alpha\mathcal{L}_1(\boldsymbol{x}_\mathrm{A},\boldsymbol{x}_\mathrm{B}; \theta_\mathrm{A}) + (1-\alpha)\mathcal{L}_2(\boldsymbol{x}_\mathrm{A}; \theta_\mathrm{A}) - \beta\mathcal{L}_3(\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{E}; \theta_\mathrm{A}), \tag{C.4}$$

where α and β are user-defined parameters, weighting each loss component. In the next, we describe in detail each term and its rationale.

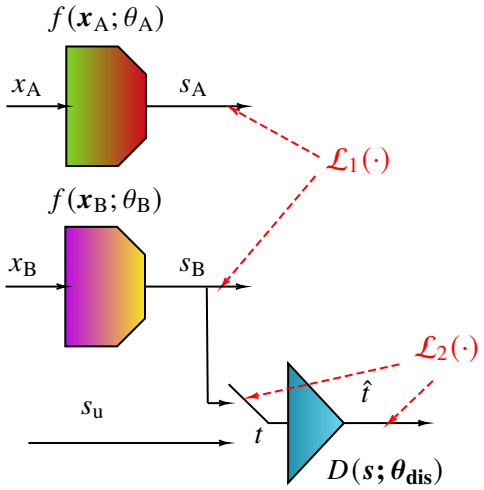


Figure C.1: Training architecture of the raw key extraction functions, $f(x;\theta_A)$ and $f(x;\theta_B)$, where each subsequent block is paired to respective loss.

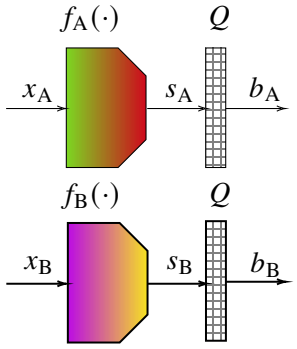


Figure C.2: Raw key extraction exploitation for Alice and Bob, where the raw key extractors are paired with a uniform quantizer, Q.

C.3.1 Reciprocity Enhancement

The first loss term, $\mathcal{L}_1(\cdot)$, concerns the reciprocity enhancement. Our goal is to have a raw key extractor function that retains the correlation between Alice's and Bob's feature vectors and thus to extract the matching information between Alice and Bob. Thus, we define

$$\mathcal{L}_1(\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}; \theta_\mathrm{A}) = \|f(\boldsymbol{x}_\mathrm{A}; \theta_\mathrm{A}) - f(\boldsymbol{x}_\mathrm{B}; \theta_\mathrm{B}))\|^2 \ . \tag{C.5}$$

C.3.2 Randomness

The aim of the second loss term, $\mathcal{L}_2(\cdot)$, is to force the raw key extractor output, $\tilde{\boldsymbol{s}}_\mathrm{A}$ (or $\tilde{\boldsymbol{s}}_\mathrm{B}$), to be uniformly distributed. Without loss of generality, we will consider the space as dominion $[-1, 1]^M$. We remark that, during the training, the output is a vector of M scalar values.
As discussed in [115], this also means that the raw key extrators should be indistinguishable from the output of a random source drawing sequences from $\boldsymbol{s}_\mathrm{u} \sim \mathcal{U}([-1, 1]^M)$. In order to do so, we introduce a third NN, the discriminator, modeled by the function $D(\boldsymbol{s}, \theta_\mathrm{dis})$, that is a binary classifier must distinguishing the key extractor outputs $\boldsymbol{s}_\mathrm{A}$ from the target distribution samples $\boldsymbol{s}_\mathrm{u}$.
Formally, let us introduce the binary tag t, such that, for the generic discriminator input $\boldsymbol{s}$ with output $\hat{t}$, it holds

$$t = \begin{cases} 0 & \text{if } \boldsymbol{s} = \tilde{\boldsymbol{s}} \\ 1 & \text{if } \boldsymbol{s} = \boldsymbol{s}_\mathrm{u} \end{cases} \ , \ \hat{t} = \begin{cases} 0 & \text{if } h(\boldsymbol{s}, \theta_\mathrm{dis}) \geq \lambda \\ 1 & \text{if } h(\boldsymbol{s}, \theta_\mathrm{dis}) < \lambda \end{cases} \ , \tag{C.6}$$

where λ is a user-defined parameter, tuned to match the desired false alarm probability, i.e., the probability $P(\hat{t} = 1 | t = 0)$.
The discriminator is a binary classifier with the cross-entropy as a loss function

$$\mathcal{L}_2(\boldsymbol{x}_\mathrm{B}; \theta_\mathrm{B}) = t \log D(\boldsymbol{s}, \theta_\mathrm{dis}) + (1 - t) \log(1 - D(\boldsymbol{s}, \theta_\mathrm{dis})). \tag{C.7}$$

C.3.3 Information Leakage

The loss term $\mathcal{L}_3(\cdot)$ measures the relation between Alice's and Eve's extracted bit sequences. Indeed, we want the key extractor to extract a key as uncorrelated as possible to one of Eve.
Inspired by [119], we consider the loss function

$$\mathcal{L}_3(\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{E}; \theta_\mathrm{A}) = \|f(\boldsymbol{x}_\mathrm{A}; \theta_\mathrm{A}) - f(\boldsymbol{x}_\mathrm{E}; \theta_\mathrm{E})\|^2 \ . \tag{C.8}$$

Finally, notice that the same process can be repeated by substituting $\boldsymbol{x}_\mathrm{B}$ to $\boldsymbol{x}_\mathrm{A}$ in (C.9). Still, this is not necessary, since, thanks to the reciprocity loss $\mathcal{L}_1(\cdot)$, we have $f(\boldsymbol{x}_\mathrm{A}; \theta_\mathrm{A}) \approx f(\boldsymbol{x}_\mathrm{B}; \theta_\mathrm{B})$.

C.4 Numerical Results

In this Section, we report the performance of the proposed raw key extractors. In particular, first we will detail the used datasets, and next we will report the actual performance.

C.4.1 Dataset

The simulated dataset includes feature vectors of length $K = 4$ and raw key extractors with $M = 2$ channels. Next, each channel is quantized using a uniform quantizer Q, with $L = 2^b$ levels for $b \in \{1, 2, 3, 4\}$ bit, in the interval $[0, 1]$.

For each parameter choice, we used datasets containing 10^5 observations. Then, 60% of the sample vectors were used for training, 15% for validation, and 25% for testing.
More in detail, the Gaussian dataset $\mathcal{X}_{\mathrm{G}} = (\boldsymbol{x}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{B}}, \boldsymbol{x}_{\mathrm{E}})$. This dataset may model several scenarios, e.g., the received signal amplitude at $K = 4$ sufficiently spaced apart antennas in a MIMO system. More in detail, we consider each agent observation to have been standardized and $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I_4})$ with $i \in (\mathrm{A}, \mathrm{B}, \mathrm{E})$. On the other hand, Alice, Bob, and Eve's observations are statistically correlated. In particular $\mathbb{E}[\boldsymbol{x}_{\mathrm{A},k}\boldsymbol{x}_{\mathrm{B},k}] = \rho_{\mathrm{AB}}$, $\forall k$ while $\mathbb{E}[\boldsymbol{x}_{\mathrm{A},k}\boldsymbol{x}_{\mathrm{E},k}] = \mathbb{E}[\boldsymbol{x}_{\mathrm{B},k}\boldsymbol{x}_{\mathrm{E},k}] = \rho_{\mathrm{AE}}$, $\forall k$.

C.4.2 NN Architectures

The encoder NN consists of five fully connected layers of sizes $4-3-3-2-2$, respectively. The inputs are just clamped to the first layer, which has an identity activation function. The following three layers have an leaky rectified linear unit (LeakyReLU) activation function, whereas the last one has a linear activation. The encoder output layer is composed of two neurons, thus it outputs a vector $\tilde{\boldsymbol{s}}_{\mathrm{A}} \in \mathbb{R}^2$. Then, by using the uniform quantizer $\boldsymbol{Q}$, we will draw from each vector composed by b bits, with $b = 2, 4$, and 6 bit.
The decoder has an architecture that mirrors the encoder's: it has five fully connected layers of sizes $2-2-3-3-4$, respectively, where the first one has an identity activation as it serves only to clamp and take data in input, the next three layers are activated with a LeakyReLU, and the last one has a linear activation.
The discriminator NN is composed of four fully-connected layers of sizes $2-500-500-1$, respectively. The first layer has an identity activation function, the two hidden layers are activated with a LeakyReLU function, while the output layer has a sigmoid activation function.
In the loss function (C.4), after experimental investigation, the factors that weight each loss contribution were chosen as $\alpha = 0.1$ (fixed), while β changed dynamically during training as

$$\beta = \beta_0 \left(1 - \frac{1}{1 + \exp(-3n_{\mathrm{ep}}/N_{\mathrm{ep}})}\right), \tag{C.9}$$

where β_0 is the initialization value, N_{ep} is the total number of training epochs, and n_{ep} is the index of the current training epoch.
in (C.9), with $\beta_0 = 1.05$ and $N_{\mathrm{ep}} = 500$. The learning rate was fixed to $\ell = 0.02$.

C.4.3 Performance Results

Fig. C.3 reports the mutual information achieved by Alice and Bob as a function of ρ_{AB} for several values of quantization bits per channel, b. As discussed in Section C.3, $I(\boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}})$ measures both the Alice and Bob agreement and the randomness of each raw key sequence. As expected for all the considered parameters values, $I(\boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}})$ grows with both ρ_{AB} and b.
Next, Fig. C.4 reports instead the lower bound $C_{\mathrm{sk}}^{\mathrm{low}}$ as a function of the attacker correlation, for $\rho_{\mathrm{AB}} = 0.8$. This turn, as ρ_{AE}, and thus also the information of Eve about the extracted key, grows as the secret key capacity reduces. On the other hand, again $C_{\mathrm{sk}}^{\mathrm{low}}$ also grows with b, but the gains appear to reduce as b increases, hinting at the presence of a saturation value.

C.5 Conclusion

This paper proposes a novel strategy for PKG in 5G and B5G, where Alice and Bob train a pair of NN that acts as a raw key extractor using an input the observations from a near field wireless channel. Next, during inference and the advantage distillation step of the PKG, the agent will use their own key extractor pair with a traditional quantizer to extract the binary raw key, which will then be fed to the information reconciliation

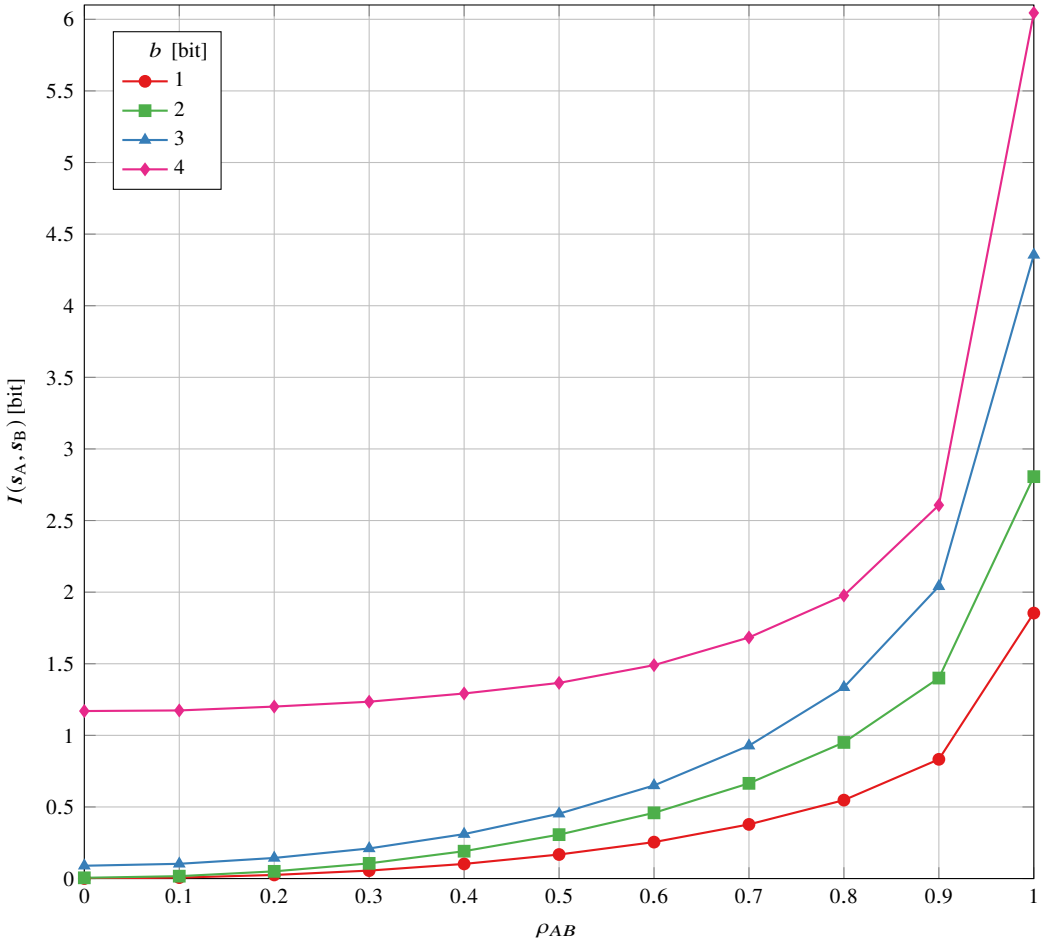


Figure C.3: Mutual information $I(s_A, s_B)$ obtained from the AWGN dataset, considering extractors with $M = 2$ channels, a uniform quantizer with b bits per channel and $\rho_{AB} = 0.8$.

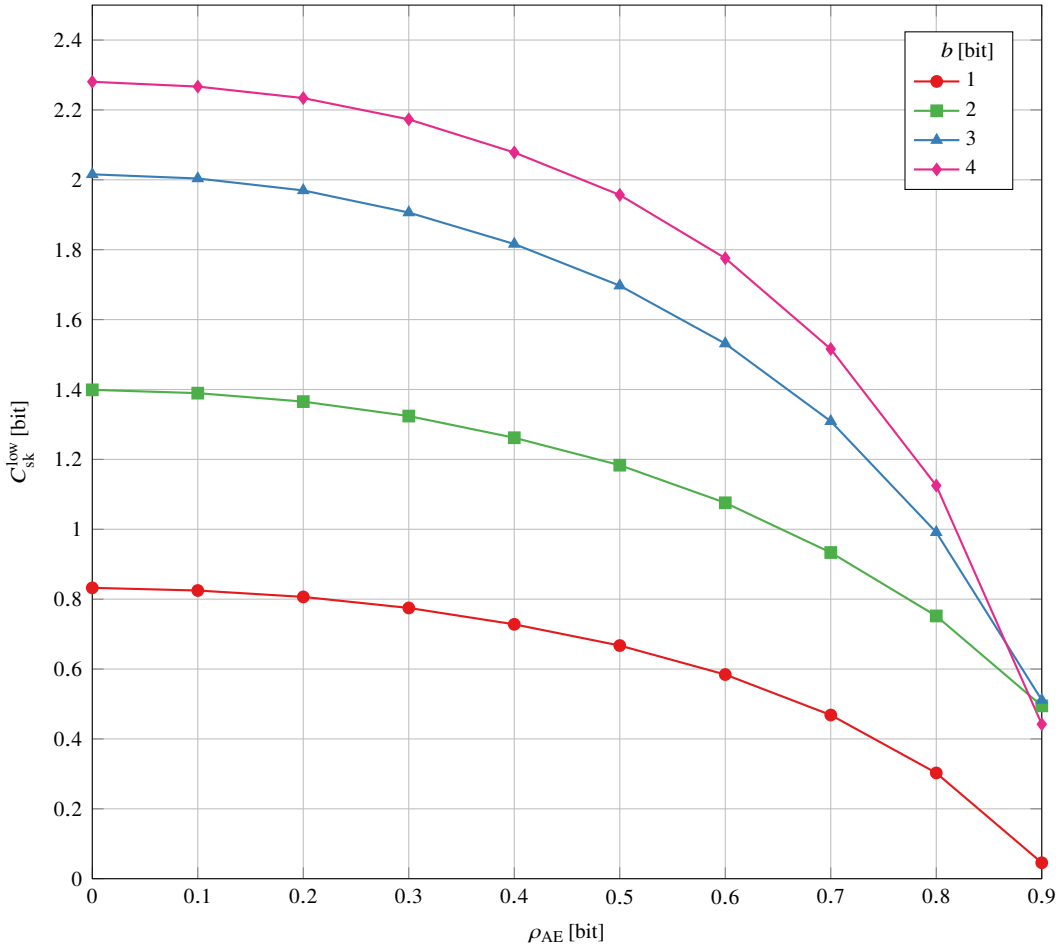


Figure C.4: C_{sk}^{low} obtained from the AWGN dataset, considering extractors with $M = 2$ channels, a uniform quantizer with b bits per channel and $\rho_{AB} = 0.8$.

and privacy amplification blocks. We designed a training procedure where each block and respective loss match the terms of the (lower bound on the) secret key capacity. In particular, the overall loss is composed of three terms: the first measures the reciprocity between Alice and Bob extracted sequence, the second the randomness, and the latter the information leakage to the eavesdropper Eve. Performance results highlight the effectiveness of the proposed approach considering various values of correlation of both legitimate and non-legitimate observation.

Appendix D

Adversarial Attacks on ISAC Systems

To be submitted.
by Mattia Piana and Stefano Tomasin

Abstract:Integrated sensing and communication (ISAC) is expected to be a key enabler for next-generation networks, posing unprecedented security issues. In this paper, we evaluate the security of ISAC systems under adversarial maximum likelihood (ML) attacks. In particular, we have Alice and Bob cooperating to perform bistatic sensing of the environment. As the scatterers are located in different regions, Alice and Bob can obtain a coarse estimation of the scatterers' locations by classifying to which area the received signal belongs. On the other hand, the attacker Trudy aims at disrupting such a procedure by properly designing her transmitting beamformer to fool Bob, and make him estimate a target region. We evaluate the effectiveness of the proposed attack via numerical simulations.

D.1 Introduction

Integrated sensing and communication (ISAC) systems emerge as a cornerstone technology for the sixth-generation era, seamlessly incorporating sensing functionality into wireless networks as a native capability [120]. The object's localization capabilities of such a technology are of crucial interest, as the ability to monitor physical factors is crucial for optimizing the network's performance, enhancing security, driving automation, and carrying out other vital tasks [121]. Still, this technology poses unprecedented security and privacy issues [122]. In the literature, spoofing attacks are of particular interest [123, 124], as the attacker can perform beamforming and disrupt the sensing phase of ISAC systems, and often maximum likelihood (ML) strategies are adopted for tackling such issues [125].
In this paper, we have Alice and Bob cooperating to perform bistatic sensing of the environment. In particular, the scatterers are located in different regions; thus, by determining from which region the received signal was scattered, we can obtain a coarse estimation of the scatterers' location. On the other hand, the attacker Trudy aims at disrupting such a procedure by properly designing her transmitting beamformer to fool Bob, and make him estimate a target region. In particular, the contributions are as follows:

- We model a realistic ISAC channel, using a geometrical channel model that takes into consideration the scatterer's location.
- We train a standard convolutional neural network (CNN) to classify the received signal into the area where the scatterers are located.
- We design a projected gradient descent (PGD) attack that Trudy can perform to induce her desired classification area, taking into account the required transmitting power.

- We numerically evaluate the attack, demonstrating its effectiveness

The paper is organized as follows: in Section D.2 we present the system model, in Section D.3 we perform a security analysis on the proposed attack, in Section D.4 we present the numerical results of the simulations and in Section D.5 we draw the main conclusions.

D.2 System Model

In this paper, the objective of Alice and Bob is to attain a coarse estimation of the scatterers' location using ML, in a bistatic fashion. On the other hand, Trudy aims at attacking this process by impersonating Alice and transmitting malicious signals specifically crafted so that the estimated scatterer location is not the true one, but a target one. In particular, Alice, Bob, and Trudy are multiple input multiple output (MIMO) devices transmitting signals using uniform linear arrays (ULAs). The transmitter is equipped with N_{T} antennas and sends the sensing signal $\boldsymbol{x} = \mathbf{W}\boldsymbol{s} \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$, where $\mathbf{W} \in \mathbb{C}^{N_T\times N_T}$ is the precoding matrix at the transmitter and $\boldsymbol{s} \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$ is the unit power source signal.
Each cluster of scatterers, of dimension L, is located in one of N possible squared areas $a_n, n = 1, \dots, N$, which are centered in position $\boldsymbol{p}_n$ and have side S, as depicted in Fig. D.1. The receiver Bob, equipped with N_{R} antennas, thus receive L signals due to the scatterers, whose angle of arrivals (AoAs) are contained in the vector

$$\boldsymbol{\theta}_n = [\theta_n^{(1)}, \dots, \theta_n^{(L)}] \tag{D.1}$$

while the angle-of-departures (AoDs) from the transmitter to the scatterers are contained in the vector

$$\boldsymbol{\delta}_n = [\delta_n^{(1)}, \dots, \delta_n^{(L)}] \,. \tag{D.2}$$

The received signal by Bob at time m is then

$$\boldsymbol{y}^{(m)} = \boldsymbol{Z}^{(m)}\boldsymbol{x} + \boldsymbol{w}^{(m)} \,, \tag{D.3}$$

where $\boldsymbol{Z}^{(m)} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{T}}}$ is the channel matrix. With $\boldsymbol{\beta}_A(\theta) = \frac{1}{\sqrt{A}}[1, e^{j\pi\sin\theta}, \dots, e^{j\pi(A-1)\sin\theta}] \in \mathbb{C}^{A\times 1}$, $A = \{N_{\mathrm{R}}, N_{\mathrm{T}}\}$ the steering vector operator, $\alpha_l^{(m)}$ as the complex channel gain accounting for pathloss and propagation delay and $\boldsymbol{w}^{(m)}$ is the Gaussian noise, we write the channel $\boldsymbol{Z}^{(m)}$ as

$$\boldsymbol{Z}^{(m)} = \sum_{l=1}^{L} \alpha_l^{(m)} \boldsymbol{\beta}_{\mathrm{R}}(\theta_n^{(m,l)}) \boldsymbol{\beta}_{\mathrm{T}}^T(\delta_n^{(m,l)}) \,. \tag{D.4}$$

Note that we can decompose the channel (D.4) as

$$\boldsymbol{Z}^{(m)} = \boldsymbol{H}^{(m)}\boldsymbol{G}^{(m)} \,, \tag{D.5}$$

where $\boldsymbol{H}^{(m)} \in \mathbb{C}^{N_R\times L}$ is the channel between the scatterers and Bob at time m, while $\boldsymbol{G}^{(m)} \in \mathbb{C}^{L\times N_T}$ is the channel between the transmitter and the scatterers. We also define the signal-to-noise ratio (SNR) as

$$\mathrm{SNR} = \frac{\mathrm{Tr}(\boldsymbol{Z}^{(m)H}\boldsymbol{Z}^{(m)})}{\mathbb{E}(\boldsymbol{w}^{(m)H}\boldsymbol{w}^{(m)})} = \frac{\mathrm{Tr}(\boldsymbol{Z}^{(m)H}\boldsymbol{Z}^{(m)})}{N_{\mathrm{R}}\sigma_w^2} \,. \tag{D.6}$$

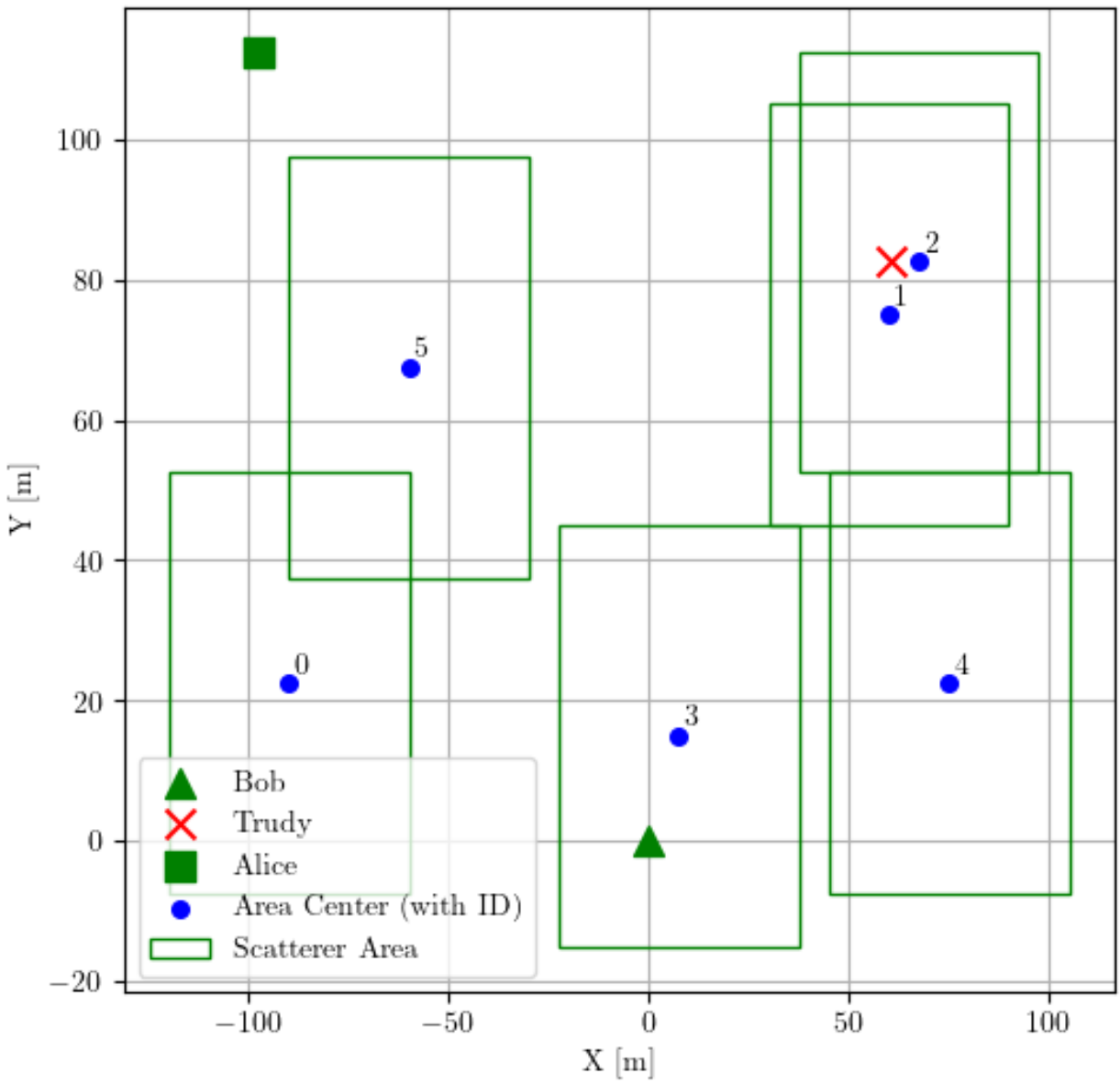


Figure D.1: System Model

D.2.1 Problem and Dataset Description

To obtain a coarse estimation of the scatterers location, Bob has available a labeled dataset $\mathcal{D} = \{(\boldsymbol{y}^{(m)}, a^{(m)})\}$ of M_S samples, where we recall the index m refers to different times instants, so that it has different scatterer positions. Bob then trains a ML model $f_{\boldsymbol{w}}$ to classify the received signal $\boldsymbol{y}^{(m)}$ to the corresponding class $a^{(m)}$. After that, Bob obtains the scatterer positions by mapping the estimated class $\hat{a}^{(m)} = a_n$ to the corresponding area centroid $\boldsymbol{p}_n$.

