Robust and Efficient Federated Learning for IoT Security

HAN WANG
The widespread adoption of Internet of Things (IoT) devices has led to substantial progress across various industrial sectors, including healthcare, transportation, and manufacturing. However, these devices also introduce significant security vulnerabilities because they are often deployed without adequate security measures, making them susceptible to cyber threats. Meanwhile, the rapid evolution of Artificial Intelligence (AI), specifically in the fields of Machine Learning (ML) and Deep Learning (DL), brings convenience and advantages to the community of IoT security. AI-driven solutions can process extensive data from IoT devices and networks, facilitating the identification of intricate and dynamic threats that may go unnoticed through conventional security methods. Nevertheless, typical ML models require a substantial volume of centralized datasets for training, which may conflict with the principles outlined in the GDPR. Recently, Federated Learning (FL) has emerged as a promising decentralized learning paradigm that enables participants to collaboratively train models without sharing private data. However, FL also brings new challenges.

The contributions of this dissertation are presented through six research papers, which address identified shortcomings and challenges of FL and ML. Initially, a comprehensive landscape study is conducted to understand available ML technologies thoroughly. A novel approach to device fingerprinting and identification is proposed to fingerprint and identify IoT devices through the application of FL. Through this work, several limitations of FL and research challenges are identified. To begin with, the challenges of non-IID and imbalanced data are addressed by proposing adaptive data rebalancing techniques in a peer-to-peer FL setup. Subsequently, a communication-efficient and robust federated aggregation rule is proposed to secure the learning process in the FL setup. Furthermore, when the Intrusion Detection System (IDS) detects anomaly records, they are shared as vulnerability alerts with the Cyber Threat Intelligence platform, which is enhanced by the proposed ML-based functionalities to automate threat processing. Lastly, an in-vehicle IDS is analyzed in the context of the automotive use case for its resilience against adversarial attacks.

The overall contribution of this dissertation enhances the aggregation methodology within FL, emphasizes its adaptability in addressing diverse critical scenarios to tackle IoT security challenges, and reinforces ML models to confront adversarial AI challenges. Given that FL is still in its early stages, with numerous unresolved challenges in IoT security, these enhancements and contributions are timely in paving the way for future advancements and providing a clearer path forward.

**Keywords:** Internet of Things, Federated Learning, Machine Learning, Intrusion Detection System, Communication Efficiency, Robustness, Adversarial AI, Device Fingerprinting, Device Identification, Cyber Threat Intelligence

Han Wang, Department of Information Technology, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

© Han Wang 2023

ISSN 1651-6214
ISBN 978-91-513-1895-0
URN urn:nbn:se:uu:diva-511774 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-511774)
To my mother and father
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the publishers.
Additional Peer-Reviewed Papers

In addition to the papers included in this dissertation, I have co-authored the following papers during my PhD studies.

• Alfonso Iacovazzi, İsmail Bütün, Han Wang, and Shahid Raza. “Towards Cyber Threat Intelligence for the IoT”. In: The 4th International Workshop on Security and Reliability of IoT Systems (2023)

Contents

Part I: Dissertation Summary ................................................................. 15

1 Introduction .......................................................................................... 17
  1.1 Research Challenges .................................................................... 19
  1.2 Methodology ............................................................................... 21
  1.3 Contributions .............................................................................. 22
  1.4 Dissertation Structure ................................................................ 24

2 Background ............................................................................................ 27
  2.1 Machine Learning and Deep Learning ..................................... 27
  2.2 Peer-to-Peer Federated Learning ............................................. 28
  2.3 Adversarial AI ............................................................................ 29
  2.4 Cyber Threat Intelligence .......................................................... 30

3 Summary of Papers ............................................................................... 33
  3.1 Paper I ...................................................................................... 33
  3.2 Paper II ...................................................................................... 34
  3.3 Paper III ..................................................................................... 35
  3.4 Paper IV ..................................................................................... 36
  3.5 Paper V ...................................................................................... 37
  3.6 Paper VI ..................................................................................... 38

4 Related Work ......................................................................................... 41
  4.1 Machine Learning for IoT Security ........................................ 41
  4.2 Data Rebalancing on Imbalanced and non-Independent and
  Identically Distributed (IID) Data ................................................. 43
  4.3 Communication-Efficient Federated Learning ...................... 44
  4.4 Model Robustness Against Adversarial Attack in FL ........... 45

5 Conclusions and Future Work .............................................................. 47
  5.1 Conclusion ............................................................................... 47
  5.2 Future Work .............................................................................. 48

Summary in Swedish .................................................................................. 51

Bibliography ............................................................................................. 53
Acknowledgements

Personal Acknowledgements

I want to express my gratitude to my main supervisor, Prof. Shahid Raza, whose guidance has played an important role in my development as an independent researcher. His unwavering support has been invaluable throughout my academic journey. I would also like to thank my co-supervisor, Professor Thiemo Voigt, for motivating and encouraging me to participate in various academic activities and build valuable connections.

I want to acknowledge the significant contributions of Dr. Luis Muñoz-González, who provided both technical expertise and mental support during our two collaborations. His optimistic encouragement inspired me to have a positive mindset. Special thanks to Dr. Luis Barriga. His willingness to share professional expertise and life experiences during our project collaborations made us good business trip pals. I also want to thank Dr. Muhammad Zaid Hameed, Dr. David Eklund, and Dr. Alfonso Iacovazzi for their constructive advice on different expertise. I thank Prof. Alina Oprea, Dr. Arash Vahidi, Dr. Rolf Blom, and Seonghyun Kim, who shared their experiences and made the research possible.

I am grateful to my lunch pals: Rikard Höglund, Marco Tiloca, Simon Bouget, and Nicolas Harrand, with whom I’ve had many interesting discussions over lunch and from whom I’ve learned valuable life lessons beyond my research. I appreciate my current and former PhD study pals at Connected Intelligence, the Cybersecurity unit at RISE, and Uppsala University: Anum Khurshid, Joel Höglund, Martin Gunnarsson, Wenqing Yan, and others for their peer support. I thank my colleagues at the Cybersecurity unit, Mazaher Kianpour, Gustav Mideus, Apostolos Pyrgelis, Martin Bergling, Erik Mickols, and Stewart Kowalski, who have created a supportive and enriching research environment.

I deeply thank my dear boyfriend, Albert. Your willingness to lend a listening ear during my moments of frustration and celebrate my achievements has made all the difference.

