ROBUST-6G

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

Deliverable D3.2
Initial Report on 6G Trustworthy and Sustainable AI Architecture and Requirements for Integrating Selected XAI Measures

Co-funded by the European Union

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:	31/03/2025	Version:	1.0
Project reference:	101139068	Call:	HORIZON-JU-SNS-2023
Start date of project:	01/01/2024	Duration:	30 months

Document properties:

Document Number:	D3.2
Document Title:	Initial Report on 6G Trustworthy and Sustainable AI Architecture and Requirements for Integrating Selected XAI Measures
Editor(s):	Arsenia Chorti (ENSEA), Ozgul Ayyildiz (GOHM)
Authors:	Enrique Tomás Martínez Beltrán (UMU), Manuel Gil Pérez (UMU), Fernando Torres Vega (UMU), Leyli Karaçay (EBY), Betül Güvenç Paltun (EBY), Ömer Tuna (EBY), Bartlomiej Siniarski (UCD), Geetika Arora (UCD), Chamara Sandeepa (UCD), Thulitha Senevirathna (UCD), Farah Abed Zadeh (UCD), Giovanni Perin (UNIPD), Nikolaos Pappas (LIU), Eunjeong Jeong (LIU), Marios Kountouris (EUR), Ioannis Pitsiorlas (EUR), Nour Jamoussi (EUR).
Contractual Date of Delivery:	31/03/2025
Dissemination level:	PU
Status:	Final
Version:	1.0
File Name:	ROBUST-6G D3.2_v1_0

Revision History

Revision	**Date**	**Issued by**	**Description**
0.1	01.02.2025	ROBUST-6G WP3	Initial draft with ToC
0.2	15.02.2025	ROBUST-6G WP3	First draft with threat assessment and prevention
0.3	05.03.2025	ROBUST-6G WP3	Second draft with 6G key technical enablers and selected cases
0.4	10.03.2025	ROBUST-6G WP3	First complete draft
0.5	20.03.2025	ROBUST-6G WP3	Second draft after internal review
0.6	26.03.2025	ROBUST-6G WP3	Final complete draft after internal review
1.0	31.03.2025	ROBUST-6G WP3	Final version

Abstract

The rapid evolution of mobile networks toward 6G is driven by the need for ultra-reliable, high-performance, and intelligent communication systems. As AI becomes increasingly embedded in network security and management, ensuring its trustworthiness, sustainability, and transparency is paramount. This deliverable addresses these challenges by proposing AI-driven security mechanisms that align with the principles of robust, explainable, and energy-efficient security solutions.

A key focus of this work is **Trustworthy AI**, which ensures that AI-based security mechanisms uphold privacy, fairness, robustness, and resilience against adversarial threats. The growing reliance on AI for intrusion detection, threat mitigation, and automated decision-making in 6G networks necessitates models that are transparent, interpretable, and free from biases. Additionally, **Sustainable AI** is a critical component, as 6G networks demand energy-efficient security solutions that optimize computational resources without compromising protection. AI-driven mechanisms must be designed to minimize power consumption while maintaining strong defenses against evolving cyber threats.

Furthermore, **Explainable AI (XAI)** is explored as a means to enhance transparency in AI security applications. By integrating XAI principles, security solutions become more interpretable, allowing stakeholders to understand, audit, and trust AI-driven decisions.

This deliverable builds upon prior research by outlining methodologies for integrating these principles into AI security architectures for 6G networks. It provides a structured approach for designing AI models that balance performance, security, and energy efficiency while ensuring explainability. The findings serve as a foundation for further research and development, guiding the transition from conceptual frameworks to real-world implementation. By establishing AI as a core enabler of secure and resilient 6G systems, this work contributes to the long-term sustainability and trustworthiness of next-generation network infrastructures.

Keywords

Adversarial threats, Privacy threats, Explainability threats, Threat assessment, Threat prevention, Distributed Learning, AI/ML

Disclaimer

Funded by the European Union. The views and opinions expressed are however those of the author(s) only and do not necessarily reflect the views of ROBUST-6G Consortium nor those of the European Union or Horizon Europe SNS JU. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

The transition to 6G networks introduces unprecedented opportunities and challenges in security, trustworthiness, and sustainability. This deliverable, D3.2, addresses these challenges by focusing on integrating Explainable AI (XAI) into the ROBUST-6G architecture, ensuring AI-driven security mechanisms are transparent, accountable, and energy-efficient.

The document builds on prior research (D3.1), which identified key AI-related security threats, to define architectural requirements and methodologies for implementing trustworthy AI security in 6G networks. It explores three foundational pillars:

1. **Trustworthy AI for 6G Security -** Establishes security, privacy, robustness, and fairness in AI-based security mechanisms.
2. **Sustainable AI for Energy-Efficient Security -** Optimizes AI-driven security solutions to minimize resource consumption while maintaining robust protection.
3. **XAI-Based Explainability and Transparency -** Enhances trust and interpretability in AI-driven security decisions, ensuring compliance with regulatory and ethical standards.

The deliverable outlines security risks in AI/ML training for 6G, focusing on adversarial attacks, model poisoning, vulnerabilities in federated learning, and potential biases in AI-driven security decisions. It proposes innovative security mechanisms leveraging XAI to mitigate these threats by improving the interpretability and accountability of AI models. Additionally, it discusses proactive defense mechanisms, such as adversarial training, secure model aggregation techniques in federated learning, and AI-driven anomaly detection frameworks.

Beyond security, the document highlights the need for resilience and reliability, ensuring 6G networks can self-recover and dynamically allocate resources to mitigate failures and cyber threats. It also explores sustainability and energy efficiency, incorporating energy harvesting, optimized network deployment, and green communication protocols to minimize environmental impact. Moreover, AI-driven optimization plays a crucial role in automating decision-making, enhancing network efficiency, and enabling adaptive resource allocation. The framework further supports heterogeneous network integration, ensuring seamless interoperability among satellite, terrestrial, and edge computing infrastructures.

Furthermore, this deliverable sets the foundation for subsequent research (D3.3) by defining conceptual and technical frameworks for AI-driven security integrations, emphasizing real-time threat detection and adaptive security responses. The ultimate goal is to enhance resilience, reliability, and scalability in 6G security operations while balancing performance, sustainability, and ethical considerations.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Acronyms and abbreviations

Term	Description
3GPP	Generation Partnership Project
ADMM	Alternating Direction Method of Multipliers
AHE	Additive Homomorphic Encryption
AI	Artificial Intelligence
AIaaS	AI-as-a-Service
CP	Conformal Prediction
CPCF	Conformal Prediction Confidence Factor

DFL	Distributed/Decentralised Fl
DNN	Deep Neural Network
DoH	DNS Over HTTPS
DoS	Denial Of Service
DP	Differential Privacy
EHFL	EHFL Based On Probabilistic Decision Rules Energy-Harvesting FL
FGSM	Fast Gradient Sign Method
FHE	Fully Homomorphic Encryption
FL	Federated Learning
GPU	Graphics Processing Unit
HE	Homomorphic Encryption
HFL	Hierarchical FL
IDS	Intrusion Detection Systems
IEEE	Institute Of Electrical And Electronics Engineers
LIME	Local Interpretable Model-Agnostic Explanations
NF	Network Functions
NN	Neural Network
NRF	Network Repository Function
NWDAF	Network Data Analytic Functions
O-RAN	Open Radio Access Network
PHE	Partial Homomorphic Encryption
PMP	Programmable Monitoring Platform
R2L	Remote To Local
RBAC	Role-Based Access Control
RNNs	Recurrent Neural Networks
S-CL	Security Control Layer
SaaS	Security-as-a-Service
SDN	Software Defined Network
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SLA	Service Level Agreement
SNNs	Spiking Neural Networks
SP	Symposium On Security And Privacy

U2R	User To Root
VAEs	Variational Autoencoders
vAoI	Version Age of Information
W3C	Web Consortium
XAI	Explainable AI
ZSM	Zero-Touch Security Management
ZTNA	Zero-Trust Network Access

1 Introduction

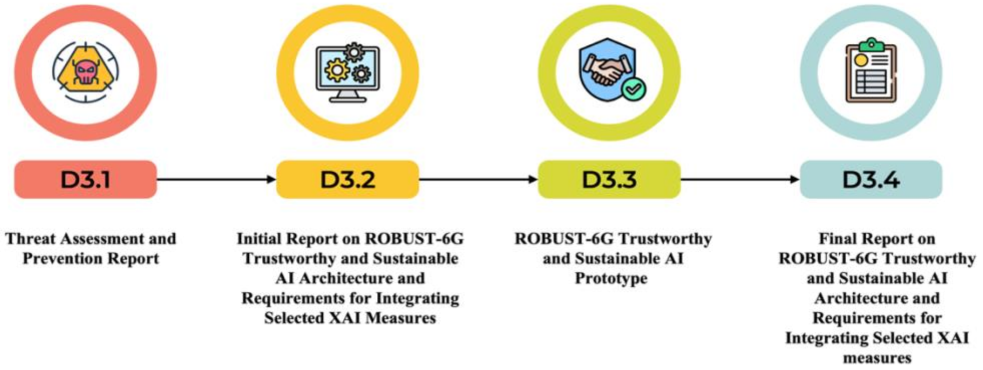
1.1 Motivation, objectives, and scope

The transition toward 6G networks is driven by the increasing demands of hyper-connected society, offering advanced network capabilities and greater computational power. While the 6G landscape represents a significant paradigm shift with mobile networks becoming deeply integrated into all aspects of life, this integration gives rise to certain challenges that should be addressed to ensure robust, secure and efficient network operations. Especially the need for trustworthy, sustainable and transparent (explainable) AI architecture to support upcoming digital infrastructures. This need originates from a number of key observations.

- Sustainability as a core design principle: 6G networks must support energy-efficient AI-driven security mechanisms that minimize resource consumption while maintaining strong protection against cyber threats.
- Trustworthiness as a non-negotiable requirement: AI-based security solutions must ensure privacy, robustness, fairness, and resilience against adversarial threats.
- Increasing role of AI in network operations: AI is integral to automating security decisions, intrusion detection, and real-time threat mitigation, necessitating reliability and transparency.

To address these challenges, this deliverable is structured around three foundational pillars.

- **Trustworthy AI for 6G Security**: This pillar focuses on establishing and maintaining high standards of security, privacy, robustness and fairness in AI implementations. Given the increasing dependency on AI for intrusion detection, threat mitigation and decision making, it is necessary to enhance the reliability, transparency and fairness of AI models deployed in 6G networks.
- **Sustainable AI for energy-efficient security mechanisms**: Energy efficiency becomes an important challenge to address with 6G networks supporting a remarkable level of connectivity and computation. This pillar ensures the need for low-power AI-driven security solutions, optimizing resource usage.
- **XAI based explainability and transparency for AI-driven security**: The security solutions should have the ability to understand, interpret, and audit the decision-making process. This can be achieved by utilizing Explainable AI (XAI) that would ensure that the security mechanisms are accountable, transparent and adaptable.


D3.1
Threat Assessment and Prevention Report
D3.2
Initial Report on ROBUST-6G Trustworthy and Sustainable AI Architecture and Requirements for Integrating Selected XAI Measures
D3.3
ROBUST-6G Trustworthy and Sustainable AI Prototype
D3.4
Final Report on ROBUST-6G Trustworthy and Sustainable AI Architecture and Requirements for Integrating Selected XAI measures

Figure

1. Progression of Work Package 3 deliverables: From Threat Assessment to Final Evaluation

This deliverable serves as the foundation for integrating AI-driven security mechanisms into 6G by defining key principles and methodologies for Trustworthy AI, Sustainable AI, and XAI-based security solutions. ***While all three pillars are introduced, the primary focus of this deliverable is the integration of XAI into the ROBUST-6G architecture.*** As presented in Figure , D3.1 served as a critical foundation for the later deliverables by providing an in-depth threat assessment and prevention strategies for AI-driven security challenges in 6G networks. This deliverable identifies key AI-related security threats, offering insights into vulnerabilities and necessary countermeasures. The findings in D3.1 directly informed the design of D3.2 by shaping the security objectives and highlighting areas where XAI could enhance trust and resilience. Additionally, the AI security solutions and prevention mechanisms outlined in D3.1 provided the initial direction for developing trustworthy AI architectures, which later became the focus of D3.2.

D3.2 lays the groundwork for D3.3 by defining the conceptual and technical framework for integrating trustworthy and sustainable AI security mechanisms into 6G networks. It proposes several solutions from each of the three pillars (Trustworthy AI for 6G Security, Sustainable AI for energy-efficient security mechanisms and XAI-based explainability and transparency for AI-driven security) and establishes a structured approach to securing AI/ML training and execution processes while incorporating energy-efficient techniques. Furthermore, D3.2 introduces a set of requirements for implementing XAI measures to enhance transparency and reliability in AI-driven security solutions. These requirements will act as the blueprint for D3.3, which will transition from theoretical foundations and preliminary work to the actual development, prototyping and integration of security functionalities based on the principles introduced in D3.2.

The key difference between D3.2 and D3.3 is that while D3.2 focuses on defining the architecture and requirements for integrating XAI, D3.3 moves toward the practical implementation of these concepts. D3.2 presents the theoretical underpinnings and initial specifications, ensuring that AI-driven security measures are aligned with principles of trustworthiness and sustainability. In contrast, D3.3 translates these concepts into working prototypes, specifying the mechanisms and technologies used to develop AI-driven security functionalities. D3.2 is, therefore, a preparatory stage that informs the design, while D3.3 is the realization of those designs through tangible security solutions.

D3.4 builds upon both D3.2 and D3.3 by conducting a final assessment of the feasibility and resilience of the security measures developed and prototyped. It evaluates the impact of these mechanisms on AI accuracy and sustainability within the 6G ecosystem, determining how effectively the proposed solutions enhance security without introducing excessive complexity or inefficiencies. By integrating findings from D3.2 and D3.3, D3.4 provides a comprehensive analysis of the strengths and limitations of the implemented XAI-based security measures. Additionally, it establishes a refined/improved and more informed set of requirements for future XAI integration feasibility, ensuring that the developed security functionalities can be effectively adapted and deployed in real-world 6G networks.

1.2 Document structure

This deliverable is organized into four key technical chapters, each addressing critical aspects of securing AI/ML training for 6G networks while aligning with the three foundational pillars of Trustworthy AI, Sustainable AI, and XAI-driven security. The following chapters provide a comprehensive analysis of security challenges, propose innovative AI-driven security mechanisms, and outline the role of XAI in improving AI security transparency and accountability.

Chapter 2 examines the security challenges in securing AI/ML training within the ROBUST-6G Trustworthy and Sustainable AI framework**.** Unlike D3.1**,** which provides a broad assessment of adversarial and privacy-related risks in AI/ML for 6G, this chapter contextualizes these challenges within the evolving 6G architecture, focusing on how distributed intelligence, federated learning, and edge computing introduce new security complexities that require targeted mitigation strategies. Extending the discussion in D3.1, this chapter analyses how identified threats manifest in AI/ML training environments, particularly in 6G-specific scenarios. While D3.1 classifies adversarial attacks such as poisoning and evasion, this chapter assesses their impact on federated learning and distributed AI models, highlighting necessary defences. Similarly, privacy risks such as model inversion and membership inference attacks are revisited to reflect their implications in decentralized AI training, a perspective not extensively covered in D3.1. The chapter also explores security risks arising from edge computing, collaborative AI training, and multi-stakeholder trust issues, emphasizing mitigation approaches suited for ROBUST-6G, thereby bridging the gap between theoretical threat assessments and practical security solutions for 6G AI deployments.

Chapter 3 presents an in-depth exploration of security measures designed to enhance the trustworthiness, robustness, and resilience of AI-driven security mechanisms in 6G. The focus is on mitigating threats such as adversarial attacks, model poisoning, and vulnerabilities in federated learning. It gives an overview of key components that ensure the security, reliability, and resilience of AI-based security mechanisms in 6G. This section also investigates how attackers can manipulate AI services and discusses existing mitigation strategies, analyses poisoning attacks that manipulate training data to degrade AI model performance and discusses an Alternating Direction Method of Multipliers (ADMM)-based approach to ensure secure and decentralized federated learning (FL) in 6G. The next subsection explores how Explainable AI (XAI) improves AI transparency and enhances user trust, discusses how to ensure fairness in AI-driven security models, reducing bias in 6G security solutions, and analyses how XAI techniques can enhance the detection and prevention of adversarial threats to AI models in 6G networks. Additionally, it introduces XAI-based approaches to Intrusion Detection Systems (IDS) and their role in 6G security.

Given the increasing energy demands of AI training and inference, **Chapter 4** explores energy-efficient security mechanisms to ensure sustainable AI integration into 6G networks. It describes an alternative optimization approach using the ADMM to optimize energy consumption in Spiking Neural Networks for security tasks. This chapter also discusses AI solutions designed to optimize power consumption and reduce computational overhead while ensuring strong security and introduces novel methodologies for optimizing computational resource allocation while maintaining security and trustworthiness in federated learning environments.

Chapter 5 focuses on the integration of XAI into security mechanisms for enhanced transparency, accountability, and interpretability in AI-driven security models for 6G. It proposes a framework integrating XAI principles into 6G security to ensure clear model decision-making and transparency. It also discusses methods to assess the effectiveness of XAI-enabled security solutions in 6G environments. Additionally, the last subsection outlines best practices and regulatory considerations for implementing XAI-based security frameworks, ensuring alignment with SNS objectives and broader 6G standardization efforts.

2 Challenges in Securing AI/ML Training for 6G

While D3.1 presented a broad threat assessment, including adversarial and privacy-related risks to AI/ML in 6G, this deliverable shifts the focus toward the specific challenges in securing AI/ML training within the ROBUST-6G Trustworthy and Sustainable AI framework. This chapter is necessary to contextualize security challenges within the evolving 6G architecture, emphasising how distributed intelligence, federated learning, and edge computing introduce new security complexities that require targeted mitigation approaches.

Unlike D3.1, which classifies and evaluates AI security threats, this chapter extends this discussion by exploring how these threats manifest in AI/ML training environments. The chapter systematically analyses adversarial attack risks, privacy concerns, and decentralized security challenges in AI/ML model training, particularly within 6G-specific scenarios. For example, while D3.1 identifies adversarial attacks like poisoning and evasion, this deliverable assesses their impact on federated learning and distributed AI models, detailing how AI models must be designed to withstand these evolving threats. Similarly, privacy risks in AI models, such as model inversion and membership inference attacks, are re-evaluated in this deliverable to reflect their implications in decentralized AI training—something that D3.1 does not address in depth.

Additionally, the chapter plays a crucial role in establishing the security architecture by detailing trustworthy and sustainable AI frameworks. It not only defines mitigation strategies but also provides a foundational reference for subsequent deliverables (D3.3 and D3.4), ensuring a comprehensive and integrated approach to AI-driven security in 6G networks. The security mechanisms explored in this deliverable are not just theoretical but provide practical design considerations for implementing trustworthy and sustainable AI in 6G, ensuring that the AI/ML training pipeline remains resilient and energy-efficient. Moreover, this deliverable introduces security-by-design principles, ensuring that the AI models developed in D3.3 (the prototype phase) inherit robust security features from the start.

2.1 Risks of Adversarial Attacks on AI/ML Models in 6G Environments

The increased reliance on AI in 6G networks introduces a broad attack surface, exposing critical components to adversarial manipulations. These vulnerabilities arise primarily due to fundamental weaknesses in AI/ML models—such as overfitting, lack of robustness to input perturbations, poor generalization, and limited explainability—which adversaries can exploit to compromise system integrity and performance. As outlined in ROBUST-6G Deliverable D3.1, these inherent limitations make AI/ML models susceptible to a range of attacks, including **evasion, poisoning, inversion,** and **extraction attacks**. Originally observed in domains like computer vision and NLP, adversarial threats have now emerged as a critical concern in 6G use cases. In this context, AI-driven functionalities—such as intelligent network slicing, resource management, and anomaly detection—become potential targets, leading to severe consequences like service disruption, network misconfiguration, and privacy violations.

Adversarial attacks in 6G can be classified based on their objectives, techniques, and attack surfaces, as comprehensively analyzed in ROBUST-6G Deliverable D3.1, Section 3.1 ("Adversarial Threats"). This section outlines how vulnerabilities in AI/ML models—such as their sensitivity to data distribution shifts, overfitting, and lack of robustness to perturbations—contribute to the emergence of these threats. Among the most critical are **data poisoning attacks**, where adversaries inject manipulated samples into training datasets, causing models to learn corrupted patterns. In federated learning scenarios, even partial poisoning by a few malicious clients can significantly degrade global model accuracy. Attackers can target data used for functions like spectrum allocation or anomaly detection, causing AI models to make faulty decisions. Another notable threat is **evasion attacks**, which involve manipulating inputs at inference time to trigger misclassification. The Fast Gradient Sign Method (FGSM), discussed in D3.1 (Section 3.1.2), exemplifies how imperceptible perturbations can cause AI-based intrusion detection systems to fail, allowing malicious traffic to pass undetected. These attack vectors illustrate the pressing need to address AI model vulnerabilities in the security architecture of future 6G networks.

Beyond these conventional adversarial techniques, more sophisticated attacks, such as **model inversion** and **extraction** pose severe threats to AI integrity in 6G. Attackers can leverage adversarial techniques to infer sensitive information about training data or extract proprietary AI models deployed in 6G networks. Model inversion attacks exploit the outputs of a model to reconstruct input data, posing privacy threats in AI-driven biometric authentication systems. Model extraction attacks, on the other hand, enable adversaries to replicate AI models used in network security functions, by repeatedly querying them and analysing the outputs. For instance, if a 6G telecom operator uses an AI-based IDS (potentially based on a proprietary model) to detect anomalous traffic patterns indicative of cyberattacks, an attacker could send a variety of inputs (traffic samples) to the system, some benign, some malicious, and observe the model's responses. Over time, the attacker can build a (local) copy of the model that behaves similarly. With this replica, they (attackers/adversaries) can test various attack strategies offline until they find ones that evade detection, and then use them to bypass the real network's security mechanisms. Open Radio Access Network (O-RAN) architectures further exacerbate these risks by introducing AI-based decision-making for network optimization. However, this openness also makes them susceptible to adversarial attacks. For example, attackers can manipulate AI-driven xApps and rApps that control resource allocation, leading to service degradation or unfair spectrum allocation. These vulnerabilities highlight the need for robust adversarial defences in AI-driven O-RAN environments.

The risks and the consequences of adversarial attacks on AI models in 6G are profound, affecting both security and network performance. Service disruptions can occur when AI-driven radio resource management is compromised, leading to inefficient spectrum allocation, congestion, and degraded Quality of Service (QoS). Security breaches become more prevalent when evasion attacks successfully bypass AI-based anomaly detection systems, allowing malicious traffic or unauthorized users to infiltrate the network. Privacy violations are another major risk, as model inversion techniques enable attackers to extract sensitive user data from AI-driven authentication and encryption systems. Additionally, implementing real-time adversarial defences in 6G networks introduces significant computational costs, potentially impacting network efficiency. To better understand the scope and potential impact of these threats, Table 1 provides a structured risk assessment of key adversarial attacks in 6G environments, evaluating each based on severity and likelihood grounded in current research.

Attack Type	Description	Severity	Likelihood	Justification
Data Poisoning	Injection of malicious data into training datasets, causing AI models to learn incorrect behaviors.	High	Medium	In federated learning scenarios, poisoning a subset of data can significantly degrade global model performance. Given the decentralized nature of 6G networks, detecting and mitigating such attacks is challenging.

Evasion Attacks	Crafting inputs during the inference phase to mislead AI models into incorrect classifications or predictions.	High	High	Techniques like the Fast Gradient Sign Method (FGSM) can subtly alter inputs, causing misclassifications in AI models. In 6G networks, this can lead to misallocation of resources or failure to detect intrusions, directly impacting network performance and security.
Model Inversion	Extracting sensitive information about training data by querying the AI model, potentially compromising user privacy.	Medium	Low	While model inversion poses significant privacy risks, executing such attacks requires substantial access and computational resources, making them less common. However, as 6G networks handle more personal data, the incentive for such attacks may increase.
Model Extraction	Duplicating a proprietary model by observing its outputs, leading to intellectual property theft and potential deployment of compromised models.	Medium	Medium	The open nature of some 6G components, like Open RAN, may expose models to observation. Attackers can exploit this to replicate models, undermining competitive advantages and introducing malicious variants.
Backdoor Attacks	Introducing hidden functionalities into AI models that activate under specific conditions, allowing unauthorized control or access.	High	Low	Embedding backdoors during the training phase requires significant access, but if successful, these attacks can provide persistent unauthorized control over network functions, posing severe security risks.
Adversarial Perturbations	Adding subtle, often imperceptible changes to inputs that cause AI models to err, affecting tasks like signal classification and channel estimation.	High	High	Given the reliance on AI for real-time decisions in 6G, such perturbations can degrade performance, leading to issues like misallocation of resources or failure in detecting anomalies. The ease of crafting these perturbations contributes to their high likelihood.

Table 1 Risk Assessment of Adversarial Attacks in 6G Networks

To counter these threats, several defence strategies have been proposed to mitigate adversarial risks in 6G AI/ML models. One of the most widely studied approaches is adversarial training, which enhances model robustness by incorporating adversarial examples into the training dataset. This method allows AI models to learn how to resist perturbations, but it also increases computational complexity and may not generalize well to unseen attack types. Robust optimization techniques, such as certified defences and randomized smoothing, improve AI model resilience against adversarial perturbations. Certified defences provide mathematical guarantees that a model's predictions will remain unchanged within a certain range of input perturbations, while randomized smoothing adds noise to inputs and relies on probabilistic averaging to make the model's output more stable and resistant to small adversarial changes. These approaches ensure that small perturbations in input data do not significantly alter model predictions. Another promising avenue is AI-driven threat detection, which utilizes anomaly detection systems to monitor input data distributions and detect deviations. These systems rely on invariant-based adversarial detection, feature-based anomaly scoring, and ensemble detection techniques to filter out adversarial inputs. Furthermore, as 6G moves toward integrating quantum technologies, adversarial threats may evolve.

As adversarial attack techniques evolve, future techniques must address several challenges. Developing efficient real-time defences is crucial, as current adversarial defence mechanisms impose high computational costs. Research should focus on lightweight, adaptive defences that can operate in real-time 6G environments.

Enhancing AI model explainability is another key area of focus, as XAI techniques can improve model interpretability, making it easier to detect and mitigate adversarial manipulations. Securing AI-driven O-RAN architectures should also be a priority, ensuring the integrity of intelligent network management systems. Additionally, blockchain-based federated learning can provide a decentralized security framework for AI training, preventing adversarial manipulation in collaborative learning environments.

Adversarial attacks in AI/ML pose significant privacy threats in 6G networks, which rely on AI-driven decision-making, edge computing, and federated learning. Attackers can exploit adversarial techniques to manipulate models, extract sensitive user data, and compromise authentication systems. For example, model inversion attacks can reconstruct private training data, while adversarial perturbations can bypass security mechanisms, leading to privacy leaks in authentication and traffic analysis. These vulnerabilities highlight the need for robust defences against adversarial threats in 6G AI-powered applications.

2.2 Privacy Concerns in AI Models

Beyond security threats, adversarial attacks in AI-driven 6G systems also introduce serious privacy risks, making it crucial to address both aspects together. This dual challenge arises as organizations increasingly rely on AI to handle sensitive data, making privacy and security critical considerations. AI systems depend on large volumes of data to train algorithms and make predictions or decisions. This data may contain personal details and other identifiable information which can be processed by AI systems without adequate consent or transparency. For example, an AI trained on multiple datasets might unintentionally combine information in a way that reveals personal details or identifies individuals. Using inference, the prediction results of the model can also be used to infer sensitive information, such as linking user identity with sensitive attributes or reconstructing the sensitive training data. A real-world example of such privacy risk was highlighted during early 5G deployments in South Korea, where researchers demonstrated that location inference attacks could be launched by exploiting 5G handover mechanisms, revealing precise user movements over time. Another example involved telecom operator usage analytics, where AI-driven optimization of service quality was found to process subscriber behavioral patterns without adequate anonymization, raising concerns under GDPR about re-identification risks. These incidents reflect how privacy vulnerabilities in 5G—if left unaddressed—can propagate into more complex, AI-integrated 6G networks.

To mitigate such risks, it is important to implement confidentiality protection such as encryption, access controls, and data anonymization to secure data both at rest and in transit against external threats, as well as apply data minimization strategies to reduce the volume of data collected and processed. Furthermore, the risk of internal threats, such as insider threats and data leakage, need to be considered.

According to [ETSI+24], there are AI-specific properties of privacy which refer to the unique privacy concerns that arise in the context of artificial intelligence systems. These properties include data privacy, algorithmic transparency, privacy by design, user control, and accountability. AI systems rely on large volumes of data, which may include sensitive personal information. Ensuring data privacy is essential, and safeguards must be in place to comply with privacy regulations during the collection, storage, and use of personal data. Due to the complexity of AI algorithms, understanding decision-making processes can be challenging. Algorithmic transparency ensures individuals can comprehend how their data is used and verify that AI decisions are fair and unbiased. AI systems should integrate privacy measures from the outset, ensuring that personal data collection is minimal, security measures are implemented, and privacy policies are clear and accessible. Individuals should have control over their personal data, with the ability to access, modify, or delete it as needed. AI systems should provide clear options for users to manage their data and respect privacy preferences. Organizations developing and deploying AI systems must be accountable for protecting individuals' privacy, taking responsibility for breaches and implementing measures to prevent future incidents.

ML models can be classified based on whether the learning task is centralized or distributed. In centralized learning, a single entity stores and processes all training data, and a model is trained on gathered data. In this type of learning, a single entity has access to all data, which arise privacy concerns. In distributed/federated learning, each participant trains its local model using local data and shares the model parameters with other participants to build a shared model. Although federated learning provides inherent privacy, studies show that protecting sensitive client training data could not be ensured solely by keeping the client training data local. The key roles in FL include central server and local clients. The adversary can compromise the central server and/or some or all local clients. In [LYM+22], it is stated that before the model is trained, malicious local clients may disturb the integrity, confidentiality, and availability of data and degrade the model. During the model training phase, a malicious local client could also manipulate its data, model gradients, and parameters.

As a result, if adversaries manage to compromise local clients, they can disrupt the integrity of the training dataset or the model, thereby degrading the performance of the global model. Additionally, the malicious central server may also carry out passive or active inference attacks and by having access to intermediate model update parameters during federated learning, might reveal information related to the private training data of clients. After the model is trained, the global model is deployed onto local client devices, regardless of their participation in the training process. The gradients leakage attacks, membership inference attacks, attribute inference, and model inversion attacks are types of data leakage attacks that have been studied in distributed learning environments.

2.3 Challenges Posed by the Decentralized and Highly Dynamic Nature of 6G

As 6G builds upon the foundation of 5G with advancements in ultra-low latency, massive connectivity, and pervasive intelligence, its decentralized and dynamic nature introduces new security and privacy risks. Unlike traditional centralized models, where data is collected and processed in a secure data center, 6G relies heavily on federated learning and edge AI, where data remains distributed across multiple edge nodes and user devices. This increases exposure to adversarial attacks, such as data poisoning and model inversion attacks, where malicious actors can manipulate training data or extract sensitive information from models. Second, network heterogeneity and dynamic topology introduce vulnerabilities. 6G networks are expected to support various devices, ranging from autonomous vehicles to IoT sensors, each with varying computational capabilities and security standards. The dynamic nature of these networks, characterized by frequent topology changes due to device mobility and network slicing, makes it challenging to establish consistent security policies and trust mechanisms. Third, resource constraints at the edge limit the implementation of robust security mechanisms. Unlike centralized AI/ML systems that operate in data centers with ample computational resources, edge devices in 6G networks often have limited processing power, storage, and energy. This restricts the ability to employ advanced cryptographic techniques, secure enclaves, or continuous authentication, thereby increasing exposure to attacks such as model stealing and inference-time adversarial attacks. Fourth, trust and authentication in decentralized learning become more complex. Traditional AI/ML models rely on a trusted central authority to validate training data and model updates. In 6G, the absence of such a central entity necessitates using blockchain-based solutions or zero-trust architectures to ensure the integrity and authenticity of model updates. However, these solutions introduce their challenges, including scalability, latency, and the need for consensus mechanisms that can operate efficiently in highly dynamic environments. Lastly, adversarial AI and emerging attack vectors pose a significant challenge. As AI becomes more integral to 6G network management and optimization, attackers can exploit vulnerabilities in ML models through adversarial examples, evasion attacks, or backdoor injections. The decentralized nature of 6G makes detecting and mitigating these threats more difficult, as model updates and training processes are distributed across multiple untrusted nodes.

Addressing these challenges requires a multi-faceted approach, including developing robust federated learning frameworks, lightweight cryptographic techniques, adaptive security policies, and AI-driven anomaly detection systems. In parallel, standardization bodies such as ETSI, 3GPP, and ITU-T have been actively evolving their frameworks to address these emerging risks. For instance, ETSI ISG PDL (Permissioned Distributed Ledger) explores integrating blockchain and decentralized identity management to improve trust in collaborative AI systems. Similarly, 3GPP SA3 is advancing security architectures for edge and AI-native networks, introducing concepts like zero-trust architectures and secure federated learning models. The ITU-T Focus Group on Autonomous Networks (FG-AN) is also outlining trustworthiness metrics and decentralized control models to guide AI system governance in highly dynamic 6G environments. These evolving standards aim to formalize key aspects of AI-driven, decentralized infrastructures—including model update validation, distributed trust management, and privacy-preserving mechanisms—thereby enabling secure, interoperable, and scalable deployments. Without proactive security measures and aligned standardization efforts, the decentralized and dynamic nature of 6G could undermine the reliability and trustworthiness of AI/ML applications in next-generation networks.