D.2.2 CNN Architecture

We employ a lightweight CNN, with three convolutional layers and a single final layer for classification.

D.3 Security Analysis

In this section, we present the attack analysis that the attacker can perform.

D.3.1 Attacker Model

We assume the attacker Trudy has access to the trained model $f_{\boldsymbol{w}}$ as well as the channel $\boldsymbol{Z}^{(m)}$ from her position to Bob at each time m.

D.3.2 Attack Strategy

To break the localization procedure, Trudy exploits her knowledge of the model $f_{\boldsymbol{w}}$ and the channel between her location and Bob $\boldsymbol{Z}^{(m)}$ to perform a targeted PGD attack [126], and craft a channel $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{W}$ via the beamforming matrix $\boldsymbol{W}$ to reliably induce a target class $\tilde{a}_n$ when the true class is a_n. Still, due to hardware limitations, we assume a power constraint of $P_{\max}$ on the beamforming matrix, i.e., $\|\boldsymbol{W}\|_F \leq P_{\max}$, where $\|\cdot\|_F$ is the Frobenius norm. In detail, the attack procedure works as follows:

1. Select a source and target class, namely a_n and $\tilde{a}_n$ respectively
2. Select a maximum perturbation ϵ on her channel $\boldsymbol{Z}^{(m)}$, such that
$$\|\tilde{\boldsymbol{Z}} - \boldsymbol{Z}^{(m)}\| \leq \epsilon \tag{D.7}$$
3. Perform a PGD and find the target channel $\tilde{\boldsymbol{Z}}$
4. Find the optimal beamformer $\boldsymbol{W}^\star$ by projecting the target channel $\tilde{\boldsymbol{Z}}$ onto the feasible solutions, taking into consideration the power constraint $P_{\max}$. In detail, the optimization problem for Trudy is:
$$\begin{aligned} \boldsymbol{W}^\star = \arg\min_{\boldsymbol{W}} \quad & \|\tilde{\boldsymbol{Z}} - \boldsymbol{Z}^{(m)}\boldsymbol{W}\| \\ \text{s.t.} \quad & \|\boldsymbol{W}\|_F \leq P_{\max}. \end{aligned} \tag{D.8}$$

D.4 Numerical Results

We simulated the scenario in Fig. D.1, where there are $N = 5$ squared areas with size $S = 30\,\text{m}$. We used a number of antennas of $N_T = N_R = 4$, a maximum transmitter power by Trudy of $P_{\max} = 40\,\text{dB}$. The target area for Trudy is the region $\tilde{a}_n = 2$, when the true region is $a_n = 1$.

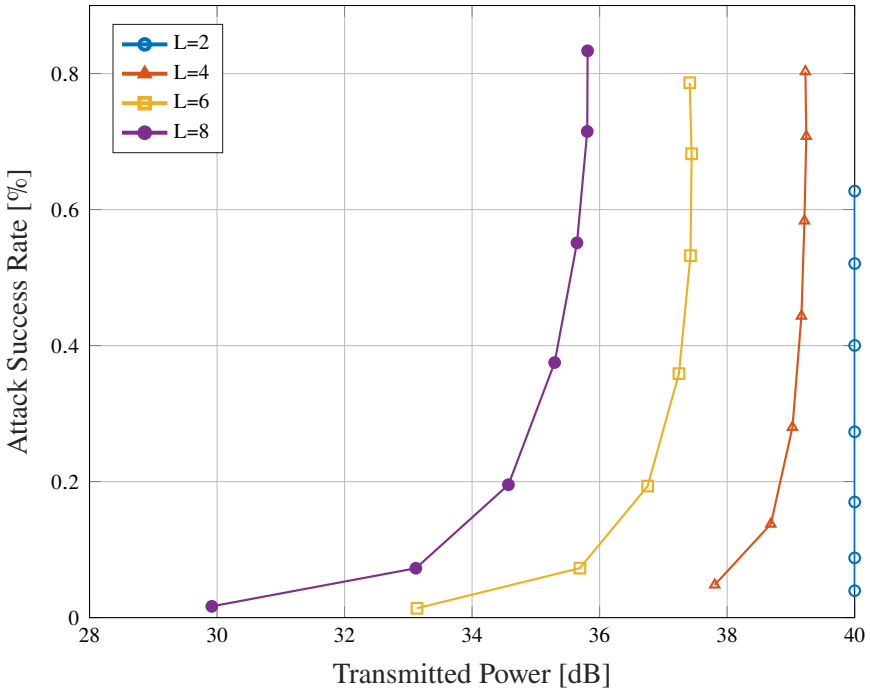


Figure D.2: Attack Success Rate as a function of the transmitted power, for different numbers of scatterers L.

D.4.1 Attack Success Rate VS Number of Scatterers

We see from Fig. D.2 that by increasing the transmitted power P_{T}, Trudy can perform more effective attacks, reaching an attack success rate, i.e., an accuracy on the target class, of $\geq 80\%$ when the number of scatterers $L \geq 4$. We also notice that the performance is extremely dependent on the number of scatterers L: in fact, the greater L, the easier it is for Trudy to find a beamforming matrix $\boldsymbol{W}$ to fool Bob. This effect relies on the rank of the cascaded channel $\boldsymbol{Z}^{(m)}$: in fact from (D.4), if $L \geq N_R, N_T$ then $\boldsymbol{Z}^{(m)}$ becomes invertible, thus it easier for the attacker to solve (D.8). This effect is particularly evident with $L = 2$: in that case $\boldsymbol{Z}^{(m)}$ has at most two non-zero eigenvalues, thus the optimal beamformer saturates at $P_{\max} = 40\,\mathrm{dB}$. Note also that in that case, multiple solutions are available to the attacker: for each maximum perturbation ϵ in (D.7), the attacker can find the beamformer that respects the power constraint. Another solution would be to modify the PGD algorithm by directly taking into account the power constraint into the solution $\tilde{\boldsymbol{Z}}$, and this is left for future works.

D.4.2 Attack Success Rate VS SNR

In Fig. D.3, we observe that as the SNR increases, the required power for Trudy decreases, yet the results remain very similar. This effect can be justified by the fact that when the channels in input to the PGD are less noisy, it is easier for the algorithm to find suitable channels to fool Bob's model.

D.5 Conclusion

In this paper, we applied tools of adversarial ML to disrupt the localization procedure common in ML based ISAC systems. In particular, the attacker Trudy, by performing a PGD attack on the model trained for

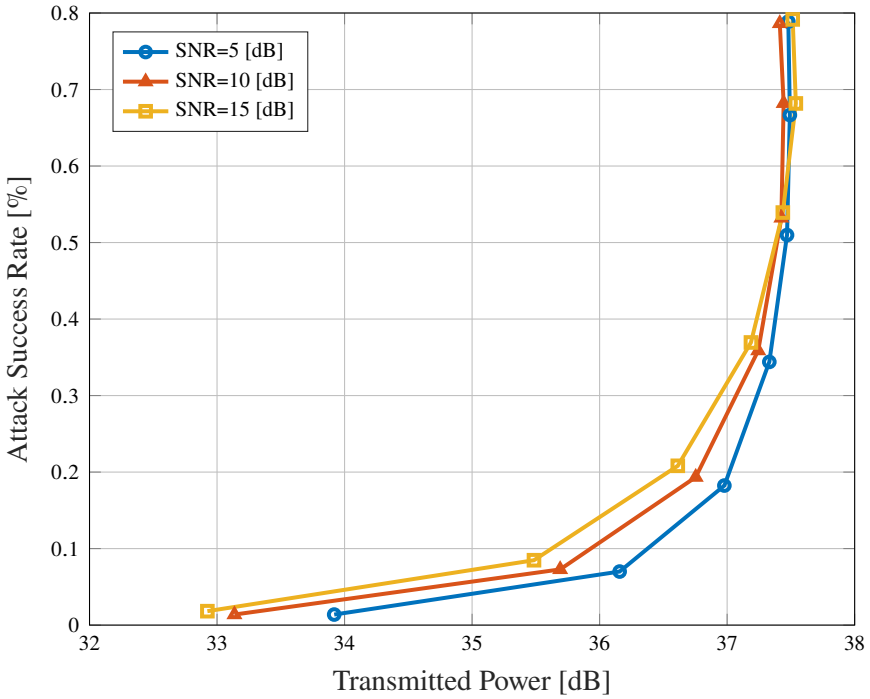


Figure D.3: Attack Success Rate as a function of the transmitted power, for different numbers SNR levels.

scatterers' localization, can effectively design her beamformer to break the localization procedure, achieving a success rate of $\geq 80\%$ with ~ 36 dB of transmitted power.

Appendix E

Bounds on the information leakage of short packet wiretap codes

(V. Bioglio, L. Luzzi, paper in preparation, to be submitted to ISIT 2026*.)*

E.1 Background and motivation

Wiretap coding techniques allow to transmit confidential information without the use of secret keys in the presence of passive adversaries at rates up to the secrecy capacity, as long as an asymmetry in the channel quality between the legitimate receiver and the adversary can be guaranteed.
In the asymptotic setting where the blocklength tends to infinity, secrecy capacity achieving coding schemes have been developed, notably by employing polar codes [127], which are an attractive solution since they are already part of the 5G New Radio standard.
However, for practical applications requiring short packets or low latency, it is important to obtain non-asymptotic bounds for the secrecy rate of wiretap codes in finite blocklength.

E.2 Proposed methodology

Building on the theoretical breakthrough by Polyanskiy, Poor and Verdù in the analysis of finite-length channel coding rates [128], Yang, Schaefer and Poor [58] proved tight bounds on the optimal second-order coding rate over discrete memoryless channels (DMCs) and Gaussian wiretap channels. In [58], the information leakage is measured in terms of the total variation distance (TVD) between the joint distribution of the secret message M and the eavesdropper's observation Z^n, and an ideal distribution in which M is uniformly distributed and independent of Z^n:

$$S(\mathsf{M}|\mathsf{Z}^n) = \mathbb{V}(p_{\mathsf{M}\mathsf{Z}^n}, p_{\mathsf{M}}p_{\mathsf{Z}^n}). \tag{E.1}$$

We consider the physically degraded wiretap channel depicted in Figure E.1. We assume that Bob's channel $W_b : \mathcal{X} \to \mathcal{Y}$ of transition probabilities $p_{\mathsf{Y}|\mathsf{X}}$ and Eve's channel $W_e : \mathcal{X} \to \mathcal{Z}$ of transition probabilities $p_{\mathsf{Z}|\mathsf{X}}$ are both symmetric. Furthermore, we assume that both channels are binary-input, i.e. $\mathcal{X} = \{0, 1\}$.
For this model, the secrecy capacity (in bits per channel use) is given by the differences of the capacities of Bob and Eve's channels: $C_s = C_b - C_e$. In finite blocklength, [58, Theorem 13] showed that for $\epsilon + \delta < 1$, the maximal secrecy rate $R^*(n, \epsilon, \delta)$ for blocklength n, and average error probability ϵ under the secrecy constraint

$$S(\mathsf{M}|\mathsf{Z}^n) \leq \delta \tag{E.2}$$

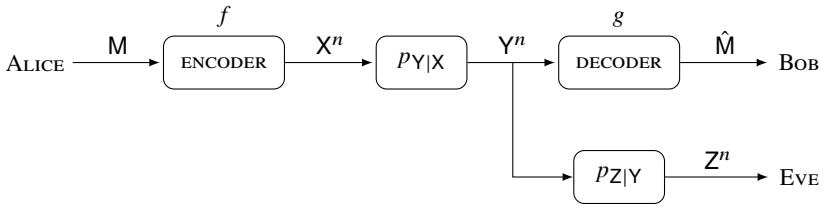


Figure E.1: The degraded wiretap channel.

is upper bounded by

$$R^*(n,\epsilon,\delta) \leq C_s - \sqrt{\frac{V_c}{n}} Q^{-1}(\epsilon+\delta) + O\left(\frac{\log n}{n}\right), \tag{E.3}$$

where C_s is the secrecy capacity, Q denotes the Q-function, and

$$V_c = \sum_{x\in\mathcal{X}} p^*_{\mathsf{X}}(x) \left(\sum_{y,z} p_{\mathsf{ZY}|\mathsf{X}}(z,y|x) \left(\log \frac{p_{\mathsf{ZY}|\mathsf{X}}(z,y|x)}{p_{\mathsf{Z}|\mathsf{X}}(z|x) p_{\mathsf{Y}|\mathsf{Z}}(y|z)} \right)^2 - \mathbb{D}\left(p_{\mathsf{ZY}|\mathsf{X}=x} || p_{\mathsf{Y}|\mathsf{Z}} p_{\mathsf{Z}|\mathsf{X}=x}\right)^2 \right). \tag{E.4}$$

In the expression above, p^*_{X} is the uniform distribution, which is the unique distribution that maximizes $\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z})$ in the degraded symmetric case.
In our previous work [57], we investigated the secrecy performance of wiretap code constructions based on polar codes in the special case where Bob's channel is error-free and Eve's channel is a binary erasure channel (BEC). We now extend our analysis to more general degraded wiretap channels.

Polar coding scheme We consider the general wiretap polar coding scheme in [127] for blocklength $n = 2^m$. The input set $[\![1;n]\!]$ is partitioned into three disjoint subsets $\mathcal{A} \cup \mathcal{R} \cup \mathcal{B}$, where $|\mathcal{A}| = k$ and $|\mathcal{R}| = r$. Intuitively, $\mathcal{B}$ corresponds to the indices of bit-channels that are bad for Bob and for Eve; $\mathcal{A}$ to bit-channels that are good for Bob but bad for Eve; and $\mathcal{R}$ to bit-channels that are good for Bob and for Eve.
Given the confidential message $\mathsf{M}^k \in \{0,1\}^k$, the input U^n of the polar encoder is defined by setting $\mathsf{U}_{\mathcal{B}} = \mathbf{0}^{n-k-r}$ (frozen bits), $\mathsf{U}_{\mathcal{A}} = (\mathsf{U}_{i_1}, \ldots, \mathsf{U}_{i_k}) = \mathsf{M}^k$ and $\mathsf{U}_{\mathcal{R}} = \mathsf{V}^r$ a vector of uniformly random bits. We denote the corresponding polar codeword by $\mathsf{X}^n = G_n \mathsf{U}^n$, where G_n is the polarization transform [129].

Bound for the average error probability Recall that Bob's block error probability under SC decoding is upper-bounded by the sum of Bhattacharyya parameters of Bob's bit-channels in $\mathcal{G} = \mathcal{A} \cup \mathcal{R}$ (i.e. the bit-channels that are not frozen) [130]:

$$P_e \leq \sum_{i\in\mathcal{A}\cup\mathcal{R}} Z(W_b^{(i)}) \tag{E.5}$$

When Bob's channel W_b is a BEC, the parameters $Z(W_b^{(i)})$ can be computed recursively [129]. For a general channel W_b, in order to estimate the upper bound (E.5), we will use Tal and Vardy's algorithm [59, Algorithm D].

Bounds for the leakage Given a symmetric channel $W : \mathcal{X} \to \mathcal{Y}$ with transition probability $p_{\mathsf{Y}|\mathsf{X}}$, let p_{X} denote the uniform input distribution over $\mathcal{X}$ and $p_{\mathsf{Y}} = p_{\mathsf{Y}|\mathsf{X}} \circ p_{\mathsf{X}}$ the corresponding output distribution. We define the TVD of the channel W as

$$T(W) = \mathbb{V}(p_{\mathsf{XY}}, p_{\mathsf{X}} p_{\mathsf{Y}}).$$

One can show that the following bounds hold for the leakage in total variation distance:

$$S(\mathsf{M}^k|\mathsf{Z}^n) \overset{(1)}{\leq} \frac{1}{2} \sum_{i \in \mathcal{A} \cup \mathcal{B}} T(p_{\bar{\mathsf{Z}}^n \bar{\mathsf{U}}_{[\![1;i-1]\!] \cap (\mathcal{A} \cup \mathcal{B})} | \bar{\mathsf{U}}^i}) \overset{(2)}{\leq} \frac{1}{2} \sum_{i \in \mathcal{A} \cup \mathcal{B}} T(W_e^{(i)}). \tag{E.6}$$

Computation of Bound 2 Bound 2 in equation (E.6) shows that the average TVD (E.1) of the wiretap code is upper bounded by the sum of the TVDs of the eavesdropper's bit-channels $W_e^{(i)} : \{0,1\} \to \mathcal{Z}^n \times \{0,1\}^{i-1}$ corresponding to the positions of the bits $i \in \mathcal{A} \cap \mathcal{B}$.
When Eve's channel W_e is a BEC, these TVDs can be computed recursively, since $T(W_e^{(i)}) = 1 - Z(W_e^{(i)})$. For general channels, there is no closed form expression for the TVDs of the bit-channels $W_e^{(i)}$, and their exact recursive computation is unfeasible since the cardinality of the output alphabet grows exponentially with i. For channel coding applications, Tal and Vardy [59] proposed a low-complexity algorithm to approximate the bit-channels with an upgraded or degraded version of themselves, with output alphabet of cardinality smaller than a chosen threshold 2μ, by performing suitable merge operations on the output symbols. We use this algorithm in order to evaluate (E.6) for more general channels such as the binary symmetric channel (BSC) and the binary input additive white Gaussian noise (BI-AWGN) channel.
In particular, we focus on the *upgrading merge* in order to obtain an upper bound for the TVDs. We apply [59, Algorithm B] with parameter 2μ to the eavesdropper's channel W_e. Let $\tilde{W}_e^{(i)}$ be the output of the upgrading algorithm corresponding to the bit-channel $W_e^{(i)}$, for $i = 1, \ldots, n$. Since $W_e^{(i)}$ is degraded with respect to $\tilde{W}_e^{(i)}$, we have $T(W_e^{(i)}) \leq T(\tilde{W}_e^{(i)})$.

Computation of Bound 1 We are able to numerically evaluate Bound (1) in equation (E.6) by Monte-Carlo simulation only in the case where Eve's channel W_e is a BEC, similarly to [57].

Wiretap code design We propose a simple algorithm to choose the sets $\mathcal{A}, \mathcal{R}, \mathcal{B}$ in the polar coding scheme so that $P_e \leq \epsilon$ and $S(\mathsf{M}^k|\mathsf{Z}^n) \leq \delta$. First, the Bhattacharyya parameters for Bob's bit-channels $W_b^{(i)}$ are estimated and sorted in increasing order; the set $\mathcal{G} = \mathcal{A} \cup \mathcal{R}$ is chosen as the largest possible set information set of "good bit-channels" (in terms of Bhattacharyya parameters) such that the bound (E.5) on the error probability is smaller than ϵ. Alternatively, a Monte-Carlo bound on the error probability can be used if $\epsilon > 10^{-7}$. Subsequently, the TVDs for Eve's bit-channels $W_e^{(i)}$ are estimated and stored in increasing order; the set $\mathcal{A}$ is chosen as the largest possible subset of $\mathcal{G}$ such that the sum of the TVDs of the bit-channels $W_e^{(i)}$ for $i \in \mathcal{A} \cup \mathcal{B}$ is smaller than 2δ (Bound 2 in (E.6)). Alternatively, if Eve's channel is a BEC, Bound 1 may be computed by Monte Carlo simulation, as explained previously.

E.3 Experimental results and analysis

Binary Erasure Wiretap Channel Figure E.2 shows the lower bounds on the achievable secrecy rate (Bounds 1 and 2) for polar codes in the case where W_b and W_e are BECs with erasure probabilities when $p_b = 0.05$ and $p_e = 0.4$, under the average error probability constraint $\epsilon = 0.01$ and secrecy constraint $\delta = 0.1$.

Binary Symmetric Wiretap Channel Figure E.3 shows the lower bound on the achievable secrecy rate (Bound 2) when both W_b and W_e are binary symmetric channels with transition probabilities $p_b = 0.05$, $p_e = 0.3$ respectively, $\epsilon = 0.01$ and $\delta = 0.1$. The bound is obtained through Tal and Vardy's upgrading merge approximation of bit-channels with parameter $\mu = 64$.

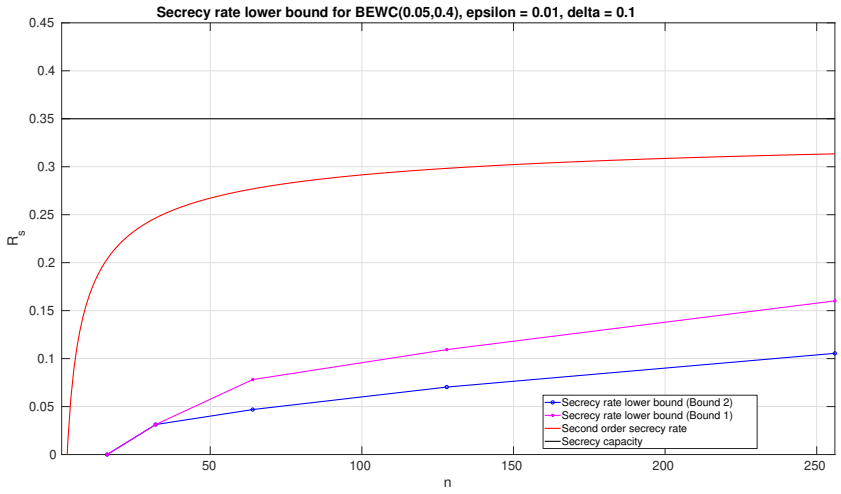


Figure E.2: Comparison of the lower bounds on the achievable secrecy rate for polar codes with the second-order approximation secrecy rate in (E.3) over a degraded binary erasure wiretap channel with parameters $p_b = 0.05$ and $p_e = 0.4$, under the average error probability constraint $\epsilon = 0.01$ and secrecy constraint (E.2) with $\delta = 0.1$.