Above all, I want to express my deepest gratitude to my loving parents for their unwavering support, encouragement, and sacrifices to help me achieve my academic and professional goals. This journey has been made possible because of their unconditional love.
Funding Acknowledgements
The following projects have funded the research in this thesis: EU H2020 ARCADIANIoT (GA No. 101020259), EU H2020 CONCORDIA (GA No. 830927), and RISE KP.

Han Wang
Stockholm, September 2023
List of Acronyms

AI    Artificial Intelligence
ARP   Address Resolution Protocol
CAN   Controller Area Network
CTI   Cyber Threat Intelligence
CoAP  Constrained Application Protocol
DDoS  Distributed Denial of Service
DBSCAN Density-Based Spatial Clustering of Applications with Noise
DL    Deep Learning
DTLS  Datagram Transport Layer Security
EU    European Union
ECU   Electronic Control Unit
FL    Federated Learning
FedAvg Federated Averaging
FGSM  Fast-Gradient Sign Method
GDPR  General Data Protection Regulation
IDS   Intrusion Detection System
IoT   Internet of Things
IoC   Indicator of Compromise
IID   Independent and Identically Distributed
JSON  JavaScript Object Notation
LSTM  long short-term memory
LLC   Logical Link Control
ML    Machine Learning
MQTT  MQ Telemetry Transport
MISP  Malware Information Sharing Platform
RBM   Restricted Boltzmann machine
SGD   Stochastic gradient descent
SVM   Support Vector Machines
SMOTE Synthetic Minority Oversampling Technique
STIX  Structured Threat Information Expression
TCP   Transmission Control Protocol
UDP   User Datagram Protocol
WSN   Wireless Sensor Network
Part I:
Dissertation Summary
1. Introduction

Nowadays, the IoT is prevalent in households, buildings, and even entire cities. IoT refers to a network in which various objects, including smart devices and imaginable items, can communicate and exchange data autonomously. However, IoT devices can be targeted by various cyber threats, spanning from Distributed Denial of Service (DDoS) attacks and malware infections to data privacy breaches. These malicious activities can lead to a spectrum of consequences, ranging from the unauthorized extraction of sensitive data to the commandeering of devices for participation in botnets.

To address these issues encompassing detection, analysis, and recovery, IDS are commonly employed. IDS monitors network traffic to identify anomalous behaviors that may indicate potential cyber-attacks. They can be configured to identify specific attack types, such as DDoS incidents or malware infiltrations, and can enact automated measures to thwart attacks or isolate affected devices. Through the implementation of IDS, organizations and service providers can effectively safeguard their IoT networks against an array of cyber risks, ensuring the integrity and security of both devices and data.

AI, primarily ML, has become an increasingly popular approach for developing IDS for IoT networks by analyzing a vast volume of traffic data generated in IoT networks [21]. Unlike rigidly programmed or rule-based approaches, AI has the ability to utilize abundant IoT data to deduce valuable insights and forecast future actions effectively. However, this potential also brings about substantial privacy concerns. The ethical use of AI is therefore inevitable for the success and widespread use of AI.

The relationship between cybersecurity and AI can be one of adversaries or allies, signifying that AI has the potential to either launch assaults or safeguard IoT systems. When AI is employed to tackle security issues in the realm of IoT, current approaches suggest the accumulation of all IoT data for learning purposes, employing advanced ML techniques within a centralized cloud setting. From a security and privacy perspective, repetitive data exchange between the cloud and IoT networks can result in potential data leaks, delays, and congested network traffic.

In light of emerging privacy regulations like the ePrivacy Directive and the General Data Protection Regulation (GDPR) in the European Union (EU), there is a growing emphasis on sharing only essential data, promoting privacy-by-default and privacy-by-design principles. The regulation hinders performing all AI operations in cloud environments rather than encouraging bringing ML right into IoT devices or at the edge of IoT deployments and sharing only processed data needed to run a service in a cloud environment.
FL emerges as a promising framework for privacy-persevering within distributed systems [29]. It enables multiple participants to collaboratively train a learning model while safeguarding their training datasets. A central node or aggregator manages the learning process, all without direct access to the participants’ individual datasets [37]. More technical details about FL are described in Chapter 2.

The scope of FL extends across various domains, including the realm of IoT networks. For instance, within an organization’s buildings, the participants engaged in an FL task could be the gateways where IDSs are deployed. In this scenario, FL proves advantageous for creating a global model that amalgamates information regarding both normal and abnormal traffic from diverse IDSs, all without necessitating data sharing or centralized database management. Subsequently, attacks exclusively observed in certain IDS can be learned by the global model, thereby enabling all participants in the FL task to reap the benefits of enhanced accuracy in attack detection. This stands in contrast to cases where IDSs are trained or re-trained in isolation.

FL boasts numerous advantages; nonetheless, its constraints can impede its suitability for implementation within IoT networks. When applying FL to an IoT network, often within a Wireless Sensor Network (WSN) context, the constraints of limited connectivity and bandwidth between distinct devices and components pose a challenge in ensuring each participant maintains a connection to the central aggregator. IoT networks inherently exhibit heterogeneity due to the variations in physical and environmental conditions among devices. This diversity results in the data not adhering to the principle of being IID. Furthermore, the frequent exchange of updated models between the aggregator and clients brings substantial communication costs. Unfortunately, there are no perfect solutions established for these issues.

On the other hand, as the utilization of ML applications on IoT devices increases, numerous potential concerns and vulnerabilities have surfaced. The rise of increasingly sophisticated attackers has led to attempts to manipulate the results and models produced by ML systems to achieve their own objectives further. These malicious actions targeting ML models are called adversarial AI within ML terminology. For instance, attackers may subtly influence the training dataset over an extended period to distort outcomes, a phenomenon referred to as a poisoning attack. This concern is particularly pronounced within IoT networks due to the frequent inter-device communication and diverse device characteristics. Consequently, safeguarding against adversarial attacks targeting deployed ML models becomes crucial for the effectiveness of ML systems deployed in IoT networks.

The overall theme of the research conducted as part of this dissertation has been strengthening FL framework to make it applicable for solving security problems in IoT networks and consolidating ML models to make it robust against adversarial AI.
1.1 Research Challenges

The core focus of my research lies in leveraging AI/ML to address issues on IoT security and, concurrently, establishing mechanisms to imbue AI/ML techniques with enhanced security and privacy measures. Bringing ML to the edge of the IoT network is rapidly emerging as a prominent trend. This idea offers several merits: decentralization, optimal resource utilization, diminished latency, and rapid responsiveness. The paramount advantage of employing AI at the edge of IoT networks lies in the retention of data within personal or industrial domains while concurrently harnessing the benefits of ML to address challenges related to performance enhancement and IoT security.