3 Trustworthy AI for 6G

3.1 Components for Enhancing AI/ML Security and Resilience in 6G

3.1.1 Adversarial attacks against AI-as-a-Service and defence-EBY

Adversarial attacks have the potential to undermine the security of AI-driven networks, creating significant risks, especially in areas like telecommunications, where security is crucial. This contribution focuses on inference query-based Black-Box attacks that may target AI/ML models deployed in an AI-as-a-Service (AIaaS) framework and introduces a defence mechanism to counter these attacks. Our approach, as presented in Figure 2, enables the model owner to assess the model's uncertainty (aleatoric uncertainty) during predictions and use this information to adjust the model weights, resulting in more accurate predictions. AIaaS provides AI services and functions to the users, eliminating the need for them to build and maintain their own AI infrastructure [HEX-20]. AIaaS providers can offer pre-built models accessible through APIs (Application Programming Interfaces), allowing external users to interact with deployed models. The AI agents process inference requests from the users, provide predictions or reports as a service. However, deployed models are vulnerable to attacks like model evasion and model inversion, where adversaries manipulate inputs to influence the decisions of the model. These risks highlight the need for continuous monitoring of inference outputs from AI agents and prompt retraining when required.

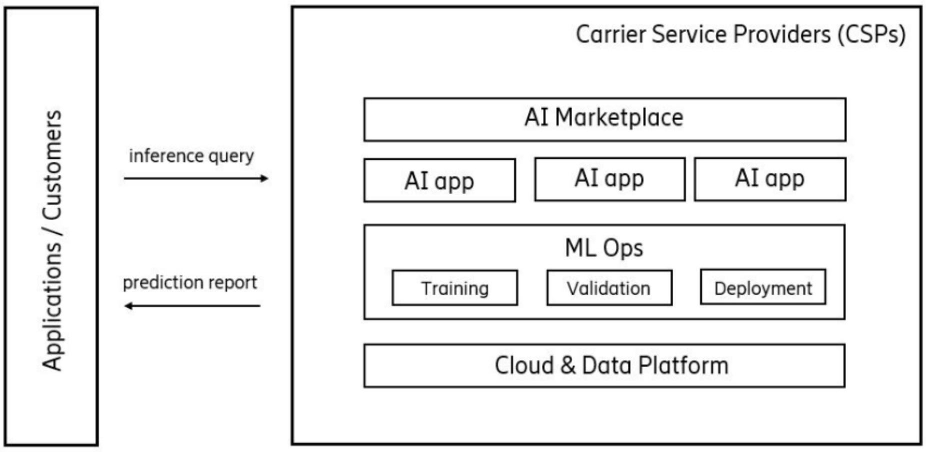


Figure 2. An overview of the AIaaS framework

Due to the inherent vulnerabilities in deep neural network (DNN) models, defending against adversarial attacks is a challenging task. These models can produce unexpected results because they are highly sensitive to even slight changes in the input data. Adversarial attacks are typically categorized into two groups based on the level of knowledge of the attacker. White-Box Setting is the one in which the attacker has complete access to the details of the model, such as its architecture, weights, and hyper-parameters. In contrast, a Black-box setting occurs when the attacker has no knowledge of the deployed model. Our primary focus here is on attacks in a Black-Box setting, although we partly address certain White-Box attacks that have a similar impact on model decisions as Black-Box attacks.

Our proposed solution is a defence mechanism designed to mitigate inference queries based Black-Box attacks (evasion attacks) during the inference phase of the AI/ML model. The goal is to protect against scenarios where the probability of a misclassified class is nearly equal to the probability of the actual class, indicating high uncertainty in the model predictions. The impact of our method is particularly significant against adversarial samples that are close to the decision boundary of the model. This is typical in Black-box attack scenarios, where the attacker lacks access to model details and, as a result, cannot fine-tune adversarial perturbation to cause confidently incorrect prediction. Some White-box adversarial attacks, such as Deepfool and Carlini&Wagner (with default setting), also fall into this category. Our solution enables the model owner to assess the uncertainty estimates of the model during predictions. For cases with high uncertainty, we suggest adjusting the model weights in a direction to minimize quantified uncertainty (aleatoric uncertainty). This adjustment acts as a one-time update specific to the suspicious input. After updating, the model makes a prediction using the modified weights. Once the prediction for the suspicious input is completed, the model owner reverts to the original model. Figure 3 illustrates the detailed implementation of our solution during a Black-Box Adversarial Attack and the algorithm details are provided in Algorithm 1.

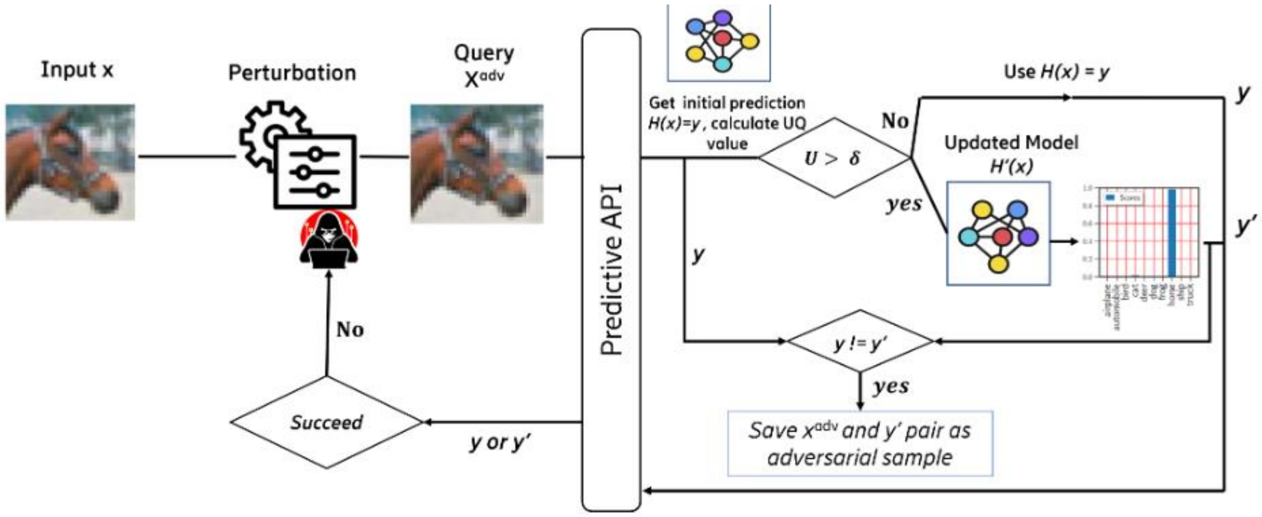
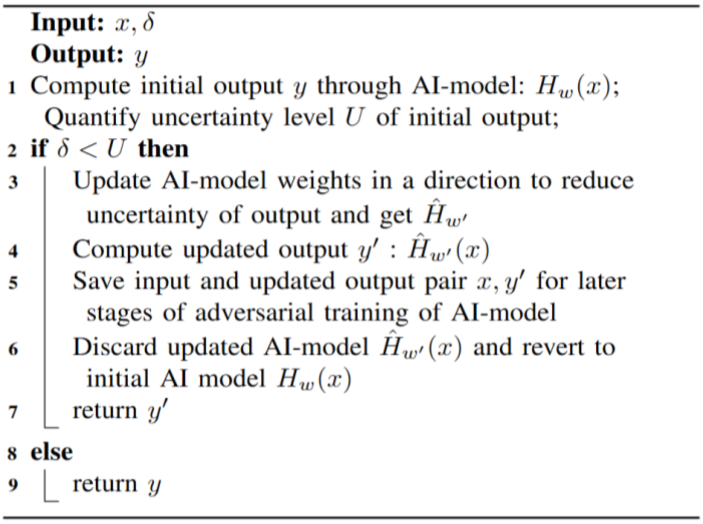


Figure 3. Detailed implementation of the proposed solution during a BlackBox Adversarial Attack


```
Input: $x, \delta$
Output: $y$
Compute initial output $y$ through AI-model: $H_w(x)$;
  Quantify uncertainty level $U$ of initial output;
if $\delta < U$ then
    Update AI-model weights in a direction to reduce
      uncertainty of output and get $\hat{H}_{w'}$
    Compute updated output $y'$ : $\hat{H}_{w'}(x)$
    Save input and updated output pair $x, y'$ for later
      stages of adversarial training of AI-model
    Discard updated AI-model $\hat{H}_{w'}(x)$ and revert to
      initial AI model $H_w(x)$
    return $y'$
else
    return $y$
```

Algorithm 1. The proposed algorithm to defend against inference queries based Black-Box attacks

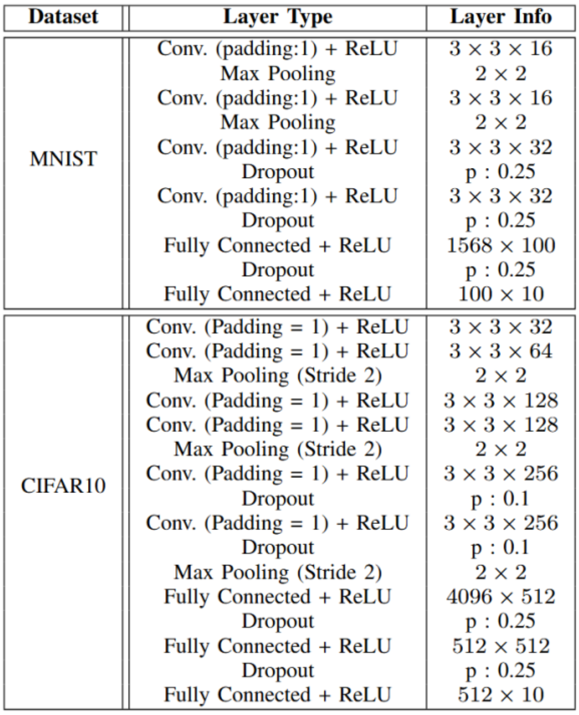
For the uncertainty quantification, we adopted the efficient approach originally proposed by Gal et al. [GG-16] and later enhanced by Kwon et. al. [KWK+20]. Gal et al. demonstrated that a DNN model with inference time dropout is equivalent to a Bayesian approximation of a Gaussian process. Inference time dropout acts as an ensemble approach, where in each single ensemble model, the different neurons are dropped out in network layers based on the dropout ratio during prediction. The overall uncertainty in the prediction is estimated by calculating the variance of the probabilistic feed-forward Monte Carlo (MC) dropout sampling at prediction time. Subsequently, Kwon et al. introduced an alternate method to quantify both epistemic and aleatoric uncertainty in classification models. In their approach, the prediction variance consists of two components: one representing aleatoric uncertainty and the other representing epistemic uncertainty. Let $H_{\hat{\omega}}$ represents the neural network model with parameters $\hat{\omega}$, and k denote the number of output classes. The model prediction for any test sample x, given the weights of the model, is denoted as $p(y|x, H_{\hat{\omega}})$, where $y \in \mathbb{R}^k$. The formulation for their method is provided below:

$$var_{p(y|x_1 H_\omega)}(y) = \mathbb{E}_{p(y|x_1 H_\omega)}(y^{\otimes 2}) - \mathbb{E}_{p(y|x_1 H_\omega)}(y)^{\otimes 2}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[diag\{p(y|x_1 H_{\hat{\omega}_t})\} - p(y|x_1 H_{\hat{\omega}_t})^{\otimes 2}\right]$$

$$+\frac{1}{T}\sum_{t=1}^{T}\left[p\left(y|x_1 H_{\hat{\omega}_t}\right)-\hat{p}\left(y|x_1 H_{\hat{\omega}_t}\right)\right]^{\otimes 2}$$

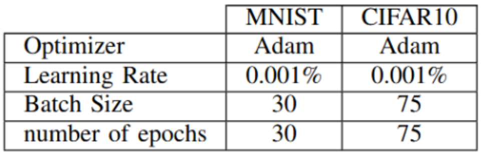
Where,

$$\hat{p}(y|x, H_{\hat{\omega}_t}) = \textstyle\sum_{t=1}^{T} p(y|x, H_{\hat{\omega}_t}) \text{ and } y^{\otimes 2} = yy^T$$



Dataset	Layer Type	Layer Info
MNIST	Conv. (padding:1) + ReLU	$3 \times 3 \times 16$
	Max Pooling	2×2
	Conv. (padding:1) + ReLU	$3 \times 3 \times 16$
	Max Pooling	2×2
	Conv. (padding:1) + ReLU	$3 \times 3 \times 32$
	Dropout	p : 0.25
	Conv. (padding:1) + ReLU	$3 \times 3 \times 32$
	Dropout	p : 0.25
	Fully Connected + ReLU	1568×100
	Dropout	p : 0.25
	Fully Connected + ReLU	100×10
CIFAR10	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 32$
	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 64$
	Max Pooling (Stride 2)	2×2
	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 128$
	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 128$
	Max Pooling (Stride 2)	2×2
	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 256$
	Dropout	p : 0.1
	Conv. (Padding = 1) + ReLU	$3 \times 3 \times 256$
	Dropout	p : 0.1
	Max Pooling (Stride 2)	2×2
	Fully Connected + ReLU	4096×512
	Dropout	p : 0.25
	Fully Connected + ReLU	512×512
	Dropout	p : 0.25
	Fully Connected + ReLU	512×10

Table 2 CNN model architecture

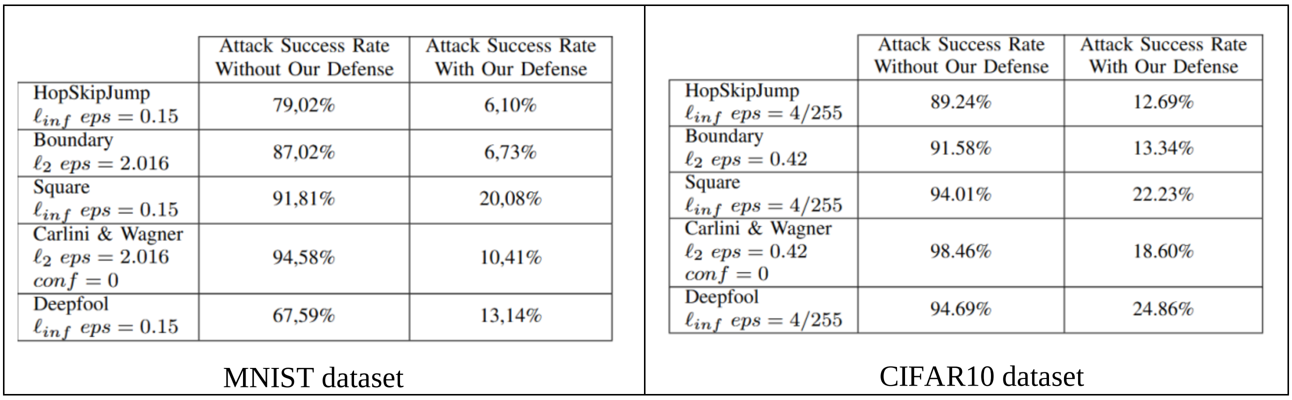


	MNIST	CIFAR10
Optimizer	Adam	Adam
Learning Rate	0.001%	0.001%
Batch Size	30	75
number of epochs	30	75

Table 3 Hyperparameters

Validation and results

We conducted our experiments by training two distinct CNN models using MNIST (Digit) and CIFAR-10 datasets, achieving accuracy rates of 99.44% and 82.43%, respectively. The architectures of the CNN models and the hyperparameters used during training are outlined in Table 2 and Table 3. Next, we applied a variety of attack types to each test sample in order to generate their adversarial counterparts. The selected set of attacks includes both Black-Box attack and White-Box attack algorithms. We then tested those adversarial samples back to the target models, performing these tests with and without our defence mechanism enabled. For the one-time model update in Step 3 of Algorithm 1, we used the Adam optimizer with a learning rate of 0.001. Regarding the selection of δ, we recommend evaluating the uncertainty values of the model for both correct and incorrect predictions on the test dataset, calculating the mean for each, and then averaging the two. Accordingly, for MNIST, we set δ = 0.0125 and for CIFAR10, we set δ = 0.0114. The results of our experiments are provided in Table 4.



	Attack Success Rate Without Our Defense	Attack Success Rate With Our Defense
HopSkipJump ℓ_{inf} $eps = 0.15$	79,02%	6,10%
Boundary ℓ_2 $eps = 2.016$	87,02%	6,73%
Square ℓ_{inf} $eps = 0.15$	91,81%	20,08%
Carlini & Wagner ℓ_2 $eps = 2.016$ $conf = 0$	94,58%	10,41%
Deepfool ℓ_{inf} $eps = 0.15$	67,59%	13,14%

MNIST dataset

	Attack Success Rate Without Our Defense	Attack Success Rate With Our Defense
HopSkipJump ℓ_{inf} $eps = 4/255$	89.24%	12.69%
Boundary ℓ_2 $eps = 0.42$	91.58%	13.34%
Square ℓ_{inf} $eps = 4/255$	94.01%	22.23%
Carlini & Wagner ℓ_2 $eps = 0.42$ $conf = 0$	98.46%	18.60%
Deepfool ℓ_{inf} $eps = 4/255$	94.69%	24.86%

CIFAR10 dataset

Table 4 Experimental results

The results show that our proposed solution significantly enhances the robustness of AI models, which might be deployed in the cloud and offered as-a-service manner. Additionally, the performance of our method is higher in inference queries based Black-Box attacks compared to White-Box attacks. Lastly, we evaluated the impact of our method on the natural (clean) performance of the model. By testing the model on all the clean test samples from MNIST with our method enabled, we achieved an accuracy of 99.37%, which is nearly identical to the original performance of the model. Note that, for samples where the model is exposed to an adversarial sample but still makes the correct prediction through our proposed one-time model update operation, the model owner has the opportunity to store the adversarial sample to be used in later stages of adversarial training. This approach enhances the robustness of the model without requiring additional resources to generate these adversarial samples in advance.

Use within ROBUST-6G System Design

The AI model lifecycle management component in the "trustworthy and sustainable AI service layer" is responsible for managing a machine learning model, covering all the stages and activities that a model undergoes throughout its life, from data collection to model deployment, monitoring, and eventual decommissioning. The proposed method, which is encapsulated in Enhanced AI component of the "trustworthy and sustainable AI service layer", contributes to the ROBUST-6G architecture in a way that it proposes additional steps in MLOps pipeline to enhance the robustness of the model. Figure illustrates the proposed updates to the MLOps pipeline to enable our proposed solution and Figure 5 shows the envisaged placement of proposed solution in ROBUST-6G architecture.

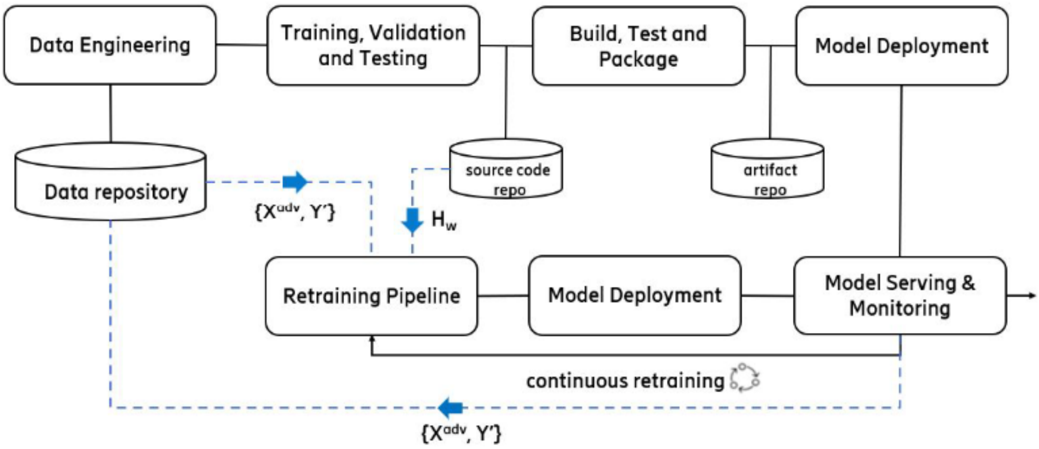


Figure 4. Proposed updates to the MLOps Pipeline to enable our proposed solution

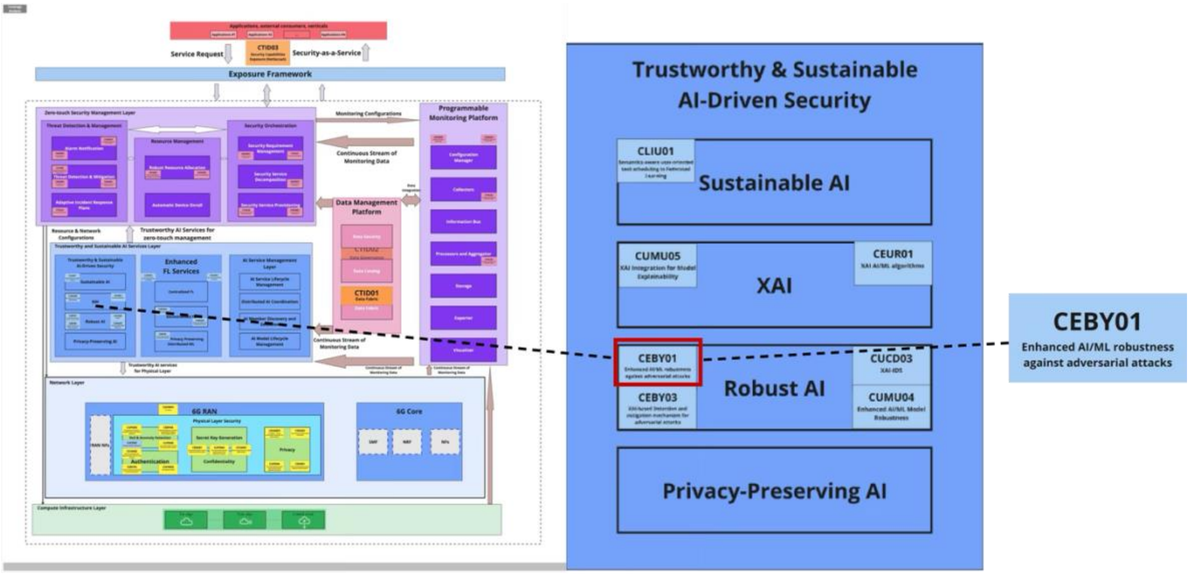

Service Request
Security-as-a-Service
Exposure Framework
Monitoring Configurations
Programmable Monitoring Platform
Data Management Platform
Continuous Stream of Monitoring Data
Enhanced FL Services
Network Layer
6G RAN
6G Core
Compute Infrastructure Layer
Trustworthy & Sustainable AI-Driven Security
CLIU01
Sustainable AI
CUMU05
XAI Integration for Model Explainability
CEUR01
XAI AI/ML algorithms
XAI
CEBY01
Enhanced AI/ML robustness against adversarial attacks
CUCD03
XAI-IDS
CEBY03
Robust AI
CUMU04
Enhanced AI/ML Model Robustness
Privacy-Preserving AI
CEBY01
Enhanced AI/ML robustness against adversarial attacks

Figure 5. Representation of Black-Box Adversarial Attacks Mitigation on ROBUST-6G Architecture

3.1.2 Poisoning attacks

FL enables distributed model training across multiple clients while preserving data privacy. However, FL is vulnerable to poisoning attacks, where adversaries manipulate the learning process to degrade model performance or introduce malicious behaviours. These attacks can be categorized into two main types:

1. **Data Poisoning Attacks** – The attacker manipulates training data to influence the model's decision boundary. This includes label flipping, where class labels are altered, or backdoor attacks, where subtle changes in input patterns cause the model to misclassify specific instances. Data poisoning typically requires modifying a large portion of training data to be effective.

2. **Model Poisoning Attacks** – Instead of altering the dataset, the attacker directly manipulates the local model updates before sending them for aggregation. This approach can be more subtle and efficient as it does not require access to training data. Model poisoning can introduce backdoors, embed biases, or degrade performance for specific target instances while maintaining overall model accuracy to avoid detection.

While data poisoning relies on corrupting the dataset, model poisoning attacks pose a greater challenge in FL systems, as they can be executed with minimal effort and evade standard defences. Advanced poisoning techniques, such as those leveraging XAI, can make these attacks even harder to detect and mitigate.

Distributed FL poisoning attack and defence: In collaborative learning environments like Distributed/Decentralised FL (DFL), an adversary can extract the model from a neighbouring target client, inject a poisoning attack, and subsequently return the compromised model to the target. The work in [ML+24] employs Layer-wise Relevance Propagation (LRP) to introduce algorithmic bias, thereby amplifying group unfairness in FL models. LRP is an Explainable AI technique that backpropagates a model's prediction through the network to assign relevance scores to input features, showing which parts contributed most to the output. It helps interpret complex models like deep neural networks by highlighting important input elements. However, the approach in [ML+24] does not address privacy attacks or decentralised FL settings, nor does it propose a solution to mitigate these issues. Moreover, while cryptographic techniques such as homomorphic encryption-based schemes [YDK+24] can safeguard the privacy of individual client updates, they are ineffective against poisoning attacks, as encrypted model parameters cannot be directly analyzed by the receiver before aggregation [ZZW+22]. The enhanced model-sharing flexibility in DFL further exacerbates the risk of poisoning attacks, making them more severe and likely in DFL systems.

LRP-based Model Poisoning: We introduce **ROAM** - Relevance-Oriented Attack Mechanism, a novel poisoning attack that exploits the LRP-based XAI technique to embed a subtle and targeted backdoor directly into the model. The goal is to design an attack that is both highly effective and difficult to detect, increasing its success rate while minimizing traceability.

In a DFL system, an honest target peer is likely to carefully inspect a received model before aggregation, potentially using unknown defensive mechanisms. Therefore, the attacker must introduce minimal modifications to the model parameters to avoid detection. The attack is designed to be highly adaptable, allowing precise targeting of specific data instances while ensuring the overall model performance remains largely unaffected. Moreover, it requires minimal effort, as the adversary injects the poisoning backdoor directly into their local model.

Attack Methodology:

The attack manipulates the model during training by leveraging Layer-wise Relevance Propagation (LRP) to identify the k most critical neurons at a specific layer l for a chosen data instance x_s from dataset d_s. By selectively modifying these neurons, the attacker embeds a backdoor that subtly degrades the model's performance for the targeted data instance or property. Once the compromised model is shared with a peer in a DFL environment, it covertly undermines the learning process while remaining inconspicuous, bypassing detection mechanisms and selectively sabotaging specific tasks.

Attack Procedure:

1. At a specific iteration n, apply LRP to identify key neurons in a target client's model, focusing on a selected data instance at a chosen layer.
2. Introduce targeted perturbations to these identified neurons, embedding the backdoor.

This approach ensures that the attack remains subtle yet highly effective, making detection and mitigation significantly more challenging in decentralized learning environments.

Figure provides an overview of the simplified attack process. Here, we send an original data and a perturbed data using Counterfactual Explanations and apply LRP and obtain the relevance scores of a particular selected NN layer for both data instances. Then, we obtain the difference in the LRP scores for the two instances and extract the key neurons that are important in misclassifying the input. Next, we can amplify by increasing the weights/gradients associated with those neurons to enhance the poisoning attack intensity.

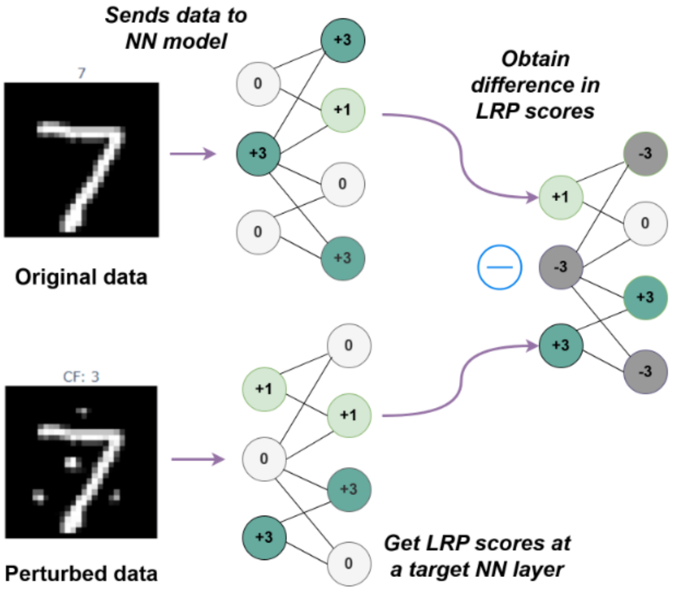


Figure 6. Unique LRP score derivation for original vs. perturbed data

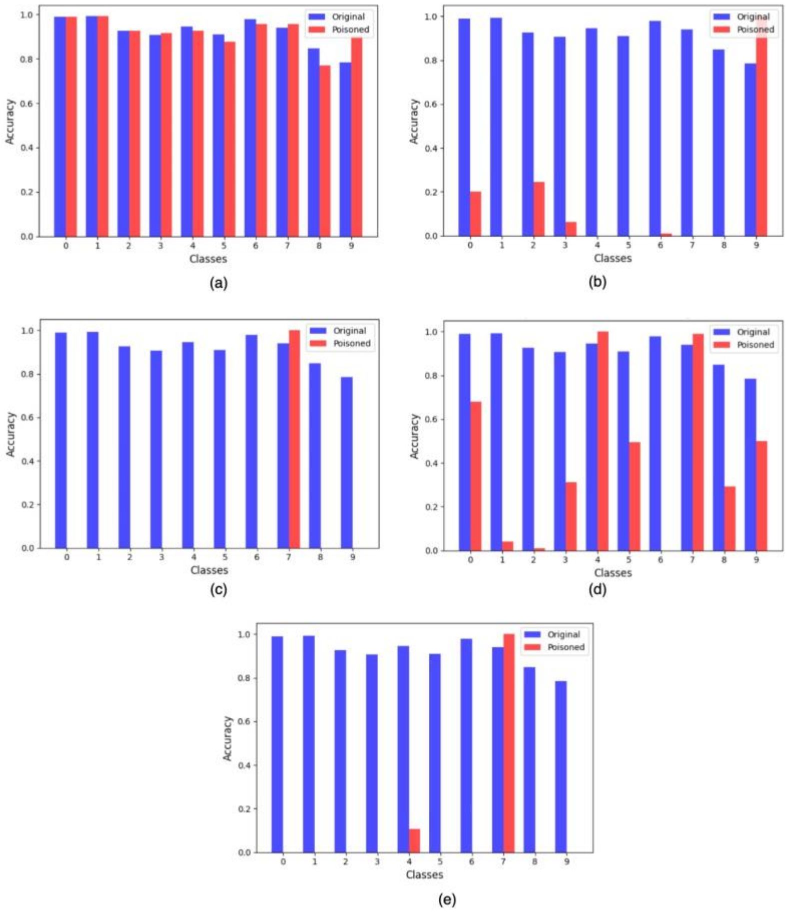
Algorithm 2 outlines the detection process, which computes unique LRP scores for a target instance using a machine learning model M. The algorithm begins by calculating the LRP scores for x_t, represented as r_t. It then retrieves a set of similar instances X_s belonging to the same class as x_t and computes their corresponding LRP scores, denoted as R_s. Next, the algorithm determines the correlation coefficients ρ_i between r_t and each r_i in R_s . To eliminate noise and irrelevant correlations, it computes the **z-scores** z_i based on the mean μ_ρ and standard deviation σ_ρ of the correlation values ρ_i. Finally, only significant correlations—identified using a predefined threshold z_{th}—are retained.

```
Input: Target model $\mathcal{M}$, target instance $x_t$, z-score threshold $z_{\text{th}}$, similar dataset $\mathcal{X}_s = \{x_1, x_2, \ldots, x_n\}$
Output: Unique LRP scores $\mathbf{r}_{\text{unique}}$ for $x_t$
function GetUniqueLRP($\mathcal{M}, x_t, z_{\text{th}}, \mathcal{X}_s$)
    Compute LRP scores for the target instance: $\mathbf{r}_t = \text{LRP}(\mathcal{M}, x_t)$
    Initialize $\mathcal{R}_s \leftarrow \emptyset$        ▷ Set of LRP scores for similar instances
    for each $x_i \in \mathcal{X}_s$ do
        Compute LRP scores: $\mathbf{r}_i = \text{LRP}(\mathcal{M}, x_i)$
        Add $\mathbf{r}_i$ to $\mathcal{R}_s$: $\mathcal{R}_s \leftarrow \mathcal{R}_s \cup \{\mathbf{r}_i\}$
    end for
    Compute correlations: $\rho_i = \text{corr}(\mathbf{r}_t, \mathbf{r}_i), \forall \mathbf{r}_i \in \mathcal{R}_s$
    Compute z-scores for correlations: $z_i = \frac{\rho_i - \mu_\rho}{\sigma_\rho}$, where $\mu_\rho = \frac{1}{n}\sum_{i=1}^{n}\rho_i$ and $\sigma_\rho = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\rho_i - \mu_\rho)^2}$
    if $|z_i| \geq z_{\text{th}}, \forall i$ then        ▷ Retain significant correlations
        $\mathbf{r}_{\text{unique}} \leftarrow \mathbf{r}_t$
    else
        $\mathbf{r}_{\text{unique}} \leftarrow -1$        ▷ Flag as conflicting relevance
    end if
    Return $\mathbf{r}_{\text{unique}}$
end function
```

Algorithm 2. Derive Unique LRP Scores for Target Model

For ROAM poisoning attacks, perturbations can be directly introduced into the identified critical neurons to compromise the model.

Validation and Results



(a) (b) (c) (d) (e)

Figure 7. Comparison of the ROAM poisoning attacks with the data poisoning

Poisoning Impact: The impact of data poisoning was compared with our LRP-based model poisoning attacks to evaluate their effectiveness. In this analysis, we applied targeted data poisoning by modifying all training dataset labels to the target classes 9 or 7, ensuring that a poisoned client consistently produces the same output regardless of the input. However, when this client was aggregated within a system of five clients, the poisoning effect was largely mitigated by the influence of the remaining models, as illustrated in Figure (a). The results indicate that data poisoning exhibits a stronger impact when most of the clients are compromised. As shown in Figure (b), when four out of five clients were poisoned, the attack significantly influenced the model. In contrast, our LRP-based poisoning requires only a single poisoner to achieve a similar effect. With a noise threshold of +0.2 applied to the weights of 20 neurons, a single poisoned model was sufficient to degrade performance, as shown in Figure (c). Even when the noise threshold was reduced to +0.02, it still lowered accuracy across most classes, even with just one poisoned client (Figure (d)). Additionally, coordinated LRP-based poisoning with four attackers (Figure (e)) produced results comparable to data poisoning. However, LRP-based poisoning is significantly more efficient, as it does not require additional data for training, making it a lower-cost yet highly effective alternative to conventional data poisoning techniques.