Binary-input Gaussian Wiretap Channel Finally, we consider the case where Alice uses Binary Phase Shift Keying (BPSK) modulation, and Bob and Eve observe the output of an AWGN channel with noise variance σ_b^2 and σ_e^2 respectively. We note that the second-order bound (E.3) for the secrecy rate still holds for this channel, with

$$V_c = \sum_{x\in\mathcal{X}} \frac{1}{2}\left(\iint p_{\mathsf{ZY}|\mathsf{X}}(z,y|x)\left(\log \frac{p_{\mathsf{ZY}|\mathsf{X}}(z,y|x)}{p_{\mathsf{Z}|\mathsf{X}}(z|x)p_{\mathsf{Y}|\mathsf{Z}}(y|z)}\right)^2 dydz - \mathbb{D}\left(p_{\mathsf{ZY}|\mathsf{X}=x}||p_{\mathsf{Y}|\mathsf{Z}}p_{\mathsf{Z}|\mathsf{X}=x}\right)^2\right). \tag{E.7}$$

In fact, although [58, Theorem 13] is stated for DMCs, it also holds for channels with finite input and continuous output, since only the finiteness of the input is required in the proof.
Figure E.4 shows the lower bound on the achievable secrecy rate (Bound 2) for $\sigma_b^2 = 0.2$, $\sigma_e^2 = 2$, $\epsilon = 0.01$ and $\delta = 0.1$, obtained through Tal and Vardy's approximation with parameter $\mu = 64$.

E.4 Conclusions and limitations

Using the channel approximation algorithm in [59], we are now able to evaluate Bound 2 in (E.6) for general channels, while evaluating Bound 1 remains an open problem.
Our numerical results confirm the fact that although they asymptotically achieve the secrecy capacity, wiretap schemes based on polar codes are suboptimal in terms of secrecy rate in finite blocklength. As already noted in [57], this is due to their suboptimal finite length scaling. The back-off from the secrecy capacity must be taken into account for practical implementation.
Designing optimal wiretap schemes with low encoding and decoding complexity remains a challenging open problem.

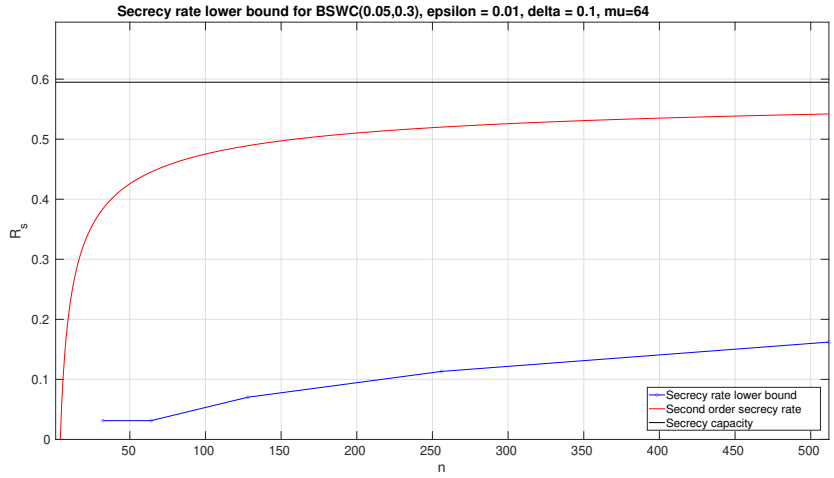


Figure E.3: Comparison of the lower bound on the achievable secrecy rate for polar codes over a semideterministic wiretap channel with the second-order approximation secrecy rate in (E.3) over a degraded wiretap channel when the main channel and eavesdropper's channel are BSCs with parameters $p_b = 0.05$, $p_e = 0.3$ under the secrecy constraint (E.2) with $\delta = 0.1$.

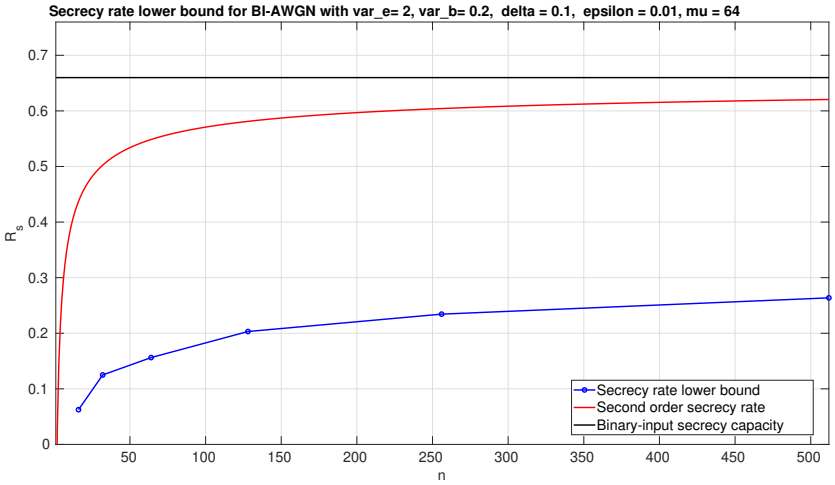


Figure E.4: Comparison of the lower bound on the achievable secrecy rate for polar codes with the second-order approximation secrecy rate in (E.3) over a degraded wiretap channel when the main channel and eavesdropper's channel are BI-AWGN with variance $\sigma_b^2 = 0.2$ and $\sigma_e^2 = 2$ respectively, under the average error probability constraint $\epsilon = 0.01$ and secrecy constraint (E.2) with $\delta = 0.1$.

Appendix F

Position-Based Cross-Layer Authentication For Industrial Communications

Submitted to IEEE International Conference on Communications (ICC 2026).
by Mattia Piana, Ali Hossary, and Stefano Tomasin

Abstract: We consider a robot (Alice) moving in an industrial environment while transmitting messages to nearby endpoints through fixed access-points (APs). An intruder robot (Trudy) aims at transmitting malicious messages to the endpoints, impersonating Alice. We aim at detecting Trudy transmissions by comparing the expected position of the transmitter with two estimates of it obtained from a) the channel state information (CSI) estimated on the signals received by the APs, and b) the traffic information in the network. Such estimates are obtained with CNN and support vector regressor (SVR) models along with Kalman filters to exploit the trajectory evolution. Numerical results obtained using the DICHASSUS dataset confirm the effectiveness of our proposed solution.

F.1 Introduction

With the dawn of Industry 4.0, artificial intelligence (AI), the Internet of Things (IoT), and robotics are gaining much interest to improve efficiency, productivity, and quality [131]. In such a context, new information and communication (ICT) systems are used to support entire supply chains [132], increasing the attack surface to malicious devices aiming to disrupt the industrial infrastructure, [133]. Authenticating transmitters in such networks is a crucial task to ensure the integrity of the transmissions. To this end, different strategies can be adopted, from conventional cryptographic schemes to novel quantum cryptography [134], to lightweight physical-layer security mechanisms, [135]. Focusing on the latter, the literature offers different strategies to authenticate transmitters directly at the physical layer (see [30] for an exhaustive survey). Among them, advancements in ML CSI-based localization techniques can improve authentication performance, [102, 103]. Concerning cross-layer solutions, different strategies can be adopted. The first is to design hybrid protocols that combine physical-layer-based with conventional key-based authentication schemes. In [136], the authors use physical-layer authentication (PLA) as a prehemptive security mechanism to alleviate the authentication burden at the core network. A similar philosophy is implemented in vehicular networks, where, after a prior upper-layer authentication, vehicles are re-authenticated based on position-dependent keys extracted at the physical layer [137]. The second provides detectors that combine information from both layers. For instance, [138] combines SNR and packet-error-rate to build an generalized likelihood ratio test (GLRT)-based authenticator, assuming a Gaussian error on both features. Similarly, [139] exploits the correlation of routing protocols, physical, and link layer data in a multihop wireless mesh network environment and

compares the classification performance obtained with support vector machine (SVM), decision trees, and Bayesian networks. For a survey of cross-layer authentication approaches, see also [140].
In this paper, we design an authentication protocol that fuses information coming from the physical and upper layers in an industrial network context. Alice is a robot, moving on a factory floor and communicating to neighboring endpoints via several APs, under the supervision of a maximum a posteriori probability (MAP), that monitors the physical-layer along with the traffic information. An intruder robot (Trudy) aims at transmitting malicious messages to the endpoints, impersonating Alice. We aim at detecting Trudy transmissions by comparing the expected position of the transmitter with two estimates of it obtained by a) the CSI estimated on the signals received by the APs and b) the traffic information in the network. In particular, a CNN is trained to estimate the transmitter position from the CSI, while a SVR uses the traffic information in the network again to infer the robot position. The two predicted positions are then fed into Kalman filters to refine them with prior estimates of the trajectory, and the refined position estimation is lastly compared with the expected legitimate one. If the three positions are close enough, the message is considered authentic; otherwise is rejected as fake.
In comparison with existing literature, the proposed solution has several novel features. Notably, we fuse information from various layers to create position information, which is then compared using a statistical test. This provides a better understanding of the detector's behavior. Furthermore, we exploit the temporal correlation of the information using the well-established Kalman filter. Finally, continuous learning techniques are adopted to adapt the model to changes in the environment.
The contributions of this work are as follows:

1. The fusion of information coming from different layers passes through a common estimate of the device position rather than as a mixed input to an ML model.
2. The refinement of the estimated positions by Kalman filters to take into account the temporal evolution.
3. A cross-layer-detector (CLD), a lightweight ML framework constituted either by CNN or a combination of CNN and SVR, that estimate the transmitter position using CSI and connectivity data, respectively.
4. A continuous learning strategy based on fine-tuning is developed to take into account the scenario changes.
5. The performance assessment of the proposed solution on both synthetic and real-world data.

The rest of the paper is organized as follows. In Section F.2, we introduce the system model, in Section F.3, we introduce the proposed ML framework and the attacker model, in Section F.5, we show the performance of the model, and in Section F.6, we draw the main conclusions.

F.2 System Model

We consider a system where two single-antenna mobile robots, Alice and Trudy, move on a factory floor over a predefined path at variable speeds. We assume that the trajectory of T positions $\boldsymbol{p}_t$ Alice visits at the discrete times $t = 1, \ldots, T$, is known by the network and determined before deployment, depending on the specific task to be carried out. This is a typical scenario in smart factories, where robots travel through the factory and cooperate with static tools to carry out specific tasks [141].
In her movement, Alice communicates with other endpoints via multiple APs. While the APs are assumed to be static, the end-points can move. Let $\mathcal{K} = \{k_1, \ldots, k_N\}$ be the set of N endpoints Alice can communicate with and $\mathcal{A} = \{a_1, \ldots, a_L\}$ the set of L APs. Each AP a is equipped with a variable number of antennas M_a. Let us denote with $\boldsymbol{z}(k)$ the position of the endpoint $k \in \mathcal{K}$.
Fig. F.1 shows an example of two trajectories and positions of APs and endpoints.

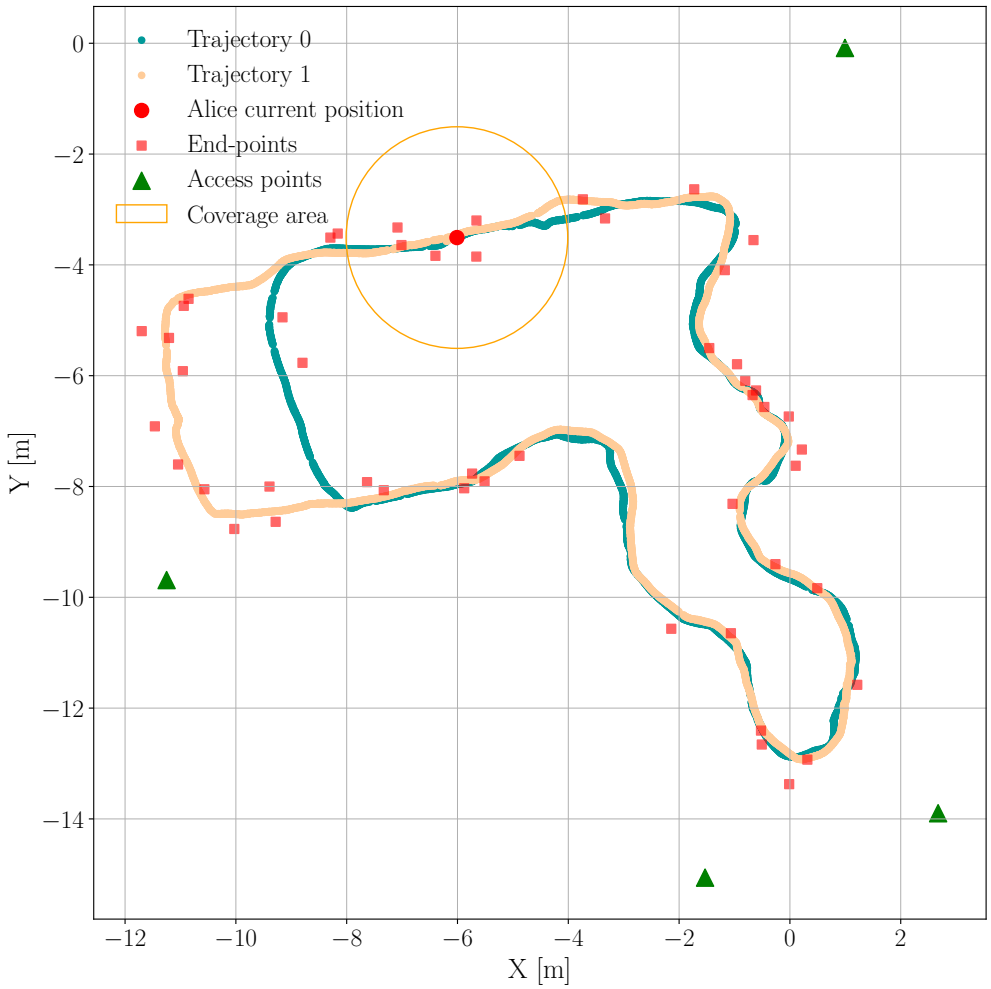


Figure F.1: Alice trajectories and APs positions from [142], with $N = 50$ generated endpoints within $D_{\mathrm{R}} = 0.5$ m from Alice path.

F.2.1 Traffic Model for Static End-points

At time t, the possible final destinations of Alice's messages are the subset of *reachable endpoints* $\mathcal{S}(\boldsymbol{p}_t)$ taken from $\mathcal{R}(\boldsymbol{p}_t, R)$, the set of endpoints located in a circle centered around Alice's current position $\boldsymbol{p}_t$, with radius R. In formulas, the set of reachable endpoints at time t is then

$$\mathcal{S}(\boldsymbol{p}_t) = \{k : k \in \mathcal{K}, z(k) \in \mathcal{R}(\boldsymbol{p}_t, R)\}\,. \tag{F.1}$$

Alice sends packets to the set of *active* endpoints, which is a subset of $\mu \in \{0, \ldots, |S(\boldsymbol{p}_t)|\}$ endpoints taken from $S(\boldsymbol{p}_t)$. Here, we assume that such a subset is obtained by taking uniformly at random the points from the set of reachable endpoints. Referring to Fig. F.1, the reachable endpoints are the 5 ones within the circle, but only μ are active. For example, if $\mu = 2$ then two endpoints picked at random are active, if $\mu = 5$ then all endpoints within the circle are active.

F.2.2 Network Monitoring

We assume the existence of an MAP that monitors the traffic as well as the physical layer information in the network.

Traffic Monitoring The APs collect informtion about the traffic and sent it to the MAP. In particular, such information at time t is represented by the vector $\boldsymbol{v}_t \in [0,1]^N$ with binary entries, where each v_n, $n = 1, \ldots, N$, is 1 if the endpoint n is active, and it is 0 otherwise.

Channel Monitoring We assume that transmissions use orthogonal frequency-division multiplexing (OFDM) signals, constituted by N_S subcarriers over a band B. Upon the transmission of pilots by a mobile robot, AP a estimates at time t the matrix of $M_a \times N_\mathrm{S}$ complex baseband equivalent channel gains for all subcarriers and antennas, obtaining

$$\boldsymbol{Y}_t^{(a)} = \boldsymbol{H}_t^{(a)} + \boldsymbol{W}_t^{(a)}\,, \tag{F.2}$$

where $\boldsymbol{H}_t^{(a)} \in \mathbb{C}^{M_a \times N_\mathrm{S}}$ is the channel matrix and $\boldsymbol{W}_t^{(a)} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ is the thermal noise matrix whose enties are i.i.d Gaussians with zero mean and variance σ^2. The APs then applies a subcarrier averaging on C consecutive subcarriers to reduce the dimensionality of the estimated channels without sacrificing useful information. The processed CSI matrix is denoted as $\hat{\boldsymbol{H}}_t^{(a)}$, has dimension $M_a \times \frac{N_\mathrm{S}}{C}$, and represents compactly the spatial and frequency information.
Along with the channel, each AP a keeps track of the SNR at each antenna, which is denoted as $\gamma_t^{(a,i)}$, $i = 1, \ldots, M_a$. The AP then computes the SNR coefficient $\gamma_t^{(a)} = \sum_{i=1}^{M_a} \gamma_t^{(a,i)}$. At each time t, the processed CSI matrix $\hat{\boldsymbol{H}}_t^{(a)}$ and the SNR coefficient $\gamma_t^{(a)}$ are transmitted to the MAP.

F.2.3 Attacker Model

Another robot, Trudy, acts as an intruder into the considered system. Her purpose is to transmit messages to the endpoints that are confused as coming from Alice. We assume that Trudy can follow the same Alice trajectory, and she knows Alice's current position.

F.3 Cross-layer Anomaly Detection for Static End-points

Our target is to detect Trudy transmissions by identifying anomalies in the network. To this end, we propose an anomaly detection system that exploits the knowledge of the network conditions both at the physical and

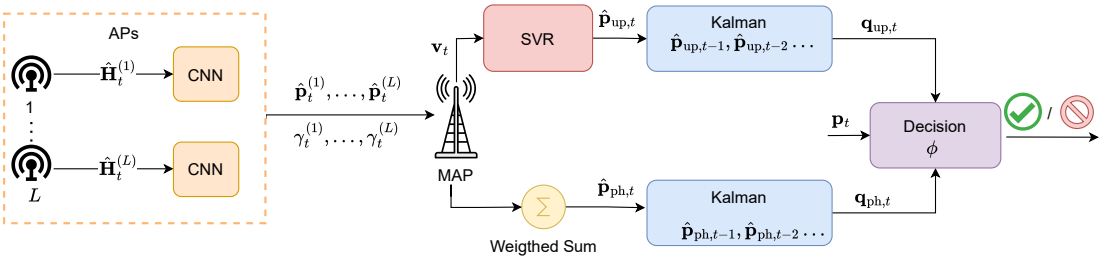


Figure F.2: Cross-layer authentication solution.

upper layers. We frame this task as a one-class classification problem, in which, given features belonging to class $\mathcal{H}$, we have to decide whether they are legitimate, i.e., $\mathcal{H} = \mathcal{H}_0$: Alice is transmitting, or not, i.e., $\mathcal{H} = \mathcal{H}_1$: the transmitter is Trudy. Note that one-class classification problems are particularly challenging as they do not use attacker samples in the training, thus they are robust to multiple attacks.
First, we describe such features that are used to detect anomalies, and then we detail the proposed solution. Lastly, we discuss possible attacks by Trudy against the designed solution.

Detection Features The features that our framework uses to detect anomalies are time series data, where at each discrete time t, the available features are: the traffic vector $\boldsymbol{v}_t$, the SNR coefficient of each AP $\gamma_t^{(a)}$, and the CSI information $\hat{\boldsymbol{H}}_t^{(a)}$ from each AP. This information is then processed either directly at the APs or at the MAP, as explained in the following. Since the active endpoints belong to the reachable set $\mathcal{S}(\boldsymbol{p}_t)$ defined by (F.1), the traffic patterns $\boldsymbol{v}_t$ implicitly encode spatial information. On the other hand, the CSI is affected by local reflection and scattering phenomena, which also carry information on the location of the transmitter.

Detection Mechanism Our detection mechanism is composed of different steps shown in Fig. F.2, and here outlined:

1. *Channel Estimation and Processing*: in this phase, each AP estimates the channel and processes it as explained in Section F.2.2.

2. *Coarse Position Estimation*: each AP a obtains estimate $\hat{\boldsymbol{p}}_t^{(a)}$ of the transmitter position from the estimated CSI $\hat{\boldsymbol{H}}_t^{(a)}$, using a CNN.

3. *Local Information Forwarding*: each AP a forwards to the MAP the received message, the estimated position $\hat{\boldsymbol{p}}_t^{(a)}$, and the SNR coefficient $\gamma_t^{(a)}$.

4. *Fine Position Estimation*: the MAP performs a weighted average of the positions estimated by all the APs. The weights are the SNR coefficients. Thus, defining $\gamma_{\text{tot}} = \sum_{a\in\mathcal{A}} \gamma_t^{(a)}$, the estimated position at the physical layer is
$$\hat{\boldsymbol{p}}_{\text{ph},t} = \frac{1}{\gamma_{\text{tot}}} \sum_{a\in\mathcal{A}} \gamma_t^{(a)} \hat{\boldsymbol{p}}_t^{(a)} . \tag{F.3}$$

5. *Traffic-based Position Estimation* The MAP also obtain estimate $\hat{\boldsymbol{p}}_{\text{up},t}$ of the transmitter position using upper-layer data, i.e., the binary vectors $\boldsymbol{v}_t$, via SVR, the ML model that predicts the position from the traffic data.

6. *Time Interpolation and Classification*: the predicted positions $\hat{\boldsymbol{p}}_{\text{ph},t}$ and $\hat{\boldsymbol{p}}_{\text{up},t}$ are generally correlated in time. We can capture such a correlation by applying a Kalman filter to the estimated positions and obtain the final estimates $\boldsymbol{q}_{\text{ph},t}$ and $\boldsymbol{q}_{\text{up},t}$. With $\boldsymbol{p}_t$ being the expected position of Alice,

$$e_{\text{ph}} = \|\boldsymbol{q}_{\text{ph},t} - \boldsymbol{p}_t\| \text{ and } e_{\text{up}} = \|\boldsymbol{q}_{\text{up},t} - \boldsymbol{p}_t\| \tag{F.4}$$

the two prediction errors, then the predicted class $\hat{\mathcal{H}}$ is

$$\hat{\mathcal{H}} = \begin{cases} \mathcal{H}_0 \text{ if } e_{\text{ph}} \leq \phi \wedge e_{\text{up}} \leq \phi, \\ \mathcal{H}_1, \text{otherwise.} \end{cases} \tag{F.5}$$

In other words, both predictions from the physical and the upper layer should stay at a maximum distance ϕ from the legitimate one in normal conditions.

We employ two different components to estimate the position from the physical-layer and upper-layer features for efficiency, complexity, and robustness. Indeed, each AP shares with the MAP only the local predicted position rather than the whole measured CSI (two real numbers against $M_a \times \frac{N_S}{C} \times 2$), thus improving the efficiency. Moreover, reducing the input dimensionality requires a less complex system, with less data and training time, thus reducing the complexity. Lastly, a fully centralized solution would have created a single point of failure at the MAP. In the proposed solution, instead, the APs can predict the transmitter position (with lower precision) from the CSI data, even in the case of MAP failure, thus they would still be able to detect anomalies. This makes the system more robust.
In the next Section, we explain the architectures of the two models, namely CNN and SVR, as well as the Kalman filter design.

F.3.1 Mechanisms Components' Design

We now provide the details for the design of the various components of the proposed authentication solution.

CNN We first recall that each AP employs a CNN to predict the transmitter position starting from CSI data. We consider CNNs as they are a good fit for this task [143, 144]. Indeed, we can interpret the estimated CSI $\hat{\boldsymbol{H}}_t^{(a)}$ as an image with two channels (real and imaginary part) and of dimension $M_a \times \frac{N_S}{C}$. The proposed CNN comprises a feature extraction module (i.e., a sequence of convolutional layers and pooling operations) to extract important features for the task, and a fully connected module. The details of the architecture are summarized in Table F.1. The model is tailored to exploit both the spatial and frequency-domain structures present in the CSI tensor.

SVR The MAP employs SVR to estimate Alice's position $\hat{\boldsymbol{p}}_{\text{up},t} \in \mathbb{R}^2$ from the observed traffic patterns $\boldsymbol{v}_t \in \{0,1\}^N$ at time t. SVR extends the support vector machine paradigm to continuous-valued function approximation [145]. In particular, given N_{sp} labeled training samples $\mathcal{D} = \{(\boldsymbol{v}_t, \boldsymbol{p}_t)\}_{t=1}^{N_{\text{sp}}}$, we decompose the two-dimensional problem into two independent scalar regression tasks, performed by models $f_{\mathbf{w}_x}$ and $f_{\mathbf{w}_y} : \mathbb{R}^N \to \mathbb{R}$. Each model minimizes the ϵ-insensitive loss function to solve the problem

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C_{\text{reg}} \sum_{t=1}^{N_{\text{sp}}} \max(0, \|f(\boldsymbol{v}_t) - \boldsymbol{p}_t\| - \epsilon)\,, \tag{F.6}$$

where $f(\boldsymbol{v}_t) = \mathbf{w}^T \boldsymbol{v}_t + b$ and $\mathbf{w} = \mathbf{w}_x$ or $\mathbf{w}_y$, depending on whether we are using x or y coordinate, respectively. The coefficient C_{reg} controls the regularization strength, and ϵ defines the loss tolerance margin. To capture the non-linear relation between traffic and position vectors, we used the radial basis kernels (RBFs) and solved the kernel/dual form of (F.6). We refer to [145, Chp. 6] for more details.