The initial preliminary challenge or research question to be addressed within this dissertation is the examination of current state-of-the-art methods for tackling IoT security. The second question to investigate involves understanding how these methods are implemented in a decentralized manner to uphold privacy objectives effectively. To answer these research questions, we conducted a feasibility study (Paper I) aimed at exploring this research domain. Our examination of current privacy-preserving implementations of ML models revealed that while there are opportunities to train ML models on edge devices, traditional ML models do not perform as effectively as DL models. Consequently, FL emerges as a critical enabler for training DL models while preserving privacy.

Research Challenge 1, RC1: As the era of IoT emerges, encompassing billions of interconnected devices with diverse attributes from various manufacturers, the task of identifying unauthorized connected devices becomes increasingly challenging. This challenge becomes even more critical when unauthorized devices infiltrate vital infrastructure, capable of causing disruptions to operational processes. Device fingerprinting and identification methods primarily necessitate in-depth data analysis, rendering them data-centric and tailored to specific tasks, which makes these approaches not generic to the different use cases.

On the other hand, creating a competent DL-based module for device fingerprinting and identification necessitates substantial computational capabilities, often leading to the practice of transmitting data to the cloud via its services. Nevertheless, this approach introduces susceptibilities like potential information exposure and data misuse. Therefore, the research challenges are how to develop a DL-based module for generating a lightweight fingerprint of IoT device that is easy to be further used for device identification, and how this module can be trained by taking the advantages of FL.

Research Challenge 2, RC2: In general, the effectiveness of a model hinges on the training dataset, which ideally should be IID across network clients in FL. However, in practical scenarios, the gathered data diverges notably
among devices due to user preferences and local conditions variations. This aspect becomes especially pertinent in the context of IoT anomaly detection, given that the types of attacks or anomalies observed by each device can differ substantially. Combining this with the fact that anomalies usually constitute a minute fraction of the overall training set presents a formidable challenge for training distributed ML algorithms.

Zooming in on a more confined perspective, one challenge is that on-device datasets are often non-IID, a characteristic that has been demonstrated to undermine the performance of models [37]. This decline in performance is projected to be even more pronounced in instances of imbalanced datasets, where the model inclines towards favoring well-represented classes, leading to a bias. On a broader scale, the global model typically leans towards patterns presented by the majority of clients, potentially stifling the patterns originating from less prominent clients [53]. Therefore, the research challenges here are how to rebalance the local training datasets on each device and withhold data confidentiality, and at the same time, to mitigate the model degradation caused by non-IID nature of data, a common issue in FL setup.

**Research Challenge 3, RC3:** Although FL guarantees the preservation of local private data from being disclosed to other clients, it often entails disseminating complete model updates to all clients in each training iteration. This aspect can become particularly noteworthy when employing extensive DL models, resulting in substantial communication overhead. The high communication cost is a critical drawback of standard FL caused by sending model updates among participating nodes. Along with the size of the model, number of iterations, and number of participating nodes, FL becomes impractical when communication is costly [29, 45]. Moreover, in certain IoT deployments, the requirement for a central node to coordinate the learning procedure might prove unfeasible. This is especially pertinent for wireless sensor networks where connectivity and bandwidth among diverse devices and components can be constrained. The research challenges include reducing communication overhead to accelerate the FL training process and how the clients cooperate without coordination by a central aggregator.

**Research Challenge 4, RC4:** The general aggregation rule for standard FL training has proved that it may not always guarantee privacy and is robust enough to defend against adversarial manipulations by malicious clients [36]. In the FL setup, attackers can potentially compromise one or more clients, enabling them to manipulate the local training datasets or directly the model parameters exchanged with the aggregator. This manipulation is aimed at influencing the behavior of the FL algorithms during the training phase and further making models misclassify during the inference phase. These malicious actions are referred to as adversarial AI within the realm of ML terminology, and more precisely, they are known as poisoning attacks [11]. The research
challenge lies in how to robust the DL-based IDS to make it able to defend against the adversarial attacks in FL setup.

Research Challenge 5, RC5: Certain technologies in automotive systems possess inherent security vulnerabilities, which could enable an attacker to command vehicle components, including essential functions like braking and engine control. ML-based methods, particularly DL, have been integrated to identify potentially malicious actions originating from Electronic Control Unit (ECU) operating on the Controller Area Network (CAN) bus, resembling a complex IoT network. However, it’s imperative to ensure that these ML techniques remain resilient against adversarial attacks. Nonetheless, the scope of considering the in-vehicle IDS as a potential target of an attack has been explored by only a limited number of prior studies. This prevailing gap raises uncertainties about the efficacy of currently deployed IDSs in effectively countering adversarial attacks. *The research challenge here is how the ML-based IDS employed in vehicles reacts to adversarial attacks and which ML model is most resilient to these attacks.*

Research Challenge 6, RC6: Cyber Threat Intelligence (CTI) plays a pivotal role in aiding organizations to pinpoint and counter potential cyber assaults. CTI furnishes invaluable insights into emerging threats and vulnerabilities, empowering organizations to formulate well-informed strategies for mitigating risks. However, CTI typically originates from diverse origins and arrives in varying formats, ranging from network logs and security reports to social media streams and clandestine online forums. This multiplicity of sources and formats engenders non-uniformity and discord in the data, presenting a formidable obstacle to seamless integration and comprehensive analysis. Conventional methodologies for CTI management tend to be time-intensive, prone to errors, and resource-demanding. Consequently, there is an imminent demand for more streamlined and efficacious solutions. *The research challenge here is how to eradicate the need for human involvement and achieve the automation of CTI processing and management with the assistance of ML methodologies. Moreover, designing a delicate integration of ML-based components and the CTI platform presents an additional challenge within this realm of study.*

1.2 Methodology
The research methodology adopted for my dissertation is mostly experimental. It encompasses primarily three stages: (i) exploring the research challenge to formulate a research question, (ii) designing a potential solution, and (iii) validating the proposed solution through experimental assessment using relevant
metrics. The exploration of the problem domain aimed to enhance a fundamental understanding of the topic and culminated in identifying well-defined research questions. The phase of designing potential solutions commenced by delineating the confines of the research scope and making reasoned design choices. The conclusive phase of the applied research methodology involves experimental validation.