Defence Strategy: The defence for the attack can be made by considering the existing approaches such as *FoolsGold* [FYC+18], or *SHERPA* [SSW+24] where either the model parameters or XAI-based model outputs can be analysed to detect poisoning approaches. Especially, if a large noise level added to the weights of fewer nodes by the LRP-based poisoning, the anomalous models can be more accurately identified by model parameter analysis techniques like *FoolsGold*, as the variation of these parameters between benign and poisoned client is clearly anomalous. However, a more challenging situation can be observed when the noise range is applied over smaller quantities for higher number of nodes, where model prediction analysis-based techniques like *SHERPA* can be used to obtain a reasonable detection accuracy, yet the techniques like *Krum* [BEG+17] or *FoolsGold* may not detect them properly. We observed the detection accuracy of 50% for *Krum*, 24% for *FoolsGold*, and 64% detection accuracy for *SHERPA*. Thus, detection of such approaches may be feasible with XAI-based techniques like *SHERPA*.

Use within ROBUST-6G System Design

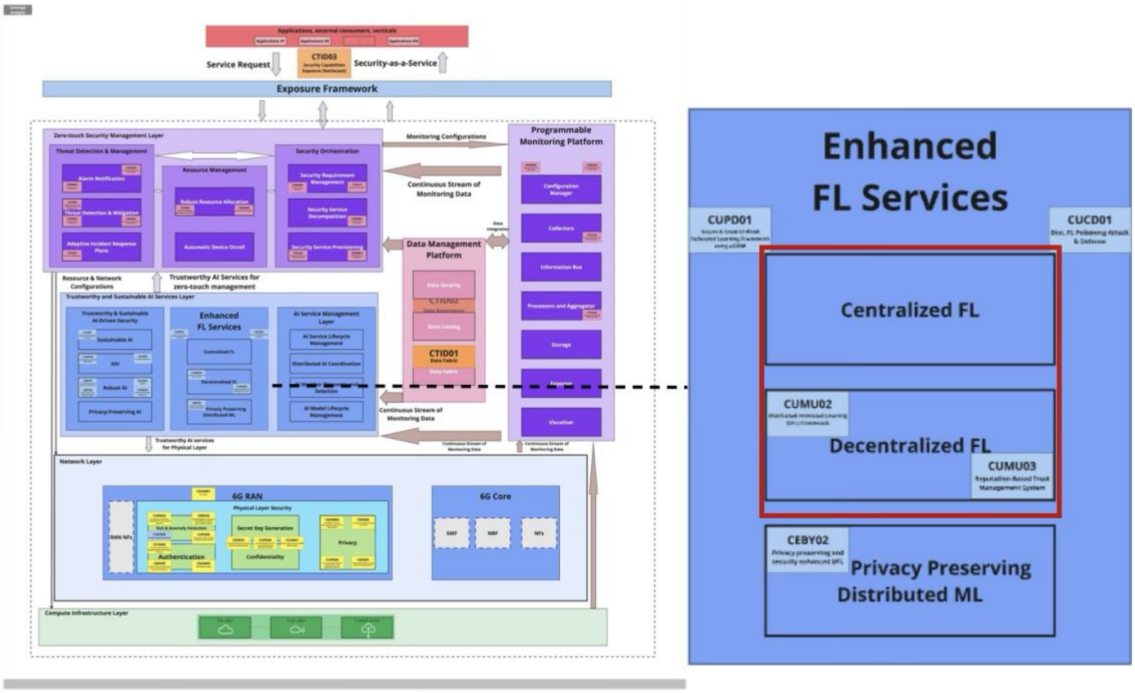

Applications, external consumers, verticals
Service Request
CTID03
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Alarm Notification
Threat Detection & Mitigation
Adaptive Incident Response Plans
Resource Management
Robust Resource Allocation
Automatic Device Onroll
Security Orchestration
Security Requirement Management
Security Service Decomposition
Security Service Provisioning
Monitoring Configurations
Continuous Stream of Monitoring Data
Programmable Monitoring Platform
Configuration Manager
Collectors
Information Bus
Processors and Aggregator
Storage
Exporter
Visualizer
Data integration
Data Management Platform
Data Security
CTID02
Data Catalog
CTID01
Data Fabric
Resource & Network Configurations
Trustworthy AI Services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Trustworthy & Sustainable AI-Driven Security
Sustainable AI
XAI
Robust AI
Privacy-Preserving AI
Enhanced FL Services
Centralized FL
Decentralized FL
Privacy Preserving Distributed ML
AI Service Management Layer
AI Service Lifecycle Management
Distributed AI Coordination
AI Model Lifecycle Management
Continuous Stream of Monitoring Data
Trustworthy AI services for Physical Layer
Network Layer
6G RAN
Physical Layer Security
RAN NFs
Authentication
Secret Key Generation
Confidentiality
Privacy
6G Core
SMF
NRF
NFs
Compute Infrastructure Layer
Enhanced
FL Services
CUPD01
CUCD01
Dist. FL Poisoning Attack & Defense
Centralized FL
CUMU02
Decentralized FL
CUMU03
Reputation-Based Trust Management System
CEBY02
Privacy preserving and security enhanced DFL
Privacy Preserving
Distributed ML

Figure 8. Integration of the poisoning attacks and defenses in ROBUST-6G architecture

To ensure the security of AI/ML training in 6G networks, the proposed ROAM poisoning attack model and poisoning defence strategies will be integrated to the enhanced FL services, as shown in Figure , where both Centralised and Decentralised FL models can be evaluated with the poisoning adversarial attacks for adversarial robustness testing and apply poisoning defence algorithms. The solutions will be provided as an independent and modular API interface which can be used as part of the testing process of use cases.

3.1.3 Secure and Decentralized Federated Learning Framework using ADMM

ADMM is a dual optimization method for convex constrained optimization based on the iterative optimization along directions defined by different variables, coupled with the update of the Lagrange multipliers enforcing constraints [BPC+11]. This method is universally known to be one of the fastest and theoretically solid convex optimization algorithms. However, its heuristic application to non-convex problems (e.g., non-convex objectives, non-convex constraints or mixed-integer programming) showed empirically that good approximate solutions can be retrieved fast, although with no guarantees to find the global optimum.

Recently, the ADMM has been used to enhance federated learning [ZL+23]. The authors propose FedADMM, an alternative to FedAvg and its enhancements where not only the aggregation rule is changed but also the local step. The clients do not perform stochastic gradient descent via backpropagation but solve an ADMM step that amounts to a local unconstrained optimization subproblem. Still, this requires an estimation of the gradient performed via forward pass of one (or multiple) batch(es) of data through the model. The method shows comparable or better performance in terms of accuracy concerning several popular benchmarks but is significantly faster in many contexts.

In the next paragraphs, we propose a different approach where not only the federated learning aggregation rule is tackled via ADMM but also the training of the model itself. In contrast to existing approaches, our original contribution is that of solving a *model-based* optimization problem that mimics the dynamics of a neural network, like previously done in [TBX+16], and at the same time inserting the model in a decentralized or split learning framework.

We propose an original model-based formulation of an optimization problem that can be tackled via consensus optimization and, among other algorithms, the ADMM. Let $i \in \{1, \dots, N\}$ be the client index and $l \in \{1, \dots, L\}$ be the neural network layer index. We define three sets of optimization variables, namely, W_l^i, z_l^i, and a_l^i, the weights, pre-activations, and post-activations at layer l and client i , respectively. We also need the (layer-specific) activation function h_l. A multi-layer perceptron model can be defined by the following optimization problem that seeks to minimize the loss $\ell(z_L, y)$dependent on the model output and the target labels, joint with the first two sets of constraints that define the model dynamics.

$$\begin{aligned} \min_{W_l, a_l, z_l} \quad & \ell(z_L, y) \\ \text{s.t.} \quad & z_l^i = W_l^i a_{l-1}^i \\ & a_l^i = h_l(z_l^i) \\ & W_l^i = W_l^j \end{aligned}$$

The third set of constraints, namely, $W_l^i = W_l^j \ \forall\, i \neq j$, is called the *consensus* constraint and ensures that the weights for each layer and for every couple of clients in the network are equal. Notably, the formulation also allows for a certain amount of personalization if this constraint is made soft (e.g., with an l_2-norm penalty added to the objective).

In this setting, like in classic FL, the dataset is split among the different clients. Every sample is linked with a different variable z and a , and clients do not need to share those variables. For the consensus constraint, instead, clients need to share a part of the information of the weights and of the Lagrange multipliers with 1-hop neighbors. This exposes the system to possible leakage of information and model inversion attacks. We will discuss possible countermeasures in the last paragraph of this section.

The ADMM updates of an approximate formulation relaxing part of these constraints can be found in closed form for several possibilities of the activation function, among which the commonly used ReLU is particularly simple if solved via an if-else logic.

Split learning with ADMM

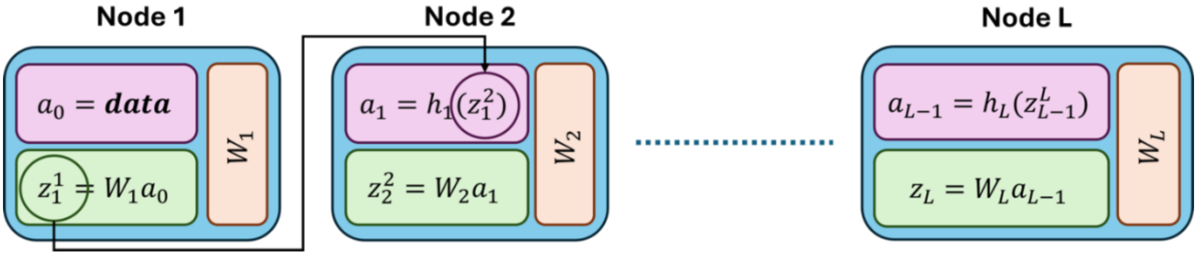


Figure 9. Split learning diagram

The case of split learning is instead a completely different scenario from a system perspective, but it can be tackled in a very similar way thanks to the ADMM framework. Using the same notation of the previous FL setting, now our goal is to study a system where the training is carried out collaboratively and layer-wise. That is, nodes possess only a (block of) layer(s) of the neural network and propagate the information through clients in an ordered way. For simplicity, we assume that each client only owns a single layer. The dataset is kept by the node that trains the input layer. The setting is depicted in Figure 9.

The associated optimization problem closely resembles the one of FL, as the model dynamics are unvaried. However, two details change: i) the owners of the variables, and ii) what variables form the consensus constraint.

$$\begin{aligned} \min_{W_l, a_l, z_l} \quad & \ell(z_L, y) \\ \text{s.t.} \quad & z_l^i = W_l a_{l-1} \\ & a_l = h_l(z_l^{i+1}) \\ & z_l^i = z_l^{i+1} \end{aligned}$$

Here, single copies of the model weights and post-activations are present, hence, there is no need to keep track of what other clients are doing in other layers. However, node $i+1$ needs the pre-activation value of i to propagate the value of the post-activation. Hence, the consensus constraint is, in this case, given by $z_l^i = z_l^{i+1}$. Depending on what client "owns" the activation function, the consensus constraint could also be put on the post-activations.

Remark 1. Notably, the two frameworks could also be combined in a setting where multiple nodes own the dataset and train the first (block of) layer(s) and following nodes train the rest of the neural network.

Remark 2. It is important to notice that, in standard SGD-derived FL, the objective function that is optimized is the (weighted) sum of the local cost function of clients. Since neural networks are highly non-convex, it is not guaranteed in general that the global minimum of the weighted sum of the local function is the same as the original global function supposing that a single entity owns the entire dataset. This problem is known as the objective inconsistency of FL and is particularly evident when the data distribution is non-iid across clients. *The formulation presented in these sections instead optimizes the global cost function, subject to model equality constraints, and therefore does not have any objective inconsistency.*

Use of differential privacy (DP) and homomorphic encryption

While the specific problem of training a neural network via model-based ADMM coupled with decentralized and/or split learning has never been tackled before with the proposed method to the best of our knowledge, the study of additional privacy and security mechanisms to the decentralized optimization with the ADMM is well established. For instance, in [ZMW+18] the authors propose the coupling of homomorphic encryption to the exchanged messages between peers. Homomorphic encryption (HE) enables additions and multiplications in the encrypted space without needing to decrypt the message, hence making unintelligible for attackers the content of the message (in our context, weights, pre-(post-)activations, and Lagrange multipliers). The addition of HE could significantly enhance the level of privacy and security of the system.

In the study [HGG+19] the authors focus on distributed learning via ADMM, a problem very close to ours, and show that the convergence of the ADMM is guaranteed even in the presence of differential privacy (DP).

DP consists in the addition of noise (Gaussian or Laplacian) to the updates to make them useless for attackers. It has been shown that, over multiple iterations, the convergence properties of the ADMM are kept even in the presence of DP.

Use within ROBUST-6G System Design

The distributed and split learning solution via ADMM described above will contribute to the development of the decentralized FL block as shown in Figure 10, which is useful for the trustworthy and sustainable AI service layer. The solution can be integrated into the ROBUST-6G architecture as a method to perform decentralized learning (local and aggregation rules), e.g., within UMU's DFL emulator. Decentralized learning is especially useful for the applications of the WP4, which must provide full automation and decentralized control of the network.

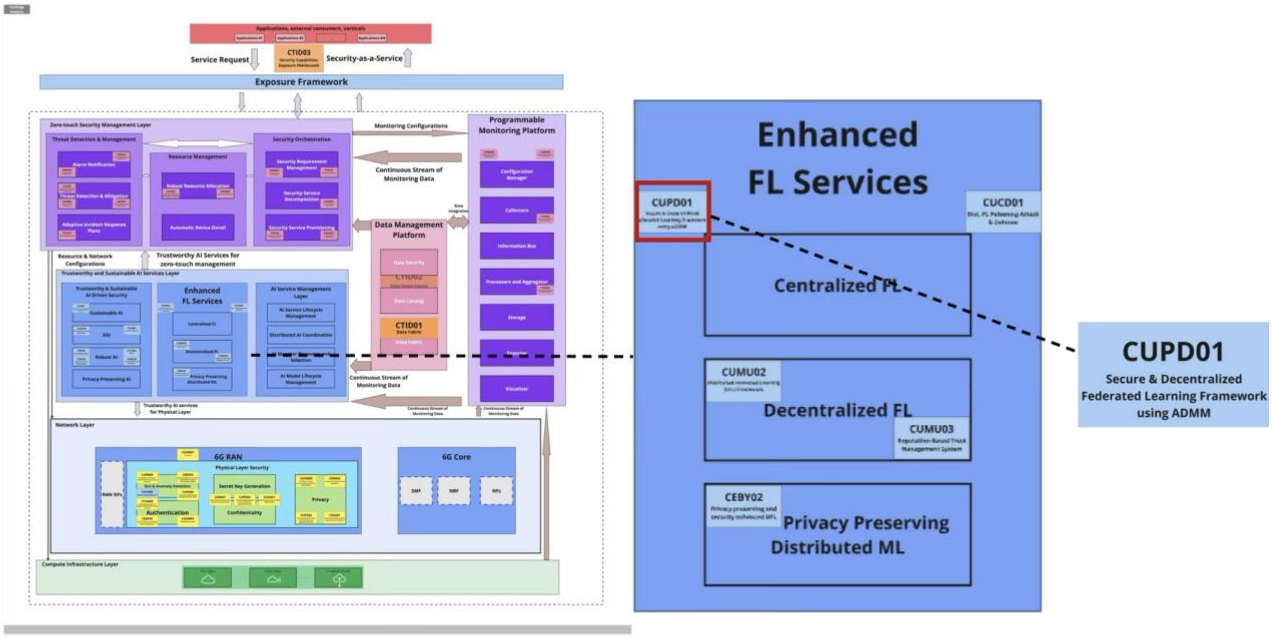

Applications, external consumers, verticals
Service Request
CTID03
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Alarm Notification
Threat Detection & Mitigation
Adaptive Incident Response Plans
Resource Management
Robust Resource Allocation
Automatic Device Onroll
Security Orchestration
Security Requirement Management
Security Service Decomposition
Security Service Provisioning
Monitoring Configurations
Continuous Stream of Monitoring Data
Programmable Monitoring Platform
Configuration Manager
Collectors
Information Bus
Processors and Aggregator
Storage
Exporter
Visualizer
Data Integration
Data Management Platform
Data Security
CTID02
Data Catalog
CTID01
Data Fabric
Resource & Network Configurations
Trustworthy AI Services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Trustworthy & Sustainable AI-Driven Security
Sustainable AI
XAI
Robust AI
Privacy-Preserving AI
Enhanced FL Services
Centralized FL
Decentralized FL
Privacy Preserving Distributed ML
AI Service Management Layer
AI Service Lifecycle Management
Distributed AI Coordination
AI Model Lifecycle Management
Continuous Stream of Monitoring Data
Trustworthy AI services for Physical Layer
Network Layer
6G RAN
RAN NFs
Physical Layer Security
Authentication
Secret Key Generation
Confidentiality
Privacy
6G Core
SMF
NRF
NFs
Compute Infrastructure Layer
Enhanced
FL Services
CUPD01
CUCD01
Dist. FL Poisoning Attack & Defense
Centralized FL
CUMU02
Decentralized FL
CUMU03
Reputation-Based Trust Management System
CEBY02
Privacy preserving and security enhanced DFL
Privacy Preserving
Distributed ML
CUPD01
Secure & Decentralized
Federated Learning Framework
using ADMM

Figure 10. Integration of Distributed and Split Learning Solution to the ROBUST-6G Architecture

3.2 Privacy Preserving AI Mechanisms for 6G Networks

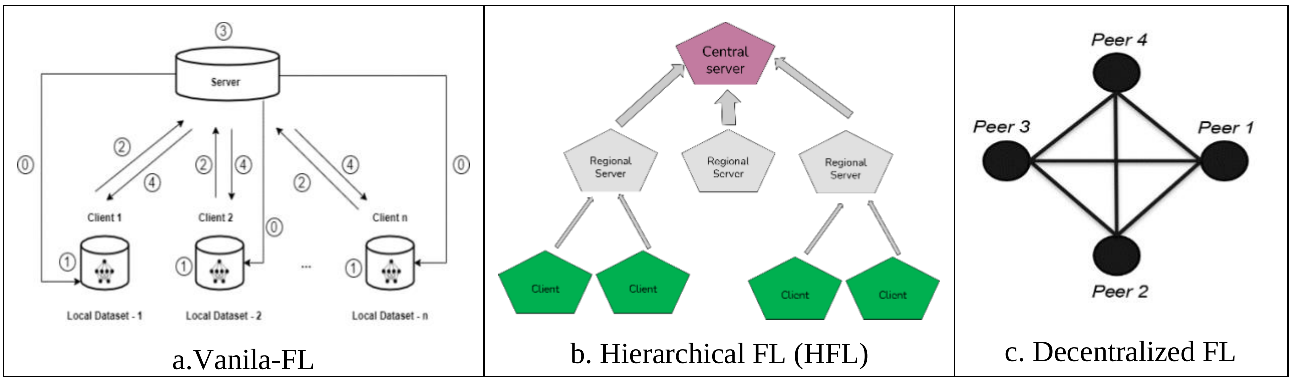
3.2.1 Privacy Preserving and Security Enhanced Federated Learning

Federated learning is a collaborative machine learning approach designed with privacy in mind, where multiple clients work together to build a global model. Each client performs local model training on its own data and then sends only the updated model parameters (or gradients) to a central server. This method offers privacy benefits, as it allows clients to collaborate without sharing their raw data directly. However, federated learning is still vulnerable to sophisticated privacy attacks, which can be conducted either by the aggregator such as Deep Leakage from Gradient attacks [ZLH+19], or by the participants such as poisoning attacks, inference attacks, model inversion attacks, etc [ZZW+22]. To mitigate these privacy risks, techniques such as differential privacy, secure aggregation protocols such as masking, secure multi-party computation, homomorphic encryption, and functional encryption are employed to conceal individual local model updates from the server while still allowing the aggregation of results. These protocols ensure that the server only receives aggregated model updates and cannot access the individual updates sent by the clients. While this enhances privacy, it creates a challenge where the server is unable to detect certain types of security threats, such as poisoning or backdoor attacks, which target the model during training. Since the server cannot analyse the individual local updates due to privacy protection, it is unable to identify abnormal patterns or inconsistencies that may arise from these attacks. This limitation underscores the need for solutions that balance both privacy and security in federated learning. Although there are a few solutions addressing both privacy and security in the literature in recent years, they have some drawbacks such as requiring two non-colluding servers, heavy cryptographic operations, or peer-to-peer communication topology.

One of the secure aggregation protocols that is used in FL is Homomorphic Encryption (HE), which is a type of encryption that enables third parties to perform arithmetic operations directly on encrypted data, or ciphertexts, without needing access to the plaintext. This property makes HE particularly valuable for privacy-preserving machine learning (PPML), especially in domains where the protection of sensitive data is a priority. In FL, homomorphic encryption ensures that the client model updates are encrypted before being sent to the server. The server could only aggregate the received encrypted model updates. Even if an attacker (or a trusted but curious server) intercepts the encrypted model updates, it will not be able to obtain the original data, because the secret keys for decryption are owned by clients. There are different types of HE. Partial Homomorphic Encryption (PHE) supports only specific types of operations, such as addition or multiplication, on encrypted data. In this context, the most basic form of PHE allows either addition or multiplication to be performed on encrypted values, but not both simultaneously. This limitation means that PHE can be used in applications where the computations involve only one type of arithmetic operation. For instance, Additive Homomorphic Encryption (AHE), which is a subtype of PHE, supports only addition operations on encrypted data. This is particularly useful in scenarios where adding encrypted data from different sources is necessary, but the confidentiality of the data needs to be preserved throughout the process. Fully Homomorphic Encryption (FHE) is another type of HE that supports a broader range of operations, allowing both addition and multiplication (and even more complex operations) to be performed on encrypted data. This advancement opens more possibilities for secure and privacy-preserving computation in fields like machine learning, where data privacy is a critical concern. There are different schemes for each type of HE. The reason we have different schemes in HE is because each scheme has a different focus and is designed to address different trade-offs between security, efficiency, and application requirements. As an example, in FHE, CKKS (Cheon-Kim-Kim-Song) and BGV (Brakerski-Gentry-Vaikuntanathan) serve different needs based on the nature of the data (real-valued vs. integer), the required precision of the result (approximate vs. exact), and the performance considerations.

In HE, both public and private keys are essential for performing operations on encrypted data. The public key is used to encrypt the data, while the private key is needed for decryption. These keys can be generated by a trusted authority or using Multi-party computations to ensure their integrity and security. However, the way these keys are distributed and managed across the different parties, varies depending on the trust model in place. In an FL scenario where the goal is to protect clients from a potentially malicious server, the central server may attempt to recover training data from the model updates it receives. In this setup, an explicit trust might be assumed among the clients, where the encryption/decryption keys are shared, but the server does not have access to any of these keys and is therefore blinded to the model updates. However, the trust model can vary in different scenarios. For instance, in federated learning, involving participants might be competing companies unwilling to let others gain insight into their proprietary training data. In this case, the secret keys should remain exclusive to each participant and not be shared between them.

In ROBUST-6G, as it is illustrated in Figure different scenarios for distribution of nodes in FL framework including Vanilla-FL, hierarchical FL, and decentralised FL are taken into consideration and we are at the stage of developing solutions to enhance both privacy and security for these FL scenarios separately.



a.Vanila-FL

b. Hierarchical FL (HFL)

c. Decentralized FL

Figure 11. Different FL scenarios

In Vanila FL, access to intermediate model updates from clients through malicious server may disclose sensitive information about local dataset of each client. Additionally, malicious client can disrupt the integrity of the training dataset or the model updates, thereby degrading the performance of the global model.

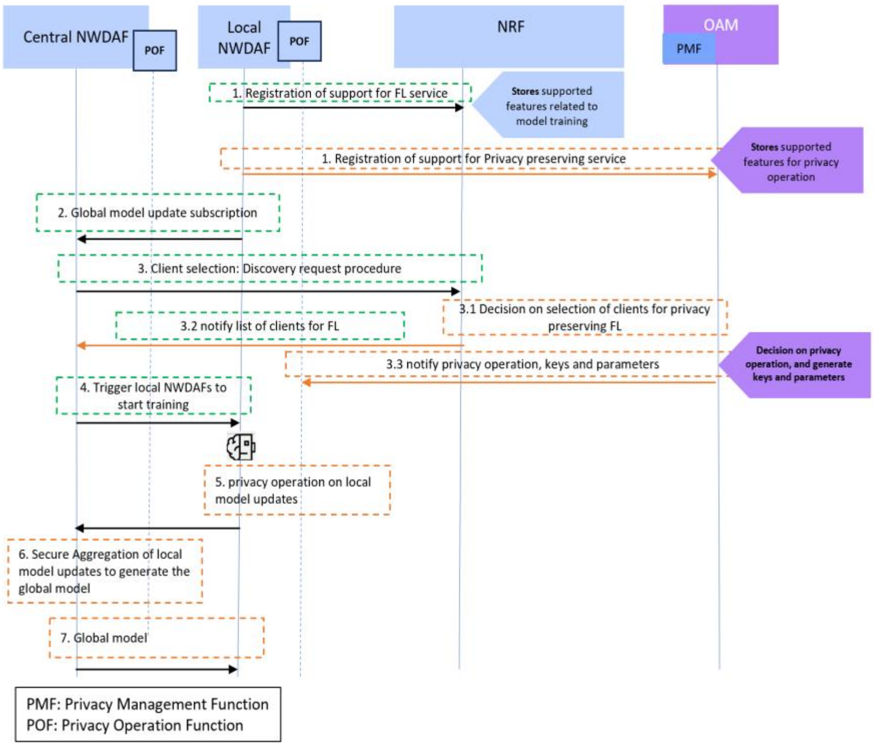
Hierarchical FL (HFL) is an approach where the training process is structured in multiple layers, involving central server, regional servers, and clients. HFL introduces an additional layer of communication, which can provide a few advantages. HFL may improve scalability by decreasing the amount of data that global server must handle at once and allowing more clients to participate in the learning process. HFL, can offer an extra layer of protection for privacy as the regional server can mask individual client updates before they are transmitted to the global server. This reduces the risk of sensitive data leakage in case of an attack. In case of failure at the global server, by enabling decentralized model aggregation, the regional servers can take over and generate the model. Thus, increases the robustness of the system and supports fault tolerance. However, HFL still suffers from privacy leakage by analyzing uploaded model parameters from clients or regional servers [SSZ+21]. Also, the malicious client or regional server can disturb the integrity of the training data or the model updates, leading the performance of the global model to be degraded.

On the other hand, DFL is a network architecture that removes the need for a central server, allowing clients to communicate directly with each other, leading to substantial savings in communication resources [YWS+24]. In DFL, the typical single point of failure problem found in traditional Vanila FL is mitigated. Since there is no central server, there is no single node whose failure can disrupt the entire learning process. This increases system robustness and makes the network more resilient to failures. Also, access to all clients' model update parameters by one entity is restricted depending on the selected topology. Despite these advantages, DFL may not be immune to the threat of privacy and poisoning attacks. Malicious clients can still inject harmful updates into the system, which could corrupt the global model and degrade its performance. Additionally, it can gain information about other user’s training data from the shared model update parameters. While decentralization helps address some vulnerabilities inherent in centralized systems, the challenge of protecting the system from malicious actors remains, and the design of DFL systems must include robust mechanisms to detect and mitigate these types of attacks.

Architectural Components

It is important to understand how the modules responsible for privacy-preserving operations, including Privacy-preserving AI and Privacy-preserving distributed ML in the AI service layer are interacting with other components in the architecture to provide privacy-preserving AI service. In Figure 12, the sample flow diagram for privacy-preserving federated learning between Network Data Analytic Functions (NWDAFs) in the 5G network is presented. NWDAF serves two main purposes. First, it acts as a service consumer, collecting data from Network Functions (NFs), who act as service providers. Second, it processes the data and provides analytics and predictions as a service provider to other NFs using Analytics and Prediction Exposure procedures. NWDAFs, which may be geographically distributed, may be willing to use other NWDAFs analytics to enhance their service, without sharing their model with others. They may benefit from FL service to be applied among different NWDAFS; however, for this service the privacy needs to be enhanced since model update parameters may disclose sensitive information about training data of each NWDAF. To apply FL process in privacy preserving manner, as illustrated in Figure 8, in step 1, the local NWDAFs register with Network repository Function (NRF) and ask for FL service. They provide features related to model training, such as current traffic load, if they have a Graphics Processing Unit (GPU) or which types of models they can train. Also, they are registered with OAM and informed which privacy operations they can use. In step 2, Local NWDAFs subscribe to central NWDAF, which is responsible for aggregating the model updates and generating the global model. In step 3, central NWDAF sends client selection discovery request to NRF to select appropriate clients for the privacy-preserving FL process. NRF may decide on the client list with the help of OAM and notify the candidate list to the central server. OAM, notify local NWDAFs about decided privacy operations, required key and parameters. In step 4, the central server triggers local NWDAFs to start training. At step 5, each local NWDAF starts training its own model and the Privacy Operation Function (POF), which can be integrated inside or outside NWDAF, applies privacy operation on the local model update. The privacy-preserved local model updates are shared with the central server at step 6, who securely aggregates them and shares the global model with the local NWDAFs. In this flow, NRF and OAM are responsible for coordinating the FL process and aligning privacy operations among clients.

While FL is already a well-established approach for distributed machine learning, decentralized FL might be an emerging extension that could become increasingly relevant in the 6G context as well. Privacy-preserving functionality is crucial in ensuring that local model updates are protected during this collaborative learning process. For the integration of privacy-preserving AI service in the ROBUST-6G architecture, multiple edge devices (e.g., smartphones, IoT devices) or network functions may request this service from privacy-preserving AI/Distributed ML according to the requirement and the use case, and the Security Orchestration in the Zero-touch service management layer could be responsible for generating, managing and distributing the required keys and parameters to support privacy-preserving functionality.


Central NWDAF
POF
Local NWDAF
POF
NRF
OAM
PMF
1. Registration of support for FL service
Stores supported features related to model training
1. Registration of support for Privacy preserving service
Stores supported features for privacy operation
2. Global model update subscription
3. Client selection: Discovery request procedure
3.1 Decision on selection of clients for privacy preserving FL
3.2 notify list of clients for FL
3.3 notify privacy operation, keys and parameters
Decision on privacy operation, and generate keys and parameters
4. Trigger local NWDAFs to start training
5. privacy operation on local model updates
6. Secure Aggregation of local model updates to generate the global model
7. Global model
PMF: Privacy Management Function
POF: Privacy Operation Function

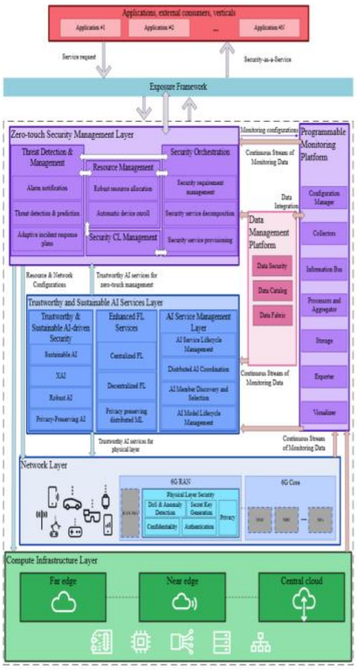

Applications, external consumers, verticals
Application #1
Application #2
...
Application #N
Service request
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Alarm notification
Threat detection & prediction
Adaptive incident response plans
Resource Management
Robust resource allocation
Automatic device enroll
Security CL Management
Security Orchestration
Security requirement management
Security service decomposition
Security service provisioning
Monitoring configurations
Continuous Stream of Monitoring Data
Data Integration
Programmable Monitoring Platform
Configuration Manager
Collectors
Information Bus
Processors and Aggregator
Storage
Exporter
Visualizer
Data Management Platform
Data Security
Data Catalog
Data Fabric
Resource & Network Configurations
Trustworthy AI services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Trustworthy & Sustainable AI-driven Security
Sustainable AI
XAI
Robust AI
Privacy-Preserving AI
Enhanced FL Services
Centralized FL
Decentralized FL
Privacy preserving distributed ML
AI Service Management Layer
AI Service Lifecycle Management
Distributed AI Coordination
AI Member Discovery and Selection
AI Model Lifecycle Management
Continuous Stream of Monitoring Data
Trustworthy AI services for physical layer
Continuous Stream of Monitoring Data
Network Layer
6G RAN
Physical Layer Security
DoS & Anomaly Detection
Secret Key Generation
Privacy
Confidentiality
Authentication
6G Core
Compute Infrastructure Layer
Far edge
Near edge
Central cloud

Figure 12. Sample Flow Diagram for Privacy-Preserving Federated Learning between NWDAFs in 5G network

3.2.2 Distributed Federated Learning Framework and Reputation Based Trust Management

A fully DFL framework has been developed in ROBUST-6G to eliminate the reliance on a central entity and align with the decentralized and autonomous characteristics of next-generation 6G networks. This paradigm will enable decentralized AI/ML while preserving data privacy, without sharing raw data among the multiple nodes to train the model and exclusively sharing model updates to converge towards a final, global AI model.

This framework enhances scalability, fault tolerance, and security by leveraging decentralized communication and decentralized model aggregation while ensuring seamless and trustworthy federated learning.