Table F.1: CNN Specifications.

Layer	Parameters
Input	Input channels: 2 (real, imaginary)
Conv2d + ReLU	Output channels: 16, kernel 3×3
Conv2d + ReLU	Output channels: 32, kernel 3×3
MaxPool2d	kernel 2×2
Conv2d + ReLU	Output channels: 64, kernel 3×3
MaxPool2d	kernel 2×2
Flatten	–
Fully Connected + ReLU	N. neurons: 64
Fully Connected + ReLU	N. neurons: 64
Output Layer	Output dim.: 2

Kalman Filter As mentioned above, the Kalman filter refines the estimates obtained with the CNN and SVR to smooth the behavior over time. The *state* of the system at time t is vector $\boldsymbol{x}_t = [p_{x,t}, p_{y,t}, v_{x,t}, v_{y,t}]^T$, which contains the positions and velocities on the x-y axes. The transition matrix is

$$\boldsymbol{F} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{F.7}$$

which describes the dynamic of the system linking the state $\boldsymbol{x}_t$ to the previous one, as

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}\,, \tag{F.8}$$

where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ is the model noise vector, representing the uncertainty on the model dynamics. The *measurement* is instead $\boldsymbol{z}_t = \hat{\boldsymbol{p}}_t = [\hat{p}_{x,t}, \hat{p}_{y,t}]^T$, which for us is either the CNN or SVR predictions. The measurement is linked to the state $\boldsymbol{x}_t$ through the measurement matrix

$$\boldsymbol{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{F.9}$$

as

$$\boldsymbol{z}_t = \boldsymbol{M}\boldsymbol{x}_{t-1} + \boldsymbol{w}'\,, \tag{F.10}$$

where $\boldsymbol{w}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ is the measurement noise. The Kalman filter is tuned by choosing appropriate covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, for instance, choosing $\mathrm{tr}(\boldsymbol{R}) > \mathrm{tr}(\boldsymbol{Q})$ means we are trusting the model dynamics rather than the measurements. We refer to [146] for a more complete description of the problem, especially on how to combine the *a priori* estimate of the state $\boldsymbol{x}_t$ with actual measurements $\boldsymbol{z}_t$, thus estimating the *a posteriori* estimate of the state. The Kalman output is then the a posteriori estimates $\boldsymbol{q}_{\mathrm{ph},t}$ and $\boldsymbol{q}_{\mathrm{up},t}$, refined versions of the CNN and SVR outputs, respectively.

F.4 Cross-layer Anomaly Detection for Moving End-points

F.4.1 Traffic Model with Mobile Endpoints

At time t, each endpoint $k \in \mathcal{K}$ is located at position $\boldsymbol{z}_t(k) \in \mathbb{R}^2$. Unlike the previous scenario where endpoints were stationary, we now consider that endpoints can move with velocity bounded by $v_{\max}$. Specifically, the position of endpoint k evolves according to

$$\boldsymbol{z}_{t+\Delta t}(k) = \boldsymbol{z}_t(k) + (\boldsymbol{v}_k + \boldsymbol{w}_k(t)) \cdot \Delta t\,, \tag{F.11}$$

where $\boldsymbol{v}_k = (v_{x,k}, v_{y,k})^\top \in \mathbb{R}^2$ is the mean velocity vector of endpoint k, with components independently drawn from a uniform distribution in $[0, v_{\max}]$ and kept constant throughout the trajectory, and $\boldsymbol{w}_k(t) \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I}_2)$ is a zero-mean Gaussian velocity perturbation with variance σ_w^2. This model represents ballistic motion with stochastic perturbations, where endpoints maintain a roughly constant velocity with small random deviations. The noise term is scaled by the sampling period Δt, reflecting uncertainty that accumulates over the time interval.
The set of *reachable endpoints* at time t is defined as those endpoints whose current positions lie within a circle of radius R centered at Alice's position $\boldsymbol{p}_t$:

$$\mathcal{S}(\boldsymbol{p}_t, t) = \{k : k \in \mathcal{K}, \|\boldsymbol{z}_t(k) - \boldsymbol{p}_t\| \leq R\}\ . \tag{F.12}$$

However, unlike the stationary case, the set $\mathcal{S}(\boldsymbol{p}_t, t)$ now explicitly depends on time t due to endpoint mobility.
Alice sends packets to the set of *active endpoints*, which is a subset of $\mu \in \{0, \ldots, |\mathcal{S}(\boldsymbol{p}_t, t)|\}$ endpoints selected uniformly at random from $\mathcal{S}(\boldsymbol{p}_t, t)$. The active endpoints represent those currently engaged in communication with Alice.

Traffic Monitoring The traffic model is analogous to the static case; still, due to endpoint mobility, the traffic pattern $\boldsymbol{v}_t$ evolves dynamically as endpoints move in and out of Alice's communication range, requiring the positioning system to adapt continuously to maintain accurate localization.

F.4.2 Detection Mechanism for Moving Endpoints

The positioning system employs a multilayer perceptron (MLP) Regressor instead of SVR to predict Alice's position from the observed traffic patterns. The network architecture is designed to extract features from the binary traffic vector and map them to continuous position coordinates. As detailed in Table F.2, the network consists of three layers. The input layer receives the N-dimensional binary traffic vector $\boldsymbol{v}_t$ and projects it to a 100-dimensional feature space using a fully connected layer with Rectified Linear Unit (ReLU) activation. The first hidden layer further compresses the representation to 50 dimensions, also with ReLU activation. Finally, the output layer maps the learned features to a 2-dimensional position estimate $\hat{\boldsymbol{p}}_t = (\hat{x}_t, \hat{y}_t)$ corresponding to Alice's predicted coordinates. The output layer uses a linear activation to allow unrestricted position predictions across the deployment area. All layers employ L_2 regularization (weight decay) with coefficient $\lambda = 0.01$ to prevent overfitting and improve generalization to unseen traffic patterns.

Offline Training of the base model The MLP is trained offline using a dataset $\mathcal{D} = \{(\boldsymbol{v}_i, \boldsymbol{p}_i)\}_{i=1}^{M}$ of M traffic-position pairs collected during an initial calibration phase. During this phase, Alice's position $\boldsymbol{p}_i$ is known, and the corresponding traffic vector $\boldsymbol{v}_i$ is recorded. For the fixed endpoint scenario, this dataset captures the spatial relationship between Alice's position and the set of reachable endpoints. For the mobile

Table F.2: MLP Architecture for Position Estimation

Layer	Type	Input Size	Output Size	Activation
Input	Dense	N	100	ReLU
Hidden 1	Dense	100	50	ReLU
Output	Dense	50	2	Linear

endpoint scenario, the dataset includes samples collected over time, capturing the dynamic traffic patterns induced by endpoint mobility.
Training is performed using the Adam optimizer with an initial learning rate of $\alpha = 10^{-3}$. A learning rate scheduler (ReduceLROnPlateau) is employed to reduce the learning rate by a factor of 0.5 when the validation loss plateaus for 20 consecutive epochs. Mini-batch gradient descent is used with batch size $B = 64$, and training continues for up to 500 epochs with early stopping based on validation loss (patience of 50 epochs). Input features are standardized to zero mean and unit variance, and output positions are similarly normalized to facilitate stable optimization.
Upon completion of offline training, the network achieves a mean positioning error of approximately $\epsilon_{\text{base}} \approx 0.40$ m on the test set for the fixed endpoint scenario. However, in the mobile endpoint scenario, the offline-trained model experiences significant performance degradation due to the distribution shift caused by endpoint mobility.

On-the-Fly Fine-Tuning for Adaptation To address the performance degradation in the mobile endpoint scenario, we employ an on the fly fine-tuning strategy that adapts the network to the changing environment with minimal computational overhead. This approach is crucial for maintaining accurate positioning as endpoints move and alter the traffic patterns.

Layer Freezing Strategy Rather than updating all network parameters, which would be computationally expensive and prone to catastrophic forgetting, we adopt a transfer learning approach. Specifically, we freeze the parameters of the input and hidden layers, treating them as a fixed feature extractor. Only the output layer parameters $\boldsymbol{W}_{\text{out}}$ and bias $\boldsymbol{b}_{\text{out}}$ are updated during online fine-tuning. This design choice is motivated by the observation that the lower layers learn general spatial features that remain relevant despite endpoint mobility, while the output layer must adapt to map these features to the new position-traffic relationship.

Fine-Tuning Procedure When a new measurement is available at time t with the corresponding traffic vector $\boldsymbol{v}_t$ and ground truth position $\boldsymbol{p}_t$, and this measurement is marked as legitimate, the finetuning procedure takes place. As outlined in Alg. F.3, given the output from the frozen layers $\phi(\boldsymbol{v}_t)$, the model predicts the position $\hat{\boldsymbol{p}}_t$ which is used to update the output layer.
The fine-tuning learning rate is set to $\beta = 0.01$, which is higher than the offline training rate to enable rapid adaptation. The number of gradient steps per sample is $J = 2$, ensuring a balance between adaptation quality and computational efficiency. It is important to note that for each sample t, the positioning error is computed *before* fine-tuning on that sample. Specifically, the model predicts $\hat{\boldsymbol{p}}_t$ using weights learned from samples $1, \ldots, t-1$, then the error $\|\hat{\boldsymbol{p}}_t - \boldsymbol{p}_t\|$ is recorded. Subsequently, the model is fine-tuned on $(\boldsymbol{v}_t, \boldsymbol{p}_t)$, which improves predictions for future samples $t+1, t+2, \ldots$. This protocol ensures that the evaluation reflects true online learning performance, where the model must predict before observing the ground truth label.

Figure F.3: On-the-fly fine-tuning of output layer

1: Extract features: $\boldsymbol{h}_t = \phi(\boldsymbol{v}_t)$ using frozen layers
2: **for** $j = 1$ to J **do**
3: Compute prediction: $\hat{\boldsymbol{p}}_t = \boldsymbol{W}_{\text{out}}\boldsymbol{h}_t + \boldsymbol{b}_{\text{out}}$
4: Calculate loss: $\mathcal{L}_t = \|\hat{\boldsymbol{p}}_t - \boldsymbol{p}_t\|^2$
5: Update weights:

$$\boldsymbol{W}_{\text{out}} \leftarrow \boldsymbol{W}_{\text{out}} - \beta\nabla_{\boldsymbol{W}_{\text{out}}}\mathcal{L}_t$$
$$\boldsymbol{b}_{\text{out}} \leftarrow \boldsymbol{b}_{\text{out}} - \beta\nabla_{\boldsymbol{b}_{\text{out}}}\mathcal{L}_t$$

6: **end for**

where β is the fine-tuning learning rate.

F.4.3 Attack Strategies

To attack, Trudy generates upper-layer data and/or transmits data as she was in a position at a distance $D_{\max}$ with respect to Alice's position. We consider different attack strategies adopted by Trudy.

- *Wrong-Physical-Wrong-Traffic (WPWT)*: In this attack, Trudy both transmits data and generates traffic at a distance $D_{\max}$ from the expected one. In other words, neither the traffic nor her signal is compatible with the expected position.
- *Wrong-Physical-Correct-Traffic (WPCT)*: Trudy generates the correct traffic data with respect to Alice's position, but it is transmitted from the wrong position.
- *Correct-Physical-Wrong-Traffic (CPWT)*: Trudy transmits from the legitimate expected position, but the upper-layer data is anomalous. This attack is interesting because it can break the protocols based on the CSI only like [102, 103].

F.4.4 Security Analysis

Our framework efficiently combines information from different layers to improve channel-based authentication [30], which relies solely on physical layer features. While our protocol can be extended to multiple devices, this increases the complexity of the CNN and SVR models. However, transfer-learning techniques can be applied to maintain a simple yet effective framework.

F.5 Numerical Results

In this section, we describe the data we used to validate our framework and evaluate the security performance in terms of false alarm (FA) probability, i.e., the probability that Alice is misled to Trudy

$$P_{\text{fa}} = \mathbb{P}(\hat{\mathcal{H}} = \mathcal{H}_1 | \mathcal{H} = \mathcal{H}_0)\,, \tag{F.13}$$

and the misdetection (MD) probability, i.e., the probability that Trudy is misled to Alice

$$P_{\text{md}} = \mathbb{P}(\hat{\mathcal{H}} = \mathcal{H}_0 | \mathcal{H} = \mathcal{H}_1)\,. \tag{F.14}$$

The FA/MD probabilities against the attacks, as well as the localization performance for different ML strategies and scenario parameters, are explained in the following subsections.

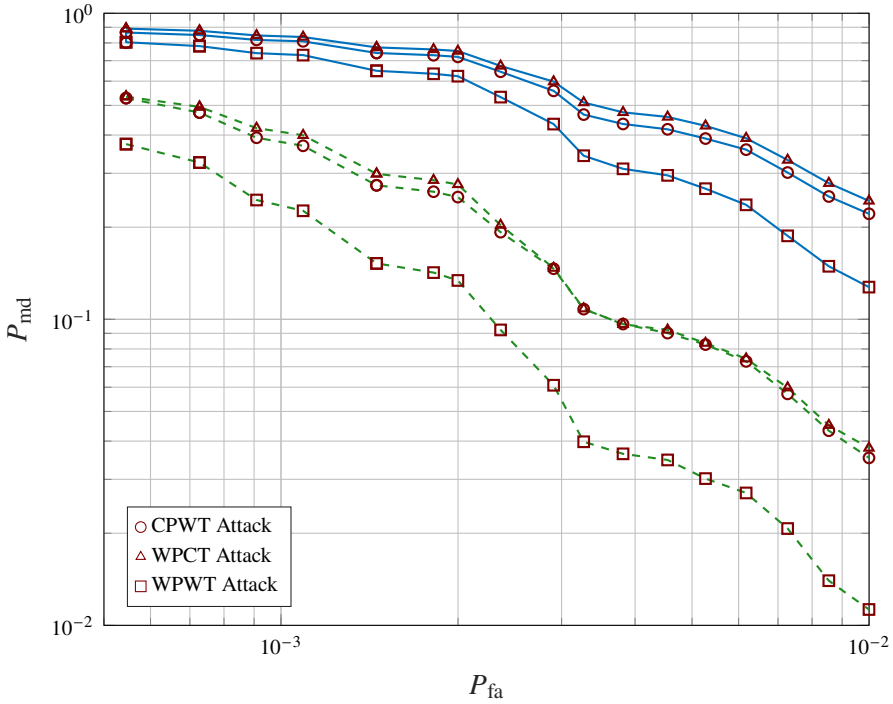


Figure F.4: DET curve of CLD against the WPWT (squares), WPCT (triangles) and CPWT (circles) attacks and at different Trudy distances $D_{\max} = 1.5$ m (solid-blue line) and $D_{\max} = 2$ m (dashed-green line).

F.5.1 Dataset Description

To validate the proposed framework, we use experimental data for the physical-layer features and artificial data for the upper-layer features.

Physical-Layer Data For the physical data, we resorted to the DICHASUS dataset [142], which contains experimental physical-data features obtained when a movable transmitter (Alice) continuously transmits OFDM-modulated pilot symbols when moving on a defined closed track in multiple rounds. Each OFDM symbol comprises $N_{\text{S}} = 1024$ subcarriers over a bandwidth $B = 50$ MHz at 1.272 GHz. The APs are $L = 4$, fixed, with arrays of different numbers of antennas, namely $M_1 = 8, M_2 = M_3 = 5$, and $M_4 = 6$. The $\sim 2.7 \times 10^4$ data points are labeled with the ground-truth positions of the transmitting device $\boldsymbol{p}_t$. Fig. F.1 shows the considered scenario, which includes two Alice trajectories as well as APs positions from [142].

Upper-Layer Dataset Concerning the upper-layer data, we randomly generated the positions of N end-points within a distance D_{R} of Alice's track. Fig. F.1 shows $N = 50$ randomly generated endpoint positions within distance $D_{\text{R}} = 0.5$ m, as well as the coverage area.

F.5.2 Performance With Static End-points

FA/MD Probabilities Against Various Attacks We begin our analysis by showing the behavior of our framework against the WPWT, WPCT, and CPWT attacks for various Trudy distances $D_{\max}$. Fig. F.4 shows the detection error tradeoff (DET) of the proposed authentication mechanism, obtained by varying the decision threshold ϕ.

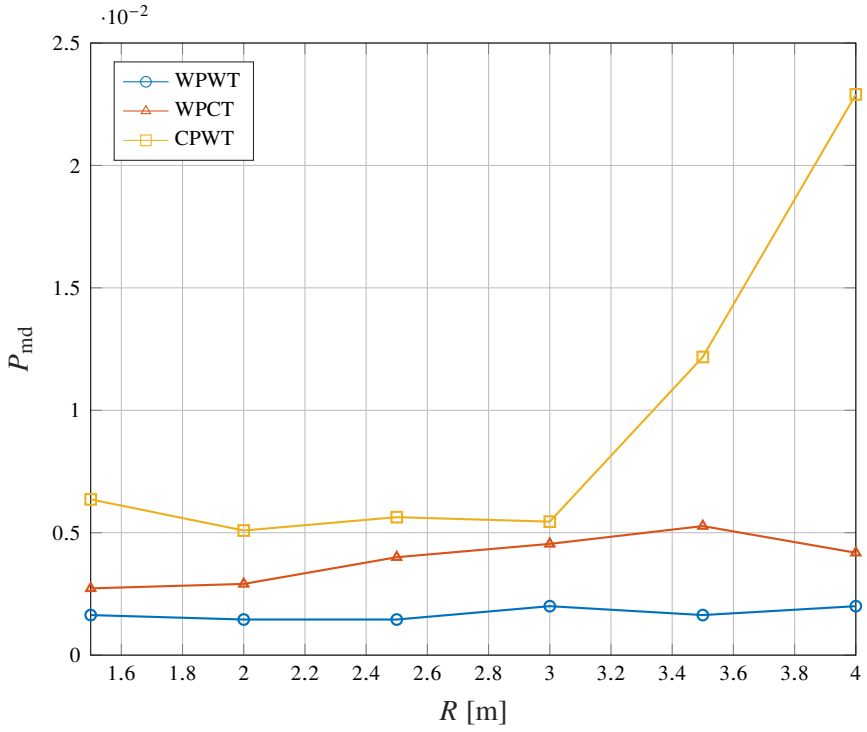


Figure F.5: MD probability as a function of the coverage radius R with $P_{\text{fa}} = 0.1$, for various attack strategies.

We note that, as $D_{\max}$ increases, the system becomes more capable of detecting anomalies, as expected, since the position recovered from the transmissions becomes more inconsistent with the expected position. Among the attacks, the least effective is WPWT, as it contains anomalies both at the physical and upper layers. Comparing the WPCT and CPWT attacks, we notice that their effectiveness depends on the localization accuracy, which in turn highly depends on the system conditions. In fact, we see in Fig. F.5 that the CPWT attack is, in general, more sensitive to the coverage radius R, as increasing the coverage radius too much reduces the localization accuracy (see Fig. F.6, right), thus Trudy can exploit that.
Lastly, we observe that our framework is also effective against the CPWT attack, which would otherwise remain undetected in [102] and [103], thus we demonstrate the superiority of our protocol over state-of-the-art CSI-only based protocols.

Position RMSE We compare four regression approaches for traffic-based localization to find the best model for our sparse binary input: random forest as a robust ensemble baseline for binary features, XGBoost for its good handling of sparse data and feature interactions, SVR with RBF kernels for non-parametric non-linear spatial mapping, and neural networks (NNs) for their universal approximation capabilities. This selection spans tree-based, kernel-based, and neural paradigms, enabling comprehensive evaluation of which learning framework best captures the implicit position-traffic relationship encoded in $\boldsymbol{v}_t$ through the reachable set $\mathcal{S}(\boldsymbol{p}_t)$. Fig. F.6 (left) shows the root-mean square error (RMSE) of the position estimation for various regression models, varying the number of active endpoints μ. The coverage radius is fixed at $R = 2$ m and the total number of endpoints is $N = 50$. All models consistently improve the localization accuracy as the number of active endpoints increases, reflecting the enhanced spatial information available. SVR achieves the best overall performance, closely followed by XGBoost and NN. Notably, the performance gap between models narrows substantially as the number of endpoints increases.
Fig. F.6 (right) illustrates the impact of the coverage radius R and total number of endpoints N on SVR

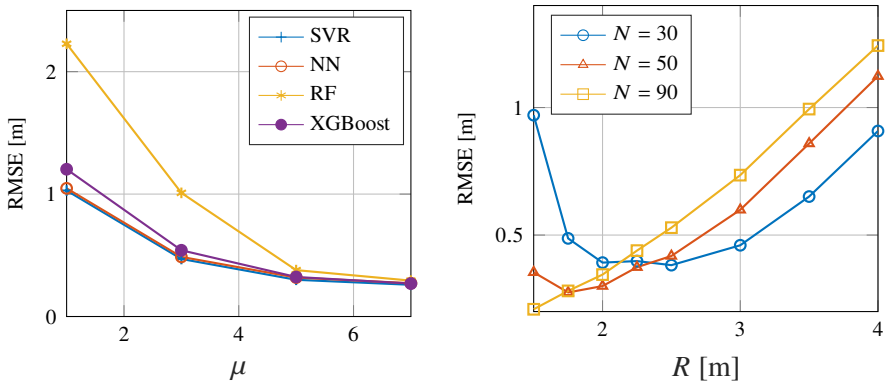


Figure F.6: RMSE of different regression models vs the number of active endpoints μ (left) and RMSE of SVR against the coverage radius R and number of endpoints N (right).

localization accuracy, with maximum active endpoints fixed at $\mu_{\max} = 5$.
Note that, beyond a certain R, whose value depends on the number of endpoints N, the localization error increases monotonically across all configurations. This is because having too many reachable endpoints makes the active endpoint information uncorrelated with the actual transmitting position: a radius covering the whole trajectory makes upper-layer information useless for position estimation. On the other hand, having a low R might lead to insufficient spatial diversity when few endpoints populate the reachable set $\mathcal{S}(\boldsymbol{p}_t)$, as for $N = 30$. These findings suggest a sweet spot on the coverage radius extension that needs to be handled with care.

F.5.3 Performance with Moving Endpoints

Impact of Endpoint Mobility Figure F.7 illustrates the impact of endpoint mobility on positioning accuracy. Without any adaptation mechanism, the positioning error increases nearly linearly with receiver velocity, reaching approximately 4 meters at $v_{\max} = 3$ mm/s. This severe performance degradation occurs because the offline-trained model learns a fixed mapping between traffic patterns and positions, which becomes invalid as receivers move and alter the spatial relationship between Alice's position and active endpoints. In contrast, the online fine-tuning strategy maintains consistently low positioning error across all velocity regimes. Even at the highest velocity of $v_{\max} = 3$ mm/s, where the non-adaptive model fails with 4 m error, the adaptive model maintains approximately 0.35 m error—comparable to the baseline performance with stationary receivers.

F.6 Conclusions

In this paper, we introduced CLD, a cross-layer ML framework for detecting anomalies using data obtained from the physical and upper layers. We demonstrate the superiority of CLD over state-of-the-art detection protocols based only on physical-layer data. Our framework can effectively detect anomalies with FA/MD probabilities of about 10^{-2} when the attacker transmits and generates traffic at a distance of only about 2 m from the expected one.

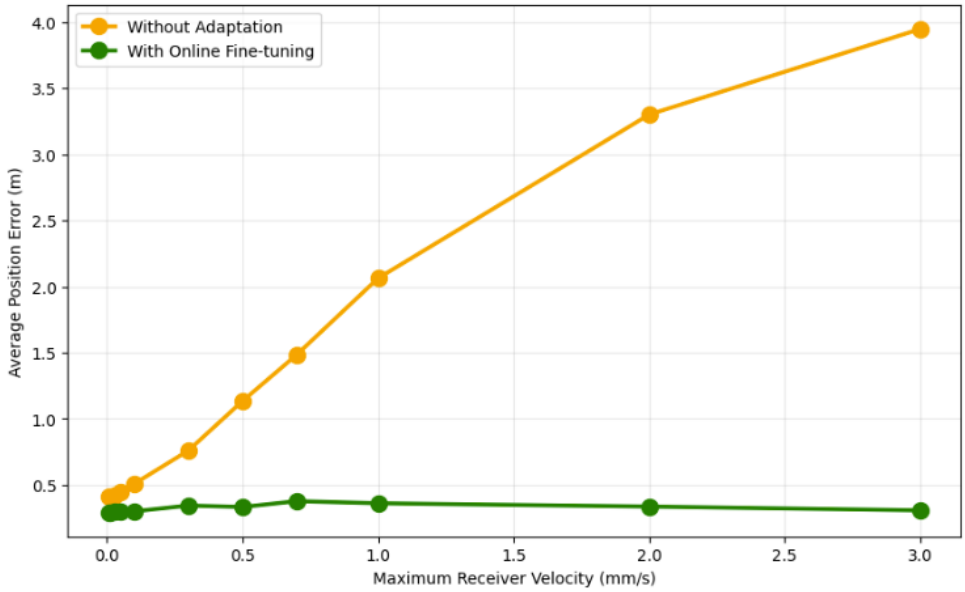


Figure F.7: Impact of receiver mobility on positioning accuracy. The orange curve shows severe performance degradation without adaptation, while the green curve demonstrates that online fine-tuning maintains consistent accuracy across all velocity regimes. Configuration: $N = 50$ receivers, $R = 2$m coverage radius, 5000 test samples.