1.3 Contributions

My dissertation’s primary contribution is enhancing the FL framework to render it suitable for addressing security issues within IoT networks while concurrently reinforcing ML models to bolster their resilience against adversarial AI. We also studied real-world use cases, pinpointing specific scenarios where the advancements outlined in this dissertation find practical utility. Aligning with our overall objective, we tackled the challenges formulated in Section 1.1, designated as RC1-6, across six research papers. The fundamental achievements of our contributions are outlined in this section.

**FL Application of Device Fingerprinting and Identification**

After conducting a detailed theoretical analysis of the state-of-the-art ML techniques outlined in Paper I, we introduce FL4IoT in Paper II to address RC 1. A dual-phase strategy is designed for the purpose of device fingerprinting and identification. This approach can be applied within both centralized and federated contexts. Our methodology entails the development of an unsupervised learning-based model tasked with generating vectorized fingerprints by examining a device’s traffic patterns. Due to the vectorized nature of the generated fingerprints, they possess a relatively lightweight profile, facilitating their storage on edge devices for subsequent utilization. Subsequently, we propose a supervised learning model catering to device identification. The fingerprints generated previously play a central role in the model’s training process for device identification. Another advantage of our proposed approach is its ability to distinguish spoofed devices’ traffic from their original traffic profiles.

**Federated Data Rebalancer**

To address RC2, we have first introduced an innovative technique for data rebalancing named P2PK-SMOTE, which is strategically applied within peer-to-peer FL for anomaly detection (Paper III). This approach, P2PK-SMOTE, has been meticulously crafted to address the challenges posed by imbalanced and non-IID data within IoT networks. It revolves around the concept of sharing sophisticated synthetic data points generated by diverse clients engaged in the FL process. P2PK-SMOTE improves upon the conventional Synthetic Minority Oversampling Technique (SMOTE) method by employing multiple nearest neighbors to compute linear interpolation to create synthetic data.
points, enhancing their unpredictability and suitability for sharing. Through a comprehensive evaluation, we showcased the effectiveness of this novel approach by conducting a comparative analysis against three state-of-the-art data re-balancing solutions across four distinct IoT scenarios.

Communication-Efficient and Robust Federated Aggregation
After the data rebalancer for addressing RC2 was established in Paper III, we continued to introduce an efficient and robust weighting scheme, called SparSFA, to facilitate efficient communication and defend against adversarial attacks in peer-to-peer FL. Through the designed framework proposed in Paper IV, we addressed RC3 and RC4. To reduce communication costs, we improve TopK sparsification by adding momentum to the residual, which helps to increase performance and stability during training. We further proposed weighting the clients according to five metrics (data size, data variance, connectivity, model similarity, and model divergence) to emphasize the contribution of benign clients and restrict the bad impact from adversarial clients.

SparSFA’s performance was further investigated against four different attacks, and the extensive empirical evaluation showed the robustness and efficiency of our proposed design. It outperformed the other four state-of-the-art robust aggregation rules and two communication-efficient methods. We showed that SparSFA is capable of maintaining its robustness in any network topology, which highlights its ability to adapt different network configurations in the peer-to-peer environment and its applicability to the imbalanced dataset as it can be easily employed in combination with P2PK-SMOTE, the federated data rebalancer proposed in Paper III.

Resilience of ML-based IDS to Adversarial AI
Through the study in Paper V, we addressed RC 5. We have pointed out the limitations and weaknesses of certain ML- or DL-based IDS. This analysis is conducted by subjecting selected models to rigorous scrutiny, specifically, those tailored for automotive networks. Our evaluation exposes the vulnerabilities and shortcomings inherent in these models, as evidenced by their susceptibility to adversarial samples. Additionally, we delve into the transferability of such adversarial samples, revealing that even when generated from a simpler foundational model, they can (with certain limitations) effectively evade the targeted IDS. This underscores the notion that attackers need not possess intricate knowledge of the IDS architecture to perform impactful attacks. Following a comprehensive examination of these findings, we investigate a discourse regarding the suitability of current solutions for deployment within contemporary vehicles.

Automated CTI processing and management
To address RC6, we have presented MAS-CTI in Paper VI, a comprehensive CTI processing system that leverages ML models to enhance threat analy-
sis, IoC classification and clustering in the context of CTI management. We addressed three main problems, threat ranking, event type classification, and IoC correlation, using state-of-the-art ML/DL algorithms and techniques. We also introduced a policy-based method for assigning levels of confidentiality to IoCs, ensuring data privacy and secure sharing in collaborative scenarios.

1.4 Dissertation Structure

This dissertation is divided into two parts: Part I presents a comprehensive summary of the work and is further divided into several chapters. Chapter 2 describes the background information about the technological foundations of this dissertation. Chapter 3 summarizes the six peer-reviewed papers that constitute the core of my work. Chapter 4 contains a summary of related works, and finally, Chapter 5 draws the concluding remarks and the discussion of the future work. Part II contains a reprint of the six papers included in this dissertation.
Figure 1.1. This figure illustrates the combined outcomes of this dissertation’s contributions. The dissertation primarily focuses on technologies related to FL and ML models designed for IDSs that operate on IoT edge devices.
2. Background

This chapter provides the background information required for comprehending and achieving the results of this dissertation. The information covers a brief introduction to ML and DL techniques, the basic concept of conventional FL and peer-to-peer FL, adversarial AI, and some technologies that are widely used with CTI.

2.1 Machine Learning and Deep Learning

ML is a subset of AI that empowers computers to learn and improve from experience without explicit programming. It revolves around the idea of training algorithms to recognize patterns in data, enabling them to make accurate predictions or decisions based on new, unseen information. Instead of relying solely on explicit instructions, ML algorithms utilize statistical techniques to generalize from existing data and adapt their behavior accordingly.

The process begins with feeding the algorithm a training dataset consisting of input data paired with corresponding desired outputs. The algorithm adjusts its internal parameters through iterative learning, gradually improving its ability to make accurate predictions. Various types of ML approaches exist, including supervised learning, where the algorithm learns from labeled examples; unsupervised learning, which focuses on identifying patterns in unlabeled data; and reinforcement learning, where the algorithm learns through trial and error interactions.

DL is a subset of ML that focuses on using artificial neural networks to model and understand complex patterns in data. It is inspired by the structure and functioning of the human brain’s neural networks. What sets deep learning apart is its utilization of deep neural networks, composed of multiple layers of interconnected nodes or artificial neurons.