Architectural Components

The framework is designed with a modular and scalable architecture that enables fully decentralized training while ensuring security, reliability, and efficiency. The architecture is composed of key components that facilitate the management of the federated learning process, govern inter-node communication, and enable secure model aggregation without a central coordinating entity. These components work together to support privacy-preserving model training, optimize resource utilization, and mitigate risks associated with adversarial activity in a distributed setting as shown in Figure .

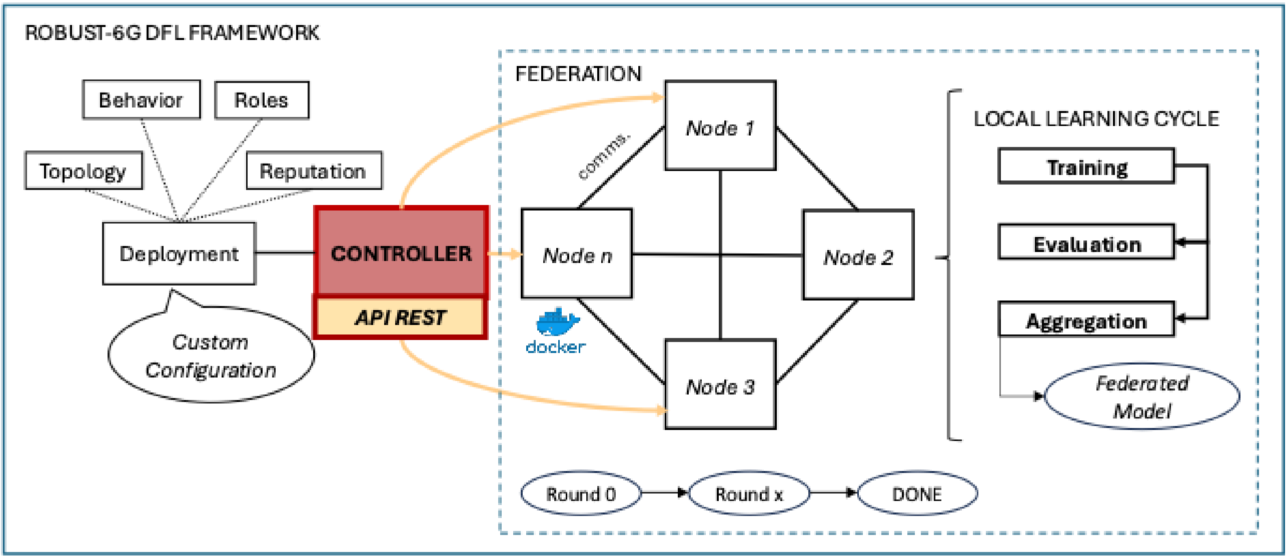


Figure 13. DFL Framework Overview

Core Framework Components

The framework consists of four fundamental components: the topology module, the deployment controller, the federated learning nodes, and the global model repository. Each of these components plays a crucial role in enabling decentralized learning while ensuring resilience against failures and adversarial threats.

The topology module defines the communication structure between participating nodes in the federation. The way nodes interact directly affects the efficiency of model aggregation and the overall performance of the learning process. The framework supports multiple topology configurations to provide flexibility for different deployment scenarios.

- Ring Topology: Each node exchanges updates with a single neighbor, forming a circular sequence. This structure reduces communication overhead but may slow down convergence.
- Fully Connected Topology: Each node communicates with all other nodes in the federation. While this approach accelerates convergence, it increases communication costs and computational overhead.
- Custom Topologies: Customizable network structures allow for topology designs tailored to specific deployment constraints and performance objectives. These include hierarchical topologies, or hybrid approaches where nodes interact based on trust scores or performance metrics.

The topology module ensures that nodes can communicate effectively while maintaining privacy and optimizing learning efficiency. The ability to define flexible communication structures is critical for adapting the framework to different network conditions and application domains.

The deployment controller is responsible for initializing the federation, orchestrating the interaction between nodes, and enforcing security policies. It defines key characteristics such as:

- The number of participating nodes.
- The communication topology.
- The machine learning model architecture and hyperparameters.
- The security mechanisms, including encryption protocols for model updates.

The controller integrates a script-based automation process, ensuring a seamless deployment of federated learning scenarios. Through a RESTful API, external applications can interact with the deployment controller to modify configurations dynamically, scale the number of participants, or adjust model training parameters.

Each federated learning node operates independently, training its model locally and participating in the distributed learning process. Unlike traditional FL setups where model updates are sent to a central server, nodes in this framework exchange updates directly with their peers based on the defined topology. Each node performs the following operations:

- Model Training: Local datasets are used to update the model parameters while preserving privacy.
- Model Evaluation: The node assesses the model’s performance using locally available validation data before sharing updates.

- Decentralized Aggregation: Instead of relying on a central server, nodes share their updates with designated peers, contributing to a global learning process.

The decentralized nature of these nodes eliminates central points of failure, enhances robustness against attacks, and ensures that learning can continue even if some nodes drop out of the federation.

A global model repository serves as a structured storage system for tracking model evolution throughout multiple training rounds. While the learning process remains decentralized, the repository provides an interface to:

- Store intermediate and final model versions for auditing and analysis.
- Enable retrieval of historical model snapshots to analyze training progression.
- Support explainability tools that assess how different models evolve over time.

This repository allows users to extract models at different stages, apply external evaluation techniques, and conduct post-training analysis to ensure fairness, robustness, and explainability.

Communication and Training Workflow

The communication and training workflow in this framework follows a decentralized approach, removing the need for a central aggregator while maintaining the efficiency of collaborative learning. The process is structured around a decentralized model update exchange mechanism, where nodes train models locally and share updates directly with their peers according to the defined topology.

1. Each node trains a local model using its private dataset, ensuring data privacy and preventing raw data from being transmitted. The model undergoes multiple iterations of training based on the node's computational capacity and data availability.
2. Before exchanging model updates, each node evaluates its locally trained model against a validation dataset. This ensures that only meaningful updates are shared within the federation. Nodes that exhibit erratic learning behavior may have their updates flagged for further analysis.
3. Instead of submitting model updates to a central server, each node communicates with designated peers based on the defined topology. The framework supports multiple **aggregation strategies** to ensure effective model convergence:
   a. **Weighted averaging (FedAvg):** Combines updates based on dataset size, ensuring that contributions from nodes with larger datasets have a greater influence.
   b. **Adaptive aggregation:** Dynamically adjusts the number of exchanged updates based on node trust scores, ensuring that only reliable updates are integrated.
4. The process of **training, evaluating, and aggregating** continues iteratively over multiple rounds until the federation reaches a predefined convergence threshold. At each round, model updates are refined through peer-to-peer exchanges, improving overall model performance without requiring centralized control.

Security and Privacy Considerations

Security and privacy are central to the effectiveness of any federated learning system. In a decentralized setup, ensuring the confidentiality, integrity, and authenticity of model updates is critical to preventing adversarial manipulation.

All communications between nodes are secured using AES and RSA encryption, ensuring that model updates remain confidential during transmission. The integration of synchronous communication mechanisms guarantees that updates are only exchanged between verified peers.

The trust management system, which is based on reputation, is a critical component for safeguarding the integrity of the training process. This system continuously evaluates the quality and consistency of each node's contributions, assigning reputation scores that reflect their reliability. Key mechanisms implemented include:

- Continuous Contribution History and Evaluation: Each node is monitored over multiple training rounds, with its contributions being recorded and analyzed to determine their positive impact on the overall process. Consistent, well-evaluated updates lead to an increase in reputation, while anomalous or erratic behaviors result in penalties.
- Reciprocal Evaluation and Dynamic Feedback: Nodes mutually share and validate their model updates. This reciprocal evaluation process helps detect deviations or manipulation attempts, allowing

the reputation scores of individual nodes to be adjusted in real time. As a result, only reliable nodes significantly influence the global knowledge generated during the federation.

RESTful API and Interaction Mechanisms

The framework incorporates a RESTful API to facilitate seamless interaction between system components, external applications, and monitoring tools. This API-driven design ensures modularity, extensibility, and real-time control over the federation.

The API follows a RESTful architecture, using standard HTTP methods (GET, POST, DELETE) for data retrieval, submission, and federation management. The following categories define the API's primary functionalities:

1. Federation Deployment: Enables the initialization of a decentralized learning federation with user-defined configurations.
2. Model Retrieval and Tracking: Allows users to extract models from different training rounds and nodes.
3. Federation Monitoring: Provides real-time training metrics, node reputation tracking, and security insights.
4. Attack Simulation: Supports controlled adversarial testing to evaluate model robustness.

Endpoint #1: Deploy a Federation

The federation deployment endpoint initializes the distributed learning environment, specifying key parameters such as:

- Topology Type: Defines the communication structure (e.g., ring, fully connected, trust-based).
- Number of Nodes: Determines the scale of the federation.
- Model Architecture: Specifies the neural network structure and hyperparameters.
- Security Settings: Configures encryption mechanisms and trust policies.

Deployment is triggered by sending a POST request to the federation controller, ensuring that all nodes are initialized with the correct settings before training begins. This automation reduces setup complexity, minimizes configuration errors, and enables reproducibility in federated learning experiments.

Endpoint #2: Extract Federated AI Models

Model extraction is crucial for tracking training progress, analyzing intermediate updates, and performing external evaluations. This endpoint provides structured access to:

- Per-node model retrieval: Allows inspection of individual node contributions.
- Round-based retrieval: Facilitates analysis of model progression over time.
- Final model access: Extracts the trained model after the completion of training rounds.

Endpoint #3: Obtain Federation Metrics

Real-time monitoring is essential for assessing training efficiency, fairness, and trustworthiness in a federated learning setup. The federation metrics endpoint provides:

- Model performance statistics: Tracks accuracy, loss, and convergence rates.
- Node-level contribution analysis: Identifies active and underperforming participants.
- Trust and reputation insights: Displays how nodes' reputation evolves over time.

Use within ROBUST-6G System Design

The DFL framework described above will contribute to the improvement and assessment of the trustworthiness of the AI models generated within the federation of nodes. It is positioned within the Trustworthy & Sustainable AI-driven Security domain, which is part of the Trustworthy & Sustainable AI Services Layer. The DFL framework is also located within this layer, more specifically in Enhanced FL Services, as shown in Figure . This framework is needed for Use Case 1, Scenario 1.

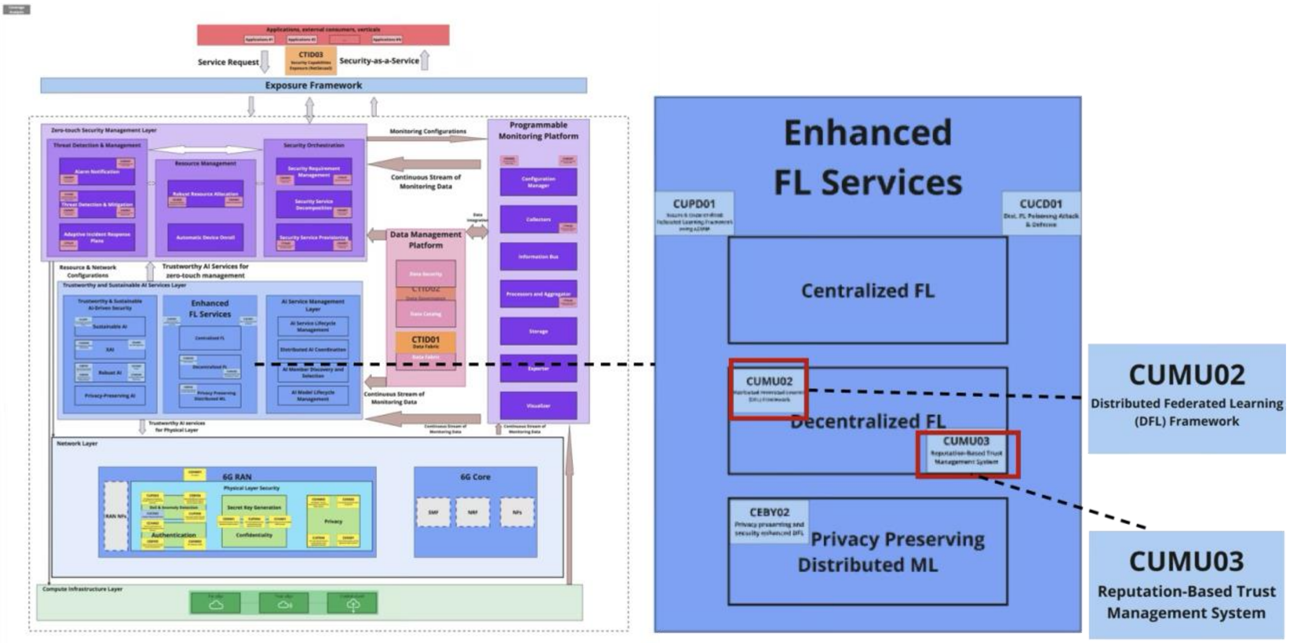

Applications, external consumers, verticals
Service Request
CTID03
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Alarm Notification
Threat Detection & Mitigation
Adaptive Incident Response Plans
Resource Management
Robust Resource Allocation
Automatic Device Onroll
Security Orchestration
Security Requirement Management
Security Service Decomposition
Security Service Provisioning
Monitoring Configurations
Continuous Stream of Monitoring Data
Programmable Monitoring Platform
Configuration Manager
Collectors
Information Bus
Processors and Aggregator
Storage
Exporter
Visualizer
Data Integration
Data Management Platform
Data Security
CTID02
Data Catalog
CTID01
Data Fabric
Resource & Network Configurations
Trustworthy AI Services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Trustworthy & Sustainable AI-Driven Security
Sustainable AI
XAI
Robust AI
Privacy-Preserving AI
Enhanced FL Services
Centralized FL
Decentralized FL
Privacy Preserving Distributed ML
AI Service Management Layer
AI Service Lifecycle Management
Distributed AI Coordination
AI Member Discovery and Selection
AI Model Lifecycle Management
Continuous Stream of Monitoring Data
Trustworthy AI services for Physical Layer
Network Layer
6G RAN
Physical Layer Security
RAN NFs
DoS & Anomaly Detection
Authentication
Secret Key Generation
Confidentiality
Privacy
6G Core
SMF
NRF
NFs
Compute Infrastructure Layer
Enhanced
FL Services
CUPD01
CUCD01
Centralized FL
CUMU02
Decentralized FL
CUMU03
CEBY02
Privacy Preserving
Distributed ML
CUMU02
Distributed Federated Learning (DFL) Framework
CUMU03
Reputation-Based Trust Management System

Figure 14. Integration of DFL Framework with ROBUST-6G Architecture

3.3 XAI Strategies for Transparent and Robust AI in 6G

3.3.1 XAI Approaches for Trustworthiness and Model Explainability

XAI has rapidly become an essential area within AI, driven by the increasing demand for transparency and accountability in AI systems [GA+19]. By offering clear explanations of decision-making processes in AI/ML models, XAI encourages trust and dependability, making AI-powered solutions more interpretable and defensible. In the context of the ROBUST-6G framework, XAI acts as a key element, ensuring that security mechanisms remain explainable, adaptable, and resilient. By incorporating XAI into the AI-based ROBUST-6G security orchestration, the project aims to create a transparent, self-evolving, and reliable security framework. This integration improves the intelligence, adaptability, and dependability of 6G networks, equipping them with strong defences against rapidly advancing cyber threats.

Some traditional ML models, such as decision trees and linear regression, are inherently interpretable, although their predictive power is limited [CPC+19]. Meanwhile, contemporary ML algorithms, notably deep learning models, have achieved outstanding performance in diverse applications. However, their opaque nature requires the incorporation of XAI techniques to explain and make sense of their sophisticated decision-making processes. On the other hand, explanation methods can be classified as model-specific and model-agnostic. Model-specific techniques are applied to specific models or groups of models, allowing for an understanding of decisions by investigating their underlying mechanisms, such as explaining coefficient weights in neural networks. Model-agnostic methods, alternatively, study the relation between input and output variables without having structure of the model, allowing for generalization across multiple models. These approaches are further separated into two types: global, which describe overall model behaviour, and local, which explain individual predictions and help to understand the reasons determining specific decisions. The two well-known model-independent explanation strategies in the literature are Local Interpretable Model-Agnostic Explanations (LIME) [RSG+16] and Shapley Additive Explanations (SHAP) [LL+17]. LIME provides localized explanations by analysing model predictions with varying input data. The procedure entails generating a new data set consisting of perturbed samples and the corresponding predictions of a black-box model. Afterward, local surrogate models are trained using this new dataset to approximate the predictions made by the underlying black box model. This surrogate model attempts to accurately approximate the prediction of the AI model within a particular local context. Alternatively, SHAP is a specialized methodology designed to provide both local and global explanations. It is based on the concept derived from cooperative game theory, the SHAP framework utilizes Shapley values as a method to understand the reasoning behind individual predictions. The values represent the mean marginal impact of each feature, where the feature values of a specific data instance can be likened to coalition members. As a result, SHAP serves the purpose of shedding light on the importance and function of each characteristic in the prediction process, incorporating a systematic and analytically rigorous method of explanation.

In the context of the ROBUST-6G, the advancement of XAI techniques not only strengthens the definition of security-related AI models but also facilitates the identification and mitigation of failures and errors. By identifying the factors that influence AI-based security decisions, XAI supports understanding, validating, and improving the results of AI-based security systems. Furthermore, the multidisciplinary nature of cybersecurity requires a different approach to XAI that considers the unique challenges and requirements of security environments. Model-specific explanation techniques are mapped to the complexity of security-focused ML models that provide a deeper understanding of the decision-making processes involved in threat detection, and IDS. However, model-independent methods provide a versatile framework for analysing relationships between input characteristics and safety-related outcomes in different ML models, providing general perspectives on safety and risk vulnerability.

Use within ROBUST-6G System Design

Figure demonstrates the incorporation of the XAI component within a ROBUST-6G architecture to support reliable and sustainable AI-driven security. The sections emphasized show where XAI components can be effectively implemented to enhance security operations, interpretability, and robustness. The XAI component is integral to the Trustworthy AI framework, supporting both Robust AI and Privacy-Preserving AI. This component facilitates better decision making and transparency in identifying adversarial attacks and ensuring resilient security defences. XAI also assists in Threat Detection & Mitigation, enhancing the precision of identifying adversarial attacks. It may contribute to adaptive incident response plans by offering interpretable insights into potential threats, refine alarm notification systems by decreasing false positives, and improve detection reliability. Additionally, XAI might help analyse and explain Denial of Service (DoS) and Anomaly Detection alerts. Enhances Authentication Mechanisms by ensuring secure access through the identification of potential adversarial authentication attempts. In general, the figure underscores the role of XAI in improving security, providing interpretability, and improving decision-making across different network security layers.

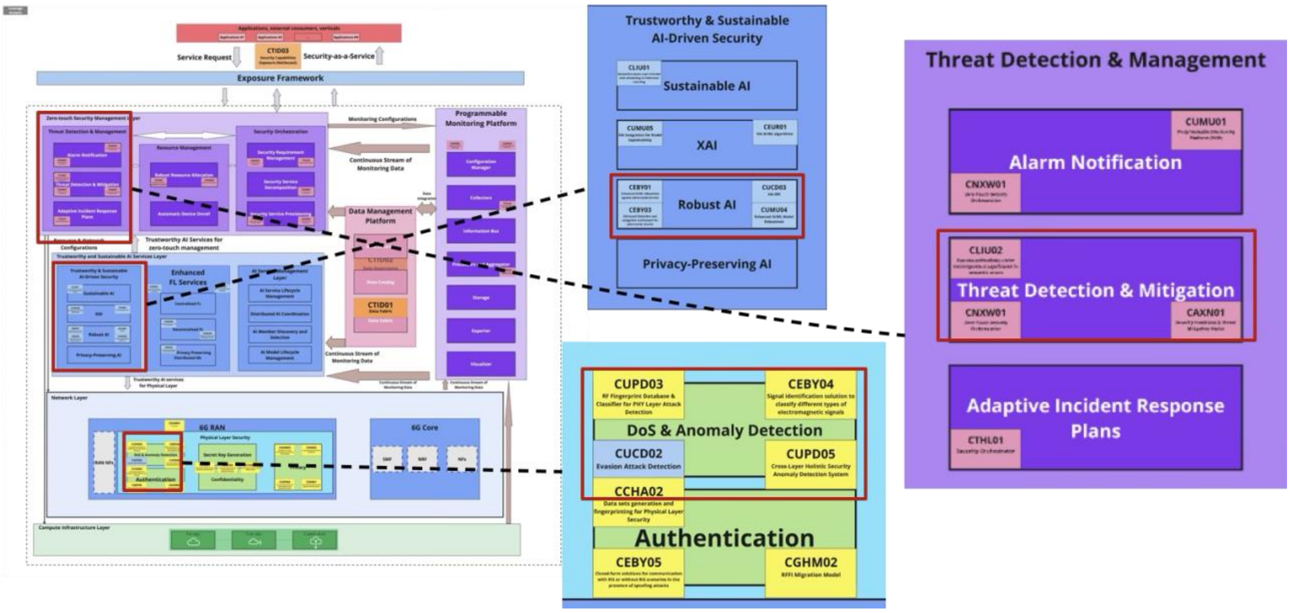

Service Request
CTID03
Security-as-a-Service
Exposure Framework
Programmable Monitoring Platform
Data Management Platform
CTID01
Enhanced FL Services
Network Layer
6G RAN
6G Core
Authentication
Compute Infrastructure Layer
Trustworthy & Sustainable AI-Driven Security
CLIU01
Sustainable AI
CUMU05
CEUR01
XAI
CEBY01
CUCD03
Robust AI
CEBY03
CUMU04
Privacy-Preserving AI
Threat Detection & Management
CUMU01
Alarm Notification
CNXW01
CLIU02
Threat Detection & Mitigation
CNXW01
CAXN01
Adaptive Incident Response Plans
CTHL01
Security Orchestrator
CUPD03
RF Fingerprint Database & Classifier for PHY Layer Attack Detection
CEBY04
Signal identification solution to classify different types of electromagnetic signals
DoS & Anomaly Detection
CUCD02
Evasion Attack Detection
CUPD05
Cross-Layer Holistic Security Anomaly Detection System
CCHA02
Data sets generation and fingerprinting for Physical Layer Security
Authentication
CEBY05
CGHM02
RFFI Migration Model

Figure 15. Incorporation of the XAI component within a ROBUST-6G architecture

3.3.2 Robust Continual Learning with Conformal Prediction

The issue of AI's trustworthiness has never been more crucial. As AI systems are deployed in sensitive and high-stakes scenarios, ensuring their reliability and preventing unintended harm is paramount. However, if an AI system experiences catastrophic forgetting—losing the ability to recall previous tasks while learning new ones—it may behave unpredictably, increasing the risk of critical errors or even harm. Therefore, tackling catastrophic forgetting is not merely a matter of improving AI performance; it is a fundamental step toward strengthening trustworthiness and ethical dimensions of AI.

Catastrophic forgetting remains a major challenge in trustworthy AI, especially in dynamic environments where machine learning models must adapt without losing previously learned knowledge. Traditional AI systems, designed for static learning, struggle in real-world applications such as healthcare, autonomous systems, and wireless communications, where models must continually learn new tasks. Existing continual learning (CL) techniques mitigate forgetting to some extent, but they often fail to quantify predictive uncertainty, leading to overconfident and unreliable decisions.

To address this, we introduce the Conformal Prediction Confidence Factor (CPCF), a metric that integrates conformal prediction (CP) into continual learning. Unlike conventional accuracy-based evaluations, CPCF quantifies how confident a model remains in previously learned tasks, allowing for a dynamic and interpretable assessment of catastrophic forgetting without requiring ground truth labels. This makes it well-suited for privacy-sensitive and real-time AI applications, where storing past data is impractical.

Our framework is based on Adaptive Conformal Prediction [AB+23], which evaluates confidence degradation as new tasks are introduced. The methodology consists of three phases: data splitting, calibration, and prediction.

1. Data Splitting: The dataset is divided into training data and calibration data, determined by a predefined calibration ratio. A higher calibration ratio improves conformal score reliability but reduces available training data.
2. Calibration Phase: We compute conformal scores by ranking softmax probabilities and identifying the minimum probability mass required to include the correct label. These scores are used to calculate the quantile threshold q_alpha which determines the confidence level for forming prediction sets.
3. Prediction Phase: At inference, prediction sets are formed for each test sample by including classes whose cumulative probability meets or exceeds q_alpha . The size of these sets quantifies confidence—shorter sets indicate high certainty, while longer sets suggest increased uncertainty and potential forgetting.

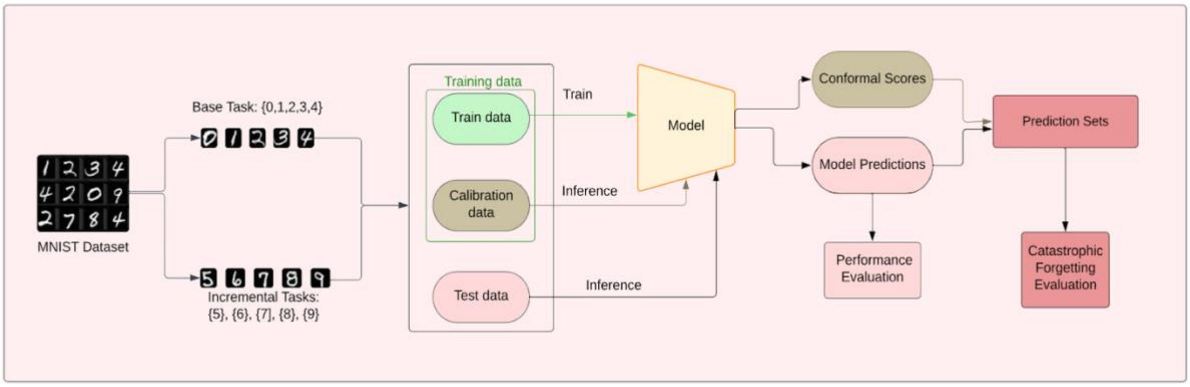


Figure 16. Framework for evaluating catastrophic forgetting in continual learning using conformal prediction. The process includes training on the MNIST dataset, computing conformal scores from calibration data, forming prediction sets, and evaluating forgetting.

The CPCF metric is computed by averaging the lengths of prediction sets for previous tasks after training on a new task. A low CPCF means strong knowledge retention, while an increasing CPCF signals forgetting and reduced confidence. This process is illustrated in Figure 16, which outlines how conformal prediction is applied in a continual learning setting.

Experimental Setup and Validation Results

To validate our approach, we use the MNIST dataset in an incremental class learning setting. A Multi-Layer Perceptron (MLP) with three fully connected layers is trained in two stages: base training on digits 0-4, followed by incremental learning where digits 5-9 are introduced sequentially. A low learning rate (2×10^{-5}) is chosen to prevent catastrophic forgetting by ensuring gradual adaptation while maintaining previously acquired knowledge.

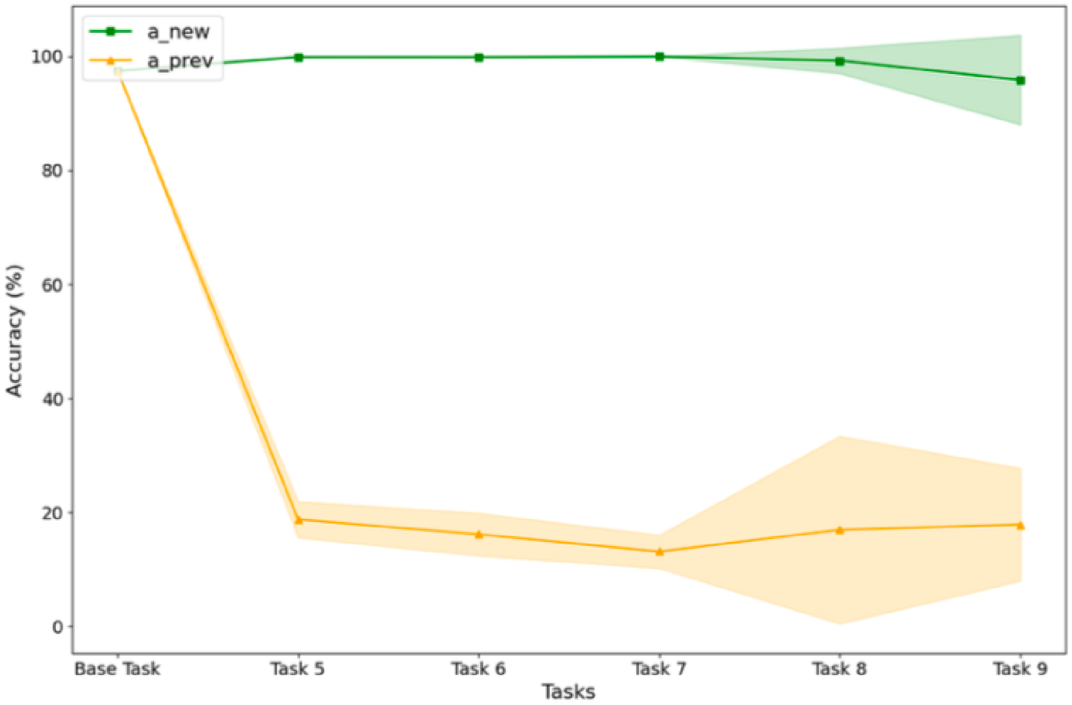


Figure 17. Illustration of Catastrophic Forgetting: Comparison of accuracy metrics for previously learned tasks (a_prev) and newly learned tasks (a_new) across incremental tasks

To assess catastrophic forgetting, we compare CPCF with the accuracy of previously learned tasks a_prev. Figure 17 illustrates how accuracy on prior tasks drops as new tasks are introduced, confirming catastrophic forgetting. Our results show a strong correlation between increasing CPCF values and declining accuracy, reinforcing CPCF as a reliable measure of forgetting.

Calibration Ratio	**Spearman Corr. (%)**	**Spearman p-value**	**Pearson Corr. (%)**	**Pearson p-value**
0.05	-51.40	1.3461×10^{-4}	-51.63	1.2416×10^{-4}
0.10	-51.75	3.5179×10^{-8}	-53.16	1.2638×10^{-8}
0.15	-51.18	1.4544×10^{-4}	-52.65	8.5758×10^{-8}
0.20	-50.68	1.7270×10^{-4}	-51.58	1.2635×10^{-4}

Table 5 Correlation between CPCF and a_prev for fixed significance level alpha ($\alpha = 0.1$) and varying calibration ratios

We further analyse CPCF's robustness by varying the calibration ratio and significance level α. As seen in Table 5 Correlation between CPCF and a_prev for fixed significance level alpha ($\alpha = 0.1$) and varying calibration ratios, CPCF remains stable across different calibration ratios, indicating its reliability regardless of how training data is proportioned. However, adjustments to α influence CPCF's sensitivity to model uncertainty. A higher α lowers the quantile threshold, leading to larger prediction sets and increased granularity in detecting uncertainty. This adaptability makes CPCF effective across diverse learning scenarios.

The ability to detect and mitigate catastrophic forgetting is essential for trustworthy AI. As AI becomes more integrated into critical applications like healthcare, autonomous systems, and wireless networks, ensuring that models retain knowledge while adapting to new tasks is key to reliability and transparency. The CPCF metric, rooted in conformal prediction, offers a dynamic and interpretable approach to measuring model confidence, providing an alternative to traditional accuracy-based evaluations. By eliminating the need for stored ground truth labels, CPCF is well-suited for real-time continual learning in privacy-sensitive environments.

Our results demonstrate that CPCF effectively captures catastrophic forgetting, aligning with accuracy-based assessments while providing deeper insights into model uncertainty. Future work could extend CPCF beyond classification tasks to regression problems, memory replay techniques, and multimodal datasets, further enhancing its utility in real-world AI applications.

3.3.3 XAI-IDS: Explainable AI-Based Intrusion Detection System

With the rapid evolution of cyber threats, modern network systems face unprecedented risks that challenge their resilience and security infrastructure. While ML-based Intrusion Detection Systems (IDS) have significantly improved threat detection, they often prioritize predictive accuracy at the expense of interpretability. In recent years, XAI techniques have gained traction across various domains for providing insights into model behavior. However, their application in IDS research remains largely limited to post-hoc explanations, focusing solely on interpreting the outputs of pre-existing models.

The integration of XAI methods, such as SHAP and LIME, presents opportunities for improved model transparency and trustworthiness; however, these methods must overcome challenges related to computational complexity, stability, and potential vulnerabilities to adversarial manipulation.

Although significant research has been dedicated to integrating XAI with IDS, most studies have primarily emphasized interpretability, transparency, and fostering user trust. Several existing frameworks predominantly apply XAI techniques to offer post-hoc explanations for black-box model predictions rather than embedding explainability directly within the IDS development process [AGR+24].

Furthermore, feature reduction approaches leveraging XAI methods have been proposed to enhance IDS efficiency. However, these methodologies typically achieve only marginal improvements in detection accuracy. [BBA+23] investigated explainability methods to identify significant features for LSTM-based models detecting DDoS attacks within CIC datasets, achieving only minor performance enhancements despite reduced feature sets. Similarly, [KKP+23] utilized a combination of SHAP, PFI, ICE, and PDP explanations (collectively SPIP) with their LSTM-based IDS. Although feature reductions were achieved through these techniques, subsequent model retraining with reduced feature subsets yielded minimal improvements, highlighting the persistent challenge of achieving meaningful performance gains and underscoring the necessity for automated analytical frameworks rather than manual feature analysis.

Despite notable progress, significant challenges persist within the XAI-integrated IDS domain. One recurring issue is balancing interpretability against predictive accuracy, as models that are inherently interpretable often compromise detection effectiveness. Moreover, scalability and generalizability remain limited across diverse network environments and varied attack scenarios. For instance, [ZRL+22] restricted their XAI integration exclusively to DNS over HTTPS (DoH) attack detection, reflecting the prevalent limitation of domain-specific solutions. Thus, a clear requirement exists for an innovative approach that inherently integrates explainability throughout the IDS modelling lifecycle, simultaneously maintaining robust predictive performance.