Appendix G

Predictive Modeling for RF Fingerprint Evolution

G.1 Experimental Testbed

The experimental validation for the predictive modeling task is conducted on a custom testbed comprising 30 IoT transmitters and a receiver array. The setup is located in a controlled indoor environment to minimize external interference while allowing for the observation of hardware-induced drift.
Figure G.1 displays the wide view of the 30-transmitter grid and receiver array. The setup ensures that all devices are subjected to identical environmental conditions (temperature, humidity) during the long-term capture.

Figure G.1: Wide view of the RF-PREDICT testbed showing the transmitter grid and receiver array.

Figure G.2 provides a detailed view of the sensors. Half of these units are powered by batteries to enable the investigation of spectral drift caused by voltage decay, while the other half utilize stable DC power as a control group.

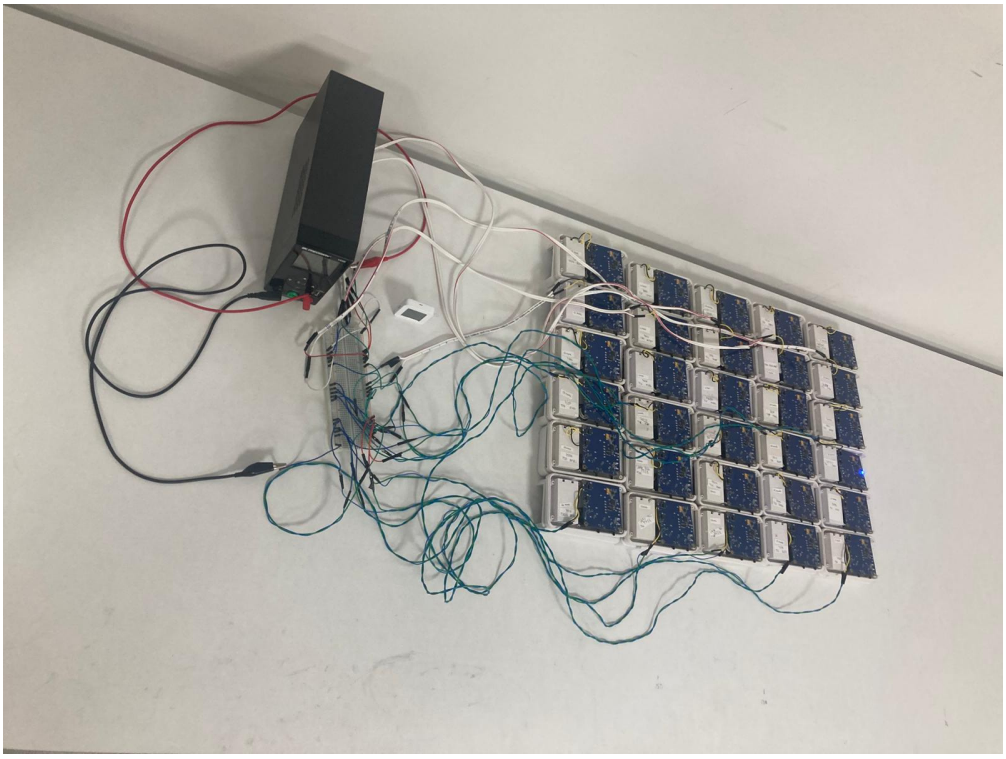


Figure G.2: Close-up of the transmitter array showing the mixed deployment of battery-powered and DC-powered units.

G.2 Data Collection Protocol

To study the "silence decay" phenomenon, transmitters are configured with varying transmission intervals. Table G.1 details the assignment of intervals to specific Device IDs.

Table G.1: RF-PREDICT: Transmission Intervals and Device Configuration

Device IDs	Interval	Device IDs	Interval	Device IDs	Interval
T01 - T02	15 Seconds	T11 - T12	15 Minutes	T21 - T22	6 Hours
T03 - T04	30 Seconds	T13 - T14	30 Minutes	T23 - T24	8 Hours
T05 - T06	1 Minute	T15 - T16	1 Hour	T25 - T26	12 Hours
T07 - T08	5 Minutes	T17 - T18	2 Hours	T27 - T28	18 Hours
T09 - T10	10 Minutes	T19 - T20	4 Hours	T29 - T30	24 Hours

G.3 Packet Structure

Each transmission utilizes a custom packet structure designed to capture telemetry data relevant to signal drift. As shown in Table G.2, fields for temperature and power level are embedded in every frame to allow for future correlation between spectral changes and physical device state.

Table G.2: Packet Structure for RF-PREDICT Dataset

Packet Feature	**Length (Bytes)**	**Description**
Preamble	4	Network Synchronization
Sync Word	4	Frame Detection
Length	1	Payload Length
Sequence Number	2	Packet tracking
Device ID	4	Unique Transmitter Identifier
RTC Timestamp	4	Real-Time Clock value
Temperature	1	On-chip temperature sensor
Power Level	1	Battery voltage level
CRC	2	Error check
Total	**23**	

Appendix H

Federated Authentication for 6G Networks

To be submitted.
by Mattia Piana, Stefano Rini, and Stefano Tomasin

Abstract: We study a scenario where several authorized devices (collectively denoted as Alice) are transmitting from several *authorized areas*. An attacker device, Trudy, is instead impersonating Alice, i.e., transmitting messages and claiming to be Alice. However, Trudy does not have access to the areas where Alice is. Therefore, we propose that multiple base stations (BSs) of a 6G network (collectively indicated as Bob) determine whether the received signal comes from the authorized areas or not, to authenticate the received messages, and determine whether they have been transmitted by Alice or not. The BSs collaborate to authenticate the transmitting device at the physical layer. This problem can also be seen as a problem of a distributed in-region location verification problem, [101].
For authentication, each AP first trains a local model that, from features extracted from the CSI estimated on the received message, makes a classification and decides about the origin region of the signal. Due to the fact that all BSs aim at recognizing the feature of the same areas, which is *seen from different viewpoints* from the BSs, the trained models should share some similarities. On the other hand, since each BS is located in a different position, the statistics of the observed features are not the same, and the classification models are also different among the APs. To exploit the commonalities and make the learning process fast, while safeguarding the differences of each model, we propose FedLoss, a novel federated-learning (FL) framework that tackles the non-IID data heterogeneity present in common wireless environments via local fine-tuning. By numerical evaluations using realistic channel models, we demonstrate the superiority of the proposed method over standard federated learning and non-federated learning algorithms.

H.1 Introduction

In recent years, FL has gained much interest as it allows different devices to collaborate on a common objective without explicitly sharing their data [147]. Each device, in fact, uses local data for local training, then uploads the model to the server for aggregation, and finally, the server sends the global model back to the participants.
Different aggregation strategies can be employed by the server. One is FedAvg [148], in which the server simply averages over the devices' updates. This is a very effective method when the data are i.i.d. across clients, but can perform poorly in the case of non-i.i.d. client data distributions [149]. The problem of non-i.i.d. data distributions is still open, and several strategies are proposed in the literature [150]. In [149], the authors propose to share a (small) dataset across the clients to regularize the training and mitigate weight divergence. In [151], the authors propose FedGroup, which employs a clustered-federated learning approach

where clusters are formed on the cosine similarity of the model updates of the nodes. The devices within each cluster then train using FedAvg. Similarly, [152] proposes to split users by performing a hierarchical clustering.
Meta-learning is another approach to tackle non-IID datasets in federated learning: a meta-learner model is trained and shared across clients, which then perform local fine-tuning on their own datasets. This approach, named Personalized FedAvg, was first proposed in [153], and then its performance and optimality analysis were treated in [154]. Finally, in [155], the authors tackle the data imbalance by re-balancing the training datasets via the synthetic generation of data points from the minority classes.
The application of federated learning in PLA is at its dawn, and to the best of our knowledge, only a few works have been done. In [156], we have an Internet of Things (IoT) scenario where multiple constrained devices collaborate to authenticate the transmitter using the FedAvg algorithm. In [157], a distributed anomaly detection system for detecting compromised devices in LoRa-enabled IIoT is presented. It exploits device-specific features such as Carrier Frequency Offset (CFO) as the device fingerprint, whose deviations from expected behaviors allow for the detection of attackers. In [158], multiple Wi-Fi routers use FedProx [159] to cooperate in localizing the transmitting device.
In this report, we study a scenario where multiple BSs collaborate to authenticate the transmitting device at the physical layer. The legitimate transmitter, Alice, can be located in different areas; thus, each BS, Bob, needs to identify the transmitting area and, by knowing the legitimate one, can authenticate Alice. Trudy, on the other hand, aims at impersonating Alice by transmitting from another area with respect to Alice, thus fooling the BSs. The main contribution of this report is FedLoss, a novel FL framework that tackles the non-IID data heterogeneity present in common wireless environments via local fine-tuning. In particular, the contributions are as follows:

1. We propose a realistic channel model, taking as baseline the 3GPP specifications.
2. We present FedLoss, the FL framework able to tackle the non-IID data distributions via local finetuning.
3. We numerically evaluate FedLoss, demonstrating its effectiveness even in the case of challenging channel conditions and scarcity of data.

The rest of the chapter is organized as follows: in Section H.2 we present the System Model, in Section H.3 we present FedLoss, in Section H.4 we discuss the numerical results, while Section H.5 draws the main conclusions.

H.2 System Model

In our system, we have E BSs that use ULAs with M antennas. Each BS aims at authenticating the received signals at the physical layer; consequently, we have multiple authenticators Bob, each denoted as b_e, $e = 1, \dots, E$. The transmitter Alice, on the other hand, is a single antenna user and can transmit her messages from multiple positions. These positions are grouped into N areas, each denoted with a_n, $n = 1, \dots, N$, and the BSs collaborate to infer to which area the received signal belongs. We assume the BSs to know the provenance area of the legitimate transmitter; thus, their estimated area can be used to authenticate it. Trudy, whose position is unknown to both Alice and the BSs, on the other hand, is a single antenna transmitter aiming at impersonating Alice.

H.2.1 Channel Model

Let us refer to $\boldsymbol{p}_e$ and $\boldsymbol{p}_j$ as the positions of base station b_e and a general transmitter, respectively.

According to the 3GPP model [160], even in the line of sight (LoS) case, when the transmitter sends a signal, the BS receives a cluster of L signals. The AoA of each ray at the BS is

$$\theta_\ell^{(e,j)} = \theta^{(e,j)} + \Delta\theta_\ell \,, \tag{H.1}$$

where $\theta^{(e,j)}$ is the main AoA from the transmitter in $\boldsymbol{p}_j$ and BS b_e, $\Delta\theta_\ell \sim \mathcal{U}(-\theta_{\text{sp}}, \theta_{\text{sp}})$ represents the intra-cluster angular spread and α is a coefficient regulating the strength of such random variation on the AoA.
We assume the transmitter sends OFDM signals, thus the received signal at the subcarrier k in the frequency domain by the BS b_e upon the transmission of the pilot symbol x from position $\boldsymbol{p}_j$ is

$$\boldsymbol{y}^{(e,j)}(k) = \boldsymbol{h}^{(e,j)}(k)x + \boldsymbol{w} \in \mathbb{C}^{M\times 1} \,, \tag{H.2}$$

where $\boldsymbol{h}_k^{(e,j)}$ is the sum of the LoS and non line of sight (NLoS) components:

$$\boldsymbol{h}^{(e,j)}(k) = \sqrt{\frac{\kappa}{\kappa+1}}\boldsymbol{h}_{\text{LOS}}^{(e,j)}(k) + \sqrt{\frac{1}{\kappa+1}}\boldsymbol{h}_{\text{NLOS}}^{(e,j)}(k) \,. \tag{H.3}$$

The LoS vector is defined as

$$\boldsymbol{h}_{\text{LOS}}^{(e,j)}(k) = \sum_{\ell=1}^{L} \gamma^{(e,j)}\boldsymbol{\beta}(\theta_\ell^{(e,j)})e^{-j2\pi k\Delta f \tau_\ell^{(e,j)}} \,, \tag{H.4}$$

where $\boldsymbol{\beta}(\theta_\ell^{(e,j)}) = \frac{1}{\sqrt{M}}[1, e^{j\pi\sin\theta_\ell^{(e,j)}}, \ldots, e^{j\pi(M-1)\sin\theta_\ell^{(e,j)}}]$ is the steering vector and Δf is the subcarrier spacing. As the rays come from the same source, in the LoS, the pathloss experienced by each ray is approximately the same and constant across L. The channel gain coefficient using the free-space pathloss formula is $\gamma^{(e,j)} = \frac{c}{4\pi d^{(e,j)} f_c}$, where c is the speed of light in air and $d^{(e,j)} = \|\boldsymbol{p}_e - \boldsymbol{p}_j\|$ is the distance between transmitter and receiver. With $\tau_0 = \frac{d^{(e,j)}}{c}$ as delay of the first ray, according to the 3GPP we, can model the delays of the intra-cluster rays as uniformly random around the first one, i.e.,

$$\tau_\ell^{(e,j)} = \tau_0 + \Delta\tau_\ell \,, \tag{H.5}$$

where $\Delta\tau_\ell = \Delta\tau'_\ell - \min\{\Delta\tau'_\ell\}_{\ell=1}^{L}$ with $\Delta\tau'_\ell \sim \mathcal{U}(0, 2c_{\text{DS}})$.
The NLoS vector components on the other hand are i.i.d zero-mean complex Gaussians with covariance matrix $\boldsymbol{R}^{(e,j)} = \gamma^{(e,j)2}\boldsymbol{I}$, i.e., $\boldsymbol{h}_{\text{NLOS}}^{(e,j)}(k) \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{R}^{(e,j)})$.
Assuming the transmission of unit power symbols, $\mathbb{E}(\|x\|^2) = 1$, the SNR is defined as

$$\text{SNR}^{(e,j)} = \frac{\mathbb{E}(\|\boldsymbol{h}^{(e,j)}(k)\|^2)}{\mathbb{E}(\|\boldsymbol{w}\|^2)} = \frac{L(\gamma^{(e,j)})^2}{\sigma_w^2}. \tag{H.6}$$

The single-input multiple-output (SIMO) channel from the transmitter in position $\boldsymbol{p}_j$ and the BS b_e is the channel frequency response (CFR) matrix:

$$\boldsymbol{H}^{(e,j)} = [\boldsymbol{h}^{(e,j)}(1), \ldots, \boldsymbol{h}^{(e,j)}(N_{\text{S}})] \in \mathbb{C}^{M\times N_{\text{S}}} \,. \tag{H.7}$$

The areas are squares of side S_{A} centered in positions $\boldsymbol{p}_n$ of a 3D reference frame, and are denoted as a_n. In Fig. H.1, we see $N = 5$ areas with side $S_{\text{A}} = 400$ m with $E = 4$ BSs.

H.2.2 Dataset Description

Each base station b_e has N_{c} samples available per area, thus a total number of $C = N_{\text{c}}N$ samples. We denote as $\mathcal{D}_e = \{(\boldsymbol{H}^{(e,i)}, a^{(i)}), i = 1\ldots, C\}$ the dataset available at the BS b_e, where $\boldsymbol{H}^{(e,i)}$ is the channel in (H.7) between the BS b_e and Alice when she is located in $\boldsymbol{p}_j$, which corresponds to area $a^{(i)} = a_n$.

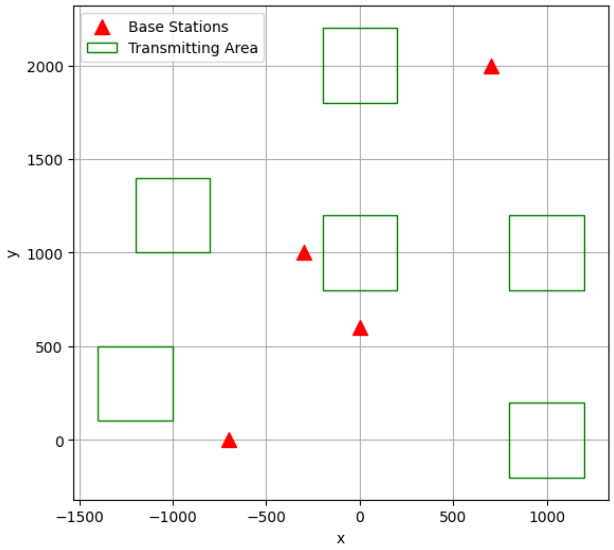


Figure H.1: Simulation scenario.

H.2.3 Attacker Model

In this paper, we assume the areas a_n from which a transmitter can send its signal to be common knowledge; thus, they are known by the BSs Bob, Alice, and Trudy. Trudy aims at impersonating Alice by transmitting a signal from one of the possible transmitting areas a_n. Yet, we assume Trudy cannot transmit from Alice's area, thus if Alice transmits from area a_n, Trudy can only transmit from area a_m , $m = 1, \dots, N$, $m \neq n$, i.e., she has available $N - 1$ areas.

H.3 Problem Definition and FedLoss

The goal of each BS b_e is to correctly classify the measured channel $\boldsymbol{H}^{(e,i)}$ to the correct area $a^{(i)}$. This is a multiclass classification problem that can be tackled with standard ML-based classifiers. Still, due to the scarcity of the available data at each BS, they need to collaborate in a federated approach, i.e., without explicitly sharing their data. Note that, as the location of the BSs is different, their datasets are intrinsically non-IID. This complicates the learning problem, as a one-model-fits-all approach proposed by standard federated learning algorithms (e.g., FedAvg) can perform poorly. Consequently, in this report, we adopt a first knowledge sharing among BSs in a FedAvg fashion, followed by a local fine-tuning so that each BS can obtain its own personalized model.
In formulas, with $\boldsymbol{\theta} = \{\theta_e\}$ the set containing the E models in which θ_e is the model for the BS b_e, the learning problem is

$$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \frac{1}{E} \sum_{e=1}^{E} \mathcal{L}\left(\mathcal{D}_e; \theta_e\right), \tag{H.8}$$

where $\mathcal{L}(\cdot, \cdot)$ is the multicalss cross-entropy loss.

H.3.1 FedLoss

The proposed FL framework works in two phases:

1. *FedAvg Phase*: In this phase, the goal is learning a single global model θ_{G} using the FedAvg algorithm. In particular, at each epoch, the global model is updated by computing

$$\Delta\theta_{\mathrm{G}} = \frac{1}{E}\sum_{e=1}^{E}\Delta\theta_e, \tag{H.9}$$

where $\Delta\theta_e$ are the local updates. Then the global model gets updated as

$$\theta_{\mathrm{G}} \leftarrow \theta_{\mathrm{G}} + \Delta\theta_{\mathrm{G}}.$$

2. *Fine-tune Phase*: Here each BS b_e fine-tunes the global model θ_{G} with its own dataset $\mathcal{D}_e$ to obtain personalized models $\theta_e^\star$. More in detail, we have that

$$\theta_e^\star = \arg\min_{\theta} \mathcal{L}\big(\mathcal{D}_e;\theta\big) + \lambda\|\theta-\theta_{\mathrm{G}}\|^2, \tag{H.10}$$

where λ is a regularization parameter that encourages θ to remain close to θ_{G}.

H.3.2 Switching Epoch

The switching epoch between Phase 1) and Phase 2) is a crucial point, as switching too early results in a global model not yet converged, while on the other hand switching too late simply wastes training time. The idea is to monitor the training loss: if the (average) training loss did not decrease enough in the last T_{E} epochs, then switch to local fine-tuning. In formulas, with $\theta_{\mathrm{G},t}$ the global model at epoch t, the BS b_e switches to local training once the following condition is met:

$$\frac{1}{T_{\mathrm{E}}}\sum_{t=t_0}^{t_0+T_{\mathrm{E}}-1}\mathcal{L}\big(\mathcal{D}_e;\theta_{\mathrm{G},t}\big) - \mathcal{L}\big(\mathcal{D}_e;\theta_{\mathrm{G},t-1}\big) < \mathcal{L}_{\min}\,. \tag{H.11}$$

H.4 Numerical Results

To validate the effectiveness of FedLoss, we compare it with three baselines, named Global, Single, and FedAvg, considering all the possible transmitting positions of both Alice and Trudy. In the Global case, there is a "virtual" single BS that has a dataset containing all the BS data. This is optimal if the datasets were formed by IID data, i.e., when the BS were located close enough to one another. In the Single case, on the other hand, all the BS train on their own dataset, even if small. This approach is supposed to work well when the BSs are far from each other; thus, sharing local information is damaging the overall performance. Finally, in FedAvg, the BS perform the FedAvg algorithm.
The scenario we used for our simulations is depicted in Fig. H.1, where $E = 4$ BSs collaborate to classify signals coming from $N = 5$ areas. The parameters we used to generate the channels are the following: a constant $\mathrm{SNR}^{(e,j)} = 6\,\mathrm{dB}, \forall e, j$, $\theta_{\mathrm{sp}} = 6\,\mathrm{deg}$ and $c_{\mathrm{DS}} = 5$ ns, from 3GPP specifications [160]. Each BS has available $N_c = 32$ datapoints per class, thus a total of $N_c E = C = 160$ samples for training.

H.4.1 Network Architecture

To validate the proposed framework, we employed a standard ResNet-18 CNN as a backbone [161], with a two-layer fully-connected head. This is a state-of-the-art architecture, which makes our findings replicable and verifiable.

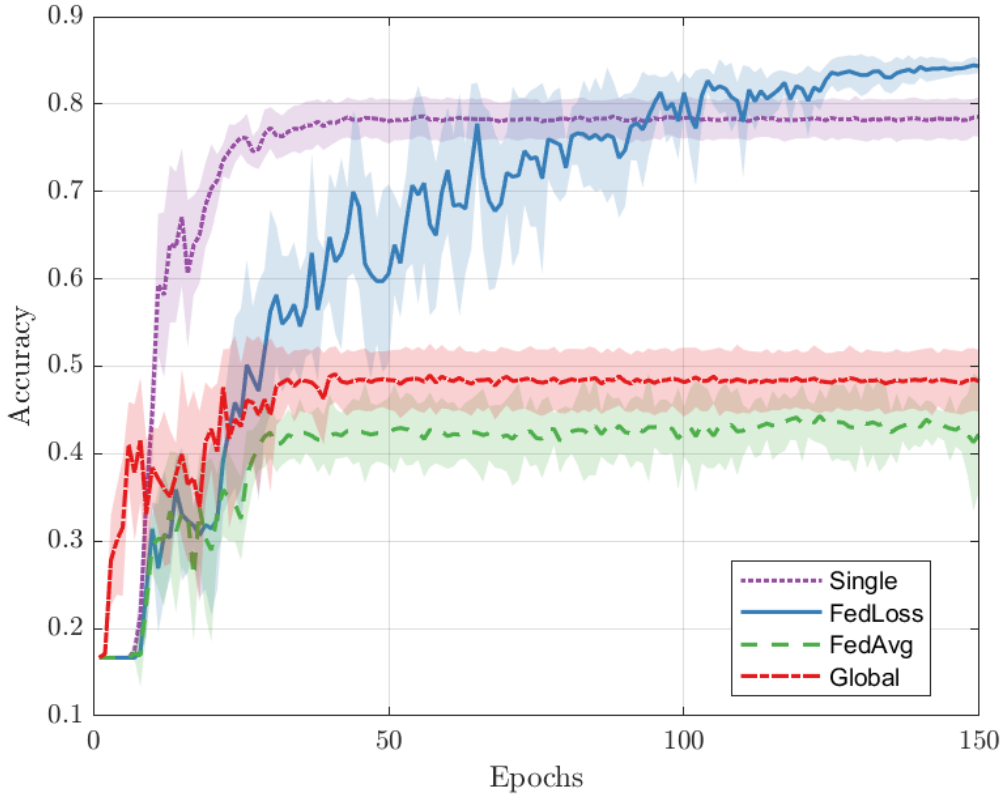


Figure H.2: Accuracy VS Epochs for Single Client, Global, FedAvg, and the proposed FedLoss.