Each layer in a deep neural network processes the data and passes it on to the next layer, with each successive layer building upon the features learned by the previous ones. There is usually an objective or loss function defined, and the overall purpose of the learning process is to minimize the defined loss function by adjusting the parameters or weights of each neuron. The algorithms automatically learn hierarchical representations of data, gradually extracting higher-level abstractions from the raw input. This enables DL models to handle vast amounts of unstructured data.
2.2 Peer-to-Peer Federated Learning

Traditional FL constitutes a decentralized structure of learning that facilitates the collaborative training of an ML model, particularly DL models. Within this framework, individual participants engage in training the model using their distinct private datasets. After this localized training, participants send their trained models to a central node, which aggregates the local models into a unified global model. This global model is then shared back with the participants [7, 28]. Federated Averaging (FedAvg) is the primary and standard aggregation rule defined in [28].

The workflow of a traditional FL algorithm is as follows: initially, the central node dispatches parameters to the participants, who utilize them to train their local models over a few epochs with their private datasets. Subsequently, the participants forward the updated parameters to the central node. Ultimately, the central node employs an aggregation rule, such as FedAvg, to merge the received parameters from the participants. The aggregated model is then transmitted back to the participants. This iterative process continues until a certain level of convergence is achieved, or a predefined maximum number of training rounds is attained [28]. Typically, the global objective function for FL can be formulated as:

\[
\min_{W \in \mathbb{R}^d} F(W) \tag{2.1}
\]

where

\[
F(W) = \sum_{k=1}^{K} \frac{|D_k|}{|D|} f_k(W) \quad \text{where} \quad f_k(w) \overset{\text{def}}{=} \frac{1}{|D_k|} \sum_{i \in D_k} f_i(W) \tag{2.2}
\]

where \( f_k(w) \) indicates the local objective function associated with client \( k \), while \( D_k \) represents the local dataset exclusively belonging to client \( k \). The notation \(|D|\) indicates the total count of data samples encompassing the entire dataset. Lastly, \( W \) represents the model parameters.

Peer-to-peer FL is a relatively new concept. It involves the removal of the central server/node that coordinates the learning process. It offers benefits such as reducing communication overhead because the updated models are only shared between connected peers. Additionally, it enhances privacy and security because the models are less influenced by compromised ones. Since there is no global model, the Equation 2.2 can be re-formulated as:

\[
F_k(W_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} F_n(W_k; D_n) \tag{2.3}
\]

where \( W_i \) denotes the parameters of the model for client \( k \), and \( N_k \) indicates the total number of neighbors that connected to client \( k \). Without the coordination of the central node, each client updates their local model by aggregating the model updates from its peers. Roy et al. [44] and Wink et al. [63] first applied peer-to-peer to FL, and proposed a simple framework in different applications.

28
2.3 Adversarial AI

ML algorithms are susceptible to several security attacks. This means that when used in a security context, an ML solution may itself become the target of attacks. Adversarial AI refers to a field within AI, particularly DL, that is applied to defend against model-targeted attacks or exploit vulnerabilities in AI systems. It includes two aspects: adversarial attacks and adversarial defense. The concept of adversarial attacks is intentional manipulations of the input data designed to deceive ML models, while the concept of adversarial defense is developing techniques to make AI models more robust against adversarial attacks by detecting adversaries, mitigating the vulnerability, or preventing the impact of adversarial attacks. Adversarial AI has significant implications across various domains, including cybersecurity and autonomous vehicles. It highlights the importance of not only building accurate AI models but also making them resilient against potential manipulations and threats.

The adversarial attacks evolve over time. These attacks overall can be categorized as including poisoning attacks where training data is maliciously modified to affect model behavior; inference attacks, which allow a third party to retrieve training data by querying the trained model; and evasion attacks where the careful selection of the input manipulates model’s output [14]. On the other hand, adversarial defense techniques include adversarial training, defensive distillation, and more. Adversarial training involves incorporating adversarial samples during the model training phase, encouraging the model to become robust against adversarial perturbations [41]. Defensive distillation includes training an auxiliary model using the logits of the primary model, aiming to smooth out the decision boundary and make it harder for attackers to craft adversarial examples [42]. However, the work in [12] has shown that defense distillation is still susceptible to adversarial perturbations.

Our research work in this dissertation focuses on three types of adversarial attacks. More details are provided in the following subsections.

2.3.1 Poisoning Attack

The goal of a poisoning attack is to introduce malicious or carefully crafted data points into the training dataset in order to compromise the performance, accuracy, or security of the resulting ML model. The attackers inject these malicious data points intending to cause the trained model to produce incorrect predictions or classifications, either by biasing the model’s learning process or by exploiting vulnerabilities in the learning algorithms. The injected data might be designed to resemble normal training samples but contain subtle alterations that cause the model to behave in an unintended and potentially harmful manner. The most well-known adversarial samples crafting technique is the Fast-Gradient Sign Method (FGSM), first introduced by Goodfellow et al. [23]. Paudice et al. [43] proposed a strategic approach
for the attackers to choose particular training samples for label flipping deliberately, guided by their influence on the model’s loss. Another kind of data poisoning attack is also proposed in [43], where the attacker adds Gaussian noise with random labels to the training dataset. Poisoning attacks can be particularly problematic when the model is deployed in critical applications such as security systems or autonomous vehicles.

2.3.2 Model Evasion Attack
While poisoning attacks center on diminishing the accuracy of an IDS prior to deployment or during the training phase, evasion attacks concentrate on diminishing accuracy after the system has been deployed, that is, during the inference phase. Evasion attacks hold specific significance within IDS implementations, as they possess the potential to enable adversaries to evade detection [14]. This typically involves learning to replicate normal behavior more effectively or capitalizing on vulnerabilities within ML models to introduce perturbations. Studies demonstrating the potential for evading IDS through exploiting vulnerabilities to perturbations have been conducted in conventional network environments [4].

2.3.3 Byzantine Attack
The Byzantine attack can be seen an untargeted model poisoning attack, especially in distributed ML and FL, that the adversary employs to send arbitrarily malicious model updates to the aggregator with the aim of compromising the learning process on the other clients [32].

A Byzantine attack can be defined as:

$$\Delta W_{t,k} = \begin{cases} * , & \text{if } k\text{-th client is adversary} \\ \nabla F(W_{t,k}) , & \text{otherwise} \end{cases} \quad (2.4)$$

where * represents an arbitrary values as the model update of client k, that is going to be shared with the others in tth iteration.