To address these gaps, this section proposes a novel methodology utilizing XAI techniques specifically for feature selection, moving beyond conventional retrospective explainability approaches. This novel approach ensures substantial and genuine enhancements in detection rates, demonstrable through comprehensive comparative evaluations before and after feature selection. The proposed method uniquely combines effective interpretability with significant model performance improvements, offering a practical balance between explainability and accuracy. Furthermore, its applicability across multiple domains, including emerging fields such as 5G and 6G networks, and various datasets with distinct features and attack types, highlights its versatility and broad relevance in the evolving threat landscape.

Framework Components

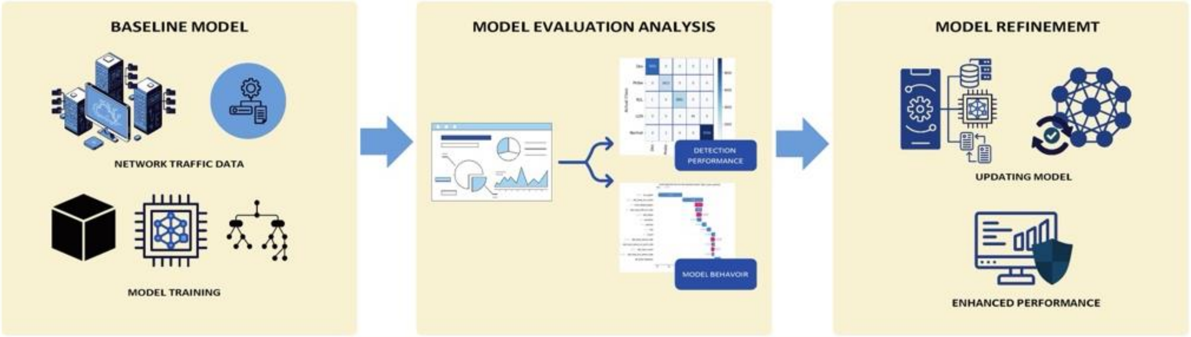


Figure 18. Proposed XAI-based Feature Refinement Pipeline in Intrusion Detection Framework

The proposed framework, as shown in Figure , focuses on securing the training and execution of AI/ML models specifically designed for intrusion detection within 6G network environments. Our methodology involves an explainable and robust Intrusion Detection System (XAI-IDS) that utilizes Explainable AI techniques not merely for post-hoc interpretation but integrally within the model training and evaluation phases. Initially, the NSL-KDD dataset is utilized due to its comprehensive structure and wide acceptance for IDS benchmarking. This dataset includes three subsets: KDDTrain+ for training, KDDTest+ for standard evaluation, and KDDTest-21 for advanced assessment. It categorizes network traffic into five classes—Normal, Probe, Denial of Service (DoS), User to Root (U2R), and Remote to Local (R2L)—capturing diverse attack patterns. Each instance comprises 41 features, subsequently expanded to 122 features following preprocessing via one-hot encoding of categorical attributes (protocol_type, flag, and service).

The core ML component employs a XGBoost classifier. The multiclass XGBoost classifier is chosen due to its robustness in managing heterogeneous and imbalanced datasets, as well as its proven performance in multiclass scenarios. XGBoost builds an ensemble of decision trees sequentially, where each tree corrects errors from the previous iteration, iteratively refining the model. This ensemble-based gradient boosting approach offers strong predictive accuracy and computational efficiency. Moreover, XGBoost's probabilistic output (multi:softprob) enables the generation of confidence scores for each prediction,

To ensure transparency and improve model performance through feature optimization,SHAP are integrated as the primary XAI method. SHAP values, derived from cooperative game theory, quantify the contribution of each feature to the prediction by calculating average marginal contributions across all possible feature subsets. This theoretical robustness distinguishes SHAP from other explainability methods, enabling precise measurement of feature impact. Moreover, SHAP efficiently identifies influential features, providing insights at both local (individual prediction explanations) and global (overall model behavior explanations) levels. By incorporating SHAP directly into the IDS modelling pipeline, this approach ensures transparency and interpretability, thus enhancing model reliability and facilitating the identification and removal of redundant or non-influential features, ultimately improving detection performance and computational efficiency.

SHAP-Based Feature Refinement

Our implementation involves several methodical steps. Initially, SHAP values are computed individually for each prediction across all classes using the TreeExplainer algorithm, which is particularly suited for tree-based models. This computation is confined strictly to the training set to prevent data leakage, ensuring a valid and unbiased evaluation of feature significance.

Subsequently, the absolute mean of the aggregated SHAP values is computed across all classes for each instance in the training dataset, yielding a comprehensive matrix of size (125957, 122). The use of absolute mean values ensures focus solely on the magnitude of feature contributions, irrespective of their directional impact, thereby accurately capturing overall feature significance.

In the next analytical step, mean SHAP values for each feature across all training instances are calculated, reducing the dataset dimensionality further into a compact vector of size (122,). This concise representation allows clear identification and ranking of global feature importance, facilitating strategic selection. Finally, only features exhibiting positive mean SHAP values—those actively contributing to the predictive capacity—

are retained. This selective process substantially reduces the feature space from 122 to 68 influential features. Figure 19 illustrates the top-ranked features in the training set, clearly indicating that the feature 'src bytes' dominates others, demonstrating its significant contribution to the overall predictive capability of the model.

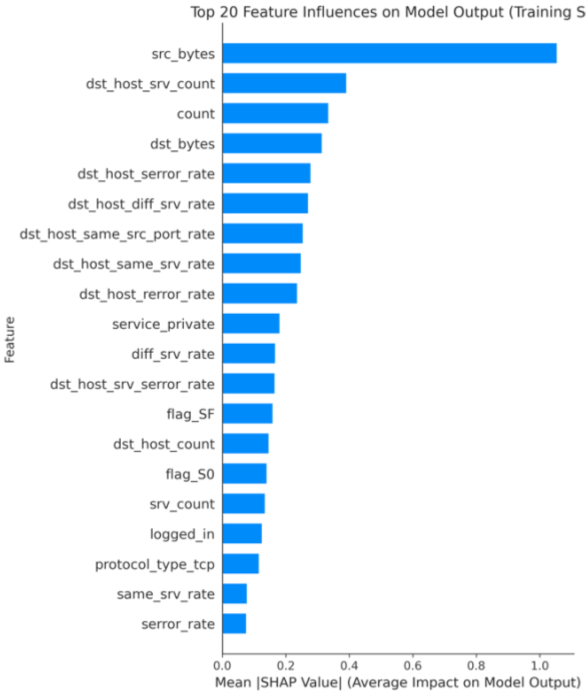


Figure 19. Top 20 features ranked by mean absolute SHAP value in the NSL-KDD training dataset. Each bar represents a feature’s average contribution to the model’s output

Upon identifying the contributing features, we employ an advanced feature selection method, SHAPRefine, to further reduce this set to 30 features. SHAPRefine adopts a selection strategy that incrementally adds features from the ranked list shown in Figure 19, evaluating model performance at each step. Figure provides a comparative analysis of three XGBoost models trained on different feature subsets, showing that selecting the top-ranked features based solely on initial SHAP values does not always yield the best performance. This discrepancy primarily arises due to hidden interdependencies and interactions among network data features, which are not immediately evident through individual feature importance rankings alone. Our advanced feature selection explicitly addresses these hidden dependencies, significantly enhancing detection accuracy and robustness. The resulting refined feature set demonstrates notable performance improvements, validated through rigorous comparative analyses conducted before and after feature reduction, thereby making the framework highly adaptable across multiple domains and diverse datasets relevant to emerging 5G and 6G network security scenarios.

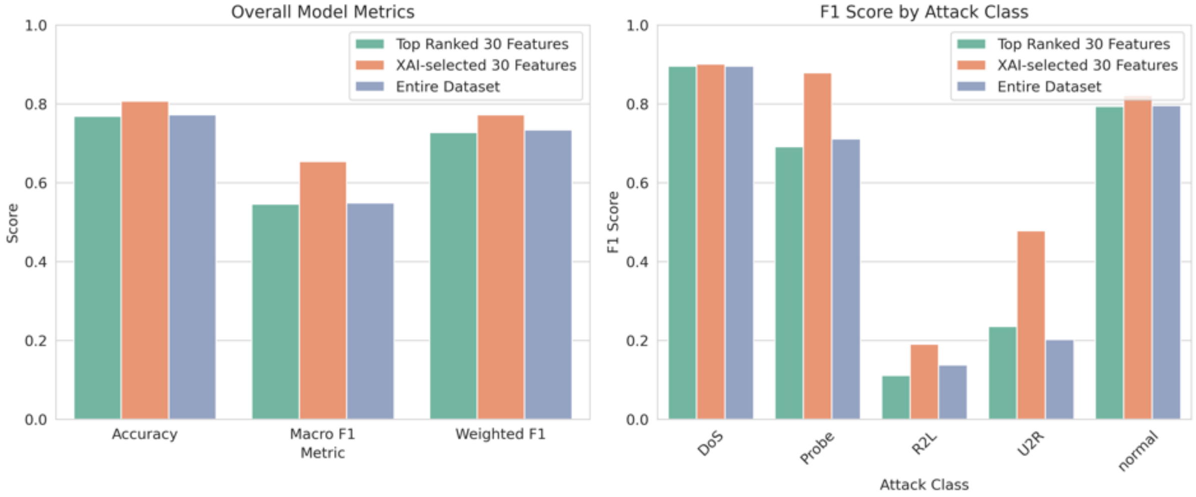


Figure 20. Comparison of three XGBoost models trained on different feature subsets. The left panel presents overall model accuracy, macro F1, and weighted F1, whereas the right panel displays F1 scores for individual attack classes.

This section presents a robust, explainable, and efficient intrusion detection methodology specifically tailored to address emerging cybersecurity challenges in 6G networks. By integrating SHAP-based explainability into the feature selection and model training pipeline, our proposed XAI-IDS framework not only enhances detection accuracy and computational efficiency but also significantly improves transparency and interpretability. Within the ROBUST-6G project's functional architecture, this methodology directly supports secure AI/ML deployment by offering adaptable, transparent intrusion detection capabilities. Furthermore, its versatility ensures seamless integration with complementary solutions developed within other work packages, such as fairness-enhancing algorithms and secure data acquisition frameworks, ultimately contributing to a comprehensive, resilient, and trustworthy 6G security ecosystem.

Use within ROBUST-6G System Design

The integration of XAI within 6G networks (as shown in Figure 21) is essential for ensuring transparency, interpretability, and trustworthiness in AI-driven decision-making processes. Conventional AI models often operate as black boxes, making it difficult for users to interpret their reasoning, which poses significant risks in high-stakes applications such as cybersecurity, autonomous systems, and network anomaly detection. This is further exacerbated by the widespread adoption of large and not-so-large language models [S+25]. Ensuring that AI models operate reliably under evolving network conditions necessitates the adoption of confidence-aware evaluation methodologies. Recent advances in XAI emphasize the importance of trust calibration, where models not only provide accurate predictions but also communicate their level of certainty and confidence. This aligns with the broader goal of trustworthy AI, which integrates fairness, robustness, and explainability into AI-driven processes. We have developed an effective method for assessing XAI in 6G is the use of latent space representations to quantify model confidence. By leveraging Variational Autoencoders (VAEs), a model can learn meaningful latent embeddings that capture the underlying structure of network data, enabling an assessment of how well new observations fit within known distributions. The Mahalanobis distance, computed in this latent space between the training data and the inference data, provides a quantitative measure of how anomalous a given input is relative to previously seen data, making it a valuable tool for assessing uncertainty in AI-driven anomaly detection systems. In our work [PAK+24] on trustworthy IDS, this approach was applied to the NSL-KDD dataset, demonstrating a 45% correlation between Mahalanobis distance in latent space and the reconstruction error. This correlation highlights the effectiveness of latent space confidence estimation in distinguishing normal from malicious network activities. By optimizing key parameters, such as the latent space dimension (optimal at 20) and KL weight (optimal at 0.25), we were able to fine-tune the model’s ability to generate reliable uncertainty estimates, enhancing the robustness of AI-driven security mechanisms in 6G.

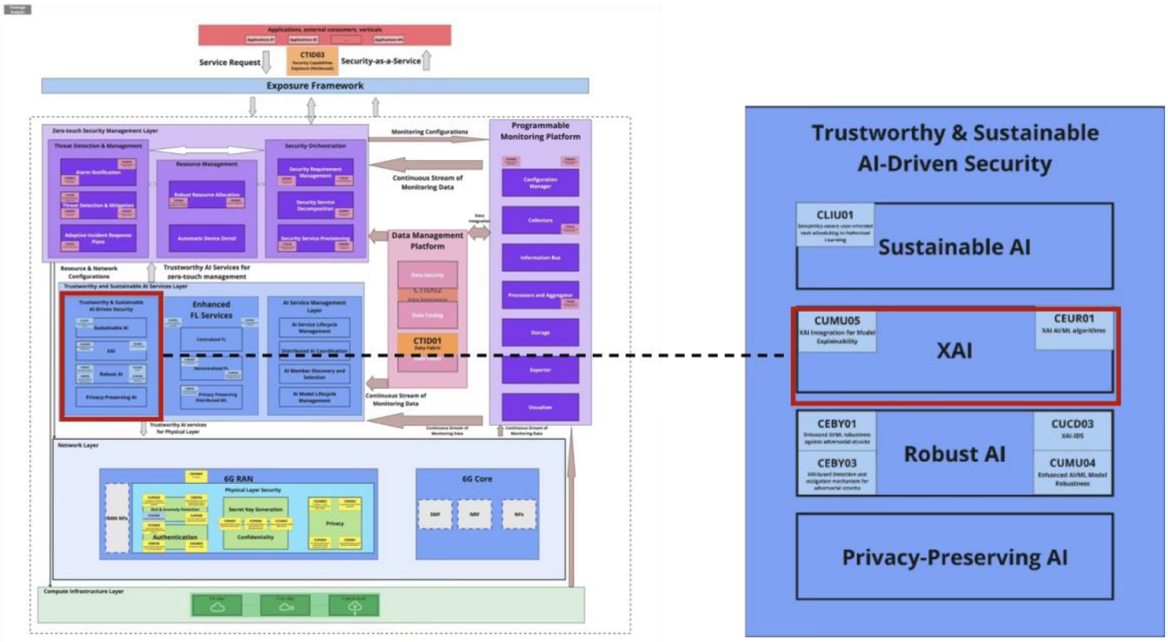

Applications, external consumers, verticals
CTID03
Service Request
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Resource Management
Security Orchestration
Monitoring Configurations
Programmable Monitoring Platform
Continuous Stream of Monitoring Data
Data Management Platform
Trustworthy AI Services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Enhanced FL Services
AI Service Management Layer
CTID01
Network Layer
6G RAN
Physical Layer Security
6G Core
Compute Infrastructure Layer
Trustworthy & Sustainable AI-Driven Security
CLIU01
Sustainable AI
CUMU05
XAI Integration for Model Explainability
XAI
CEUR01
XAI AI/ML algorithms
CEBY01
Enhanced AI/ML robustness against adversarial attacks
CEBY03
XAI-based Detection and mitigation mechanism for adversarial attacks
Robust AI
CUCD03
XAI-IDS
CUMU04
Enhanced AI/ML Model Robustness
Privacy-Preserving AI

Figure 21. Integration of Explainable AI with ROBUST-6G Architecture

Another critical metric for evaluating XAI integration in 6G is the CPCF which provides a dynamic and theoretically grounded measure of model reliability. Unlike conventional confidence scores derived, for example, from softmax outputs, CPCF uses conformal prediction to construct adaptive prediction sets that quantify uncertainty at a specified confidence level without requiring access to ground truth labels. In our experiments on continual learning, CPCF was validated as an effective indicator of catastrophic forgetting, with a strong correlation (51% to 57%) between CPCF values and the accuracy of previously learned tasks. The robustness of CPCF was further confirmed across different calibration ratios and significance levels, showing its adaptability in diverse learning scenarios. This is especially relevant in 6G applications, where AI models must operate under non-stationary environments, dynamically changing wireless conditions, and adversarial network attacks, making CPCF a valuable tool for maintaining trust and reliability in real-time AI-driven systems.

3.3.4 XAI-Based Detection and Mitigation of Adversarial Attacks

ML models are highly effective for identifying and countering threats but may also introduce novel risks. So, ensuring the confidentiality, integrity, and availability of AI models and their associated data is crucial. Threats may appear in various forms, such as adversarial attacks, data tampering, and unauthorized access to data [CAD+18]. As ML technology evolves, so do adversarial techniques, and traditional defences like adversarial training are computationally intensive and often inadequate for real-time threat detection. These approaches generally require a compromise between resilience and model efficiency, posing challenges for applications that need immediate actions. On the other hand, XAI can be seen as a novel method for identifying and mitigating adversarial attacks [APA+21]. Most of the research that employs XAI to investigate adversarial attacks has focused on image classification [KMG+20], [FBS+20]. However, addressing the cybersecurity risks introduced by adversarial AI techniques continues to pose a major challenge. Currently, security protocols are increasingly dependent on AI / ML frameworks to detect and mitigate emerging sophisticated threats. For example, IDSs are essential for analysing network activities and recognizing suspicious behaviour that might signal potential attacks. Progress in ML has improved the performance of IDSs, enabling a more effective detection of anomalies. This effectiveness is especially significant given the sheer amount of data managed in 5G and future mobile network environments. Nonetheless, attackers consistently attempt to degrade the performance of ML-based IDSs, thereby increasing the susceptibility of the network.

Robust Intrusion Detection System with Explainable Artificial Intelligence

The vulnerability discussed above is analysed in detail in this article [GFE+25] that presents a notable contribution towards the development of a comprehensive framework dedicated to the detection and mitigation of adversarial threats in IDS. The paper introduces an agnostic approach that employs XAI techniques to assess how adversarial samples affect ML model interpretations. Furthermore, a zero-touch detection method has been designed to enhance IDS capabilities, focusing on strengthening defences against network-centric attacks and advanced techniques used by adversaries to bypass detection mechanisms. By proactively addressing these security vulnerabilities, IDS security and resilience are improved. The proposed method allows IDS not only to identify new attack vectors, but also to react to potential threats, even with a limited number of features. The test environment used is the O-RAN infrastructure, chosen for its emphasis on utilizing AI/ML to optimize network functionality and capabilities [NRD+22]. O-RAN adopts 3rd Generation Partnership Project (3GPP) standards to improve interfaces and protocols, tackling the increased attack surface due to virtualization, open interfaces, and multivendor configurations. However, despite these improvements, the architecture still falls short in addressing specific vulnerabilities related to ML/AI within O-RAN components, such as the Non-Real-Time RAN Intelligent Controller (Non-RT RIC) and the Real-Time RAN Intelligent Controller (Near-RT RIC) [IL+15].

The main contributions in [GFE+25] are listed as below:

- By integrating an XAI feature into the ML-based detection framework, a novel method is developed for real-time evaluation.
- The detection model improves the understanding of the significance of features, improves the selection of crucial attributes to identify adversarial attacks, and reduces false positives.
- The concept behind the XAI feature involves understanding the behaviour of unseen data, ensuring that its distribution aligns with the standard behaviour observed during training, emphasizing the importance of assessing SHAP value distributions within the training data, a novel proposal.
- The proposed ML-based detection method with XAI integration enhances the performance of the detection system, ensuring better adaptation to changing attack patterns.
- In addition to providing detection and mitigation solutions, this approach offers a zero-touch strategy aimed at augmenting IDS functionalities, thus strengthening defences against adversarial threats.

The literature has explored numerous methods to protect IDSs from adversarial threats. For example, adversarial training, as introduced in [AKN+19], aims to enhance the identification of adversarial examples. A deep learning-based methodology, specifically designed for adversary detection, has been suggested in [NKG+21]. However, current solutions face several problems. Primarily, integrating multiple techniques during training and applying them within the ML model of IDS exposes pre-trained models to potential analysis by attackers. This exposure allows adversaries to adjust their strategies to exploit model vulnerabilities. In addition, many solutions ignore the specific attributes of attack vectors, resulting in the exclusion of essential features. The significance of the explainability for IDS is underscored in the survey [NAA+22]. Although earlier studies have advanced the explainability of IDS, a comprehensive framework combining the resilience of ML against adversarial threats, real-time functionality, and IDS robustness without human input has been lacking. Another previous study [GF+24] demonstrated that XAI is particularly effective in detecting adversaries by evaluating the explanations of pre-processed network traffic data along with the decisions made by ML-based IDSs. However, this method requires a large number of features to analyse adversarial inputs, suggesting the need for a stronger strategy for broader applications. Addressing these gaps with an integrated approach could significantly improve IDS efficiency and reliability. To address these challenges, this paper introduces a robust IDS capable of identifying and countering adversarial attacks by incorporating an XAI component. The proposed detection model operates in real time, even with a limited feature set, under both white-box and black-box conditions. Further elaboration is provided in the following sections.

An adversarial detection strategy is proposed that aims to enhance the robustness of IDS by integrating the XAI feature for real-time evaluation. This approach is designed to be model-agnostic and applicable to any IDS framework or context. The primary objective is to improve IDS by enhancing its defences against adversaries using network and adversarial techniques to evade detection. The core assumption of the framework involves the use of XAI to reveal and emphasize system vulnerabilities, allowing the effective identification of altered network traffic for the action of IDS. Figure illustrates the general structure of the proposed IDS methodology. By incorporating the XAI feature into the ML model of IDS, the system can assess its vulnerability to adversarial approaches. This assessment is based on the definition of the normal data behaviour pattern through the distribution of SHAP importance values.

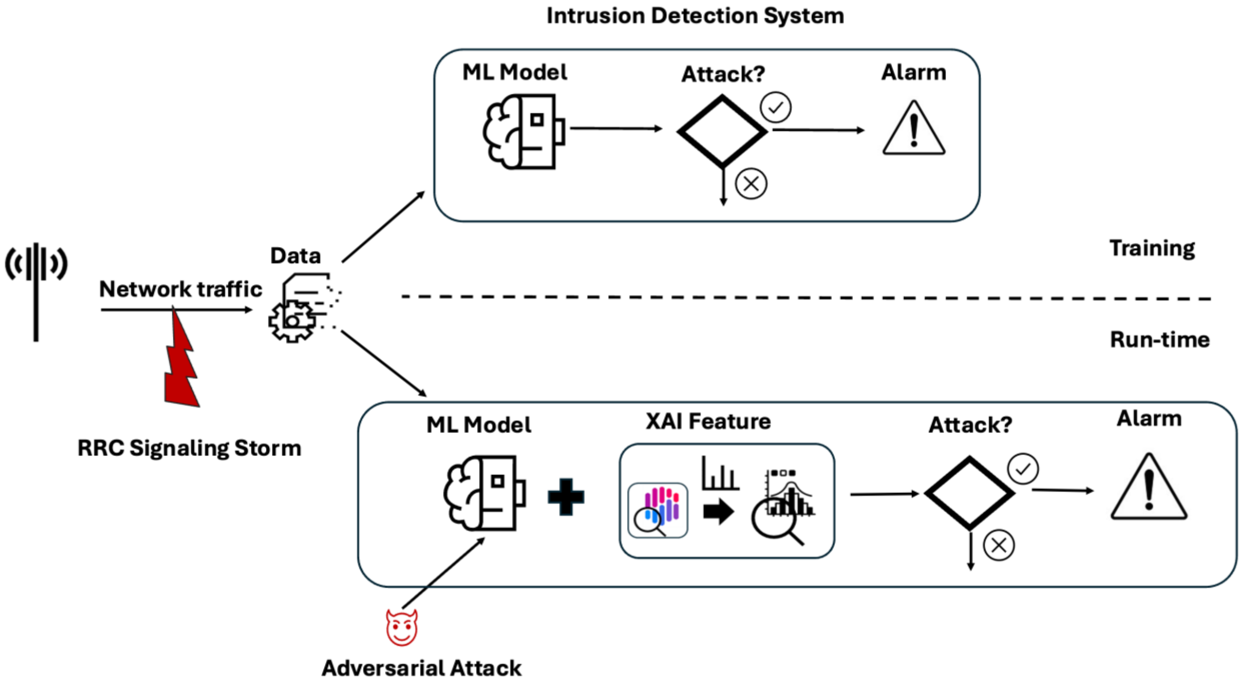


Figure 22. General structure of the problem and proposed XAI-based adversarial

A novel XAI-based adversarial detection framework is intended to identify if new data has been altered. The learning procedure is split into training and run-time stages.

In the training phase, SHAP feature importance values are collected for each input to characterize the normal behaviour of the data by analysing the distribution of these importance values.

- Let X $= \{x_1, x_2, \dots, x_n\}$ be the training data set and let f(X) be the ML model trained on X .
- The SHAP importance values for each input x_i are denoted by $S(x_i) = \{S_{1(x_i)}, S_{2(x_i)}, \dots, S_{m(x_i)}\}$ where m is the total number of features.
- Assume the distribution of SHAP values for each feature i follows a normal distribution: $Sj(x) \sim N(\mu j, {\sigma_j}^2)$ where μ_j and ${\sigma_j}^2$ represent the mean and variance computed from the training data.

During run-time, the ML model evaluates unseen data by checking whether the SHAP feature importance values align with the normal behaviour distribution observed in the training data.

- Let $X' = \{x_1', x_2', \dots, x_k'\}$ be the test data. The SHAP values for a test sample x_i are given by: $S(x_i') = \{S_1(x_i'), S_2(x_i'), \dots, S_m(x_i')\}$
- The behaviour of the test data is assessed by verifying if each $S_j(x_i')$ falls within the expected normal range: $\mu_j - \lambda\sigma_j \leq S_j(x_i') \leq \mu_j + \lambda\sigma_j$
- Where λ is a threshold parameter (e.g., $\lambda = 2$ for a 95% confidence interval under the assumption of normality).

If the SHAP values of the unseen data maintain the same distribution as the training data, x_i' is classified as **Normal**, otherwise the input x_i' is identified as an **Attack**.

Validation and Results

This section demonstrates the capability of our approach to detect and mitigate adversarial input that significantly degrades IDS performance, specifically designed for RRC signalling storm attacks in real time. XAI is used to identify and prevent the degradation of IDS performance. The proposed framework aims to identify potential adversarial attacks by detecting significant deviations in the distribution of SHAP importance values for each input in real time and to implement a simple yet effective zero-touch mitigation strategy.

There are three different experimental setups in the paper [GFE+25].

1) Initially, IDS performance is evaluated in two contexts: one with an RRC signalling storm attack intended to disrupt network traffic, and another where this attack is combined with an adversarial attack on the IDS. Various methods are used to assess the effectiveness of adversarial attacks. First, an RRC signalling storm attack is simulated, marked by excessive signalling that overloads the control plane. Then, the ability of IDS to detect is evaluated. The IDS performance is also measured in normal conditions without adversarial attacks using an auto-encoder as a baseline.

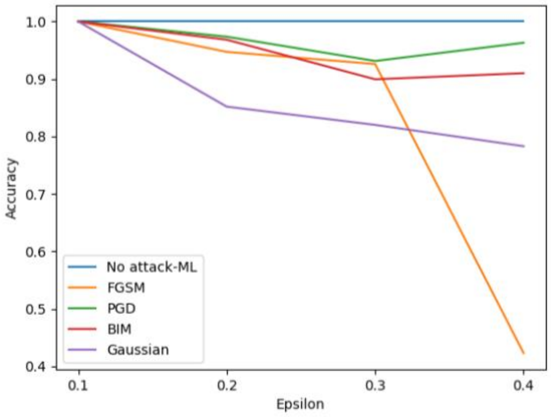


Figure 23. The accuracy of IDS under different adversarial attack method

In Figure , the accuracy of the IDS is illustrated under various adversarial attacks. The X-axis reflects the epsilon value, with larger values indicating stronger perturbations. The Y-axis displays the IDS's ability to classify inputs as normal or malicious. Without attacks, IDS maintains high baseline accuracy. As the epsilon rises, the accuracy declines, indicating the decreased efficacy of IDS against stronger perturbations. FGSM significantly reduces accuracy, highlighting the need for robust defences against gradient-based attacks. Although IDS shows improved performance over other attacks, suggesting some resilience, PGD remains partly effective, underscoring the model's partial robustness. The model demonstrates slightly improved accuracy compared to PGD, implying it manages iterative attacks like BIM more proficiently. However, the significant drop in accuracy when faced with Gaussian noise highlights its vulnerability, underscoring the importance of effective noise handling and pre-processing strategies.

2) In the second scenario, we propose an adversarial attack detection strategy that incorporates an XAI feature to assess whether the new data supplied to the IDS aligns with the normal behaviour distribution of the training data. During training, SHAP values are calculated for each input to determine their importance. Kernel density estimation is utilized to smoothly estimate the distribution from these importance values, which then defines the normal behaviour pattern for IDS data. During run-time, an autoencoder analyses the unseen data. To determine if a test input is an anomaly, a threshold is computed as a Z-Score from the established normal behaviour (derived from the training data) and verify whether the distance of unseen input exceeds this threshold.

	AE-BIM attack	Permutation	LIME	SHAP
True Negative (TN)	144	132	127	130
False Positive (FP)	0	12	17	14
False Negative (FN)	45	24	15	0
True Positive (TP)	0	21	30	45

Table 6 Confusion matrix metrics of different detections cenarios

The proposed approach is evaluated by comparing it with individual feature importance methods like LIME and permutation importance. Table 6 illustrates four confusion matrices for different detection scenarios under BIM attack using various methods. The AE-BIM attack matrix displays a high true negative rate but has difficulty detecting true positives, indicating challenges in spotting adversarial examples. The permutation matrix shows improved results with both true negatives and some true positives, suggesting a better detection of adversarial perturbations. LIME offers a balanced detection with moderate true positives and negatives, even though with some accuracy trade-offs. SHAP significantly enhances detection, achieving many true positives and fewer false negatives and positives, thus improving the ability to distinguish actual from adversarial input.

3) Lastly, a straightforward and efficient mitigation strategy examines the predicted attack instances. When unseen data fit to the same distribution as the training data (i.e. normal), it is labelled "Normal." If an input deviates from the standard distribution range and is identified as an outlier, it is classified as manipulated, and its label is adjusted accordingly.

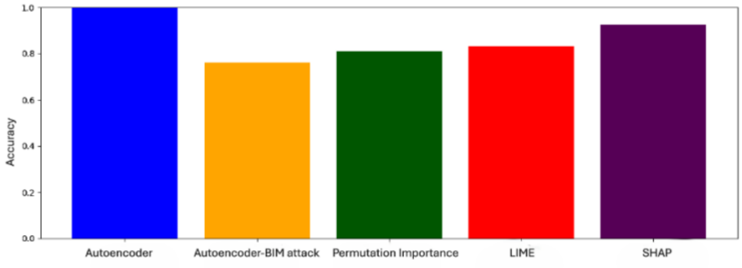


Figure 24. Method comparison with XAI-based mitigation

Figure shows the mitigation performance of the proposed method. The permutation importance approach improves accuracy, indicating the adverse effects of attacks. Although LIME aids in prediction explanation, its effectiveness is like that of the permutation method, highlighting its limitations against attacks. SHAP demonstrates the highest accuracy, showcasing its superior ability to handle adversarial instances and enhance model robustness.

In conclusion, the integration of the XAI feature into adversarial detection enhances accurate and efficient identification of adversarial actions. This approach has proven effective by enhancing IDS performance in an O-RAN setting, marking significant progress in cybersecurity improvement. The system introduces zero-touch functionality, enabling IDS to respond quickly to emerging threats and considerably mitigate risks to the network infrastructure. With the evolution of O-RAN, the collaboration of XAI and IDS will be critical in addressing new vulnerabilities. Future research and development are crucial for the advancement of XAI-enabled cybersecurity measures to maintain robust network defence against evolving cyber threats.

3.4 Fairness in AI for 6G Systems

The rise of AI-driven intelligence in 6G systems introduces not only opportunities for optimized service delivery, adaptive network management, and context-aware decision-making, but also significant ethical and societal challenges. One of the key pillars underpinning Trustworthy AI —as recognised by the European Commission’s High-Level Expert Group on AI and mirrored in projects such as HEXA-X, as well as by ROBUST-6G project is fairness. In the context of 6G, fairness refers to the equitable treatment of users, services, and applications by AI-enabled functions operating across the radio, core, and edge layers of the network. As highlighted in the HEXA-X project deliverable D6.1 ("Trustworthiness and Explainability for AI"), fairness is defined as: ***"The absence of bias, discrimination, and unfair treatment of individuals or groups in AI-based decision-making, including during the design, training, deployment, and inference phases of AI systems."***

Fairness in 6G systems must consider both technical and societal dimensions. On the technical side, AI models embedded in 6G service orchestration or network control may inadvertently encode and amplify biases present in training data or operational context. On the societal side, 6G’s role in ubiquitous connectivity and critical services (e.g., healthcare, emergency response, industrial automation) magnifies the impact of unfair decisions, potentially exacerbating digital divides or marginalising vulnerable users. Importantly, fairness is not an isolated attribute—it is interlinked with privacy, robustness, and explainability. For instance, achieving fairness may require balancing trade-offs between model performance and privacy guarantees, particularly in federated or privacy-preserving machine learning systems. This interconnectedness is acknowledged in our project’s goal of building a risk-averse yet agile resource control framework that delivers security guarantees under fairness/privacy constraints.

Furthermore, fairness in 6G must be interpreted in the multi-stakeholder and dynamic ecosystem of future networks. Unlike static systems, 6G environments evolve based on mobility, demand, and context, making real-time fairness a moving target.This creates new challenges for AI fairness auditing, adaptive mitigation mechanisms, and context-sensitive fairness definitions.

To address these challenges, some EU-funded projects such as HEXA-X, AI4Trust, and REWIRE were exploring methodologies for fair-by-design AI, focusing on bias detection, fairness constraints in model training, and feedback-driven fairness adaptation. These efforts serve as inspiration and foundation for our project's initial explorations into fairness in 6G systems.

In the following sections, we outline:

- how fairness can be framed within AI-driven 6G service provisioning
- how it interacts with privacy-preserving mechanisms such as differential privacy
- how future evaluation frameworks might incorporate fairness as a measurable and actionable criterion.