H.4.2 Accuracy VS Epochs

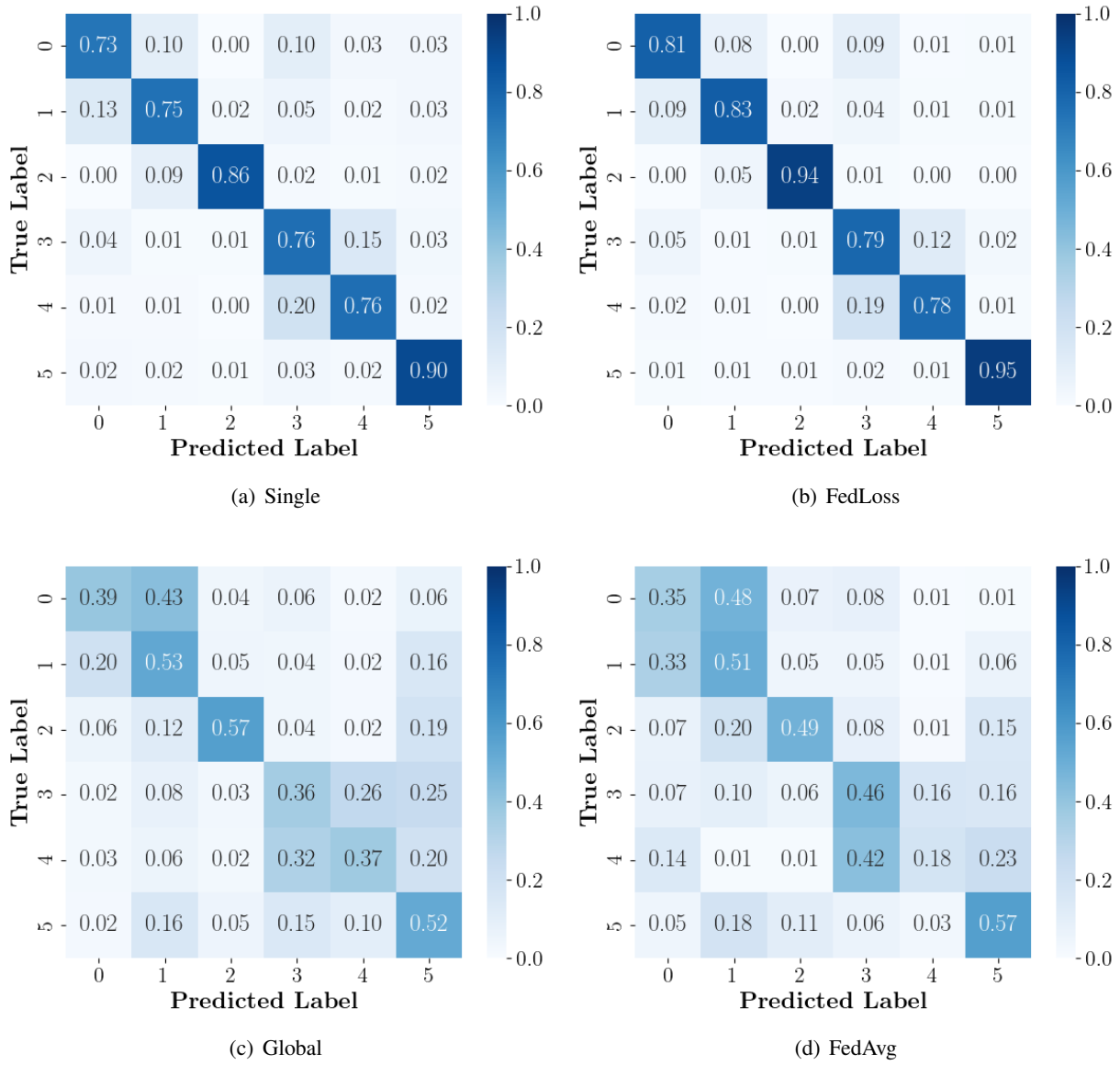
Fig. H.2 shows the average accuracy across the BSs as a function of the number of training epochs when the average distance between BSs is $D_{\text{bs}} \simeq 1.1$km. We notice that both Global and FedAvg perform poorly, as expected, since we see from Fig. H.1 that the BSs are very far from each other, leading to very non-IID datasets across the BSs. The Single performs well despite the small training dataset, but it gets outperformed by the proposed FedLoss, which achieves the best accuracy across all the methods.

H.4.3 Security Analysis

Fig. H.3 shows the confusion matrices of the proposed method FedLoss and the three baselines, averaged across BSs. These matrices are used to evaluate the security performance. In fact, assuming Alice is in a position in the area a_n and Trudy is in a position in the area a_m, we can find the probability that Trudy is misled by Alice by looking at the entry (n, m), $n \neq m$ of the confusion matrix. In other words, off-diagonal elements of the confusion represent the probability that Trudy fools the system. On the diagonal, we find the accuracy, which can be interpreted as the probability of correctly classifying Alice. By inspecting the matrices, we see the superiority of FedLoss over all the baselines.

H.5 Conclusions

In this report, we presented FedLoss, a novel federated learning framework that tackles the non-IID dataset issue common in standard federated learning algorithms by performing local fine-tuning. We numerically



(a) Single (b) FedLoss

(c) Global (d) FedAvg

Figure H.3: Average confusion matrices for Single Client, Global, FedAvg, and the proposed FedLoss.

evaluated its performance using realistic channel models in terms of accuracy, showing its superiority both over state-of-the-art federated and standard non-federated algorithms.

Appendix I

Jamming Detection in Cell-Free MIMO with Dynamic Graphs

Accepted to IEEE International Symposium on Personal, Indoor and Mobile Radio Communications 2026 (PIMRC 2026)
by Ali Hossary, Laura Crosara, and Stefano Tomasin

Abstract: Jamming attacks pose a critical threat to wireless networks, particularly in cell-free massive MIMO systems, where distributed access points and user equipment (UE) create complex, time-varying topologies. This paper proposes a novel jamming detection framework leveraging dynamic graphs and graph convolution neural networks (GCN) to address this challenge. By modeling the network as a dynamic graph, we capture evolving communication links and detect jamming attacks as anomalies in the graph evolution. A GCN-Transformers-based model, trained with supervised learning, learns graph embeddings to identify malicious interference. Performance evaluation in simulated scenarios with moving UEs, varying jamming conditions and channel fadings, demonstrates the method's effectiveness, which is assessed through accuracy and F1 score metrics, achieving promising results for effective jamming detection.

I.1 Introduction

Wireless communication increasingly adopts cell-free architectures to enhance connectivity and spectral efficiency. Cell-free MIMO relies on APs that jointly serve user equipments (UEs) without predefined cell boundaries. This paradigm shift introduces new challenges related to network dynamics and security [162]. As reliance on wireless services continues to grow, security threats have become a major concern. Wireless networks, due to the shared nature of the radio spectrum, are particularly vulnerable to jamming [163]. In MIMO wireless networks, traditional jamming detection methods rely on statistical models, which struggle to adapt to the complexities of dynamic wireless environments [85]. In contrast, deep learning (DL) techniques can be applied using a data-driven approach [164]. In [165], a jammer detection method for massive MIMO systems is proposed, utilizing unused pilots during the training phase, assuming that the jammer lacks prior knowledge of the pilot patterns. The base station detects the presence of a jammer by analyzing the received signal on these unused pilots and employing a GLRT. Recent advancements have introduced new solutions, including NNs for jamming detection [166]. DL approaches, such as CNNs, have been employed in [167,168] to analyze spectrogram images for jamming detection, outperforming conventional feature-based methods. Recent advances are tailored to the characteristics of 5G networks [31–33, 169]. In [170], a low-overhead intermittent jamming detection scheme for IoT networks is proposed, leveraging anchor nodes along with signal strength and multipath profile features. Furthermore, federated learning has been investigated for

distributed jamming detection in flying ad-hoc networks [171]. However, all these solutions are agnostic of the network structures and are not suited for cell-free communications where synchronization is looser.
When users are mobile and channel conditions vary, modeling network behavior is crucial. *Dynamic graphs* offer a powerful representation for the evolving topology of wireless networks [172], where nodes correspond to APs and UEs, and edges represent communication links based on signal strength and interference levels. To process and analyze dynamic graphs data, *graph-neural-networks (GNNs)* provides a powerful framework. Inspired by CNNs, GNNs are designed to operate on graph structures, enabling tasks such as node classification, link prediction, and other graph-related learning problems [173].
In this paper, we propose a novel framework to model cell-free massive MIMO communication, exploiting dynamic graphs to capture the time variability of the communication scenario. Then, we present a novel approach for jamming detection, leveraging dynamic graphs and GNN-based architectures. Our approach identifies jamming attacks by learning latent representations of network states and monitoring deviations from expected patterns. We evaluate the proposed method using simulations that model mobility, connectivity, and interference scenarios, demonstrating its effectiveness.
The rest of this paper is organized as follows. Section II presents the cell-free MIMO system model. Section III presents the GNN-based jamming detection framework. Section IV evaluates detection performance through simulations. Finally, Section V draws the conclusions.

I.2 System Model

We consider a cell-free massive MIMO network [174] with M APs and M UEs, focusing on the downlink transmission. Each UE is equipped with a single antenna, and each AP is equipped with N_{A} antennas. APs are static, while UEs are moving. We adopt a discrete-time model with sampling interval T, considering the network state at time instants nT, with $n \in \mathbb{Z}$. Each AP is associated with a single UE, and uses maximal ratio precoding to transmit data to its served UE, we may have more UEs than APs, but still at any given time only one UE is connected to each AP. Moreover, we account for the presence of a jammer that aims at corrupting the communication between APs and UEs. Each AP is transmitting with unitary power to each UE.

Channel Model Let $\boldsymbol{h}(k,m,n)$ denote the $N_{\mathrm{A}} \times 1$ vector of the narrowband baseband equivalent channel between the k-th UE and m-th AP at time nT. We consider a Rician fading channel; thus, the channel vector is modeled as

$$\boldsymbol{h}_{k,m}(n) = \beta\sigma_{k,m}(n) + \sqrt{1-\beta^2}\boldsymbol{g}_{k,m}(n), \tag{I.1}$$

with $\beta = \sqrt{\frac{K}{K+1}}$ a constant (and K is the Ricean K-factor) and $\boldsymbol{g}_{k,m}(n)$ being a $N_{\mathrm{A}} \times 1$ random matrix having i.i.d. zero-mean complex Gaussian entries. The variance of each entry of $\boldsymbol{g}_{k,m}(n)$ is determined by the path-loss model, which characterizes the received signal power as a function of the distance $d_{k,m}(n)$ between the k-th UE and the m-th AP at time nT, i.e.,

$$\sigma^2_{k,m}(n) = \frac{d_0^2}{d^2_{k,m}(n)}, \tag{I.2}$$

with $d_0 = 100\,\mathrm{m}$ representing the distance at which the channel has unitary variance. With $\beta = 1$ we obtain a deterministic model, while varying $\beta \in [0,1]$ we configure the randomness of the fading channel. We assume that reception is affected by additive white Gaussian noise (AWGN) with variance σ^2 per antenna.

Signal-to-noise-plus-interference Ratio The transmitter applies maximal ratio (MR) precoding to steer the transmitted signal towards the intended user, and, in the absence of jamming, a connection is established from the AP m to the UE k at time nT if the signal-to-interference-plus-noise ratio (SINR)

$$\Gamma_{k,m}(n) = \frac{||\boldsymbol{h}_{k,m}(n)||^4}{\sigma^2 + \sum_{m' \neq m} |\boldsymbol{h}_{k,m}^H(n)\boldsymbol{h}_{k,m'}(n)|^2}, \tag{I.3}$$

is above a threshols Γ_0, i.e.,

$$\Gamma_{k,m}(n) > \Gamma_0. \tag{I.4}$$

Note that the formula includes the interference from other APs.

Mobility Model We consider a system with UEs and APs ditributed within a square area of edge length L. The coordinates of each AP, indexed by m, are positioned at fixed locations that cover the area. At $n = 0$, the UEs are uniformly distributed within the square $[0, L]$. The coordinates of the position of user k at time nT are

$$\begin{aligned} x_k(n+1) &= x_k(n) + (v_{x,k} + w_{x,k}(n+1))T, \\ y_k(n+1) &= y_k(n) + (v_{y,k} + w_{y,k}(n+1))T, \end{aligned} \tag{I.5}$$

where $v_{x,k}$ and $v_{y,k}$ are the reference velocities of user k, uniformly distributed in the interval $[0, v_{\max}]$. The terms $w_{x,k}(n+1)$ and $w_{y,k}(n+1)$ are zero-mean Gaussian components with variance σ_w^2. If a user reaches the boundary of the square, its position is reset to a new location, uniformly sampled within the square, and assigned a new reference velocity. We assume that each user maintains a minimum distance $d_{\min}$ from any AP.

I.2.1 User Assignment Rule

We adopt the following rule for the assignment of UE to its serving AP. We proceed iteratively. We start with the full list of APs and UEs, and select the UE k and AP m that have the minimum distance among all pairs in the list. We assign UE k to AP m, then we remove a couple of devices from the list. The next iteration identifies the next AP-UE couple among the non-assigned APs and UEs.
Note that this procedure generates the assignment between APs and UEs, while an effective communication link (connection) between each couple is obtained only if condition (I.4) is satisfied.

I.2.2 Jammer Behavior

We consider the presence of a jammer that intermittently affects the communication between UEs and APs. Time is divided into F frames, each of duration T_{F}. Within each frame, the jammer remains active for a duration $\tau \in [0, T_{\mathrm{F}}]$. The jammer is equipped with a single antenna since its target is to disrupt any communication around it. When the jammer is *active*, the resulting SINR for a transmission from AP m to UE k at time nT becomes

$$\Gamma_{k,m}(n) = \frac{||\boldsymbol{h}_{k,m}(n)||^4}{\sigma^2 + P_J + \sum_{m' \neq m} |\boldsymbol{h}_{k,m}^H(n)\boldsymbol{h}_{k,m'}(n)|^2}, \tag{I.6}$$

where σ_{J}^2 is the jammer transmit power, $P_J = \sigma_{\mathrm{J}}^2 |S_k(n)|^2$, and $S_k(n)$ is the complex scalar channel from the jammer to UE k at time nT, according to the Rician model (I.1).

I.3 Jamming Detection By Dynamic Graph

We model the cell-free massive MIMO network as a dynamic connection graph $\{G(n)\}$, where $G(n)$ is the connection graph at time nT and T is the sampling time of the graph representation. In particular, each graph $G(n)$ has $N = 2M$ nodes (in the set $V(n)$), corresponding to both the APs and the UEs. The edges (collected in the set $E(n)$) represent the connections between APs and UEs. Specifically, an edge exists between UE k and AP m when (I.4) is satisfied. Each edge from AP m to UE k is labeled with the vector $\boldsymbol{w}_{k,m}(n) = [\alpha d_{k,m}, \zeta\gamma_{k,m}]$, where α and ζ are normalization factors that ensure proper scaling between distance and SINR values. The edge weights encode key connectivity metrics:

- *connection distance* $d_{k,m}(n)$, which defines the physical distance between an AP and a UE,
- *link quality* $\Gamma_{k,m}(n)$, quantified by the SINR, captures the reliability and performance of the communication link.

I.3.1 Jamming Detection Technique

Graph neural networks (GNNs) are neural models that capture the dependence of graphs via message passing between the nodes of graphs. In recent years, variants of GNNs such as graph convolutional network (GCN), graph attention network (GAT), and graph recurrent network (GRN) have demonstrated ground-breaking performances on many deep learning tasks [175]. We propose a novel jamming detection framework to identify jamming attacks in wireless networks, based on the dynamic graph representation. The architecture leverages the dynamic graph $\{G(n)\}$, graph convolution, and attention mechanisms to capture the distinctive patterns of connectivity disruptions caused by signal jammers.
The proposed jamming detection system consists of:

1. **Feature Extraction**, Each static graph $G(n)$ is constructed from the network topology and connectivity data between nodes.
2. **Spatial processing module (GCN layer):** Utilizes two stacked Gated Graph Convolutional layers to process each network snapshot independently and extract meaningful node-level representations (embeddings).
3. **Temporal processing module (Transformer layer):** Applies a multi-head self-attention mechanism across a sequence of graphs to detect temporal patterns that are indicative of jamming.
4. **Classification module:** Outputs a binary decision indicating whether the input sequence contains a jamming attack.

Fig. I.1 illustrates the overall architecture. The system processes sequences of K network graphs $\mathcal{G}(t) = \{G(t), G(t+1), \ldots, G(t+K-1)\}$, where each sequence represents a specific network condition over time, to provide a binary decision on whether jamming activity is present within the sequence.
The system processes a sequence of N_{steps} consecutive network graphs:

$$\mathcal{G}(n) = \{G(n), G(n+1), \ldots, G(n+N_{\text{steps}}-1)\},$$

where each $G(n)$ represents the state of the wireless network at time nT.

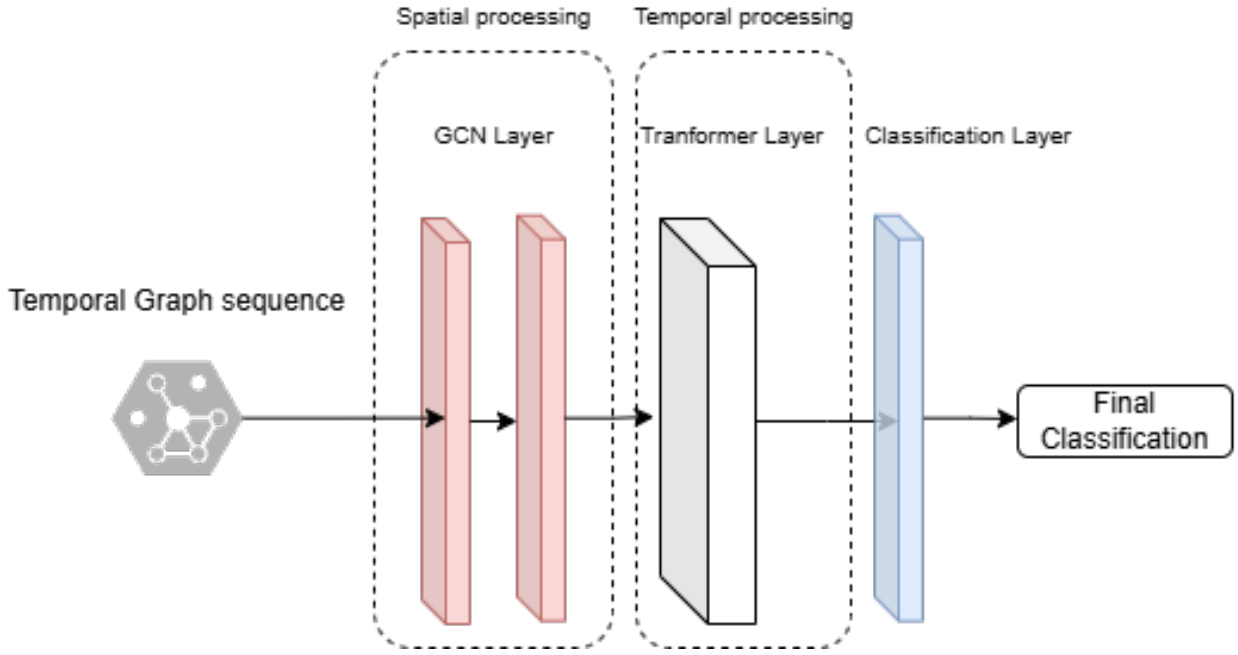


Figure I.1: Architecture of the proposed jamming detection model.

Feature Extraction and Graph Construction

Each static graph $G(n)$ is constructed from the real-time network topology and connectivity data. From each graph, we extract the following features:

- **Node-level features:**
  - *Degree centrality* $d_v(n)$: the number of connections of node v at time n;
  - *Node type* $\tau_v \in \{0, 1\}$: where 0 denotes Access Points (APs) and 1 denotes User Equipments (UEs);
  - *Position coordinates* $(x_v(n), y_v(n))$: the physical location of node v in 2D space.
- **Edge-level features:**
  - *SINR* $\Gamma_{u,v}(n)$: the signal-to-interference-plus-noise ratio between nodes u and v;
  - *Distance* $d_{u,v}(n)$: Euclidean distance between nodes u and v, computed as:

$$d_{u,v}(n) = \sqrt{(x_u(n) - x_v(n))^2 + (y_u(n) - y_v(n))^2}.$$

These features are extracted from the dynamic graph object, which stores node types, positions, and connection weights between APs and UEs. After conducting ablation experiments by selectively removing features and measuring the resulting performance, we found the above features to be the most critical for detecting jamming events.
After experimenting with removing features and measuring performance degradation, the above-mentioned features are the most impactful for the jamming detection process.

Spatial Processing Module The extracted node and edge features are fed into a Graph Neural Network to compute node embeddings. These embeddings encode both the local structure (who a node is connected to) and attributes (such as position and type). Specifically, for each node v at time n, we compute:

$$h_v(n) = \text{GNN}(G(n), \boldsymbol{\xi}_v(n)),$$

where $\boldsymbol{\xi}_v(n)$ is the feature vector of node v. The GCN aggregates information from neighboring nodes and edges, enabling each node to "learn" a summary of its local neighborhood and behavior.

Temporal Attention and Jamming Classification

The sequence of node embeddings from each graph is passed to a Transformer layer. This layer uses temporal self-attention to identify patterns across time, specifically, it can emphasize graphs that exhibit abnormal behavior (such as sudden drops in SINR or rapid topology changes) and downweight normal periods. This is essential because jamming effects may not be constant but occur intermittently across the sequence.

Classification Module The final detection is performed by a single linear layer that classifies the aggregated representation. The Transformer outputs a temporal representation $T_o(n)$, which is passed through a fully connected classification layer. The final output is the probability of jamming at the sequence level:

$$p(n) = \text{Softmax}(\text{LayerNorm}(W \cdot T_o(n))). \tag{I.7}$$

where LayerNorm denotes layer normalization, and $\boldsymbol{W}$ is the weight matrix of the classification layer. The decision is based on whether the probability of the *jammer* class exceeds a fixed threshold. This design allows the model to integrate spatial and temporal information effectively, improving robustness and interpretability in jamming detection

I.3.2 Model Training

The model is trained using the cross-entropy loss in a supervised manner using labeled datasets containing examples of nominal and jamming scenarios. During training, sequences of graph snapshots are presented to the model along with binary labels indicating the presence or absence of jamming activity. This supervised approach enables the model to learn discriminative patterns that distinguish normal network fluctuations from intentional jamming interference. The weights are optimized using the Adam optimizer, implementing early stopping when validation performance plateaus.

I.4 Numerical Results

I.4.1 Dataset Generation

To evaluate the proposed jamming detection approach, we generate a dataset of dynamic network graphs simulating wireless communications with and without jamming interference.
We consider a $L \times L$ area with $L = 1$ km, containing 5 fixed APs and 10 mobile UE nodes. The fixed APs are positioned at strategic locations covering the area: four at the corners, with coordinates $(0.2, 0.2)$, $(0.8, 0.2)$, $(0.2, 0.8)$, and $(0.8, 0.8)$, and one at the center $(0.5, 0.5)$ (all in km unit). Mobile UEs move according to a controlled random walk model with velocity components drawn from a uniform distribution in $[-v_{\max}, v_{\max}]$, where $v_{\max} = 6$ km/h. We consider $T = 1$ s and $T_F = 10$ s. Connectivity between an AP and UE is established when the SINR exceeds the threshold $\Gamma_0 = 5$ dB. The noise power is $\sigma^2 = 0.001$.

The jammer affects UEs within 0.35 km radius, and it is located in a different random position for each simulation. The number of network static graphs per sequence $\mathcal{G}(t)$ is $N = 80$.
We analyze two distinct scenarios. In the *deterministic scenario*, we set $\beta = 1$, resulting in a fixed channel matrix $\boldsymbol{h}_{k,m}(n)$. In the *fading scenario*, we set $\beta = 0$, such that $\boldsymbol{h}_{k,m}(n)$ models a Rayleigh fading channel.

I.4.2 GNN Implementation

The architecture was implemented using PyTorch and PyTorch Geometric. We used a GCN layer for each snapshot of the dynamic graph that consists of 2 Gated Graph convolution layers with 64 hidden units. The Transformer encoder consists of 4 encoder layers, each with 16 attention heads, 64 hidden units, and a feed-forward dimension of 128. We use GELU activation in the feed-forward networks and apply layer normalization with batch-first processing. Since graph sequences have inherent temporal ordering, we add learned positional encodings to capture temporal relationships. A single linear layer with an intermediate dimension of 32 is used for binary classification. The model was trained for 30 epochs using the Adam optimizer with a learning rate of 1.2×10^{-4}, weight decay of 10^{-6}, and batch size of 8. We applied a dropout of 0.03 in the Transformer layers and 0.05 overall to prevent overfitting. The dataset has 2200 dynamic graphs for each scenario, training was performed on 70% of the dataset, while 10% of the dataset was used for validation and 20% for testing.