2.4 Cyber Threat Intelligence
CTI is the data and/or useful information gathered by the Computer Emergency Response Team, a group of information security experts responsible for the protection against, detection of, and response to an organization’s cybersecurity incidents. CTI is the practice of collecting, analyzing, and sharing information about potential and existing cyber threats. It involves gathering data from various sources to gain insights into the tactics, techniques, and procedures used by malicious actors. CTI can empower organizations to issue
proactive and in-depth threat alerts (such as malicious IPs, malicious DNS, malware, and attack patterns) when their systems detect suspicious external or internal threats. Consequently, the collection, analysis, and dissemination of CTI are critical in bolstering cyber defense strategies. Nevertheless, numerous organizations encounter challenges in generating their own CTI, and even those that manage to do so often grapple with resource limitations in keeping pace with the swiftly evolving threat landscape.

2.4.1 Structured Threat Information Expression (STIX)

Structured Threat Information Expression (STIX) [56], which stands for Structured Threat Information eXpression, is an open standard language and format used for describing and sharing cybersecurity threat intelligence. It was developed to enable standardized communication and sharing of cyber threat information among organizations and security tools. It helps us easily to contribute to and retrieve information from CTI. This is made possible by its comprehensive coverage of aspects related to suspicion, compromise, and attribution through well-defined relationships alongside object identifiers. Furthermore, STIX data can be stored in a machine-readable JavaScript Object Notation (JSON) format, enhancing its accessibility, and can also be visualized graphically, which can facilitate the work of analysts. Notably, STIX is open-source software, enabling effortless integration with many other tools and products available in the market.

2.4.2 Malware Information Sharing Platform (MISP)

Malware Information Sharing Platform (MISP) [61] is an open-source system created to facilitate the exchange, retention, and examination of CTI. The administrative tasks within the MISP platform are divided into three main roles: the general administrator, organization administrator, and publisher. The general administrator holds the utmost authority in overseeing all amassed threat information and data sharing. Organization administrators are tasked with overseeing threat events specific to their respective organizations. A publisher is a general user who is allowed to publish organization events.

MISP offers a collaborative platform that enables organizations and cybersecurity professionals to collaboratively exchange actionable intelligence concerning diverse threats, vulnerabilities, and Indicator of Compromise (IoC). IoC are critical in identifying and preempting suspicious or malicious cyber operations. These IoCs function as clues or evidence that signal potential security incidents. They are generated based on various attributes associated with the subject or occurrence under investigation. IoC encompass distinct types, including network indicators (such as IP addresses), system indicators
(like file hashes), and account indicators. Typically, this information is stored in an SQL database in the MISP format.
3. Summary of Papers

3.1 Paper I


Summary

This paper delved into the latest ML techniques and assessed their suitability for enhancing IoT security at the edge. We examine the attributes of a standard IoT edge device, which possesses greater capabilities than a typical battery-operated IoT device, but remains notably limited when compared to cloud-based resources. We have selected the Raspberry Pi 3 B+ module as our baseline edge device to establish a reference point. This research aims to develop potential solutions for each cutting-edge machine learning algorithm, considering the practical implementation challenges and constraints associated with this device. Drawing from our theoretical analysis, guided by four designated metrics, we made observations specific to certain data types and edge hardware configurations.

Reflections

This paper is a preliminary study on the feasibility of the ML methods being applied on the IoT edge. Through this study, most of the state-of-the-art ML methods have been analyzed theoretically. This study marks my initial foray into deploying ML techniques on edge devices to tackle IoT security challenges. I believe the analytical approach introduced herein can offer a valuable assessment of the constraints inherent to these devices. This insight could prove instrumental during the architectural design phase, aiding in formulating effective solutions.

My Contribution

I am the main author of the paper. This work is built upon a research problem conceptualized and formulated by Shahid Raza. I proposed and designed
the analytical method. I adopted the proposed method to conduct the feasibility analysis with the assistance of Luis Barriga. I wrote the manuscript with feedback from all co-authors.

3.2 Paper II

Summary
The dual-phase system we introduced in this paper includes two stages: the generation of fingerprints and device identification by examining their traffic patterns. The resulting fingerprint, generated by our proposed system, is notably lightweight and capable of identifying devices based on factors such as device types, vendors, and product modules. The proposed system also possesses the capability to identify spoofed or compromised devices. Furthermore, its key advantage lies in the fact that the generated fingerprints go beyond mere features extracted from raw traffic traces; instead, they represent a reconstructed alternative. This intricacy poses challenges for potential attackers attempting to recover the original data associated with a target device. Simultaneously, it mitigates risks during the data aggregation process.

Reflections
This paper is considered to be the application of FL to address a real IoT security problem. Using ML to fingerprint and identify IoT devices is not new. Many existing works contributed to this task have been surveyed in this paper. However, this paper is the first work applying FL to this problem based on analyzing device network traffic data. This paper also provides insight into the feasibility of applying unsupervised learning in FL. Additionally, various limitations and research hurdles associated with FL are identified for further investigation.

My Contribution
I am the main author of this paper. Shahid Raza and Alina Oprea have identified and conceptualized the research problem. I designed the framework and algorithms with feedback from David Eklund. I implemented and performed the experimental analysis to evaluate the performance of the proposed framework. I proposed the threat model, conducted the security discussion, and wrote the manuscript with feedback from Shahid Raza.
3.3 Paper III

Summary
In this paper, we have presented a data re-balancing approach specifically tailored for addressing the challenges of imbalanced and non-IID data in IoT networks within a peer-to-peer federated learning (FL) framework, contributing to anomaly detection. This novel approach introduces the concept of sharing complex synthetic data points artificially generated by diverse clients participating in the FL process. To enhance its performance, we have extended the SMOTE by considering multiple nearest neighbors and applying linear interpolation when generating synthetic points. This modification increases the unpredictability of these points, making them more suitable for sharing among participants. We evaluated our approach using two real-world datasets and compared it with three leading re-balancing methods. Our assessment encompassed four distinct scenarios, and the empirical results substantiate its effectiveness. Remarkably, our proposed approach achieves a perfect score of 100% in both recall and precision while maintaining false negatives and false positives at almost 0%.

Reflections
This paper is regarded as a significant achievement within my dissertation, marking the initial exploration of the challenges faced in FL and introducing a rebalancing approach to tackle non-IID and imbalanced data within a peer-to-peer FL framework. Notably, this solution does not require auxiliary datasets or intermediaries for orchestrating the rebalancing process. This research gives me valuable insights into the repercussions of non-IID data on ML models, a challenge that proved particularly crucial in the context of IoT networks. This paper also has made a valuable contribution by presenting an improved alternative to conventional FL methods and addressing specific issues pertaining to all FL configurations.