3.4.1 Fairness Considerations in Privacy-Preserving AI for 6G

As AI-driven intelligence becomes foundational to 6G networks, ensuring user privacy has emerged as a critical design objective, particularly in contexts involving sensitive or personally identifiable data. Techniques such as Differential Privacy (DP) are increasingly adopted to provide formal guarantees that individual data cannot be inferred from AI model outputs. The random noise injected by DP to protect individual data often degrades model utility disproportionately for underrepresented or vulnerable groups, exacerbating existing biases [WC+25]. In essence, stringent privacy guarantees can widen accuracy gaps between majority and minority groups. For example, studies have shown that as the privacy budget (ε) is tightened (more noise), the performance disparity between well-represented and underrepresented demographics initially grows [YD+24]. This occurs because noise obfuscates the already limited information from smaller groups, causing greater relative error for those groups. If the noise becomes extremely high, the model's outputs degenerate to mostly random (uniformly poor for all groups), trivially achieving "fairness" at the cost of usefulness. Such findings underscore a fundamental trade-off: privacy safeguards can conflict with fairness, demanding careful calibration of DP mechanisms to balance individual privacy and equitable model performance.

One fairness-aware approach is to adjust the DP noise magnitude per group so that no subgroup is unduly penalized by privacy noise. Rather than apply a one-size-fits-all noise level, the perturbation can be calibrated to the characteristics of each group (e.g., their size or risk of bias). For instance, researchers have proposed stratified DP mechanisms that run separate DP operations for each demographic group and then aggregate the results [GA+24]. By tailoring noise addition within each subgroup, this technique can yield more balanced outcomes across groups without increasing the total privacy budget. Essentially, each individual still has the same DP guarantee, but the noise is injected in a context-aware manner.

Beyond static noise shaping, fairness-aware DP algorithms embed fairness goals into the model training procedure. These in-processing approaches jointly optimize for privacy and fairness, often by imposing fairness constraints or re-weighting during learning while noise is added. For example, one can formulate a constrained optimization where the model must satisfy a group fairness metric (such as equality of odds or limited disparity in error rates) subject to DP noise being added at each step. Recent work by Ding et al. [DZL+20] calibrates the DP noise in a logistic classifier by a fairness-aware objective, effectively dynamically adjusting the noise or clipping for each class feature to meet a fairness criterion. Another strategy called FairDP casts private model training as a bilevel optimization that automatically modulates the influence of each training instance based on its class and difficulty [TF+23]. In FairDP, the algorithm increases or decreases the contribution of data from certain classes (within DP-SGD's clipping and noise framework) to counteract bias, using a theoretical bias-variance analysis to guide these adjustments. This yields a self-adaptive DP mechanism that, for example, might allocate a slightly larger effective privacy budget (lower noise) to a class that saw high error, thus improving that class's accuracy while still respecting an overall privacy limit. The key idea across these methods is to intertwine fairness constraints with the DP mechanism so that the model's privacy-preserving perturbations are not blind to fairness concerns. By doing so, the algorithm actively preserves model fairness (e.g., similar error rates across groups) during training rather than treating fairness as an afterthought.

In the context of AI-driven 6G networks, fairness-aware DP becomes an enabling technology for inclusive and privacy-preserving intelligence. 6G networks will intertwine with diverse user data (from IoT sensors to personal devices) and use AI for everything from network optimization to personalized services. Ensuring privacy is non-negotiable in these scenarios–techniques like DP will be employed to protect sensitive information flowing through learning algorithms at the network's edge. However, if those privacy techniques unfairly degrade the service for certain users (e.g., users in minority demographic groups or sparsely connected regions whose data contributions are smaller), the network's goal of universal service quality would falter.

3.4.2 Fairness Aspects in Trustworthy AI for 6G

As 6G envisions a hyper-connected world powered by pervasive intelligence, the integration of AI into the network's fabric introduces complex fairness challenges. In such a dynamic and decentralized environment, AI models will make critical decisions in tasks related to security and privacy, in addition to conventional tasks, such as spectrum sharing, edge resource allocation and scheduling, and network slicing. These decisions must be not only efficient and accurate but also equitable across diverse user profiles, geographical regions, and usage scenarios. Without explicit fairness mechanisms, AI systems risk amplifying existing disparities, especially when trained on data that reflect historical biases or uneven representation – leading to disproportionate QoS for underrepresented users or devices. Thus, fairness is not just a desirable property, but a prerequisite for trustworthy AI in 6G.

In secure 6G systems, AI plays a central role in ensuring real-time threat detection, adaptive access control, and privacy-preserving communication. However, integrating AI into security-critical functions introduces new fairness challenges. As AI models automate decisions such as intrusion detection, user authentication, and anomaly classification, it is essential to ensure that these systems do not discriminate against specific users, devices, or regions based on biased training data or flawed assumptions. For example, an AI-based intrusion detection system that flags behaviour based on historical usage patterns may unjustly penalize users from underrepresented or non-conventional usage scenarios—raising serious ethical and trustworthiness concerns.

In this context, fairness becomes a critical dimension of secure AI, where trust is eroded not only by false positives or security breaches but also by inequitable treatment. Ensuring fairness in secure 6G AI systems involves auditing detection rates across different demographic or device groups, verifying that access control policies are enforced uniformly, and minimizing disparities in model confidence and false alarm rates. Moreover, fairness-aware AI design should be coupled with privacy-preserving machine learning techniques, such as federated learning or differential privacy, to protect user data while maintaining equitable treatment.

One promising direction to address fairness in AI predictions for 6G is through fair conformal prediction—a framework that provides valid confidence intervals or prediction sets for AI outputs with guaranteed coverage probabilities. Traditional conformal prediction methods ensure that the true label lies within the predicted set with a certain probability (e.g., 90%), but this guarantee typically holds in aggregate across the entire population. In heterogeneous 6G environments, this can lead to coverage disparities between different subgroups—such as rural vs. urban users or low-power vs. high-capability devices. To mitigate this, group-conditional conformal prediction or fair adaptive calibration can be used to ensure that the desired coverage levels hold across all relevant subpopulations. For instance, the system could enforce that confidence intervals for IDS maintain equal reliability regardless of user location or device type.

Another promising solution is related to an auditing perspective in fair and trustworthy AI. Ensuring fairness in AI-driven 6G systems requires not only designing equitable algorithms but also establishing mechanisms to audit and certify their behaviour post-deployment. From a trustworthy AI standpoint, AI auditing serves as an essential tool to systematically assess whether the decisions made by AI models comply with predefined fairness criteria. One powerful and statistically rigorous way to frame such audits is through hypothesis testing.

In this paradigm, fairness auditing can be modelled as a series of statistical tests where the null hypothesis typically represents the claim that the AI system behaves fairly across specified groups (e.g., equal error rates or predictive coverage), while the alternative hypothesis reflects a violation of fairness. For example, in a delay prediction model deployed across urban and rural regions in a 6G network, an auditor might test:

- H_0: The average prediction error is equal for both user groups.
- H_1: There is a statistically significant difference in prediction error between the groups.

Such hypothesis-driven auditing enables quantifiable, interpretable, and reproducible evaluations of fairness across subpopulations. It also helps in setting confidence levels (e.g., 95%) for the conclusions, allowing stakeholders to make decisions grounded in statistical significance rather than heuristic thresholds. Importantly, this framework can be extended to multivariate fairness audits, testing for interactions between user attributes (e.g., device type, location, network load) and model performance metrics.

For AI systems in 6G, which often operate under non-stationary and partially observable environments, dynamic or online hypothesis testing methods are particularly relevant. These include techniques like sequential testing, adaptive p-value correction for multiple comparisons, and change detection to identify fairness degradation over time. Moreover, coupling these tests with conformal prediction auditing can assess whether prediction sets maintain coverage parity across user groups – a key fairness concern when deploying AI in mission-critical 6G applications like autonomous transport or remote healthcare.

Finally, fairness constraints could be embedded into multi-objective optimization frameworks in the AI pipeline, balancing performance, confidence interval width, and fairness across groups. These solutions would allow 6G networks to make trustworthy, uncertainty-aware decisions while ensuring equitable service for all users—advancing both technical robustness and social responsibility in next-generation communication systems. Such aspects are studied in Task 4.4 in WP4, connecting trustworthy AI solutions designed in WP3 with risk-aware resource allocation developed in WP4.

4 Sustainable AI for Ensuring Energy-Efficiency in 6G

4.1 An ADMM-based optimizer for SNNs

Spiking neural networks

Spiking neural networks (SNNs) are a type of neural networks (NNs) mimicking more closely how the human brain works and are effective in processing time series. Interestingly, when running on dedicated hardware, SNNs show three orders of magnitude improvement in the energy-delay-product (EDP) [RBG+22] and are thus promising for reducing the inference energy at the network edge.

Computational neuroscience is the field that studies the modelling of neurons and how groups of neurons interact [GKN+14]. The neuronal dynamics are, in general, described by differential equations and require some state variables to be stored and updated according to the model during time. Among the simplest but computationally efficient models, the most used to build complex SNNs is the leaky integrate and fire (LIF) neuron [GKN+14], which will be presented in the next paragraphs concerning the problem formulation.

Currently, the scientific literature is focusing on searching for efficient training methods for SNNs. The commonly used stochastic gradient descent (SGD) cannot be directly applied to train SNNs due to a non-differentiability of the activation function (a step function). However, starting from this idea, the surrogate gradient method [NMZ+19] has been proposed, which consists in approximating the backpropagation step with a hyperbolic tangent. Nonetheless, this introduces an approximation error and the vanishing gradient problem, making the method ineffective for deeper networks.

In the next paragraphs, we introduce an alternative training method based on the ADMM [BPC+11].

Use of the ADMM to train NNs

The use of the ADMM to train NNs is rooted in the seminal paper of Taylor *et al.* [TBX+16], where the purpose of the authors was developing an ADMM optimizer to avoid the expensive execution of SGD on GPUs, exploiting instead a parallel execution on multiple CPUs. While the results were interesting and promising, the method did not gain popularity due to the increased availability of efficient and relatively cheap GPUs that made SGD and its variants the golden standard. However, more recently the method has been re-investigated and improved with the addition of a computationally efficient subroutine for the pseudo-inverse estimation and Anderson acceleration [ZBD+25]. In 2023, Bemporad investigated the use of ADMM to train recurrent neural networks (RNNs) [BEM+23], a problem more similar to the training of SNNs due to the dynamic nature of equations. In the very recent work presenting BADM [WZL+24], the authors adopt a mixed approach of ADMM and SGD where they train an NN with backpropagation but split the batches of sample through ADMM. The method outperforms the best SGD-derived approaches like ADAM and RMSprop.

Besides the fact that the recent advances in the use of the ADMM are promising also concerning the performance of the trained model, a dual method is especially interesting when applied to SNNs due to the non-differentiability of the activation function, which makes the direct application of backpropagation infeasible. In the next paragraphs, we will explain the framework developed and tailored for SNNs and present some preliminary results obtained from training a simple model with a batch of data in stochastic and deterministic fashions, respectively.

Problem formulation and SNN model

To tackle the training of SNNs via ADMM we need to first design a model-driven optimization problem that defines i) the objective function, namely, the loss function, ii) the dynamics of the LIF neurons and the relative state variables, and iii) how the LIF neurons interact, i.e., the dynamics of the neural network. This is done via the following formulation, where the variables W_l, $z_{l,t}$, and $a_{l,t}$ represent the weights at layer l and the membrane potentials and output currents (spike) at layer l and time t for each neuron, respectively. The activation function of SNNs is modelled as a Heaviside step function centered in ϑ and denoted h_ϑ, which is non-convex and, especially, non-differentiable. The hyperparameter ϑ represents the membrane potential firing threshold: When the membrane potential reaches or surpasses this threshold, the neuron emits a spike, and the potential is reset to a rest value (the threshold is subtracted). The full optimization problem encompassing points i), ii), and iii) is formulated according to Figure 21 a graphical representation of the SNN model is also depicted, linking the variables to their meaning.

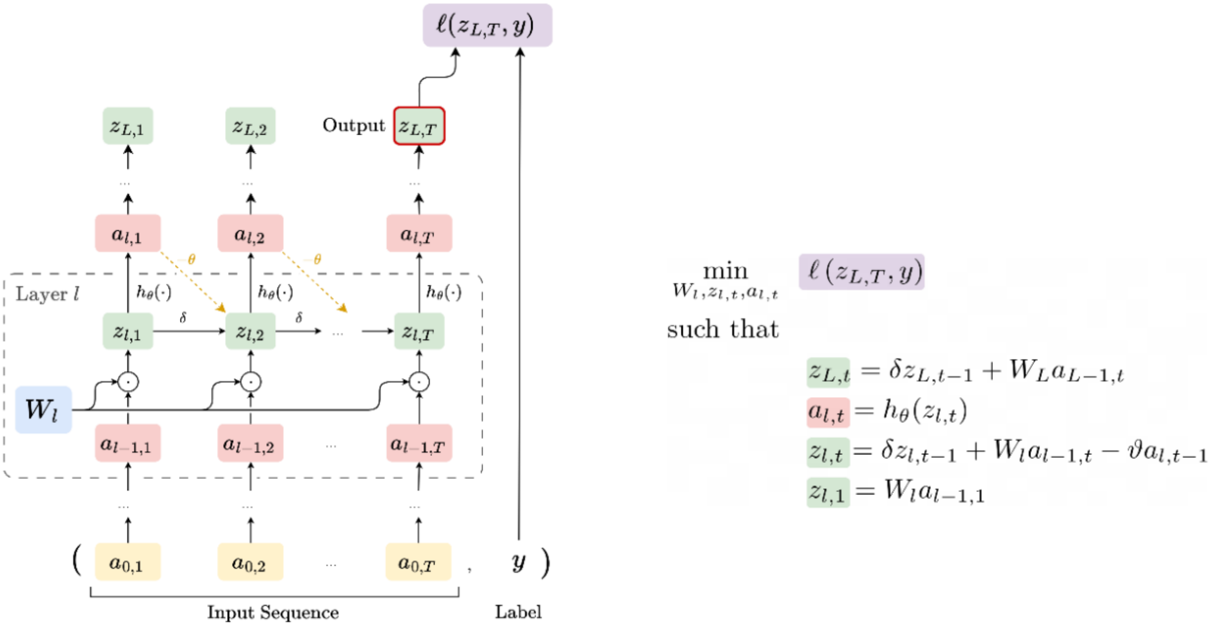


Figure 25. Graphical representation of a spiking neural network as a dynamic system (left). Problem formulation with variables having the same color of the dynamic system (right)

ADMM solution and algorithm

The problem presented in the previous paragraph is non-convex due to the multiplication between weights and neuron outputs. The use of an alternating direction method splits it into subproblems where the two variables are alternatively optimized, and this practically reaches stability. However, the presence of a second non-convexity is caused by the Heaviside step function.

For these reasons, the problem is tackled in a relaxed version via an augmented Lagrangian that adds multipliers only for the output of the SNN, which is used in the loss function. The form of the augmented Lagrangian is

$$\begin{aligned}\mathcal{L}_{\alpha,\beta}(W_l, z_{l,t}, a_{l,t}, \lambda) = \ell(z_{L,T}, y) + \langle z_{L,T} - \delta z_{L,T-1} - W_L a_{L-1,T}, \lambda\rangle + \\ + \sum_{l=1}^{L-1}\sum_{t=2}^{T} \alpha_{l,t} \left\| z_{l,t} - \delta z_{l,t-1} - W_l a_{l-1,t} + \vartheta a_{l,t-1} \right\|^2 + \\ + \sum_{t=2}^{T} \alpha_{L,t} \left\| z_{L,t} - \delta z_{L,t-1} - W_L a_{L-1,t} \right\|^2 + \\ + \sum_{l=1}^{L} \alpha_{l,1} \left\| z_{l,1} - W_l a_{l-1,1} \right\|^2 + \sum_{l=1}^{L-1}\sum_{t=1}^{T} \beta_{l,t} \left\| a_{l,t} - h_l(z_{l,t} - \vartheta) \right\|^2 \\ .\end{aligned}$$

This formulation is equivalent to relaxing the non-convex constraints via l_2-norm penalties. At this point, the partial derivatives with respect to each variable must be taken to obtain the primal updates, to which the dual update of the Lagrange multiplier must be added. These updates, according to the ADMM algorithm, must be iteratively performed until reaching convergence. The full updates are given in the following Algorithm 3, which also uses the subroutine in Algorithm 4 for z update. This update requires a special projection due to the presence of the Heaviside function.

```
for $l = 1, \dots, L-1$ do
    $W_l^+ \leftarrow \left(\sum_{t=1}^T x_{l,t} a_{l-1,t}^T\right) \left(\sum_{t=1}^T a_{l-1,t} a_{l-1,t}^T\right)^{-1}$
    for $t = 1, \dots, T-1$ do
        $a_{l,t}^+ \leftarrow \left(\vartheta^2 \mathbb{1}_{\mathcal{A}}(l) + W_{l+1}^T W_{l+1} + I\right)^{-1} [-\vartheta w_{l,t+1} \mathbb{1}_{\mathcal{A}}(l) + W_{l+1}^T (v_{l+1,t} \mathbb{1}_{\mathcal{A}}(l) + u_{L,t} \mathbb{1}_{\mathcal{B}}(l)) + \mathbb{1}(z_{l,t} - \vartheta)]$
        $z_{l,t}^+ \leftarrow \texttt{check\_entries}\left(\frac{q_{l,t} + \delta(r_{l,t+1} + \vartheta a_{l,t}^+)}{1+\delta^2}, \text{params}\right)$
    end for
    $a_{l,T}^+ \leftarrow \left(W_{l+1}^T W_{l+1} + I\right)^{-1} [W_{l+1}^T (v_{l+1,t} \mathbb{1}_{\mathcal{A}}(l) + (u_{L,t} - \frac{1}{\rho}\lambda) \mathbb{1}_{\mathcal{B}}(l)) + \mathbb{1}(z_{l,T} - \vartheta)]$
    $z_{l,T}^+ \leftarrow \texttt{check\_entries}(q_{l,T}, \text{params})$
end for
$W_L^+ \leftarrow \left(-\frac{1}{\rho}\lambda a_{L-1,T}^T + \sum_{t=1}^T x_{L,t} a_{L-1,t}^T\right) \left(\sum_{t=1}^T a_{L-1,t} a_{L-1,t}^T\right)^{-1}$
for $t = 1, \dots, T-1$ do
    $z_{L,t}^+ \leftarrow \frac{s_{L,t} + \delta r_{L,t+1}}{1+\delta^2}$
end for
$z_{L,T}^+ \leftarrow \frac{\rho s_{L,T} + y - \lambda}{1+\rho}$
$\lambda^+ \leftarrow \lambda + \rho(z_{L,T} - \delta z_{L,T-1} - W_L a_{L-1,T})$
```

Algorithm 3. ADMM Optimizer for SNNs

```
procedure CHECK_ENTRIES($z$, function cost())
    for each entry of z do
        $\Delta_1 \leftarrow \beta_{l,t}(1 - 2a_{l,t}^{n_l,m})$    ▷ Activation cost variation.
        $\Delta_2 \leftarrow J_{l,t}^{n_l,m} - \alpha_{l,t}(\vartheta - q_{l,t}^{n_l,m})^2 - \alpha_{l,t+1}(r_{l,t+1}^{n_l,m} + \vartheta a_{l,t}^{n_l,m} - \delta\vartheta)^2$ ▷ LSQ cost variation.
        if $z_{l,t}^{n_l,m} > \vartheta$ & cost($\vartheta$) $\leq$ cost($z_{l,t}^{n_l,m}$) then▷ $z$ is active but switching it off would yield a better cost.
            $z_{l,t}^{n_l,m} \leftarrow \vartheta$    ▷ Highest value to commute the Heaviside function.
        else if $z_{l,t}^{n_l,m} \leq \vartheta$ & cost($\vartheta + \varepsilon$) $<$ cost($z_{l,t}^{n_l,m}$) then    ▷ $z$ is inactive but turning it on would yield a better cost.
            $z_{l,t}^{n_l,m} \leftarrow \vartheta + \varepsilon$    ▷ Lowest value to commute the Heaviside function.
        end if
    end for
end procedure
```

Algorithm 4. Subroutine to find the z variable minimizer depending on the value of the Heaviside step function

Validation and Results

Preliminary results are presented in this section, obtained with a two hidden layers SNN composed of 32 and 64 LIF neurons, respectively. The test was conducted on a batch of data (300 samples) of the neuromorphic-MNIST (N-MNIST) dataset, where each sample has 200 time steps.

Two methods are compared: in the orange line, the version of the ADMM as given in Algorithm 3 is shown, whereas the blue line with the shaded region presents a stochastic version where the order of the updates is randomized along layer and time dimensions. The average of 10 runs is plotted together with the maximum and minimum results, represented by the shaded region. As can be seen from Figure , the stochastic version outperforms the deterministic version in terms of accuracy, as it happens for SGD vs regular GD. This is due to the non-convexity of the problem: Stochasticity helps to avoid poor local minima. It is worth mentioning that, despite the loss being slightly better for the deterministic version, the model dynamics are not respected as shown in Figure , hence producing poor accuracy in the inference phase. In Figure , the convergence characteristics of the two methods are shown. While both methods show the convergence of the residuals at the last layer (deterministic ADMM is even faster), the Lagrangian cost diverges for the orange curve after epoch 400. This proves the inability of the deterministic version to stabilize the dynamics of the SNN in the hidden layers, resulting in poorer performance.

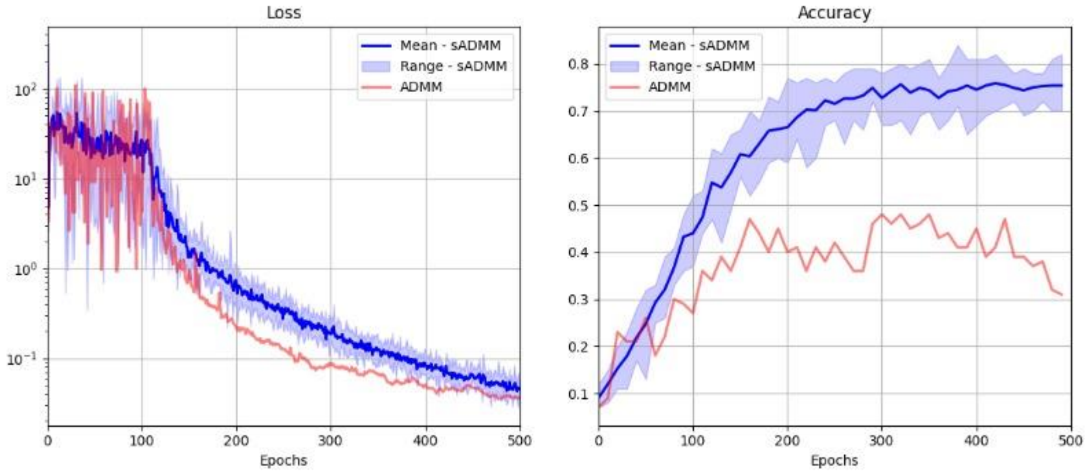


Figure 26. Loss (left) and accuracy (right) obtained by optimizing a two hidden layers SNN based on a batch of data. Comparison between a deterministic and stochastic version of the ADMM

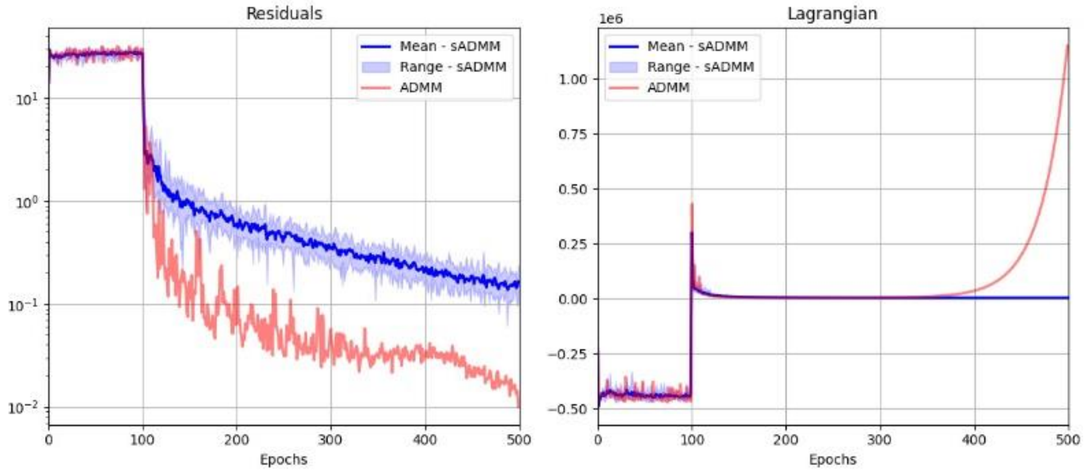


Figure 27. Convergence of the proposed method. Dual residuals of the constraints (left) and value of the associated relaxed Lagrangian (right)

Extension to decentralized and split learning

The problem formulation and algorithm presented in this section are tailored for a centralized training of SNNs where the full dataset and full neural architecture are kept within a common storage system to which the processors have access. However, the framework can be extended to two decentralized scenarios of interest according to the method presented in Section 3.1.4. Specifically, we can perform

- **Decentralized peer-to-peer training.** With the addition of constraints imposing the equality of weights learned on multiple nodes and the relative Lagrange multipliers, the decentralized training of the global model can be reached. In this setting, like in federated learning, each node holds a private dataset not to be shared with others and communicate only a part of the Lagrange multipliers to the neighbors. The convergence speed in this case depends on the communication topology.
- **Split training.** The split training setting is the case where each node holds a part of the neural architecture (e.g., a layer) and the global model is trained collaboratively. In this case duplicate variables must be created containing the useful information of the previous/following layer/block of neural network and each node must share Lagrange multipliers with the node holding the previous/following layer/block. The dataset should be kept by the node processing the input layer.

Use within ROBUST-6G System Design

The ADMM-based optimizer for the SNNs algorithm described above will contribute to the development of the spiking neural network simulator, which is useful for the trustworthy and sustainable AI services layer.

The algorithm is important because it allows energy-efficient inference with scalable models at the edge. Because of its possible extensions as decentralized learning, it can be integrated into a decentralized FL component but also used as a standalone component to perform efficient computing.

SNNs have shown particularly promising performance for time-series processing and have been applied with success at the physical layer. If dedicated hardware is used, processing can be made faster and more energy efficient, however SNNs can be also coded in field programmable gate arrays (FPGAs), a commonly used device in telecommunication networks. Hence, possible applications are foreseen in the WP5.

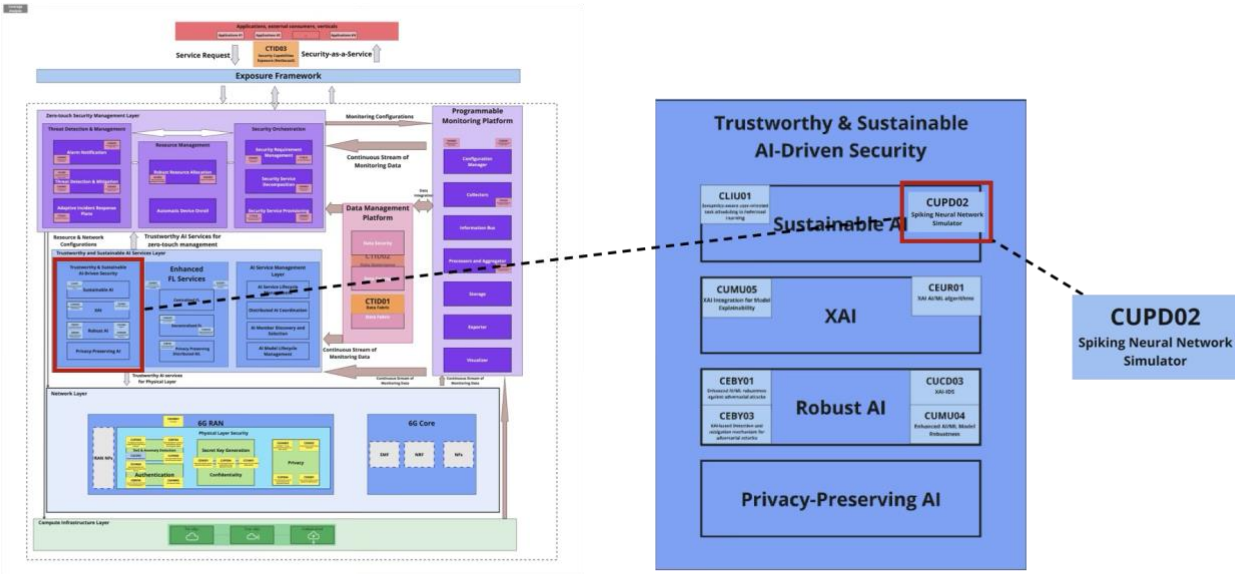

Service Request
CTID03
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Monitoring Configurations
Programmable Monitoring Platform
Continuous Stream of Monitoring Data
Data Management Platform
CTID01
Enhanced FL Services
AI Service Management Layer
Trustworthy and Sustainable AI Services Layer
Network Layer
6G RAN
6G Core
Compute Infrastructure Layer
Trustworthy & Sustainable AI-Driven Security
CLIU01
Sustainable AI
CUPD02
Spiking Neural Network Simulator
CUMU05
XAI Integration for Model Explainability
XAI
CEUR01
XAI AI/ML algorithms
CEBY01
Robust AI
CUCD03
XAI-IDS
CEBY03
CUMU04
Enhanced AI/ML Model Robustness
Privacy-Preserving AI
CUPD02
Spiking Neural Network Simulator

Figure 28. Integration of ADMM-based optimizer for SNN algorithm to the ROBUST-6G Architecture

4.2 Semantics-Aware User-Oriented Task Scheduling in Federated Learning

Federated Learning (FL) is significantly challenged by energy consumption, particularly during client-side computations, which constitute the largest portion of energy usage across all FL stages, including server-side computation, downlink communication, client-side computation, and uplink communication [YGS+23, TGF+25]. The energy efficiency of FL is primarily determined by the frequency of model aggregations, highlighting the critical need to balance FL performance and energy consumption.

This study explores the potential benefits of semantic-aware participation strategies for FL clients with limited battery resources. Particularly, the work presents investigation into the optimization of client participation strategies within EHFL. Through analysis of the interplay between battery capacity, computational cost, and charging probabilities, the potential of battery-aware and training-oriented client selection approaches to enhance energy efficiency and performance stability needs to be demonstrated. To address the client selection problem, a battery-aware client participation approach is proposed. This approach enables clients to engage in global model aggregation based on their battery levels, thereby minimizing redundant local training and transmission, and ultimately achieving satisfactory performance with significantly reduced energy consumption compared to standard FL.

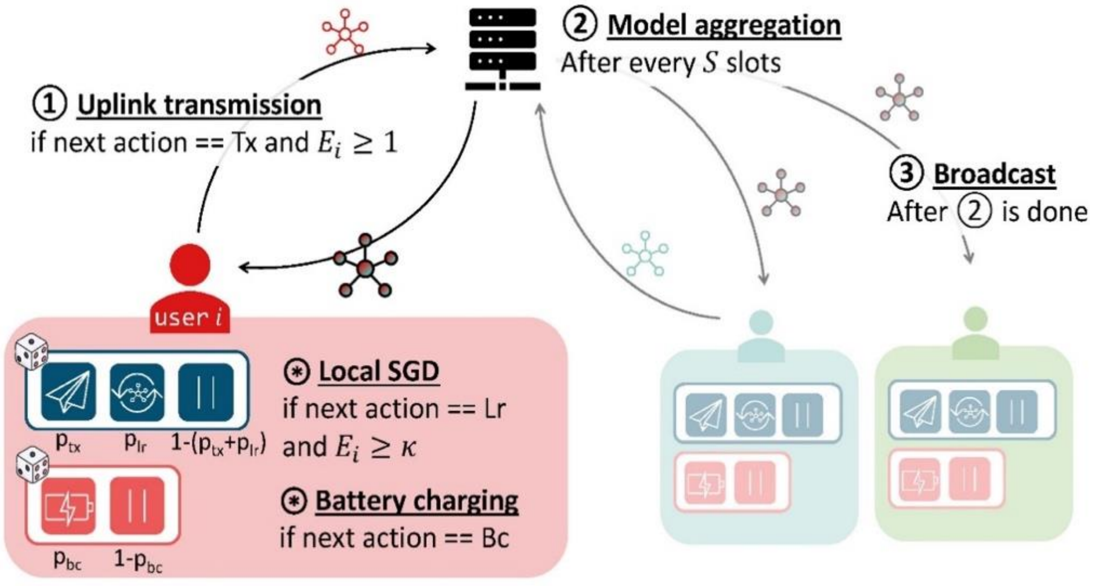


Figure 29. A Schematic View of EHFL based on Probabilistic Decision Rules

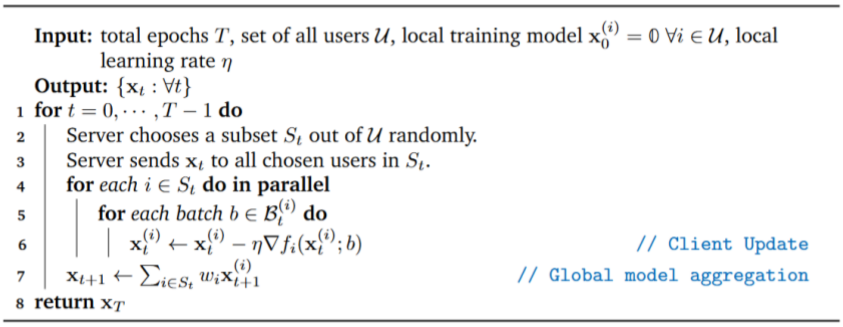
Energy-Harvesting FL (EHFL) with Probabilistic Decision Rules

In the context of IoT deployments, where sensor nodes often operate with limited energy resources, the following EHFL model can be considered. Particularly, it is applicable to scenarios in which an IoT node must decide among transmitting data, training a local model, or remaining idle based on its current battery level and probabilistic rules.