I.4.3 Performance Metrics

Let TP be the number of True Positives, TN be the number of True Negatives, FP be the number of False Positives, and FN be the number of False Negatives. The accuracy is

$$a = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{I.8}$$

F1 score is

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{I.9}$$

I.4.4 Simulation Results

This section presents a comprehensive experimental evaluation of our dynamic graph-based jammer detection system under two primary training scenarios: (1) mixed-τ training using data from all jammer persistence patterns $\tau \in \{1, 2, ..., 10\}$, and (2) $\tau = 10$ specialist training using only continuous jammer scenarios. The parameter τ represents the jammer activation frequency within each temporal sequence, where $\tau = 1$ indicates sporadic jamming (active for only 1 out of 10 timesteps), $\tau = 5$ represents moderate persistence (active for 5 out of 10 timesteps), and $\tau = 10$ denotes continuous jamming (active throughout the entire sequence). All experiments were conducted with 80-timestep sequences on cell-free MIMO networks, evaluating performance under both fading and non-fading channel conditions.

I.4.5 $\tau = 10$ Training Analysis

The $\tau = 10$ results under non-fading conditions, shown in Fig. I.2, achieved accuracy consistently above 99% across $\tau = 1-9$, and F1-scores reaching 99.8% at $\tau = 3$. However, a notable performance degradation occurs at $\tau = 10$, where accuracy drops to 97.1% and F1-score to 97.4%. This indicates that training exclusively on continuous jammer scenarios, counterintuitively, provides excellent generalization to sporadic and rhythmic jamming patterns under non-fading channels.
In contrast, the fading scenario, shown in Fig. I.3, reveals the specialist's true generalization limitations and more pronounced performance variations. While maintaining strong overall performance (accuracy range:

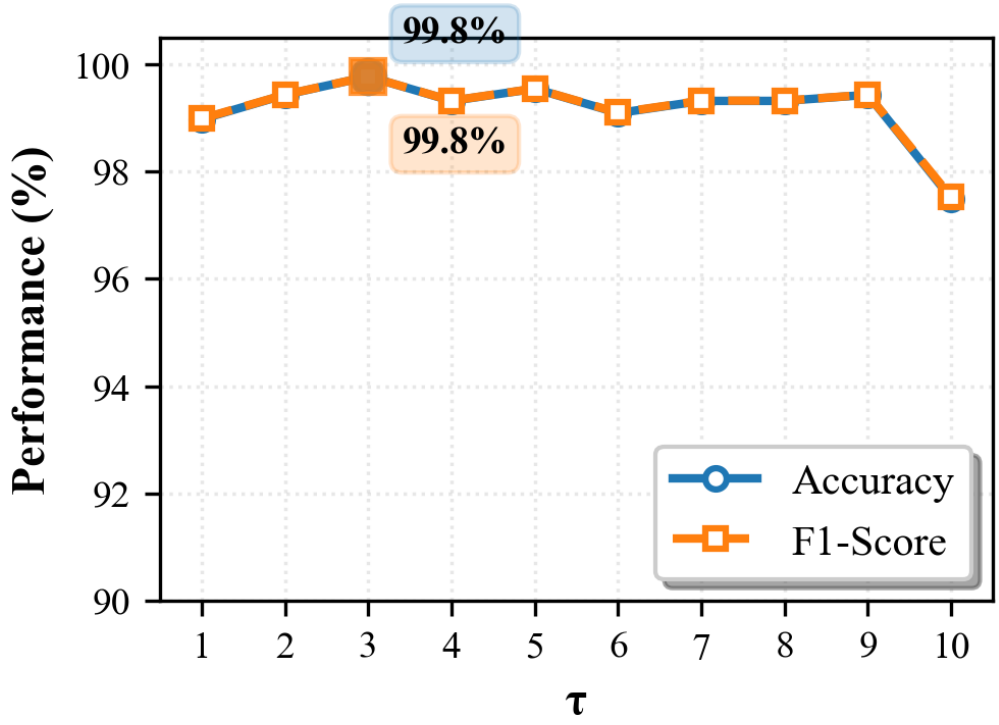


Figure I.2: Accuracy and F1 score vs τ, for the deterministic scenario. Training performed with a dataset having $\tau = 10$.

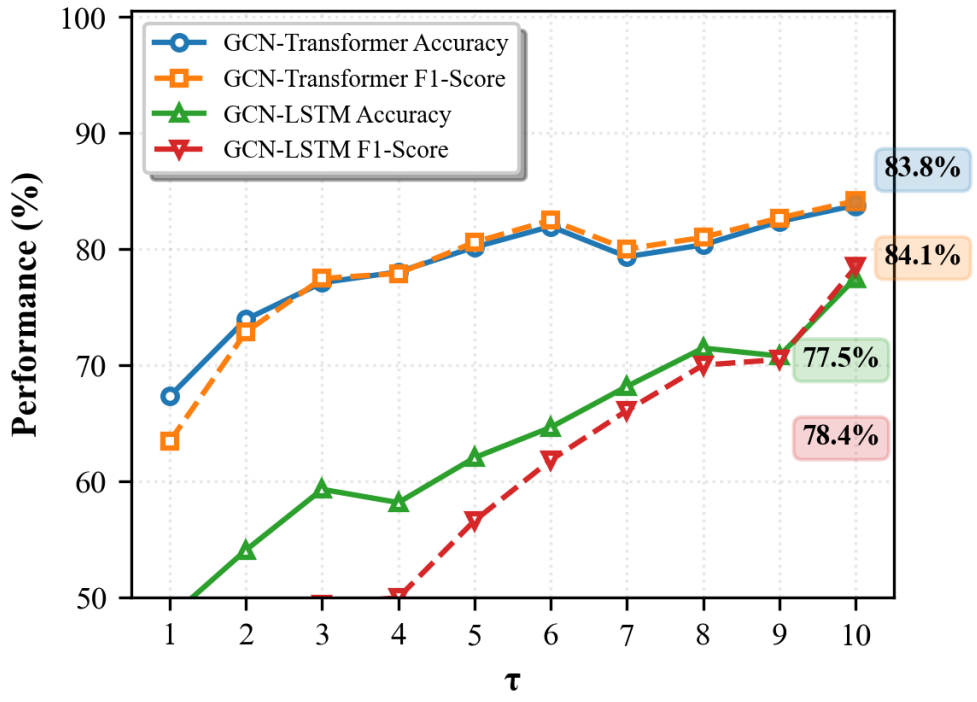


Figure I.3: Accuracy and F1 score vs τ, for the fading scenario. Training performed with a dataset having $\tau = 10$.

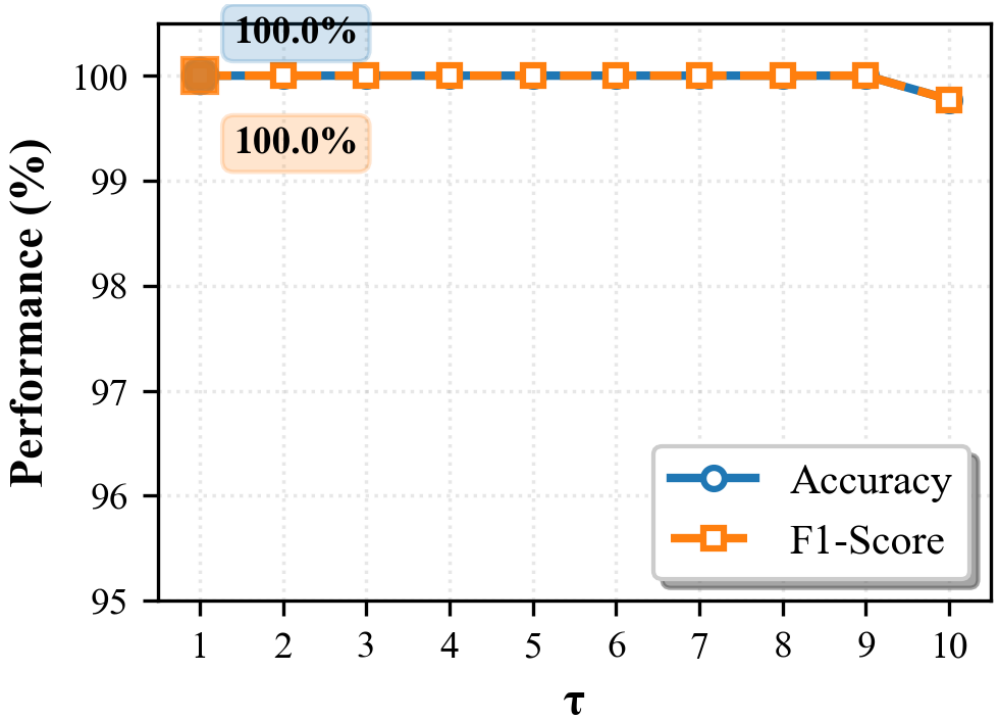


Figure I.4: Accuracy and F1 score vs τ, for the deterministic scenario. Training performed with a dataset having a mixture of attacks with different values of τ.

67.2%-83.8%), the model shows increased sensitivity to jammer persistence patterns, however, a comparison has been done on the same dataset using the known Long Short Term Memory GCN (GCN-LSTM) [34] which combines the capabilities of LSTMs to extract temporal dependencies with the feature learning power of the GCN, and as the figure shows, our model performed better in all projected jamming behaviours. The performance progression from $\tau = 1$ (67.2% accuracy) to $\tau = 8$ (83.8% accuracy) demonstrates the model's adaptation to different temporal structures, with optimal detection occurring in the rhythmic jamming domain ($\tau = 6 - 8$).

I.4.6 Mixed-τ Training Performance

Fig. I.4 shows the performance of our mixed-τ training approach under non-fading channel conditions. The model exhibit 100% accuracy across $\tau = 1 - 9$, with minimal degradation to 99.7% at $\tau = 10$.
In the fading scenario, presented in Fig. I.5, the obtained accuracy ranges from 75.6% at $\tau = 1$ to 89.7% at $\tau = 8$, before decreasing to 79.4% at $\tau = 10$. The monotonic improvement from $\tau = 1$ to $\tau = 8$ (73.2% to 89.5% F1-score) suggests that the model learns increasingly effective detection strategies as jammer persistence increases, until reaching the domain boundary at $\tau = 9 - 10$.

I.4.7 Channel Fading Effects on Detection Performance

Comparing non-fading versus fading scenarios reveals significant differences in detection robustness. Under non-fading conditions, both training strategies achieve near-perfect performance across most τ values, suggesting that the absence of channel fading provides cleaner signal characteristics that enhance jammer detection reliability. The stable channel conditions appear to preserve jamming signatures without additional noise from natural channel variations.

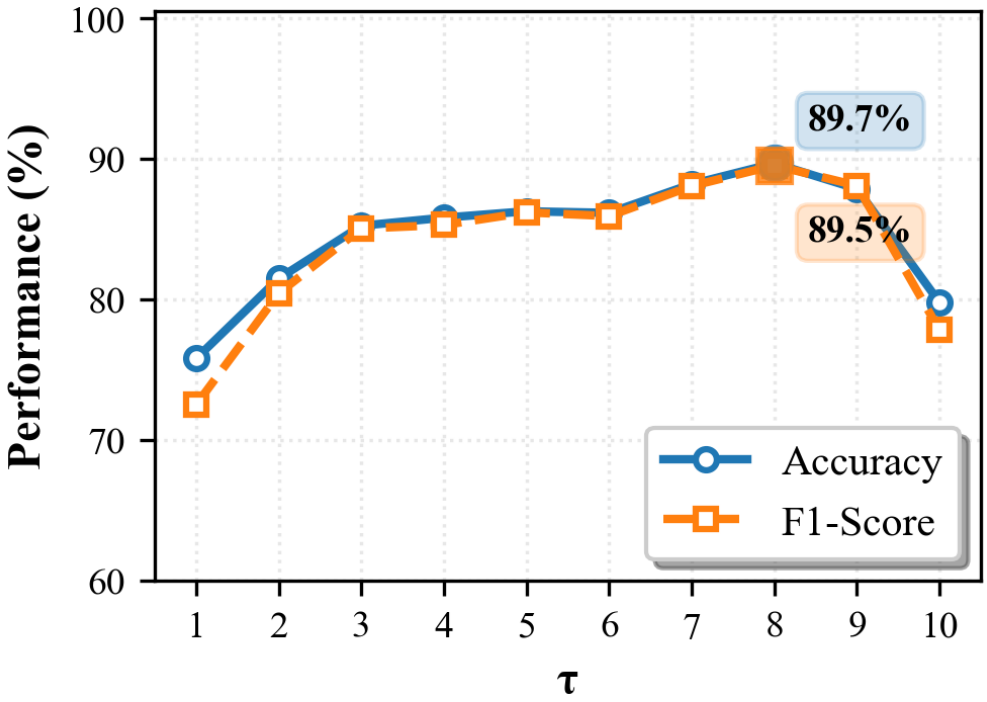


Figure I.5: Accuracy and F1 score vs τ, for the random fading scenario. Training performed with a dataset having a mixture of attacks with different values of τ.

Conversely, fading scenarios present more challenging detection environments, with performance variations of 10-15 percentage points across different τ values. This increased difficulty under fading channels indicates that channel-induced signal variations may mask jamming signatures, requiring more sophisticated detection algorithms to distinguish between fading-induced and jammer-induced signal degradations.

I.4.8 Training Strategy Effectiveness Comparison

The mixed-τ training approach demonstrates improved generalization capabilities and overall performance compared to the $\tau = 10$ specialist across both channel conditions. Under non-fading conditions, mixed-τ training achieves near-perfect performance (more than 99% accuracy) across the entire τ spectrum, while under fading conditions, it maintains reasonable performance levels (76-90% range) with more graceful degradation patterns. In contrast, the $\tau = 10$ specialist, despite showing perfect performance under non-fading conditions, exhibits significant generalization limitations under fading scenarios, with performance dropping as low as 67% at $\tau = 1$.
The mixed-τ approach's enhanced robustness across different channel conditions and jammer persistence patterns indicates that exposure to diverse jamming behaviors during training provides more generalizable feature representations. This finding supports the hypothesis that multi-domain training strategies are essential for robust jammer detection in dynamic wireless environments.

I.4.9 Baseline Shift Problem in Persistent Jamming

The performance degradation observed at $\tau = 10$ across all experimental configurations can be attributed to the fundamental baseline shift problem in persistent jamming scenarios. When jammers operate continuously, cell-free MIMO networks undergo adaptive responses. These network adaptations effectively establish a new operational baseline where continuous interference becomes the "normal" state.

I.5 Conclusions

This paper presented a comprehensive analysis of jammer detection in cell-free MIMO networks using dynamic graphs and specific graph neural network architecture, revealing insights into the effect of channel fading in the jamming detection process, and the multi-domain nature of temporal anomaly detection, in addition to this, our experimental evaluation across different jammer patterns ($\tau \in \{1, 2, ..., 10\}$) demonstrated that mixed-τ training achieves enhanced generalization compared to specialist approaches, with performance exceeding 99% under non-fading conditions and maintaining robustness above 75.6% even in challenging fading scenarios, higher than existing known models. The comparative analysis between fading and non-fading channels revealed that stable channel conditions significantly enhance detection reliability, while channel fading introduces additional complexity that degrades performance by 10-15 percentage points across all τ values.

A nice finding of this work is the identification of the baseline shift problem in persistent jamming scenarios ($\tau = 9-10$), where continuous jammer presence causes network adaptation responses that establish a new operational baseline, making traditional anomaly detection approaches ineffective. This phenomenon explains the characteristic performance degradation observed at high τ values across all experimental configurations, highlighting the need for detection strategies that can identify adaptation artifacts rather than direct interference signatures. The delineation of three distinct detection domains, namely sporadic ($\tau = 1 - 3$), rhythmic ($\tau = 4 - 8$), and persistent ($\tau = 9 - 10$), provides a theoretical framework for developing domain-specific architectures that address the unique challenges of each jammer behavior pattern.

Bibliography

[1] T. M. Pham, L. Senigagliesi, M. Baldi, R. F. Schaefer, G. P. Fettweis, and A. Chorti, "Leveraging angle of arrival estimation against impersonation attacks in physical layer authentication," 2025. [Online]. Available: https://arxiv.org/abs/2503.11508

[2] S. Skaperas and A. Chorti, "Misspecified Cramer–Rao bound for AoA estimation at a ULA under a spoofing attack," *in arxiv:2512.16735 and under review in IEEE Wireless Communications Letters*.

[3] B. Trinh-Nguyen, S. Berri, S. G. Teo, T. Truong-Huu, and A. Chorti, "High-accuracy AoA-based localization using hierarchical ML classifiers in outdoor environments," in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Taipei, TW, Dec. 2025.

[4] L. Senigagliesi, A. V. Guglielmi, M. Baldi, and S. Tomasin, "Security analysis of RIS-assisted physical-layer authentication over multipath channels," in *Proc. IEEE 17th IEEE International Workshop on Information Forensics and Security (WIFS)*, Perth, Australia, Dec. 2025.

[5] A. Mayya, A. Chorti, R. F. Schaefer, and G. P. Fettweis, "Secret key generation rates for line of sight multipath channels in the presence of eavesdroppers," in *Proc. 27th International Workshop on Smart Antennas (WSA)*. IEEE, 2024.

[6] A. Mayya, L. Senigagliesi, and A. Chorti, "Theoretical and practical analysis of secret key rates based on design parameters and channel characteristics," *under review in IEEE IoT Journal*, 2025.

[7] A. Mayya, Y. Richhariya, A. K. Boroujeni, S. Vorberg, M. Matthé, R. Vinz, L. Senigagliesi, K. Klamka, and A. Chorti, "Context-aware secret key generation demonstrator based on physical layer security," in *Proc. 2025 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2025.

[8] A. K. Angélo Passah, R. C. De Lamare, and A. Chorti, "Physical layer authentication using information reconciliation," in *Proc. 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, 2024, pp. 1–5.

[9] A. Kokuvi Angélo Passah, A. Chorti, and R. C. de Lamare, "Enhanced multiuser CSI-based physical layer authentication based on information reconciliation," *IEEE Wireless Communications Letters*, vol. 14, no. 2, pp. 544–548, 2025.

[10] A. K. Angélo Passah, R. C. De Lamare, and A. Chorti, "Adaptive CSI preprocessing for physical layer authentication," in *Proc. 2026 IEEE Wireless Communications and Networking Conference (WCNC)*, under review.

[11] A. Kokuvi Angélo Passah, A. Chorti, and R. C. de Lamare, "Channel state information preprocessing for CSI-based physical-layer authentication using reconciliation," *IEEE Transactions on Communications*, in arxiv 2512.16719 and under review in IEEE Trans. Communications.

[12] S. Gil, M. Yemini, A. Chorti, A. Nedić, H. V. Poor, and A. J. Goldsmith, "How physicality enables trust: A new era of trust-centered cyberphysical systems," *in arxiv:2311.07492 and under review in the Proceedings of the IEEE*, 2024.

[13] B. Trinh-Nguyen, S. Berri, S. G. Teo, T. Truong-Huu, and A. Chorti, "A framework for global trust and reputation management in 6G networks: Position paper," in *Machine Learning for Networking: 7th International Conference, MLN 2024, ACM Digital Library*, 2024.

[14] M. Delamou, L. Chen, E. M. Amhoud, and A. Chorti, "Enhanced physical layer authentication via robust and trustworthy sensing," in *in techriv, 10.36227/techrxiv.176403878.80301053/v1, and under review in IEEE ICC 2026*.

[15] D. Wang, L. Chen, and F. Nait-Abdesselam, "SA-SWOMP: Radar-assisted sparse channel estimation for joint sensing and communication," in *Proceedings of IEEE International Coneference on Communications, ICC 2025*, 2025.

[16] R. Khanzadeh, S. B. Fjolla Ademaj-Berisha and, L. Senigagliesi, A. Chorti, A. Springer, and H.-P. Bernhard, "Trustworthiness-aware resource allocation in network slicing via hierarchical reinforcement learning," in *under review in IEEE ICC 2026*.

[17] J. A. Zhang, M. L. Rahman, K. Wu, X. Huang, Y. J. Guo, S. Chen, and J. Yuan, "Enabling joint communication and radar sensing in mobile networks—a survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 306–345, 2021.

[18] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu, Y. Shen, F. Colone, and K. Chetty, "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[19] M. M. Şahin and H. Arslan, "Multi-functional coexistence of radar-sensing and communication waveforms," in *Proc. IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 2616–2630.

[20] A. Tabeshnezhad, A. L. Swindlehurst, and T. Svensson, "Ris-assisted interference mitigation for uplink noma," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 24, no. 2, 2023.

[21] M. Srinivasan, L. Senigagliesi, H. Chen, A. Chorti, M. Baldi, and H. Wymeersch, "AoA-based physical layer authentication in analog arrays under impersonation attacks," in *2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024, pp. 496–500.

[22] L. Sun and Q. Du, "A review of physical layer security techniques for internet of things: Challenges and solutions," *Entropy*, vol. 20, no. 10, p. 730, 2018.

[23] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. D'Oro, T. Melodia, S. Ioannidis, and K. Chowdhury, "No radio left behind: Radio fingerprinting through deep learning of physical-layer hardware impairments," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 165–178, 2020.

[24] H. Li, K. Gupta, C. Wang, N. Ghose, and B. Wang, "Radionet: Robust deep-learning based radio fingerprinting," in *2022 IEEE Conference on Communications and Network Security (CNS)*, 2022, pp. 190–198.

[25] G. Baldini, R. Giuliani, C. Gentile, and G. Steri, "Measures to address the lack of portability of the RF fingerprints for radiometric identification," in *2018 IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2018, pp. 1–5.

[26] A. Elmaghbub and B. Hamdaoui, "Lora device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability," *IEEE Access*, vol. 9, pp. 142 893–142 909, 2021.

[27] A. Mohammadian and C. Tellambura, "Rf impairments in wireless transceivers: Phase noise, cfo, and iq imbalance – a survey," *IEEE Access*, vol. 9, pp. 111 718–111 791, 2021.

[28] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7167–7176.

[29] State of IoT 2024: Number of connected IoT devices growing 13% to 18.8 billion globally. Accessed on 7 Dec., 2024. [Online]. Available: https://iot-analytics.com/number-connected-iot-devices

[30] J. Zhang, F. Ardizzon, M. Piana, G. Shen, and S. Tomasin, "Physical layer-based device fingerprinting for wireless security: From theory to practice," *IEEE Trans. on Inf. Forensics and Secur.*, May 2025.

[31] M. Varotto, S. Valentin, and S. Tomasin, "Detecting 5G signal jammers using spectrograms with supervised and unsupervised learning," in *Proc. IEEE Int. Conf. on Commun. Work. (ICC Work.)*, 2024, pp. 767–772.

[32] ——, "Detecting 5G signal jammers with autoencoders based on loose observations," in *Proc. IEEE Global Telecommun. Conf. Workshops (GLOBECOM WS)*, Dec. 2023, accepted for publication.

[33] M. Varotto, S. Valentin, F. Ardizzon, S. Marzotto, and S. Tomasin, "One-class classification as glrt for jamming detection in private 5G networks," in *Proc. 2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024, pp. 201–205.

[34] L. García-Duarte, J. Cifuentes, and G. Marulanda, "Short-term spatio-temporal forecasting of air temperatures using deep graph convolutional neural networks," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 5, p. 1649–1667, Dec 2022.

[35] R. Fuladi and B. Cicek, "Image-based frequency-domain analysis for robust DDoS detection in SDN," in *Proc. SecSoft 2025 - 7th International Workshop on Cyber-Security in Software-defined and Virtualized Infrastructures*, June 2025.

[36] M. Akbulut, B. Çiçek, and R. Fuladi, "Radio frequency fingerprint-based classification performance analysis with ML models in the presence of hardware impairments," in *Proc. 2025 33rd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2025, pp. 1–4.

[37] E. Dushku, M. M. Rabbani, J. Vliegen, A. Braeken, and N. Mentens, "Prove: Provable remote attestation for public verifiability," *Journal of Information Security and Applications*, vol. 75, p. 103448, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212623000327

[38] M. M. Rabbani, J. Vliegen, J. Winderickx, M. Conti, and N. Mentens, "Shela: Scalable heterogeneous layered attestation," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 240–10 250, 2019.

[39] E. Dushku, M. M. Rabbani, M. Conti, L. V. Mancini, and S. Ranise, "SARA: Secure Asynchronous Remote Attestation for IoT Systems," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3123–3136, 2020.

[40] M. Ambrosin, M. Conti, R. Lazzeretti, M. M. Rabbani, and S. Ranise, "Pads: Practical attestation for highly dynamic swarm topologies," in *Proc. 2018 International Workshop on Secure Internet of Things (SIoT)*, 2018, pp. 18–27.