My Contribution
I am the main author of this paper. The research problem is identified and conceptualized by Luis Muñoz-González. I designed the framework and algo-
Algorithm mathematically verified by David Eklund. I implemented and evaluated the proposed algorithm and formulated meaningful findings from the experiments. I conducted the security analysis of the proposed framework and wrote the majority of the manuscript. The above-mentioned contributions were continuously reviewed and supported by feedback from Luis Muñoz-González and Shahid Raza.

3.4 Paper IV

Summary
In this paper, we present a novel weighting scheme that is both efficient and resilient, designed to safeguard against adversarial attacks in a peer-to-peer federated FL setup. To enhance performance and training stability while minimizing communication costs, we enhance the TopK mechanism by introducing momentum to the residuals. In addition to its communication cost reduction benefits, the suggested scheme also enhances security by concealing a substantial portion of the model’s information. This concealment arises from the model’s sparsification, where only a limited subset of parameters is shared. Additionally, we put forward a client weighting strategy that incorporates five key metrics: data size, data variance, connectivity, model similarity, and model divergence. This approach aims to amplify the influence of benign clients while mitigating the adverse effects stemming from adversarial clients.

Reflections
This paper represents another significant achievement within my dissertation, marking a seamless continuation of the research journey initiated with P2PK-SMOTE. I believe the contributions of this paper to be highly significant, particularly with regard to the integration of both Paper III and Paper IV, which solidifies and enriches the framework of peer-to-peer FL. Furthermore, we extend the applicability of the peer-to-peer setup from a fixed network topology to encompass random networks and more adversarial actors are considered in this paper. The scenarios explored in this study are designed to be practical and adaptable across various domains within IoT networks.
My Contribution

I am the main author of this paper. I identified and formulated the research problem. Muhammad Zaid Hameed and I co-designed the algorithm improvement for model sparsification for communication efficiency. I and Luis Muñoz-González co-designed the approach for the robustness of the federated aggregation rule. David Eklund mathematically reviewed two proposed algorithms. I implemented the proposed scheme and carried out all the evaluations and experiments. I proposed the threat model and security discussion reviewed by Luis Muñoz-González and Shahid Raza. I wrote most of the manuscript with comments and contributions from Luis Muñoz-González and Muhammad Zaid Hameed.

3.5 Paper V


Summary

In this paper, we conduct a comprehensive exploration of the susceptibility of four distinct ML-based intrusion detection systems, each tailored for automotive networks, to adversarial samples as a means of model evasion attack. These adversarial samples are crafted employing the FGSM. Our findings reveal that these adversarial samples have a detrimental effect on three out of the four solutions under investigation. Additionally, we scrutinize the transferability of these adversarial samples across different systems. We delve into the performance of detection and the success rate of attacks following the incorporation of adversarial samples into the training process. After a meticulous analysis of these outcomes, we engage in a discourse regarding whether the existing solutions have reached a level of maturity suitable for deployment in contemporary vehicles.

Reflections

This research urges researchers to consider the limitations of ML-based models and the overall reliability of using such models. This study delves into a crucial facet of ML inference: the potential for manipulated data samples to escape detection by IDSs. Our discoveries have substantiated the notion of the transferability of adversarial perturbations, demonstrating that an adversarial perturbation, trained using the same adversarial model with only a few
iterations, can effectively elude detection by various IDSs. We conduct this investigation in a high-stakes scenario.

My Contribution
I am the second author of this paper. Arash Vahidi initially identified and formulated the research problem. I, together with Shahid Raza and Alfonso Iacovazzi, supervised Ivo Zenden, who was a Master’s thesis student. I co-designed the research direction and methodology with Alfonso Iacovazzi. Ivo Zendon implemented and conducted the experiments. All authors contributed to writing the manuscript.

3.6 Paper VI

Summary
In this paper, we introduce an enhanced iteration of the widely used CTI platform, MISP, harnessing the capabilities of ML for CTI operations. Our primary focus is addressing three fundamental challenges within the CTI domain. This includes prioritizing and ranking received IoCs based on factors like their severity, likelihood of exploitation, and potential impact; classifying the IoCs according to the type of attacks or threats they represent; and aggregating IoCs exhibiting similar characteristics into distinct clusters. Furthermore, in response to IoC confidentiality concerns, we investigate the utilization of FL for event identification. We have subjected the models to extensive evaluation using three publicly available CTI datasets, with the outcomes affirming the potential of ML models to augment the processing and analysis of CTI.

Reflections
The foremost achievement of this paper lies in its formulation of the system architecture, which constitutes a fusion of a widely recognized CTI platform with ML-based modules. This research explores four distinct ML techniques: feature extraction strategies, learning-to-rank algorithms, federated event classification, and clustering. Each of these tasks can stand as an independent research challenge. This paper adeptly delineates a guiding and comprehensive roadmap on how to effectively employ these techniques within the CTI domain, elucidating their potential contributions to the enhancement of CTI processing and management.
My Contribution

I am the main author of the paper. The research idea was initially formulated by Shahid Raza, and I expanded the scope of the research and conceptualized the research problem. I co-designed the framework with Alfonso Iacovazzi. I designed and implemented the strategy for feature extraction from different data formats. Seonghyun Kim implemented the module of threat ranking, and I implemented two modules for federated event classification and IoC clustering. I further integrated these three modules with the MISP platform. I conducted most of the evaluations and experiments. All authors collaborated in writing the remaining manuscript.
4. Related Work

This chapter contains an overview of the related work to put the results of the dissertation in context within the state-of-the-art in the included research fields and relevant target industry development. The related work is categorized into four broad domains: (i) Machine Learning for IoT Security, (ii) Data Rebalancing on Imbalanced and non-IID Data, (ii) Communication-Efficient Federated Learning, and (iii) Model Robustness Against Adversarial Attack in FL.

4.1 Machine Learning for IoT Security

ML-based Intrusion Detection System

An IDS is employed in IoT networks to detect malicious or suspicious events. Typically, these types of IDS are anomaly-based IDS, which means they are trained on the standard/normal behavior of the system. They monitor traffic in IoT networks to identify abnormal behaviors and leverage IoT network data to better defend against attacks [9]. Generally, DL techniques such as Autoencoders [25, 27, 51] or Restricted Boltzmann machine (RBM) [15] are used for feature selection, after which more standard ML techniques such as Support Vector Machines (SVM) and Decision Trees are employed for the actual classification [2]. Unsupervised learning techniques have also been utilized in developing anomaly-based IDSs. Banerjee et al. [5] employ Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a region-based clustering method, to detect anomalies; Guo et al. [24] and Doshi et al. [16] apply K-Dimensional Trees; Stewart et al. [55] use K-Means to detect events such as security threats. These ML models have been theoretically analyzed from the perspective of their feasibility for application at the IoT edges in our contribution in Paper I.