Initially, this study examines EHFL where transmission, local model training, and battery charging are governed by predefined probabilities. At each time slot, a user decides its next action, subject to the following constraints: users can charge one battery unit per slot but cannot simultaneously train a local model and transmit a message. A probability distribution determines the selection of one of three tasks: transmission (Tx), local model training (Lr), or remaining idle. Battery charging (Bc) events follow a separate probability distribution.

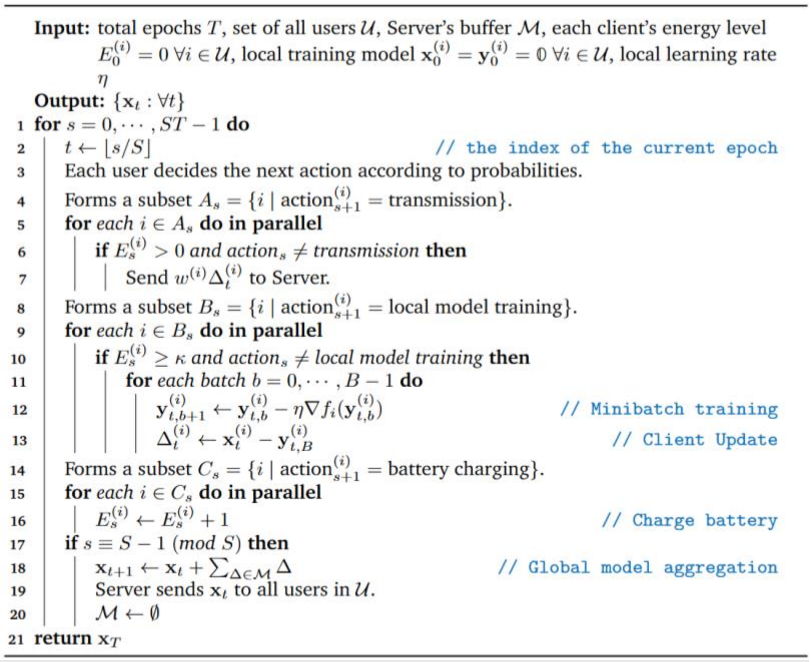
Battery levels are a fundamental constraint: users with zero battery level cannot perform any task. Actions resulting in negative battery levels are rejected. Transmission and battery charging each require one time slot (lasting a few milliseconds), whereas local model training requires k time slots, consuming k times the energy of a single transmission.

Once a user has a local update, its next action is transmission, provided it has sufficient battery. After every S slots (one epoch), the server aggregates all available updates in its buffer. Before initiating a new epoch, the server broadcasts the global model to all clients, consistent with standard FL systems.



```
Input: total epochs $T$, set of all users $\mathcal{U}$, local training model $\mathbf{x}_0^{(i)} = \mathbb{0}\ \forall i \in \mathcal{U}$, local
       learning rate $\eta$
Output: $\{\mathbf{x}_t : \forall t\}$
for $t = 0, \cdots, T-1$ do
    Server chooses a subset $S_t$ out of $\mathcal{U}$ randomly.
    Server sends $\mathbf{x}_t$ to all chosen users in $S_t$.
    for each $i \in S_t$ do in parallel
        for each batch $b \in \mathcal{B}_t^{(i)}$ do
            $\mathbf{x}_t^{(i)} \leftarrow \mathbf{x}_t^{(i)} - \eta \nabla f_i(\mathbf{x}_t^{(i)}; b)$        // Client Update
    $\mathbf{x}_{t+1} \leftarrow \sum_{i \in S_t} w_i \mathbf{x}_{t+1}^{(i)}$        // Global model aggregation
return $\mathbf{x}_T$
```

Algorithm 5. Vanilla Federated Averaging



```
Input: total epochs $T$, set of all users $\mathcal{U}$, Server's buffer $\mathcal{M}$, each client's energy level
       $E_0^{(i)} = 0 \; \forall i \in \mathcal{U}$, local training model $\mathbf{x}_0^{(i)} = \mathbf{y}_0^{(i)} = \mathbb{0} \; \forall i \in \mathcal{U}$, local learning rate
       $\eta$
Output: $\{\mathbf{x}_t : \forall t\}$
for $s = 0, \cdots, ST - 1$ do
    $t \leftarrow \lfloor s/S \rfloor$                                  // the index of the current epoch
    Each user decides the next action according to probabilities.
    Forms a subset $A_s = \{i \mid \text{action}_{s+1}^{(i)} = \text{transmission}\}$.
    for each $i \in A_s$ do in parallel
        if $E_s^{(i)} > 0$ and $action_s \neq transmission$ then
            Send $w^{(i)} \Delta_t^{(i)}$ to Server.
    Forms a subset $B_s = \{i \mid \text{action}_{s+1}^{(i)} = \text{local model training}\}$.
    for each $i \in B_s$ do in parallel
        if $E_s^{(i)} \geq \kappa$ and $action_s \neq local\ model\ training$ then
            for each batch $b = 0, \cdots, B - 1$ do
                $\mathbf{y}_{t,b+1}^{(i)} \leftarrow \mathbf{y}_{t,b}^{(i)} - \eta \nabla f_i(\mathbf{y}_{t,b}^{(i)})$        // Minibatch training
                $\Delta_t^{(i)} \leftarrow \mathbf{x}_t^{(i)} - \mathbf{y}_{t,B}^{(i)}$                  // Client Update
    Forms a subset $C_s = \{i \mid \text{action}_{s+1}^{(i)} = \text{battery charging}\}$.
    for each $i \in C_s$ do in parallel
        $E_s^{(i)} \leftarrow E_s^{(i)} + 1$                          // Charge battery
    if $s \equiv S - 1 \; (mod\ S)$ then
        $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \sum_{\Delta \in \mathcal{M}} \Delta$           // Global model aggregation
        Server sends $\mathbf{x}_t$ to all users in $\mathcal{U}$.
        $\mathcal{M} \leftarrow \emptyset$
return $\mathbf{x}_T$
```

Algorithm 6. Energy Harvesting FedAvg with probabilistic decision rules

Training-Oriented Client Strategy for EHFL

Achieving satisfactory performance necessitates a certain number of transmissions and local model training sessions. The convergence rate of Federated Averaging (FedAvg) on non-IID data is proven to be one $O\left(\frac{1}{BT}\right)$ [LHY+20], where T represents the number of uplink communications and B is the number of training batches per client. User-centric participation strategies are essential for enabling clients to make independent participation decisions.

During the battery charging process, users must decide how to allocate their energy to enhance the stability of collaborative learning and improve KPIs. A client may prioritize a specific task (e.g., training or transmission) based on the "outdatedness" of its reference model, aligning with the concept of the version Age of Information (vAoI) [HPY+24].

The probabilistic decision rules are compared with a training-oriented client strategy for EHFL. The convergence rate of FL is inversely proportional to the total number of local training events. The objective is to maximize the number of local model training occurrences while ensuring battery charging follows a Bernoulli distribution. This strategy prioritizes local training whenever possible. Each client i trains its local model as soon as its battery level E_i reaches the minimum requirement for local training, denoted by κ. If an update has not been sent, the client performs uplink transmission as soon as $E_i > 0$.

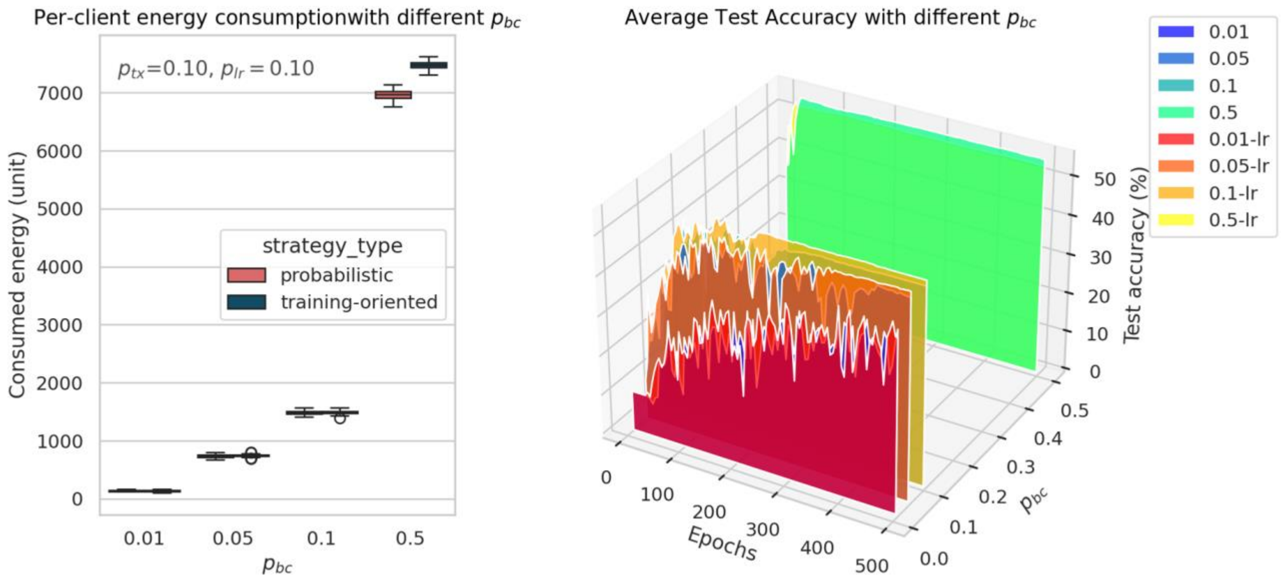


Figure 30. (Left) Boxplot of energy consumption per user for different battery charging probabilities (0.01 to 0.5). (Right) Average test accuracy for different battery charging probabilities

As the probability of battery charging events (p_{bc}) increases, users consume more energy due to frequent recharging, enabling more local model training. Conversely, a scarcity of battery charging leads to two key issues: first, slower convergence due to postponed transmission or training; and second, significant performance fluctuations affecting both average and individual test accuracy trends.

When applying training-oriented strategies, the overall trend remains consistent. However, these strategies enable clients to reach the saturation point faster than probabilistic decision rules. Despite this, prioritizing local training does not mitigate performance fluctuations. Given the experiment's IID user assumption, all users exhibit similar trends despite variations in battery levels and learning progress. Thus, performance oscillations are not solely attributed to uneven battery supply; rather, client participation must also account for unstable communication and computation patterns arising from intermittent energy harvesting.

Cyclic Client Participation in EHFL

Cyclic client participation can achieve faster asymptotic convergence rates compared to vanilla FedAvg [MMR+17] with uniform client participation, under suitable conditions [CSJ+23]. Selecting clients based on participation frequency can accelerate convergence at a rate of $O\left(\frac{1}{VT}\right)$, where V is a constant determined by data heterogeneity and partial client participation.

Beyond faster convergence, cyclic participation offers additional advantages. By grouping clients and enabling consecutive participation, the first group immediately transmits local updates, while subsequent groups have additional time for local training preparation. Grouping clients based on their training likelihood (in descending order) facilitates more efficient energy utilization and stabilizes performance trends.

The primary objective of cyclic client participation in EHFL is to minimize performance fluctuations by controlling the average number of participants per communication round, while simultaneously reducing energy consumption by avoiding redundant client-side computations. Initial results indicate that prioritizing local training can accelerate convergence, although complete mitigation of performance fluctuations arising from intermittent energy harvesting is not achieved. Furthermore, the concept of cyclic client participation has been explored as a promising method to control participant numbers and stabilize performance, thereby reducing energy consumption and minimizing redundant computations. The simulations with cyclic participation in EHFL will be finalized, with a thorough evaluation of its impact on convergence rates and energy efficiency. Additionally, the exploration of advanced client grouping and scheduling algorithms for cyclic participation, aiming to further optimize energy utilization, is intended. Furthermore, the integration of semantic-aware participation strategies based on version Age of Information (vAoI), to guide client decisions by assessing model "outdatedness," will be pursued as a challenging yet potentially high-impact contribution. Extensive simulations and real-world experiments are intended to validate the proposed strategies and quantify their impact on various performance metrics, including convergence speed, energy consumption, and model accuracy.

Use within ROBUST-6G System Design

The Semantics-aware user-oriented task scheduling in Federated Learning will contribute to ROBUST-6G MS3.1 by suggesting optimized client scheduling policies that minimize redundant computation and transmission. Additionally, the proposed strategies will also contribute to MS3.5 as the inference models are obtained in a distributed manner in which the energy harvesting system, represented as battery level status, formulates a semi-Markov model. The framework takes an important role since it enables the clients to achieve energy efficiency by reducing the carbon footprint of distributed intelligent networks.

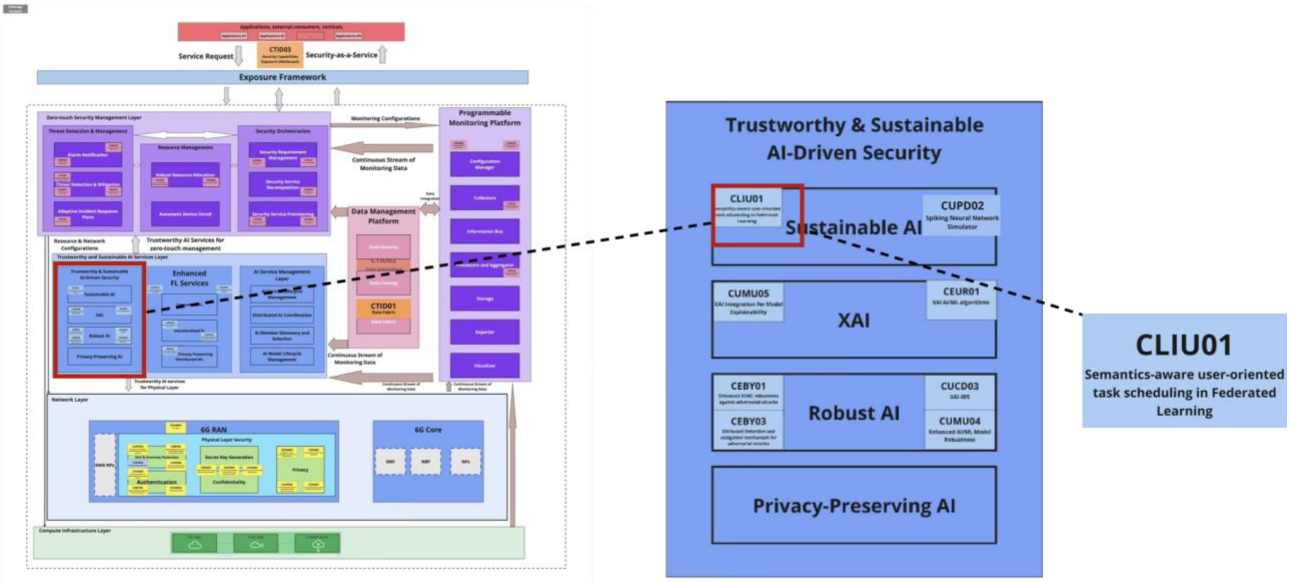

Applications, external consumers, verticals
Service Request
CTID03
Security-as-a-Service
Exposure Framework
Zero-touch Security Management Layer
Threat Detection & Management
Resource Management
Security Orchestration
Monitoring Configurations
Programmable Monitoring Platform
Continuous Stream of Monitoring Data
Data Management Platform
CTID02
CTID01
Trustworthy AI Services for zero-touch management
Trustworthy and Sustainable AI Services Layer
Trustworthy & Sustainable AI-Driven Security
Enhanced FL Services
AI Service Management Layer
Network Layer
6G RAN
6G Core
Compute Infrastructure Layer
Trustworthy & Sustainable
AI-Driven Security
CLIU01
Sustainable AI
CUPD02
Spiking Neural Network Simulator
CUMU05
XAI Integration for Model Explainability
XAI
CEUR01
XAI AI/ML algorithms
CEBY01
Enhanced AI/ML robustness against adversarial attacks
CEBY03
Robust AI
CUCD03
XAI IDS
CUMU04
Enhanced AI/ML Model Robustness
Privacy-Preserving AI
CLIU01
Semantics-aware user-oriented task scheduling in Federated Learning

Figure 31. Integration of Semantics-aware User-Oriented Task Scheduling to ROBUST-6G Architecture

5 Integrating XAI Measures into the Robust 6G Architecture

An essential aspect of ROBUST-6G is its integration of XAI, which is fundamental in achieving transparency, reliability, and real-time adaptability for security management. By incorporating XAI into the ROBUST-6G framework, security mechanisms become explainable, resilient, and automated, minimizing human involvement while enhancing network protection. For example, with XAI-driven monitoring and analysis, the Programmable Monitoring Platform (PMP) consistently collects security performance data and relays them to the AI Service Management Layer to optimize models. Security insights provided by XAI are also communicated to the Zero-Touch Security Management (ZSM) Layer and the Exposure Layer, supporting ongoing adaptation and enhancement of security protocols. For this reason, Figure is drawn to illustrate a preliminary design for embedding the XAI module within the ROBUST-6G framework. A central component of this integration is to showcase providing explainability feature of trustworthy AI with real-time monitoring and adaptability, achieved through the PMP, which continuously collects security performance data and provides feedback to the AI Service Management Layer for fulfilling given request. In addition, security insights are communicated to the Zero-Touch Security Management Layer and the Exposure Layer, allowing continuous improvements and enhancements. This design supports the ROBUST-6G vision by integrating explainable, resilient, and automated security mechanisms, reducing the need for human intervention while improving network protection against adversarial threats.

The procedure begins at the Exposure Layer, where a user sends a security service request, represented as a Robust IDS in the Figure 32. This request is then sent to the Zero-Touch Security Management Layer. Here, the Security Service Orchestration handles it and forwards it to the Security Resource Orchestration module, tasked with the dynamic allocation of security-related tasks, such as data gathering from the Infrastructure Layer. The S-CL (Security Control Layer) Management then organizes security mechanisms and initiates an AI service request, which is directed to the Trustworthy & Sustainable AI Services Layer. In the AI Service Management Layer, the AI Model Lifecycle Management addresses the request and triggers the XAI module. This module enhances security by providing explainability evaluation, compliance, transparency, robustness, and detection. Given that the main function is adversarial detection using XAI, the system initially activates the Robustness component, which then engages with the Robust AI module to identify adversarial attacks.

After the model is finalized, it is returned to S-CL Management before being sent to Domain Analytics within the Cloud Layer for deployment. The Infrastructure Layer integrates fundamental components for 6G networking and computing, comprising the 6G RAN, 6G Core, and features for the Physical Layer Security Module. Furthermore, the Virtualization Layer contains SDN Controllers that facilitate adaptable network configurations, while Domain Analytics utilizes both data and AI models to produce security insights. The data collected and optimized models are then applied through Domain Analytics and transmitted back via the Programmable Monitoring Platform, ensuring smooth, real-time security adjustments within the ROBUST-6G architecture.

5.1 Illustrating XAI Integration through a Security Service Request in ROBUST-6G

To demonstrate how XAI is embedded into the ROBUST-6G architecture, we provide a detailed walkthrough of a representative use case: a Robust IDS request initiated by a user. This example highlights how XAI facilitates transparency, real-time feedback, and automated adaptability across the various architectural layers—ranging from user-facing exposure interfaces to AI model lifecycle management and infrastructure interaction.

The Figure accompanying this section depicts a high-level architectural flow where an IDS service request triggers coordinated processes across the Exposure Framework, ZSM Layer, Trustworthy & Sustainable AI Services Layer, Infrastructure Layer, and the PMP. This flow demonstrates how explainable, trustworthy, and adaptive AI services are orchestrated to meet specific security demands in a 6G environment.

The integration of the XAI module ensures that decisions and actions taken by the AI models are interpretable, compliant, and robust—particularly in response to adversarial threats. Furthermore, the system supports continuous learning and improvement via KPI and QoS feedback loops, reinforcing the system's long-term resilience and self-optimization capabilities.

Below is a step-by-step breakdown of the process:

1. A user initiates a security service request—for example, requesting a **Robust IDS**—through the **Exposure Framework**.
2. The request is forwarded to the **ZSM**, where it is handled by the **Security Orchestration** module.
3. The **Security Orchestration** coordinates with both **Resource Management** and **Security CL Management** to begin provisioning and service preparation.
4. **Resource Management** requests necessary resources and data from the **Infrastructure Layer**, which includes components like 6G RAN/Core and the virtualized network.
5. The **Infrastructure Layer** collects relevant operational data and feedback, forwarding it to the **Domain Analytics** module for initial analysis of the requested service.
6. The **Security CL Management** uses orchestration logic to formalize the request into an **AI Service Call** and forwards it to the **AI Service Management Layer**.
7. The **AI Service Management Layer** triggers the **AI Model Lifecycle Management** process, initiating preparation of the model and needed services.
8. The **XAI Module** is activated to evaluate the model's explainability, compliance, robustness, transparency, and adversarial detection capabilities.
9. If adversarial robustness is required, the **Robustness component** engages with the **Robust AI** module to enhance the model's defensive capabilities.
10. Once the necessary parameters and requirements are collected for the requested AI service, the finalized model is sent back to **Security CL Management** for navigation toward deployment.
11. The **Domain Analytics** module (which can be placed flexibly in the architecture—here assumed to be in the cloud) deploys the model, analyzes its performance, and generates actionable insights.
12. **KPI and QoS metrics** are collected via the **Programmable Monitoring Platform (PMP)** and relayed back into the architecture.

⑬ These monitoring insights are fed into the **AI Model Lifecycle Management** for continuous model tuning and optimization.

⑭ Security insights are shared with the **ZSM Layer** to allow adaptive updates of security configurations and policy orchestration.

⑮ The final security service—enhanced with explainable AI—is exposed to the user as **Security-as-a-Service**, completing the end-to-end automated loop.

5.2 Other Security Service Requests Suitable for XAI-enabled integration:

Beyond the example of the Robust IDS, many other security services can be requested via the Exposure Framework and fulfilled through the same layered orchestration enabled by ROBUST-6G. These requests typically require coordination across AI, management, infrastructure, and monitoring components, making them ideal for full-stack flows with explainable, adaptive AI at the core.

Below is a list of potential security service requests that would follow a similar architectural flow:

1. Request for **Zero-Trust Network Access (ZTNA)** enforcement across edge devices.
2. Request for **Anomaly Detection for Encrypted Traffic** using XAI-driven inference.
3. Request for **Real-Time Threat Intelligence Integration** with dynamic policy updates.
4. Request for **Federated Learning Model Deployment** for distributed threat detection.
5. Request for **Explainability Assessment** of an already deployed AI firewall.
6. Request for **AI-Based DDoS Detection and Mitigation** tailored for ultra-low latency.
7. Request for **Adversarial Attack Simulation** to evaluate model robustness.
8. Request for **Automated Security Patch Orchestration** across the 6G Core and RAN.
9. Request for **Malware Detection as a Service** for containerized edge workloads.
10. Request for **Compliance Audit and Traceability Reporting** using XAI tools.
11. Request for **Secure Bootstrapping of IoT Devices** using lightweight trust models.
12. Request for **Dynamic Risk Score Calculation** for industrial or vehicular endpoints.
13. Request for **Trustworthy Decision Logs** of past AI security actions.
14. Request for **Security Policy Simulation and Validation** across network slices.
15. Request for **Data Provenance Tracking and Verification** for sensitive data flows.
16. Request for **Context-Aware Security Enforcement** in immersive XR environments.
17. Request for **Energy-Aware Security Analytics** in resource-constrained edge nodes.
18. Request for **SLA Violation Detection** for latency- and security-critical services.
19. Request for **Secure Collaboration Frameworks** for multi-domain service chains.

While the above list presents a diverse set of security service requests that can be triggered via the ROBUST-6G architecture, it is important to note that not all requests will follow the exact same sequence or require interaction with every architectural layer. In particular, data gathering (Step 4) from the Infrastructure Layer may not be necessary in all scenarios. For example, requests involving historical audit log retrieval, policy verification, or compliance checks might rely entirely on metadata and stored information available within the ZSM Layer or other control components, without needing fresh telemetry from the network. Similarly, the deployment location (e.g., far edge, near edge, central cloud) or the execution point of analytics may vary based on latency constraints, data locality, or use case sensitivity. However, for use cases that actively leverage XAI-particularly those involving dynamic detection, adaptation, or model retraining—the architectural flow tends to remain consistent, with ZSM coordination, AI lifecycle management, monitoring integration, and explainability evaluation being recurring elements across these flows.

To illustrate a lighter and more forensic use case within the ROBUST-6G architecture, the following flow outlines the steps involved in fulfilling a **historical audit log retrieval request**. Unlike real-time security services such as intrusion detection, this type of request does **not require data collection from the Infrastructure Layer** or model retraining. Instead, it leverages stored logs, decision metadata, and explainability modules to provide transparent insights into past security actions. The steps below demonstrate how this type of service is orchestrated efficiently through the Zero-Touch Security Management and AI Services layers:

① A user initiates a security service request through the **Exposure Framework**, requesting access to **historical audit logs** for a specific network segment, timeframe, or AI decision.

② The request is received by the **ZSM Layer**, specifically by the **Security Orchestration** module.

③ The **Security Orchestration** parses the request type, identifies it as a log retrieval task (not requiring real-time infrastructure data), and forwards it to the **Security CL Management**.

④ Since this request does not require fresh telemetry, **Resource Management is not involved**, and no data is pulled from the **Infrastructure Layer**.

⑤ The **Security CL Management** queries the internal **Security Logs Repository**, which contains audit trails of previously executed actions, decisions, and system responses.

⑥ If the request includes **explainability**, the **AI Service Management Layer** is notified to retrieve metadata on the relevant AI decisions.

⑦ The **AI Model Lifecycle Management** module accesses archived **model versions**, **inference contexts**, and **decision logs** from prior deployments.

⑧ The **XAI Module** is triggered in a **passive/explanatory mode**, providing insights into how the decisions were made (e.g., why an alert was generated or blocked).

⑨ The **Transparency** and **Compliance** components of XAI are used to generate a user-friendly explanation of the historical decisions.

⑩ No model training, robustness analysis, or detection is performed, as the purpose is purely forensic and explanatory.

⑪ The resulting audit package is passed back to the **Security CL Management**, which formats the output per policy.

⑫ The **Domain Analytics** module may be used to correlate events or summarize patterns if the request involves trend analysis or aggregated statistics.

⑬ If KPIs or trends are requested, the **Programmable Monitoring Platform (PMP)** accesses historical KPI/QoS datasets, without engaging live feedback loops.

⑭ The complete audit response is compiled and sent back through the **ZSM Layer** to be validated and securely exposed.

⑮ The final report—containing logs, AI explanations, and any analytics insights—is delivered to the user as a **Security-as-a-Service output**, fulfilling the request.

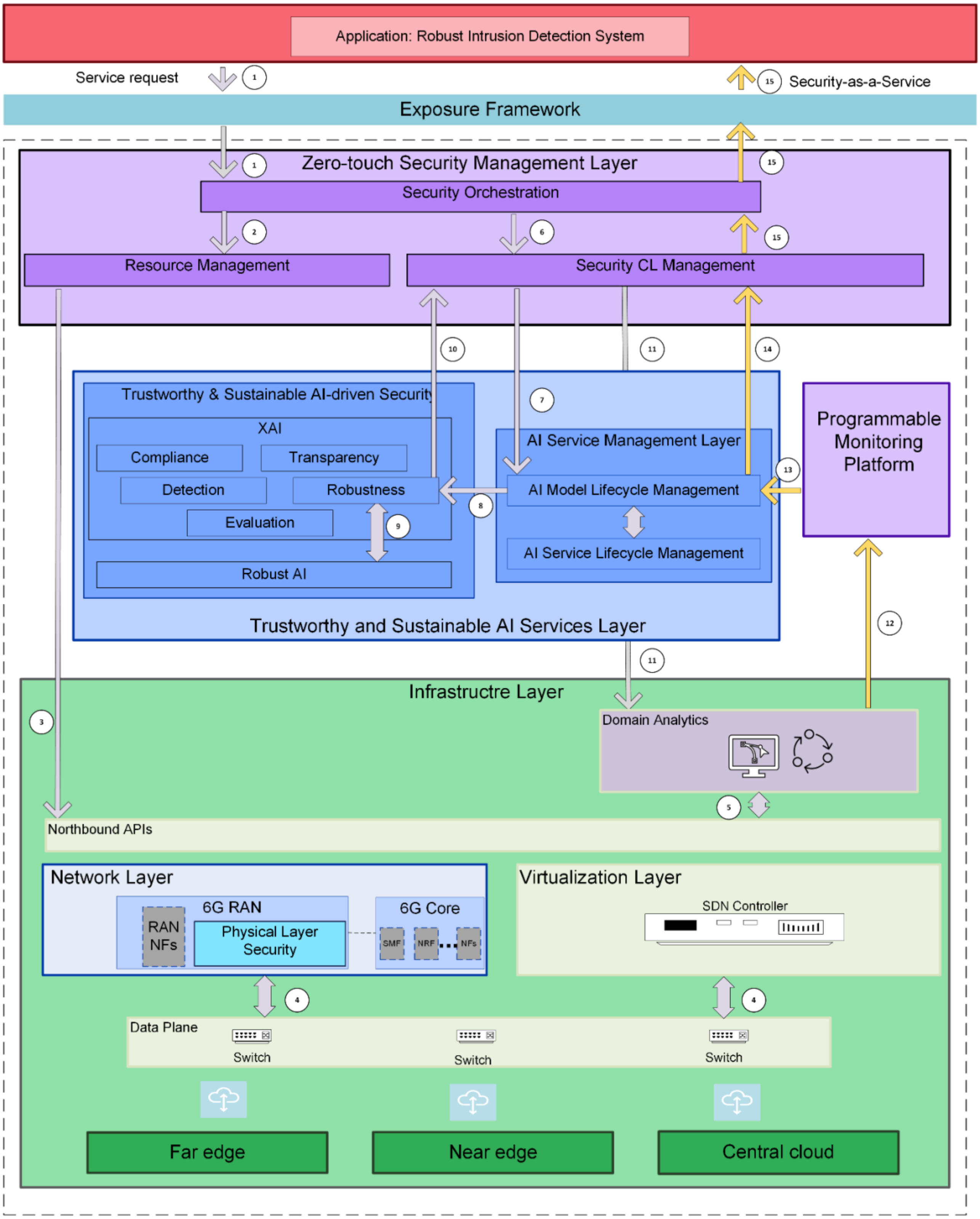


Figure 32 Initial ROBUST-6G architecture design for integration of XAI module

5.3 Requirements and Guidelines for XAI Measures for 6G Security

In 6G networks, XAI is not only essential for transparency but also plays a critical role in enhancing security, trust, and compliance. As AI-driven IDS become more sophisticated, it is imperative to establish clear requirements that ensure these systems remain resilient to adversarial attacks, aligned with legal and ethical frameworks, and accessible to diverse user groups.

The following requirements have been derived through a combination of industry best practices, regulatory guidelines (e.g., GDPR, AI Act), and security principles to ensure that XAI-IDS solutions for 6G are both trustworthy and effective. Each category of requirements addresses a critical aspect of system functionality:

- Security Requirements: Ensure resilience against adversarial attacks and unauthorized access while maintaining explainability without exposing sensitive data.
- Model Requirements: Define how explainability techniques should be integrated into the AI model's decision-making process to balance interpretability with accuracy.
- Legal and Ethical Requirements: Ensure compliance with AI regulations, privacy laws, and ethical guidelines to maintain fairness, transparency, and human oversight.
- Accessibility Requirements: Guarantee that XAI-IDS solutions are usable by all security analysts, including those with disabilities, by incorporating assistive technologies and multilingual interfaces.
- Usability Requirements: Focus on improving user experience by providing meaningful explanations, interpretability insights, and visualization tools to enhance decision-making.

By following guidelines, 6G systems can use AI to detect and counter threats while offering explanations that improve trust, compliance, and efficiency in complex network settings.