[41] I. D. O. Nunes, S. Jakkamsetti, N. Rattanavipanon, and G. Tsudik, "On the toctou problem in remote attestation," 2021. [Online]. Available: https://arxiv.org/abs/2005.03873

[42] S. Pinto and N. Santos, "Demystifying arm trustzone: A comprehensive survey," *ACM Comput. Surv.*, vol. 51, no. 6, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3291047

[43] Trusted Computing Group, *Trusted Platform Module Library Specification, Family "2.0", Level 00, Revision 01.59*, November 2019. [Online]. Available: https://trustedcomputinggroup.org/resource/tpm-library-specification/

[44] K. Eldefrawy, G. Tsudik, A. Francillon, and D. Perito, "SMART: Secure and Minimal Architecture for (Establishing Dynamic) Root of Trust." in *Proc. of the 19th Annual Network & Distributed System Security Symposium NDSS '12*, 2012.

[45] P. Koeberl, S. Schulz, A.-R. Sadeghi, and V. Varadharajan, "TrustLite: A security architecture for tiny embedded devices," in *Proc. of the 9th European Conference on Computer Systems EuroSys '14*, 2014, pp. 1–14.

[46] F. Brasser, B. El Mahjoub, A.-R. Sadeghi, C. Wachsmann, and P. Koeberl, "Tytan: tiny trust anchor for tiny devices," in *Proc. of the 52nd Design Automation Conference*, ser. DAC '15, 2015, pp. 1–6.

[47] N. Asokan, F. Brasser, A. Ibrahim, A.-R. Sadeghi, M. Schunter, G. Tsudik, and C. Wachsmann, "SEDA: Scalable embedded device attestation," in *Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15, 2015, pp. 964–975.

[48] X. Carpent, K. ElDefrawy, N. Rattanavipanon, and G. Tsudik, "LIghtweight Swarm Attestation: a Tale of Two LISA-s," in *Proc. of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIACCS '17. ACM, 2017, pp. 86–100.

[49] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. the 35th International Conference on Machine Learning*, ser. Proc. Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 531–540. [Online]. Available: https://proceedings.mlr.press/v80/belghazi18a.html

[50] T. Matsumine, H. Ochiai, and J. Shikata, "A data-driven analysis of secret key rate for physical layer secret key generation from wireless channels," *IEEE Communications Letters*, vol. 27, no. 12, pp. 3166–3170, 2023.

[51] D. Guo, J. Xiong, D. Ma, X. Liu, and J. Wei, "Physical layer secret key generation based on mutual information-driven autoencoder," *IEEE Transactions on Wireless Communications*, vol. 24, no. 10, pp. 8042–8056, 2025.

[52] Y. Du, H. Liu, Z. Shao, Y. Ren, S. Li, H. Dai, and J. Yu, "Secure and controllable secret key generation through csi obfuscation matrix encapsulation," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 12 313–12 329, 2024.

[53] R. V. Steiner and E. Lupu, "Attestation in wireless sensor networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, pp. 1–31, 2016.

[54] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[55] A. Mayya, M. Mitev, A. Chorti, and G. Fettweis, "A skg security challenge: Indoor skg under an on-the-shoulder eavesdropping attack," in *Proc. GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 3451–3456.

[56] G. Cherubin, K. Chatzikokolakis, and C. Palamidessi, "F-bleau: Fast black-box leakage estimation," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019.

[57] M. Shakiba-Herfeh, L. Luzzi, and A. Chorti, "Finite blocklength secrecy analysis of polar and Reed-Muller codes in BEC semi-deterministic wiretap channels," in *Proc. 2021 IEEE Information Theory Workshop (ITW)*. IEEE, 2021, pp. 1–6.

[58] W. Yang, R. F. Schaefer, and H. V. Poor, "Wiretap channels: Nonasymptotic fundamental limits," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4069–4093, 2019.

[59] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, 2013.

[60] S. Tomasin, T. N. M. M. Elwakeel, A. V. Guglielmi, R. Maes, N. Noels, and M. Moeneclaey, "Analysis of challenge-response authentication with reconfigurable intelligent surfaces," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 9494–9507, 2024.

[61] L. Crosara, A. V. Guglielmi, N. Laurenti, and S. Tomasin, "Divergence-minimizing attack against challenge-response authentication with irss," in *Proc. 2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2024, pp. 1986–1991.

[62] A. Ferrante, N. Laurenti, C. Masiero, M. Pavon, and S. Tomasin, "On the error region for channel estimation-based physical layer authentication over Rayleigh fading," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 941–952, 2015.

[63] G. J. Simmons, "Authentication theory/coding theory," in *Advances in Cryptology*, G. R. Blakley and D. Chaum, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985, pp. 411–431.

[64] R. Diamant, P. Casari, and S. Tomasin, "Cooperative authentication in underwater acoustic sensor networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 954–968, Feb. 2019.

[65] S. Tomasin, H. Zhang, A. Chorti, and H. V. Poor, "Challenge-response physical layer authentication over partially controllable channels," *IEEE Communications Magazine*, vol. 60, no. 12, pp. 138–144, 2022.

[66] F. Mazzo, S. Tomasin, H. Zhang, A. Chorti, and H. V. Poor, "Physical-layer challenge-response authentication for drone networks," in *Proc. IEEE Global Communications Conf.*, 2023.

[67] A. V. Guglielmi, L. Crosara, S. Tomasin, and N. Laurenti, "Physical-layer challenge-response authentication with irs and single-antenna devices," in *Proc. 2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2024, pp. 560–565.

[68] F. Ardizzon, D. Salvaterra, M. Piana, and S. Tomasin, "Energy-based optimization of physical-layer challenge-response authentication with drones," in *Proc. 2024 IEEE Globecom Workshops (GC Wkshps)*, 2024, pp. 1–6.

[69] N. Benvenuto, G. Cherubini, and S. Tomasin, *Algorithms for Communications Systems and their Applications*, 2nd ed. Wiley, 2021.

[70] H. V. Abeywickrama, B. A. Jayawickrama, Y. He, and E. Dutkiewicz, "Comprehensive energy consumption model for unmanned aerial vehicles, based on empirical studies of battery performance," *IEEE Access*, vol. 6, pp. 58 383–58 394, Oct. 2018.

[71] M. Piana, F. Ardizzon, and S. Tomasin, "Challenge-response to authenticate drone communications: A game theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4890–4903, 2025.

[72] M. Piana and S. Tomasin, "Secret key generation on aerial rician fading channels against a curious receiver," in *Proc. 2025 IEEE 26th International Workshop on Signal Processing and Artificial Intelligence for Wireless Communications (SPAWC)*, 2025, pp. 1–5.

[73] B. Çiçek and H. Alakoca, "Impact of residual hardware impairments on ris-aided authentication," in *Proc. 2024 IEEE Virtual Conference on Communications (VCC)*. IEEE, 2024, pp. 1–6.

[74] M. Sharara and S. Berri, "Multi-strategy optimization approach for location privacy and latency trade-offs in 6G networks," in *Proc. IEEE Globecom Workshops : Workshop on Enabling Security, Trust, and Privacy in 6G Wireless Systems*. IEEE, 2025, pp. 1–6.

[75] S. Alhazbi, S. Sciancalepore, and G. Oligeri, "The day-after-tomorrow: On the performance of radio fingerprinting over time," in *Proc. 39th Annual Computer Security Applications Conference (ACSAC)*, 2023, pp. 439–450.

[76] Özkan Yılmaz and M. A. Yazıcı, "The effect of ambient temperature on device classification based on radio frequency fingerprint recognition," *Sakarya University Journal of Computer and Information Sciences*, vol. 5, no. 2, pp. 233–245, 2022.

[77] C. Ayyildiz, R. Cetin, Z. Khodzhaev, T. Kocak, E. G. Soyak, V. C. Gungor, and G. K. Kurt, "Physical layer authentication for extending battery life," *Ad Hoc Networks*, vol. 123, p. 102683, 2021.

[78] Q. Xu, R. Zheng, W. Saad, and Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 94–104, 2016.

[79] D. Liao, J. Luo, J. Vevstad, and N. Pappas, "RANGAN: Gan-empowered anomaly detection in 5G cloud ran," in *Proc. 2025 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2025, pp. 1–4.

[80] Ericsson. (2022, Jun.) 5g ran explained: What, how and where next. [Online]. Available: https://www.ericsson.com/en/ran

[81] C. Kim, V. B. Mendiratta, and M. Thottan, "Unsupervised anomaly detection and root cause analysis in mobile networks," in *Proc. 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 2020, pp. 176–183.

[82] Ericsson. (2021) Building sustainable networks. Ericsson Mobility Report article. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/building-sustainable-networks

[83] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138912860700062X

[84] Z. Chen, J. Liu, W. Gu, Y. Su, and M. R. Lyu, "Experience report: Deep learning-based system log analysis for anomaly detection," *arXiv preprint arXiv:2107.05908*, 2021.

[85] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," vol. 41, no. 3, Jul. 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[86] D. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[87] T. F. Lunt, "A survey of intrusion detection techniques," *Computers & Security*, vol. 12, no. 4, pp. 405–418, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167404893900295

[88] J. P. Rouillard, "Refereed papers: Real-time log file analysis using the simple event correlator (sec)," in *Proc. 18th USENIX Conference on System Administration*, ser. LISA '04. USA: USENIX Association, 2004, p. 133–150.

[89] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner, "Rule-based anomaly pattern detection for detecting disease outbreaks," in *Eighteenth national conference on Artificial intelligence*. USA: American Association for Artificial Intelligence, 2002, p. 217–223.

[90] C. Sun, U. Pawar, M. Khoja, X. Foukas, M. K. Marina, and B. Radunovic, "Spotlight: Accurate, explainable and efficient anomaly detection for open ran," in *Proc. 30th Annual International Conference on Mobile Computing and Networking*, ser. ACM MobiCom '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 923–937. [Online]. Available: https://doi.org/10.1145/3636534.3649380

[91] A. Chawla, P. Jacob, S. Feghhi, D. Rughwani, S. van der Meer, and S. Fallon, "Interpretable unsupervised anomaly detection for ran cell trace analysis," in *Proc. 2020 16th International Conference on Network and Service Management (CNSM)*, 2020, pp. 1–5.

[92] A. Hasan, C. Boeira, K. Papry, Y. Ju, Z. Zhu, and I. Haque, "Root cause analysis of anomalies in 5G ran using graph neural network and transformer," *arXiv preprint arXiv:2406.15638*, 2024.

[93] J. Luo and N. Pappas, "On the role of age and semantics of information in remote estimation of Markov sources," *arXiv preprint arXiv:2507.18514*, 2025.

[94] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.

[95] J. Luo and N. Pappas, "Semantic-aware remote estimation of multiple Markov sources under constraints," *IEEE Transactions on Communications*, vol. 73, no. 11, pp. 11 093–11 105, 2025.

[96] A. Maatouk, S. Kriouile, M. Assaad, and A. Ephremides, "The age of incorrect information: A new performance metric for status updates," *IEEE/ACM Transactions on Networking*, vol. 28, no. 5, pp. 2215–2228, 2020.

[97] J. Luo and N. Pappas, "Minimizing the age of missed and false alarms in remote estimation of Markov sources," in *Proc. Twenty-Fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, ser. MobiHoc '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 381–386. [Online]. Available: https://doi.org/10.1145/3641512.3690161

[98] ——, "On the cost of consecutive estimation error: Significance-aware non-linear aging," *IEEE Transactions on Information Theory*, vol. 71, no. 10, pp. 7976–7989, 2025.

[99] G. Stamatakis, N. Pappas, and A. Traganitis, "Control of status updates for energy harvesting devices that monitor processes with alarms," in *Proc. 2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[100] E. Delfani, G. J. Stamatakis, and N. Pappas, "State-aware timeliness in energy harvesting IoT systems monitoring a markovian source," *IEEE Transactions on Green Communications and Networking*, vol. 9, no. 3, pp. 977–990, 2025.

[101] A. Brighente, F. Formaggio, G. M. Di Nunzio, and S. Tomasin, "Machine learning for in-region location verification in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2490–2502, 2019.

[102] R. Liao, H. Wen, F. Pan, H. Song, A. Xu, and Y. Jiang, "A novel physical layer authentication method with convolutional neural network," in *Proc. 2019 IEEE Int. Conf. on Artif. Intell. and Comput. Appl. (ICAICA)*. IEEE, 2019, pp. 231–235.

[103] S. Wang, K. Huang, X. Xu, Z. Zhong, and Y. Zhou, "CSI-based physical layer authentication via deep learning," *IEEE Wireless Commun. Letters*, vol. 11, no. 8, pp. 1748–1752, Aug. 2022.

[104] T. M. Pham, L. Senigagliesi, M. Baldi, G. P. Fettweis, and A. Chorti, "Machine learning-based robust physical layer authentication using angle of arrival estimation," in *Proc. GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 13–18.

[105] A. V. Guglielmi and S. Tomasin, "Fast iterative configuration of reconfigurable intelligent surfaces in mmWave systems," *in Proc. 2023 IEEE Global Commun. Conf.*, pp. 631–636, Dec. 2023.

[106] M. M. Selim and S. Tomasin, "Physical layer authentication with simultaneous reflecting and sensing RIS," in *Proc. 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–5.

[107] P. Zhang, Y. Teng, Y. Shen, X. Jiang, and F. Xiao, "Tag-based PHY-layer authentication for RIS-assisted communication systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4778–4792, 2023.

[108] H. Liu, L. Li, X. Tang, W. Lin, F. Yang, T. Yin, and Z. Han, "Reconfigurable intelligent surface-aided physical layer authentication with deep learning," in *Proc. 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, 2024, pp. 1–6.

[109] K. F. Masood, J. Tong, J. Xi, J. Yuan, and Y. Yu, "Inductive matrix completion and root-MUSIC-based channel estimation for intelligent reflecting surface (IRS)-aided hybrid MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 7917–7931, 2023.

[110] A. M. Sayeed, "Optimization of reconfigurable intelligent surfaces through trace maximization," in *Proc. IEEE Inter. Conf. on Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[111] P. Baracca, N. Laurenti, and S. Tomasin, "Physical layer authentication over MIMO fading wiretap channels," *IEEE Trans. Wirel. Commun.*, vol. 11, no. 7, pp. 2564–2573, 2012.

[112] M. Bloch and J. Barros, *Physical-layer security: from information theory to security engineering*. Cambridge University Press, 2011.

[113] A. Alashqar, E. O. Torshizi, R. Mesleh, and W. Henkel, "Secret key generation driven by attention-based convolutional autoencoder and quantile quantization for IoT security in 5G and beyond," *IEEE Access*, vol. 13, pp. 131 744–131 756, July 2025.

[114] T. Lu, L. Chen, J. Zhang, C. Chen, T. Q. Duong, and M. Matthaiou, "Precoding design for key generation in extremely large-scale mimo near-field multi-user systems," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 10 572–10 587, 2025.

[115] A. Giuliani, F. Ardizzon, and S. Tomasin, "ML-based advantage distillation for key agreement in underwater acoustic channels," in *Proc. of Int. Conf. on Commun. Workshops (ICC Workshops)*, 2023, pp. 703–708.

[116] T. Lu, L. Chen, J. Zhang, C. Chen, T. Q. Duong, and M. Matthaiou, "Precoding design for key generation in near-field extremely large-scale MIMO communications," in *Proc. of 2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 172–177.

[117] R. Diamant, P. Casari, F. Ardizzon, S. Tomasin, B. Sherlock, T. Corner, and J. A. Neasham, "Channel-based key generation for secure underwater acoustic communications," *IEEE Trans. Inf. Forensics Security*, vol. 24, no. 7, pp. 5678–5693, July 2025.

[118] K. Pelekanakis, S. A. Yıldırım, G. Sklivanitis, R. Petroccia, J. Alves, and D. Pados, "Physical layer security against an informed eavesdropper in underwater acoustic channels: Feature extraction and quantization," in *Proc. Underwater Commun. and Netw. Conf. (UComms)*, 2021, pp. 1–5.

[119] J. Zhou and X. Zeng, "Physical layer secret key generation for spatially correlated channels based on multi-task autoencoder," in *Proc. Int. Conf. on Intell. Comput. and Signal Proc. (ICSP)*, 2022, pp. 144–150.

[120] K. Meng, C. Masouros, A. P. Petropulu, and L. Hanzo, "Cooperative isac networks: Opportunities and challenges," *IEEE Wireless Communications*, vol. 32, no. 3, pp. 212–219, 2025.

[121] N. Jabeen, H. Lei, A. Muhammad, A. Ali, Z. U. Khan, and G. Pan, "Localization in isac: A review," *IEEE Internet of Things Journal*, vol. 12, no. 22, pp. 46 526–46 552, 2025.

[122] K. Qu, J. Ye, X. Li, and S. Guo, "Privacy and security in ubiquitous integrated sensing and communication: Threats, challenges and future directions," *arXiv preprint arXiv:2308.00253*, 2023.

[123] G. Li, C. Wang, H. Zhang, L. Jin, and N. Al-Dhahir, "Securing isac against pilot spoofing attack: A deep plug-and-play countermeasure," *IEEE Wireless Communications Letters*, vol. 14, no. 9, pp. 2813–2817, 2025.

[124] J. Li, L. Lazos, and M. Li, "Securing ofdm-based isac systems against sensing attacks," in *Proc. 2025 IEEE Conference on Communications and Network Security (CNS)*, 2025, pp. 1–9.

[125] D. K. Pin Tan, J. He, Y. Li, A. Bayesteh, Y. Chen, P. Zhu, and W. Tong, "Integrated sensing and communication in 6g: Motivations, use cases, requirements, challenges and future directions," in *Proc. 2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S)*, 2021, pp. 1–6.

[126] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[127] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, Oct 2011.

[128] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[129] E. Arikan, "Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[130] E. Arikan and E. Telatar, "On the rate of channel polarization," in *IEEE International Symposium on Information Theory*, 2009, pp. 1493–1495.

[131] M. Soori, B. Arezoo, and R. Dastres, "Internet of things for smart factories in industry 4.0, a review," *Internet of Things and Cyber-Physical Syst.*, vol. 3, pp. 192–204, 2023.

[132] S. Grabowska, "Smart factories in the age of industry 4.0," *Manage. Syst. in Prod. Eng.*, no. 2 (28, pp. 90–96, 2020.

[133] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, "Iot: Internet of threats? a survey of practical security vulnerabilities in real IoT devices," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8182–8201, Oct. 2019.

[134] S. Dhar, A. Khare, A. D. Dwivedi, and R. Singh, "Securing IoT devices: A novel approach using blockchain and quantum cryptography," *Internet of things*, vol. 25, p. 101019, Apr. 2024.

[135] P. Angueira, I. Val, J. Montalbán, O. Seijo, E. Iradier, P. S. Fontaneda, L. Fanari, and A. Arriola, "A survey of physical layer techniques for secure wireless communications in industry," *IEEE Commun. Surv. and Tut.*, vol. 24, no. 2, pp. 810–838, 2022.

[136] Y. Lee, J. Yoon, J. Choi, and E. Hwang, "A novel cross-layer authentication protocol for the internet of things," *IEEE Access*, vol. 8, pp. 196 135–196 150, Oct. 2020.

[137] M. A. Shawky, M. Bottarelli, G. Epiphaniou, and P. Karadimas, "An efficient cross-layer authentication scheme for secure communication in vehicular ad-hoc networks," *IEEE Trans. on Veh. Technol.*, vol. 72, no. 7, pp. 8738–8754, Feb. 2023.

[138] P. Hao, X. Wang, and A. Refaey, "An enhanced cross-layer authentication mechanism for wireless communications based on per and rssi," in *Proc. 2013 13th Canadian Workshop on Inf. Theory*, 2013, pp. 44–48.

[139] X. Wang, J. S. Wong, F. Stanley, and S. Basu, "Cross-layer based anomaly detection in wireless mesh networks," in *Proc. 2009 Ninth Annu. Int. Symp. on Appl. and the Internet*. IEEE, 2009, pp. 9–15.

[140] M. Mohaghegh and V. Ngo, "Cross-layer authentication and physical layer authentication in internet-of-things: A systematic literature review," in *Proc. Int. Conf. on Information Networking (ICOIN)*, 2024, pp. 473–477.

[141] S. R. Pokhrel, "Learning from data streams for automation and orchestration of 6G industrial iot: toward a semantic communication framework," *Neural Comput. and Appl.*, vol. 34, no. 18, pp. 15 197–15 206, 2022.

[142] F. Euchner and M. Gauger, "CSI Dataset dichasus-cf1x-part1: Distributed Antenna Setup in Industrial Environment, Day 2, First Part," 2022. [Online]. Available: https://doi.org/doi:10.18419/darus-3150

[143] S. De Bast, A. P. Guevara, and S. Pollin, "CSI-based positioning in massive MIMO syst. using convolutional neural networks," in *Proc. 2020 IEEE 91st Veh. Technol. Conf. (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[144] G. Kia, L. Ruotsalainen, and J. Talvitie, "A CNN approach for 5G mm wave positioning using beamformed CSI measurements," in *Proc. 2022 Int. Conf. on Localization and GNSS (ICL-GNSS)*. IEEE, 2022, pp. 01–07.

[145] V. N. Vapnik, *The nature of statistical learning theory*, 1995.

[146] Y. Kim and H. Bang, "Introduction to kalman filter and its applications," in *Proc. Introduction and implementations of the Kalman filter*. IntechOpen, 2018.

[147] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[148] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[149] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[150] E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.

[151] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Efficient federated learning via decomposed similarity-based clustering," in *Proc. 2021 IEEE Intl Conf on parallel & distributed processing with applications, big data & cloud computing, sustainable computing & communications, social computing & networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 228–237.

[152] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *Proc. 2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[153] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," *arXiv preprint arXiv:1802.07876*, 2018.

[154] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in neural information processing systems*, vol. 33, pp. 3557–3568, 2020.

[155] H. Wang, L. Muñoz-González, D. Eklund, and S. Raza, "Non-iid data re-balancing at IoT edge with peer-to-peer federated learning for anomaly detection," in *Proc. of the 14th ACM conference on security and privacy in wireless and mobile networks*, 2021, pp. 153–163.

[156] S. Wang, N. Li, S. Xia, X. Tao, and H. Lu, "Collaborative physical layer authentication in internet of things based on federated learning," in *Proc. 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 714–719.

[157] S. Halder and T. Newe, "Radio fingerprinting for anomaly detection using federated learning in lora-enabled industrial internet of things," *Future generation computer systems*, vol. 143, pp. 322–336, 2023.

[158] N. Nagia, M. T. Rahman, and S. Valaee, "Federated learning for wifi fingerprinting," in *Proc. ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 4968–4973.

[159] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[160] T. ETSI, "Study on channel model for frequencies from 0.5 to 100 ghz," *138 901 v16. 1.0, 5G*, 2020.

[161] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[162] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[163] H. Pirayesh and H. Zeng, "Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 2, pp. 767–809, Mar. 2022.

[164] Y. Zhao, Z. Nasrullah, and Z. Li, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 12, pp. 1–38, Mar. 2021.

[165] H. Akhlaghpasand, S. M. Razavizadeh, E. Björnson, and T. T. Do, "Jamming detection in massive MIMO systems," *IEEE Wireless Commun. Letters*, vol. 7, no. 2, pp. 242–245, Apr. 2018.

[166] P. Lohan, B. Kantarci, M. Amine Ferrag, N. Tihanyi, and Y. Shi, "From 5G to 6G networks: A survey on AI-based jamming and interference detection and mitigation," *IEEE Open Jour. of the Commun. Society*, vol. 5, pp. 3920–3974, Jun. 2024.

[167] Y. Li, J. Pawlak, J. Price, K. Al Shamaileh, Q. Niyaz, S. Paheding, and V. Devabhaktuni, "Jamming detection and classification in OFDM-based UAVs via feature- and spectrogram-tailored machine learning," *IEEE Access*, vol. 10, pp. 16 859–16 870, 2022.

[168] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, Mar. 2019.

[169] M. Varotto, F. Heinrichs, T. Schürg, S. Tomasin, and S. Valentin, "Detecting 5G narrowband jammers with CNN, k-nearest neighbors, and support vector machines," in *Proc. IEEE Int. Work. on Information Forensics and Security (WIFS)*, 2024, pp. 1–6.

[170] B. Upadhyaya, S. Sun, and B. Sikdar, "Machine learning-based jamming detection in wireless IoT networks," in *Proc. IEEE VTS Asia Pacific Wireless Commun. Symposium (APWCS)*, 2019, pp. 1–5.

[171] N. I. Mowla, N. H. Tran, I. Doh, and K. Chae, "Federated learning-based cognitive detection of jamming attack in flying ad-hoc network," *IEEE Access*, vol. 8, pp. 4338–4350, 2020.

[172] J. Skarding, B. Gabrys, and K. Musial, "Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey," *IEEE Access*, vol. 9, pp. 79 143–79 168, 2021.

[173] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[174] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 1, pp. 492–523, Apr. 2022.

[175] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000012