5.3.1 Security requirements

ID	Requirement description	Priority	Justification	Functional / Non-functional
SEC.RQ.1	The system should be resilient against adversarial attacks targeting explainability mechanisms.	SHOULD	Reduces risks of adversarial manipulation of feature importance scores.	Non-functional
SEC.RQ.2	The system must ensure that explainability techniques do not expose sensitive user data or attack patterns that could be exploited.	MUST	Prevents leakage of critical security insights that could benefit attackers.	Non-functional
SEC.RQ.3	The XAI-IDS must support role-based access control (RBAC) to restrict access to explanations based on user privileges.	MUST	Limits exposure of model insights to only authorized users.	Functional
SEC.RQ.4	The system should provide real-time explainability for critical alerts, ensuring timely incident response.	SHOULD	Helps security teams react quickly to threats by understanding AI-based decisions in real time.	Functional
SEC.RQ.5	The XAI-IDS must log explainability outputs to support forensic investigations.	MUST	Allows security analysts to review historical IDS explanations for compliance and threat analysis.	Functional
SEC.RQ.6	The IDS must be adaptable across different network environments, including virtualized and multi-vendor infrastructures.	MUST	Ensures robustness against adversarial attacks targeting diverse network configurations.	Functional

Table 7 Security requirements for integration of XAI in 6G

5.3.2 Model requirements

ID	Requirement description	Priority	Justification	Functional / Non-functional
MOD.RQ.1	The IDS framework should integrate XAI techniques for real-time adversarial detection and mitigation.	SHOULD	Enhances model interpretability, enabling better identification of adversarial threats.	Functional
MOD.RQ.2	The system should use explainability methods to assess the impact of adversarial samples on ML model decisions.	SHOULD	Improves understanding of feature significance and reduces false positives in IDS detection.	Functional
MOD.RQ.3	The IDS framework should support a zero-touch detection mechanism to enhance resilience against evolving attack techniques.	SHOULD	Enables proactive defence strategies without requiring manual intervention.	Functional
MOD.RQ.4	The XAI-IDS must have means integrate explainability techniques directly into the model training and evaluation phases.	MUST	Ensures transparency, interpretability, and trustworthiness beyond post-hoc explanations.	Functional
MOD.RQ.5	The system should utilize SHAP as the primary explainability method for feature importance ranking.	SHOULD	SHAP offers robust theoretical foundations for feature contribution analysis, ensuring accurate explainability.	Functional
MOD.RQ.6	The explainability methods should not significantly compromise detection accuracy.	SHOULD NOT	Balancing interpretability with predictive performance ensures an effective IDS.	Non-functional
MOD.RQ.7	The system should support confidence scores for predictions to assess classification reliability.	SHOULD	Enables risk assessment based on probabilistic model outputs.	Functional
MOD.RQ.8	The system may provide comparative performance evaluation before and after feature selection.	MAY	Demonstrates the effectiveness of feature refinement in improving model performance.	Functional
MOD.RQ.9	The system must be adaptable across different datasets and attack scenarios, including 5G/6G environments.	MUST	Ensures scalability and robustness across diverse cybersecurity challenges.	Functional
MOD.RQ.10	The system should minimize bias in feature importance rankings to ensure fairness in intrusion detection.	SHOULD	Prevents discrimination against certain types of traffic or network behaviours.	Non-functional

Table 8 Model requirements for integration of XAI in 6G

5.3.3 Legal and ethical requirements

ID	Requirement description	Priority	Justification	Functional / Non-functional
LEG.RQ.1	The system must allow human oversight and override mechanisms where security analysts can correct or refine explainability outputs.	MUST	Ensures human control over AI-driven decisions in cybersecurity.	Functional
LEG.RQ.2	The system must comply with GDPR and relevant data protection regulations regarding explainability in automated decision-making.	MUST	Ensures legal compliance and mitigates risks related to explainable AI in cybersecurity.	Non-functional
LEG.RQ.3	The XAI-IDS framework must provide an explainable reasoning mechanism for all AI-based security decisions, enabling end-users to understand the rationale behind alerts and classifications.	MUST	Ensures that individuals affected by automated IDS decisions can contest or request human intervention, aligning with GDPR.	Functional
LEG.RQ.4	The XAI-IDS must support mechanisms that allow users to access, rectify, and erase personal data captured in IDS logs where legally applicable.	MUST	Empowers individuals to maintain control over their personal data in security logs, per GDPR	Functional
LEG.RQ.5	The XAI-IDS must classify itself within the AI Act's risk categories and ensure compliance with associated legal obligations	MUST	Ensures regulatory alignment and legal compliance before deployment.	Functional
LEG.RQ.6	The XAI-IDS should provide auditability logs for AI-based decision-making processes and generate explainability reports for regulatory compliance.	SHOULD	Ensures traceability and accountability of AI-driven IDS decisions under AI Act transparency & documentation provisions.	Non-functional
LEG.RQ.7	The IDS system should implement gender-inclusive design principles in user interfaces and alert notifications if considered as needed i.e. not all systems require gender-balance.	SHOULD	Ensures that security professionals of all genders can effectively interpret alerts and system recommendations, following EU Gender Mainstreaming Guidelines.	Non-functional

Table 9 Legal and ethical requirements for integration of XAI in 6G

5.3.4 Accessibility requirements

ID	Requirement description	Priority	Justification	Functional / Non-functional
ACC.RQ.1	The XAI-IDS should provide an API for integration with other security tools (SIEMs, SOARs, etc.).	SHOULD	Enables interoperability with broader security ecosystems.	Functional

ACC.RQ.2	The system must provide accessibility features such as assistive technologies (screen readers, keyboard navigation) for analysts with disabilities.	MUST	Ensures compliance with accessibility standards and usability for all users.	Functional
ACC.RQ.3	The system should support multilingual explainability interfaces to accommodate diverse users.	SHOULD	Increases accessibility for global security teams.	Functional

Table 10 Accessibility requirements for integration of XAI in 6G

5.3.5 Usability requirements

ID	Requirement description	Priority	Justification	Functional / Non-functional
USB.RQ.1	The XAI-IDS should provide local and global interpretability insights for detected intrusions.	SHOULD	Helps analysts understand both individual and overall model decisions.	Functional
USB.RQ.2	The system must allow users to audit the decision-making process of the IDS through logs and reports.	MUST	Ensures traceability and accountability in AI-based security decisions.	Functional
USB.RQ.3	The system may integrate visualization tools for SHAP-based feature importance analysis.	MAY	Enhances user understanding and debugging of IDS decisions.	Functional
USB.RQ.4	The system must not introduce significant computational overhead that impacts real-time intrusion detection performance.	MUST NOT	Ensures that explainability methods do not slow down IDS operations.	Non-functional
USB.RQ.5	The system should include adaptive explainability techniques that adjust complexity based on user expertise (e.g., basic summaries for non-experts, detailed analysis for analysts).	SHOULD	Enhances usability by making explanations meaningful for different user roles.	Functional
USB.RQ.6	The system should provide an evaluation mechanism to assess normal vs. adversarial data behaviour.	SHOULD	Supports continuous monitoring and adaptation to new attack vectors.	Functional

Table 11 Usability requirements for integration of XAI in 6G

5.4 Design Guidelines for Integrating XAI In 6G Security Systems

While the detailed requirements for integrating XAI define specific system behaviors, design guidelines serve as a practical, high-level entry point to orient stakeholders toward building responsible, compliant, and secure systems. These guidelines were developed to:

- Translate technical requirements into actionable, role-aware principles
- Offer immediate orientation to stakeholders unfamiliar with in-depth system details
- Support early-stage planning and cross-disciplinary collaboratio
- Provide a foundation for consistently integrating explainability across the 6G security architecture

They are particularly useful before diving into requirement specifications, as they help contextualize why explainability matters in 6G security, what themes to prioritize, and how to align design intentions with functional, legal, and ethical goals. Many different stakeholders may use these guidelines, for example:

- **AI Developers & Data Scientists** can use these principles to guide model architecture, training workflows, and explanation technique selection (e.g., choosing SHAP or counterfactuals based on goal and stakeholder). Guidelines help them ensure XAI is not an afterthought, but part of the model lifecycle.
- **6G System Architects & Network Designers** can rely on these to incorporate explainability as a design goal across layers (e.g., Exposure, AI Services, ZSM). Guidelines help align high-level architectural decisions (e.g., where to place explainability modules) with trust and compliance goals.
- **Security Analysts & Operators** can use the guidelines to advocate for interpretability features that support day-to-day tasks like alert triage, incident response, and post-mortem forensics. They ensure that explainability tools are useful, understandable, and role-specific.
- **Policy Makers, Legal Teams & Ethics Boards** can Reference these as design-aware tools for assessing whether deployed systems meet ethical and regulatory expectations. Guidelines help connect abstract AI principles (e.g., human oversight, GDPR compliance) to real architectural touchpoints.
- **Vendors & Integrators** can use guidelines to evaluate whether third-party XAI solutions or modules align with broader 6G integration requirements. They enable early filtering before detailed requirement-level conformance testing.,

Each guideline is grounded in the extensive set of requirements derived in Section 5.3 , ensuring that they reflect both regulatory alignment (e.g., GDPR, EU AI Act) and technical feasibility within the 6G architecture.

5 Core Design Guidelines for XAI Integration in 6G:

#1: Secure Explainability by Design

XAI systems must be designed with built-in defences against adversarial manipulation, data leakage, and unauthorized access to explanations.

This principle ensures that explainability mechanisms are not weak links in security. Role-based access control (RBAC), explainability filtering, and resilience against feature manipulation must be embedded from the start. This protects sensitive model insights while maintaining transparency for authorized users.

#2: Embedded Explainability Throughout the Model Lifecycle

Explainability must be an integral part of the AI model lifecycle—from design and training to deployment and monitoring—enabling continuous transparency and adaptive trust.

Rather than relying on post-hoc methods, models must integrate explainability techniques during training and evaluation phases, ensuring interpretability is a core objective. Techniques like SHAP or counterfactuals should be used where appropriate, with minimal impact on detection performance.

#3: Regulatory Alignment and Human Oversight as Default

XAI-IDS must support full traceability, user control, and legal rights (access, rectification, erasure) to comply with GDPR, the EU AI Act, and emerging cybersecurity laws.

This design principle ensures that explainable systems are not only transparent but also contestable. Every AI-driven decision must be documented, justifiable, and reversible by human experts. This enhances legal accountability and ethical AI assurance.

#4: Inclusive and Accessible Explanation Interfaces

XAI systems must provide explanations that are accessible to users with varying abilities, roles, and linguistic backgrounds.

Explainability should not be a barrier—it must be democratized. Interfaces must include accessibility features (e.g., screen reader support, keyboard navigation) and offer multilingual options. This ensures inclusivity and usability across the global and distributed 6G operator landscape.

#5: User-Centric, Context-Aware Explanation Delivery

Explanations must be tailored to user expertise and operational context, supporting both high-level summaries and detailed forensic insights.

This guideline promotes adaptive explainability—giving novice users high-level rationale, while providing security analysts with in-depth interpretability tools (e.g., interactive SHAP visualizations, log-based reasoning trails). This enables trust and effective decision-making across roles.

5.5 Evaluation Metrics for XAI Integration in 6G

The integration of Explainable Artificial Intelligence (XAI) into 6G networks is essential to ensure transparency, interpretability, and trust in AI-driven decision-making. As AI becomes a foundational technology in 6G—supporting functions like security management, threat detection, traffic orchestration, and resource allocation—its decisions increasingly affect critical operations. However, many of these AI models operate as “black boxes,” making their reasoning difficult to interpret. This lack of explainability creates significant risks, particularly in high-stakes environments where compliance, accountability, and operational transparency are required.

In the context of the ROBUST-6G project, explainability is not a theoretical add-on—it is a core element of the project's vision for trustworthy, robust, and autonomous AI-driven security. As emphasized in Objective 3 of the project, XAI is critical for making AI models both operationally reliable and verifiably secure, especially under adversarial conditions. XAI metrics are thus instrumental in assessing whether the AI-based security functions integrated in ROBUST-6G meet the trust, fairness, privacy, and sustainability requirements defined in the project architecture.

To this end, evaluation metrics must capture both the technical quality of explanations and their human usability. These include:

- **Fidelity**, which assesses how faithfully an explanation reflects the underlying model's logic.
- **Interpretability**, which gauges how easily human operators (e.g., security analysts or network administrators) can understand and act upon the explanation.
- **Completeness**, referring to whether the explanation discloses all relevant factors in a decision.
- **Consistency**, which ensures similar inputs yield similar explanations—important in maintaining system trustworthiness.
- **Relevance**, which checks whether the explanation highlights the most impactful features.

These metrics support the design and deployment of AI agents in the ROBUST-6G Zero-Touch Security Management and Distributed AI-driven Security modules, where automated actions may need to be explained to human operators or certified under regulatory frameworks. XAI metrics also play a key role in the evaluation and validation of AI-based functions in Use Case 1, where explainability is applied to decentralized model generation and secure decision-making. For example, security operators may need to understand why a federated learning model flagged a particular traffic pattern as anomalous or whether a mitigation action proposed by the orchestrator is justifiable. In such contexts, XAI metrics help ensure that model outputs are both transparent and actionable.

To formalize this evaluation, ROBUST-6G leverages frameworks like REVEL (Robust Evaluation VEctorized Loca-linear-explanation), which define quantifiable metrics such as:

- **Local Concordance**: alignment of explanation with the model's local behavior.
- **Prescriptivity**: ability to identify how changes in input affect decisions.
- **Conciseness**: simplicity and clarity of explanations.
- **Robustness**: stability of explanations under small input perturbations.

These metrics will be used within the project to validate the trustworthiness of XAI-based components, especially those deployed for threat detection, privacy-preserving learning, and orchestration decision-making. Nevertheless, applying XAI metrics in a 6G setting presents unique challenges. Scalability becomes critical, as explanations must be generated for potentially thousands of edge devices and services. Real-time constraints require explanations to be low-latency and computationally efficient, especially in time-sensitive scenarios like attack mitigation. Furthermore, the diversity of users—from network engineers to external verticals—requires explanations to be adaptable in detail and format.

In summary, XAI metrics are a fundamental building block of ROBUST-6G's approach to trustworthy AI. They not only support the technical evaluation of models but also serve as enablers for secure, transparent, and accountable decision-making within the broader 6G ecosystem envisioned by the project.

5.6 Standardization and Compliance Recommendation

5.6.1 Current Standards available for XAI in 6G security.

The development of standards for Explainable AI (XAI) in 6G security is still in its early stages, with only a few working groups actively contributing to this effort. Currently, there are limited formalized standards, though organizations like IEEE, NIST, ISO, and W3C are engaged in research and discussions. The National Institute of Standards and Technology (NIST) introduced NISTIR 8312, titled "Four Principles of Explainable Artificial Intelligence," which, while not a formal standard, emphasizes ethics and human-centered AI. Similarly, the World Wide Web Consortium (W3C) published an online post in 2018 titled “Toward a Web Standard for Explainable AI?” signaling growing interest in XAI standardization.

As of March 2025, several standards and guidelines have been developed to enhance the transparency and trustworthiness of Artificial Intelligence (AI) systems through Explainable AI (XAI) methodologies. Notable among these are IEEE 2894-2024: Guide for Explainable Artificial Intelligence (XAI) Techniques and Their Evaluation, IEEE P2976: Standard for Explainable Artificial Intelligence (XAI), and European Union's Regulatory Framework for AI Systems. Following is a brief account of the aforementioned standards.

The IEEE 2894-2024: Guide for Explainable Artificial Intelligence (XAI) Techniques and Their Evaluation provides a structured approach to developing, implementing, and assessing explainable AI models. This standard covers key areas such as definitions and classifications of XAI techniques, application scenarios, evaluation metrics, and best practices for AI transparency. It establishes a technological framework for integrating explainability into AI systems, ensuring that machine learning models can provide understandable and interpretable outputs. Additionally, it outlines methods to assess explainability across different stakeholder groups, including developers, regulators, and end-users. A crucial focus of IEEE 2894-2024 is evaluating XAI methods using quantitative and qualitative metrics. The guide discusses different XAI techniques, such as feature attribution, rule-based explanations, counterfactual reasoning, and visual interpretability methods. It also provides high-level recommendations for selecting the most appropriate XAI technique based on the specific needs of AI applications across industries.

The IEEE P2976: Standard for Explainable Artificial Intelligence (XAI) [IEEE+24] aims to provide a framework for defining, developing, and evaluating explainable AI systems. This standard focuses on key areas such as defining explainability requirements, categorizing XAI methods, establishing evaluation metrics, and ensuring interoperability across AI models and systems. It sets forth mandatory and optional requirements that AI methods, algorithms, and applications must meet to be considered explainable. The standard also introduces XML schemas to facilitate the sharing and standardization of explainability-related information across different AI implementations. IEEE P2976 categorizes AI explainability into partially explainable and fully explainable methods, offering a structured approach to understanding how AI decisions are made. It defines technical and human-centered explainability requirements, ensuring that explanations are not only technically valid but also understandable to end users with different levels of expertise.

5.6.2 Gaps present in the current standards

Despite their comprehensive approaches, both IEEE P2976 and IEEE 2894-2024 have certain limitations. IEEE P2976 does not mandate specific XAI techniques for different use cases, creating ambiguity in practical implementation. It also lacks detailed guidance for domain-specific explainability challenges, such as those in healthcare and finance, where regulatory constraints vary. Furthermore, it does not extensively cover real-time XAI systems that require dynamic and adaptive explanations in critical applications. IEEE 2894-2024, while providing a structured framework for XAI, does not enforce mandatory requirements, leaving implementation choices to individual developers. It also lacks specific guidelines for real-time AI explainability, particularly in high-risk domains like autonomous systems and healthcare. Additionally, it does not fully address adversarial robustness, a crucial factor in AI security and trustworthiness.

5.6.3 Recommendations for standards

To address the limitations in existing standards, we recommend prioritizing real-time explainability, domain-specific XAI implementations, security enforcement on XAI components, and XAI based adversarial robustness. New standards should establish mandatory real-time/close to real time XAI mechanisms capable of detecting and mitigating inference-based adversarial attacks dynamically. This includes the standardization of uncertainty quantification techniques that adjust model responses based on uncertainty metrics, thereby preventing AI-as-a-Service (AIaaS) models from being exploited through adversarial queries. Furthermore, guidelines should define energy-efficient XAI mechanisms that enhance real-time model interpretability while minimizing computational overhead.

In addition, ADMM-based optimization techniques should incorporate privacy-preserving constraints to ensure secure communication and model updates while maintaining explainability. Secure aggregation techniques must also be standardized to mitigate risks associated with model inversion and poisoning attacks without compromising interpretability. Moreover, future standards should introduce explainable adversarial defense mechanisms that dynamically adapt to evolving attack vectors in real time. This should include XAI-based anomaly detection frameworks designed to enhance Intrusion Detection Systems (IDS) by improving explainable decision-making, reducing false positives, and strengthening overall system trustworthiness. With timely integration of these recommendations, future standards can effectively address ambiguities, real-time applicability, and adversarial security challenges in XAI for 6G networks, ensuring greater transparency, resilience, and regulatory compliance.

5.6.4 Benchmarking and Testing Standards for XAI in 6G Systems

Establishing benchmarking and testing standards for XAI in 6G systems is essential to ensure consistency, reliability, and regulatory compliance. Defining performance benchmarks for XAI requires the development of standardized metrics to evaluate model explainability across diverse 6G use cases, including network security, autonomous operations, and real-time decision-making. These benchmarks should assess the fidelity, interpretability, stability, and computational efficiency of XAI models, ensuring that explanations remain meaningful and actionable without introducing excessive complexity or latency. Furthermore, benchmarking should incorporate domain-specific evaluation criteria to account for the varying levels of explainability required in different sectors, such as telecommunications, finance, and healthcare.

To enhance XAI reliability, standardized validation approaches must be established to ensure consistency in interpretability methods across different AI models and applications. This includes the implementation of cross-validation techniques, adversarial robustness tests, and stress-testing frameworks that evaluate how XAI models perform under varying conditions, such as real-time inference, distributed learning environments, and adversarial settings. Additionally, validation protocols should ensure that XAI methods remain effective when applied to dynamic 6G infrastructures, where AI-driven decisions must be continuously monitored and adapted based on evolving network conditions.

Compliance testing is another critical aspect of XAI standardization, ensuring adherence to regulatory requirements related to AI fairness, bias detection, and interpretability. Testing frameworks should incorporate fairness-aware evaluation metrics that assess AI models for potential biases based on demographic, geographical, or systemic factors. Additionally, compliance protocols should require automated bias detection mechanisms that identify and mitigate discriminatory patterns in AI decision-making. Beyond fairness, interpretability testing should be mandated to confirm that AI-driven security and optimization strategies can be transparently understood and verified by stakeholders, including network operators, regulatory bodies, and end-users.

6 Conclusions

The findings in this report reinforce the necessity of a holistic approach to 6G network development, emphasizing robustness, security, and sustainability as fundamental design principles. The proposed architectural framework and enabling technologies offer a pathway toward future-proof 6G networks that can dynamically adapt to changing environmental conditions and user demands. A major takeaway from this deliverable is the integration of Explainable AI (XAI) in 6G security to enhance transparency, accountability, and trust. The adoption of XAI-based security mechanisms will ensure AI-driven decisions are interpretable, fair, and aligned with regulatory and ethical standards. Moreover, by addressing adversarial threats, federated learning vulnerabilities, and AI biases, this research paves the way for a more secure and resilient AI-enabled 6G ecosystem.

Each chapter of this deliverable contributes critical insights into different aspects of AI-driven security for 6G networks:

- Chapter 1 introduces the motivation, objectives, and scope of the report, emphasizing the need for AI-driven security in 6G. It sets the foundation for integrating XAI and energy-efficient security solutions.
- Chapter 2 identifies security threats in AI/ML for 6G, including adversarial attacks, model poisoning, and federated learning vulnerabilities. The chapter highlights risks such as evasion attacks and data poisoning that can compromise AI models in 6G networks.
- Chapter 3 explores XAI techniques such as SHAP, LIME, and saliency mapping to enhance the interpretability of AI-driven security mechanisms.
- Chapter 4 focuses on sustainable AI-driven security, emphasizing techniques such as pruning, quantization, and knowledge distillation to optimize energy efficiency while maintaining high security levels in 6G networks.
- Chapter 5 presents proactive security mechanisms, including adversarial training, secure model aggregation in federated learning, and AI-driven anomaly detection to enhance security resilience and mitigate evolving threats.

Sustainability remains a key priority in 6G development. AI-driven energy-efficient security solutions are critical to balancing performance with environmental impact. As 6G networks expand, optimizing resource allocation and minimizing power consumption through AI-based solutions will be crucial in meeting global sustainability goals. Future advancements in green AI, including energy-harvesting technologies and sustainable computing frameworks, will further contribute to the long-term viability of 6G systems.

Resilience and adaptability are also core components of 6G security operations. The ability to dynamically adjust security mechanisms in response to evolving threats is essential. Leveraging AI for real-time threat detection, proactive defense strategies, and anomaly detection will enable 6G networks to withstand emerging cyber risks while ensuring uninterrupted service availability.

Collaboration between academia, industry, and policymakers will be fundamental in realizing these advancements. Standardization efforts must align with technological innovations to ensure seamless integration of AI-driven security frameworks. ROBUST-6G actively collaborates with standardization bodies to ensure that AI-driven security solutions align with global standards, regulatory requirements, and industry best practices. This cooperation is essential for fostering interoperability, compliance, and the seamless integration of trustworthy and sustainable AI frameworks in 6G networks. Moreover, continued investment in interdisciplinary research is necessary to refine XAI methodologies, enhance adversarial defence mechanisms, and develop robust governance frameworks for AI-enabled 6G networks.

In conclusion, the success of 6G will depend on its ability to balance security, sustainability, and resilience. By leveraging XAI for transparency, optimizing AI-driven security for energy efficiency, and fostering collaboration among stakeholders, this research sets the foundation for a trustworthy, robust, and future-ready

6G network. Another crucial area for future exploration is the intersection of AI security and regulatory compliance. The deployment of AI-driven security measures in 6G must adhere to global standards on data protection, fairness, and ethical AI use. Developing standardized XAI assessment frameworks will be essential to ensuring explainability and transparency in AI-driven security decisions. The research also highlights the importance of interdisciplinary collaboration to drive innovation in AI-enabled 6G security. Close cooperation between AI researchers, cybersecurity experts, telecom industries, and policymakers will be necessary to align technological advancements with evolving security and ethical considerations.

The next phase of research (D3.3) will focus on further refining these strategies, ensuring real-world applicability and scalability for next-generation communication systems. Special emphasis will be placed on experimental validation, benchmarking security performance, and evaluating the trade-offs between security, energy efficiency, and network performance. By addressing these challenges, 6G networks can achieve a secure, resilient, and sustainable communication paradigm for the future.

7 References

[AB+23] Angelopoulos, A. N., Bates, S. (2023). *Conformal Prediction: A Gentle Introduction*. Foundations and Trends in Machine Learning, 16(4), 494–591. Now Publishers Inc. DOI: 10.1561/2200000101.

[ABC+18] Y. Adi, C. Baum, M. Cissé, B. Pinkas and J. Keshet, “Turning your weakness into a strength: Watermarking deep neural networks by backdooring”. In Proceedings of the 27th USENIX Security Symposium (USENIX Security 2018), pp. 1615-1631, August 2018.

[AGR+24] O. Arreche, T. R. Guntur, J. W. Roberts and M. Abdallah “E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection,” IEEE Access, vol. 12, pp. 23954-23988, 2024.

[AKN+19] A. Abusnaina, A. Khormali, D. Nyang, M. Yuksel, and A. Mohaisen, “Examining the robustness of learning-based ddos detection in software defined networks,” in 2019 IEEE Conference on Dependable and Secure Computing (DSC), pp. 1–8, IEEE, 2019.

[APA+21] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[BBA+23] A. Bashaiwth, H. Binsalleeh, and B. AsSadhan, “An Explanation of the LSTM Model Used for DDoS Attacks Classification,” Applied Sciences, vol. 13, no. 15, p. 8820, 2023.

[BEG+17] Blanchard, P., El Mhamdi, E.M., Guerraoui, R. and Stainer, J., 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in neural information processing systems, 30.

[BEM+23] A. Bemporad, “Training recurrent neural networks by sequential least squares and the alternating direction method of multipliers,” Automatica 156: 111183, 2023.

[BPC+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” in Foundations and Trends® in Machine learning, 3(1), 1-122, 2011.

[CAD+18] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “Adversarial attacks and defences: A survey,” arXiv preprint arXiv:1810.00069, 2018.

[CPC+19] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, p. 832, 2019.

[CSJ+23] Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 5677–5721. PMLR, 23–29 Jul 2023.

[DZL+20] Ding, J., Zhang, X., Li, X., Wang, J., Yu, R., & Pan, M. (2020, April). Differentially private and fair classification via calibrated functional mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 622-629).

[ETSI+24] ETSI. (2024-04). ETSI TR 104 225 V1.1.1: Technical report on [Securing Artificial Intelligence TC (SAI); Privacy aspects of AI/ML systems]. European Telecommunications Standards Institute. https://www.etsi.org/deliver/etsi_tr/104200_104299/104225/01.01.01_60/tr_104225v010101p.pdf

[FBS+20] G. Fidel, R. Bitton, and A. Shabtai, “When explainability meets adversarial learning: detecting adversarial examples using shap signatures,” in Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8.

[FYC+18] Fung, C., Yoon, C.J. and Beschastnikh, I., 2018. Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866.

[GA+19] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” AI magazine, vol. 40, no. 2, pp. 44–58, 2019.

[GA+24] Ghoukasian, H., & Asoodeh, S. (2024, July). Differentially private fair binary classifications. In 2024 IEEE International Symposium on Information Theory (ISIT) (pp. 611-616). IEEE.

[GF+24] B. G. Paltun and R. Fuladi, “Introspective intrusion detection system through explainable ai,” in 2024 8th Cyber Security in Networking Conference (CSNet), pp. 28–32, IEEE, 2024.

[GFE+25] B. G. Paltun, R. Fuladi, and R. El Malki “Robust Intrusion Detection System with Explainable Artificial Intelligence,” *arXiv preprint arXiv:2503.05303*, 2025.

[GG+16] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

[GKN+14] W. Gerstner, M. Kistler, R. Naud, and L. Paninski, “Neuronal dynamics: From single neurons to networks and models of cognition,” Cambridge University Press, 2014.

[GSS+25] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples", ICLR (Poster) 2015.

[HEXD5320] HEXA-X Project, D5.3: Final 6G architectural enablers and technological solutions,” https://hexa-x.eu/.

[HHG+19] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong. "DP-ADMM: ADMM-based distributed learning with differential privacy," in *IEEE Transactions on Information Forensics and Security* 15 (2019): 1002-1012.

[HPY+24] Xinyi Hu, Nikolaos Pappas, and Howard H Yang. Version age-based client scheduling policy for federated learning. In 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pages 695–699. IEEE, 2024.

[IEEE+24] "IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence," in IEEE Std 2894-2024 , vol., no., pp.1-55, 30 Aug. 2024, doi: 10.1109/IEEESTD.2024.10659410.

[IL+15] V. P. Illiano and E. C. Lupu, “Detecting malicious data injections in wireless sensor networks: A survey,” ACM Computing Surveys (CSUR), vol. 48, no. 2, pp. 1–33, 2015.

[KKP+23] M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, and A. Y. Zomaya,, “An Explainable Deep Learning-Enabled Intrusion Detection Framework in IoT Networks,” Information Sciences, vol. 639, pp. 119000, 2023.

[KMG+20] Z. Klawikowska, A. Mikołajczyk, and M. Grochowski, “Explainable AI for inspecting adversarial attacks on deep neural networks,” in Artificial Intelligence and Soft Computing. Cham, Switzerland: Springer, 2020, pp. 134-146.

[KWK+20] Kwon, Y., Won, J. H., Kim, B. J., & Paik, M. C. (2020). Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics & Data Analysis, 142, 106816.

[LHY+20] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In International Conference on Learning Representations, 2020.

[LL+17] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, vol. 30, 2017.

[LYM+22] Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., ... & Philip, S. Y. (2022). Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*.

[LZS+20] Liu, L., Zhang, J., Song, S. H., & Letaief, K. B. (2020, June). Client-edge-cloud hierarchical federated learning. In *ICC 2020-2020 IEEE international conference on communications (ICC)* (pp. 1-6). IEEE.

[ML+24] Meerza, S. I. A., & Liu, J. (2024) 'EAB-FL: Exacerbating Algorithmic Bias through Model Poisoning Attacks in Federated Learning', Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), pp. 458–466. Available at: [DOI: 10.24963/ijcai.2024/51]

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pages 1273–1282. PMLR, 20–22 Apr 2017.

[NAA+22] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities," IEEE Access, vol. 10, pp. 112392–112415, 2022.

[NKG+21] B. Nugraha, N. Kulkarni, and A. Gopikrishnan, "Detecting adversarial ddos attacks in software-defined networking using deep learning techniques and adversarial training," in 2021 IEEE International Conference on Cyber Security and Resilience (CSR), pp. 448–454, IEEE, 2021.

[NMZ+19] E. O. Neftci, H. Mostafa and F. Zenke, "Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks," in IEEE Signal Processing Magazine, vol. 36, no. 6, pp. 51-63, Nov. 2019, doi: 10.1109/MSP.2019.2931595.

[NRD+22] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent o-ran for beyond 5g and 6g wireless networks," in 2022 IEEE Globecom Workshops (GC Wkshps), pp. 215– 220, IEEE, 2022.

[PAK+24] I. Pitsiorlas, G. Arvanitakis, and M. Kountouris, "Trustworthy Intrusion Detection: Confidence Estimation Using Latent Space", in Proc. WMLC, in conjunction with 22nd Int. Symp. on Modeling and Optim. in Mobile, Ad hoc, and Wir. Networks (WiOpt 2024), Seoul, S. Korea, October 2024.

[RBG+22] B. Rueckauer, C. Bybee, R. Goettsche, Y. Singh, J. Mishra, and A. Wild, "NxTF: An API and Compiler for Deep Spiking Neural Networks on Intel Loihi," J. Emerg. Technol. Comput. Syst. 18, 3, Article 48 (July 2022), 22 pages. https://doi.org/10.1145/3501770.

[RSG+16] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144

[S+25] A. Shahid et al., "Large-Scale AI in Telecom: Charting the Roadmap for Innovation, Scalability, and Enhanced Digital Experiences", ArXiv: 2503.04184, 2025

[SSW+24] Sandeepa, C., Siniarski, B., Wang, S. and Liyanage, M., 2024, May. SHERPA: Explainable robust algorithms for privacy-preserved federated learning in future networks to defend against data poisoning attacks. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 4772-4790). IEEE.

[SSZ+21] Shi, L., Shu, J., Zhang, W., & Liu, Y. (2021, December). HFL-DP: Hierarchical federated learning with differential privacy. In *2021 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-7). IEEE.

[TBX+16] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein, "Training neural networks without gradients: A scalable ADMM approach," in International conference on machine learning (pp. 2722-2731). PMLR, June 2016.

[TF+23] Tran, K., Fioretto, F., Khalil, I., Thai, M. T., & Phan, N. (2023). Fairdp: Certified fairness with differential privacy. arXiv preprint arXiv:2305.16474.

[TGF+25] Dipanwita Thakur, Antonella Guzzo, Giancarlo Fortino, and Francesco Piccialli. Green federated learning: A new era of green aware ai. ACM Comput. Surv., February 2025. Just Accepted.

[WC+25] Wasif, D., Chen, D., Madabushi, S., Alluru, N., Moore, T. J., & Cho, J. H. (2025). Empirical Analysis of Privacy-Fairness-Accuracy Trade-offs in Federated Learning: A Step Towards Responsible AI. arXiv preprint arXiv:2503.16233.

[WZL+24] O. Wang, S. Zhou, and G. Y. Li, “BADM: Batch ADMM for deep learning,” arXiv preprint arXiv:2407.01640, 2024.

[YD+24] Yang, M., Ding, M., Qu, Y., Ni, W., Smith, D., & Rakotoarivelo, T. (2024). Privacy at a price: exploring its dual impact on AI fairness. arXiv preprint arXiv:2404.09391.

[YDK+24] Yazdinejad, A., Dehghantanha, A., Karimipour, H., Srivastava, G., & Parizi, R. M. (2024) 'A robust privacy-preserving federated learning model against model poisoning attacks', IEEE Transactions on Information Forensics and Security.

[YGS+23] Ashkan Yousefpour, Shen Guo, Ashish Shenoy, Sayan Ghosh, Pierre Stock, Kiwan Maeng, Schalk-Willem Krüger, Michael Rabbat, Carole-Jean Wu, and Ilya Mironov. Green federated learning. In Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities, 2023.

[YWS+24] Yuan, L., Wang, Z., Sun, L., Philip, S. Y., & Brinton, C. G. (2024). Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*.

[ZBD+25] E. Zeinab, G. Batista, and M. Deghat. “AA-mDLAM: An accelerated ADMM-based framework for training deep neural networks,” Neurocomputing: 129744, 2025.

[ZL+23] S. Zhou and G. Y. Li, "Federated Learning Via Inexact ADMM," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 8, pp. 9699-9708, Aug. 2023, doi: 10.1109/TPAMI.2023.3243080.

[ZLH+19] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in neural information processing systems*, *32*.

[ZMW+18] C. Zhang, A. Muaz, and Y. Wang, “ADMM based privacy-preserving decentralized optimization,” in *IEEE Transactions on Information Forensics and Security* 14, no. 3 (2018): 565-580

[ZRL+22] T. Zebin, S. Rezvy, and Y. Luo, “An Explainable AI-Based Intrusion Detection System for DNS Over HTTPS (DoH) Attacks,” in IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2339–49, 2022.

[ZZW+22] Zhao, J., Zhu, H., Wang, F., Lu, R., Liu, Z., & Li, H. (2022) 'PVD-FL: A privacy-preserving and verifiable decentralized federated learning framework', IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2059–2073.