Automotive is another common use case where IDS is employed. Modern vehicles are equipped with numerous ECUs, each responsible for overseeing specific vehicle components. These components can vary widely, from functions like controlling windows and managing the infotainment system to critical tasks such as engine and brake control. CAN is one of the widely-used technologies for communication regarding the main functionality of vehicles. In-vehicle IDS is deployed in the context of CAN networks. Song et al. [54] convert a series of CAN messages into frames. This transformation involves the extraction of CAN message IDs from 29 successive messages and their
subsequent arrangement into a matrix format. They used a convolutional neural network to implement the anomaly-based IDS, and the model’s architecture is a reduced version of the Inception-ResNet model proposed in [58]. Taylor [60] proposes another IDS based on long short-term memory (LSTM). The IDS model was trained in an unsupervised learning manner. It learned from regular CAN message data and aimed to make accurate predictions for the next message, minimizing the disparity between its predicted next input and the actual one. Our findings in Paper V show that the IDSs proposed in these works are vulnerable to adversarial attacks, and we demonstrate that this vulnerability can be mitigated by adopting adversarial training.

**ML-based Device Fingerprinting and Identification**

Over the years, device identification and fingerprinting methods have found applications in network security to facilitate device authentication. This process entails verifying the authenticity of a device’s claim about its identity. Notably, the utilization of ML has gained traction as an evolving solution to address this challenge.

Traditional ML-based approaches to feature extraction for device fingerprinting seem promising because they learn the raw data well. However, they require complex pre-processing steps and are usually designed for specific data. For example, Miettinen et al. [38] propose a set of features that represent the device fingerprint, which covers twenty-three features, including protocols used in four different layers, such as Address Resolution Protocol (ARP)/Logical Link Control (LLC) in the link layer, Transmission Control Protocol (TCP)/User Datagram Protocol (UDP) in the transport layer, and other eight specific protocols used for the application layer. Nevertheless, because of the fast growth of IoT, various protocols are released. Their approach cannot cover all the protocols, for example, Constrained Application Protocol (CoAP) and MQ Telemetry Transport (MQTT) for the application layer, or Datagram Transport Layer Security (DTLS) for the transport layer. Sivanathan et al. [52] create a testbed to monitor and collect traffic among twenty-eight IoT devices, and they utilize bag-of-word, a natural-language-processing method, to analyze the domain names and the content of the packet. This approach is designed delicately but infeasible when the traffic is encrypted, which is a common drawback for the approaches that consider the content of the packet.

DL-based approaches, in contrast, do not require feature extraction based on one’s domain knowledge. These approaches focus on designing generalized models to learn the representations or distributions that could represent the device as the fingerprint. For example, Ortiz et al. [40] propose an unsupervised-learning-based model, LSTM-embedded autoencoder, to learn the distribution from the content of the packet payload and compare the distributions of unknown devices with those that are known. However, in order to obtain the payload sequences to feed into the autoencoder, they use a snif-
fer to analyze the packet information, which is, again, infeasible if the packet is encrypted, and it also raises privacy concerns. On the other hand, Kumar et al. [31] first utilize the transmission time of a frame and inter-arrival time to construct a histogram for each frame type and assign a weight to each frame type. They treat the histogram and weight combination as the device’s fingerprint and train a neural network with these generated fingerprints to be further used for device identification. One strength of their work was that they only exploited two main features to construct the fingerprint without looking into the content of the packet. Nevertheless, their way of fingerprint construction is static but not dynamic, which means it is possible to miss the potential factors possessed by the data. Our contribution in Paper II mitigates the drawbacks mentioned, and the approach proposed is also applicable to the FL setup.

4.2 Data Rebalancing on Imbalanced and non-IID Data
Statistical heterogeneity can manifest due to the non-IID nature of data collection within IoT networks. This phenomenon can be attributed to the varying FL setups, wherein the number of data points or the data distribution significantly fluctuates among devices or clients. Such variations have the potential to degrade the model’s performance. This phenomenon has been attributed by Zhao et al. [67] to the weight divergence between FedAvg and Stochastic gradient descent (SGD) observed across different devices. They suggest storing a globally shared dataset in the cloud, characterized by a uniform class distribution. On a similar note, Wang et al. [62] design a control algorithm that gauges the trade-off, taking into account the convergence bound of gradient descent, and it learns real-time data distribution and system dynamics to minimize learning loss. FEDFMC [30] is another approach addressing model degradation on non-IID data by forking, merging, and consolidating the global model. Hybrid-FL [65] is a hybrid learning mechanism that proposed allowing a limited number of clients to upload their data to the FL server based on their client and data selection strategies. Nonetheless, these methodologies solely tackle the aspect of non-IID data and do not account for imbalanced datasets. Furthermore, a significant portion of these approaches either demand a limited quantity of private data to be uploaded to the aggregator or necessitate the involvement of mediators to reconfigure the model-sharing procedure, which requires redundant communication. Astra [17] is an example of an approach addressing imbalanced and non-IID data for FL. However, several mediators also need to collect data from grouped clients to mitigate the class-wised bias.

Data rebalancing, in general ML research domain, is not a new topic. It has been well-studied. Oversampling [10, 13, 19] and under-sampling [19, 66] are two classic techniques to address this problem. SMOTE [13] has been popular in many research works. Initially, it selects the $k$ nearest neighbors for a given data point within the minority class. Subsequently, the vector connecting one
of these neighbors to the selected data sample is scaled by a randomly chosen value ranging from 0 to 1. This scaled vector is then added to the data sample to produce a synthetic data point. Calleja et al. [10] improve SMOTE by using weighted distance to average the neighbors in positive samples to obtain the mean as the synthetic data sample. On the other hand, under-sampling is the simplest and most straightforward way to randomly remove data samples from the majority class. However, [19] points out that under-sampling may increase the variance and discard the important samples. [66] is one classic way to under-sampling. It under-samples the points in the majority class based on their distance to other points in the same class. However, the application of these techniques, including SMOTE, to FL scenarios has not been explored.

Our contribution in Paper III not only addresses the issue of performance deterioration on non-IID data but also achieves the re-balancing of local datasets within a peer-to-peer FL setup. Importantly, this re-balancing occurs without the need for a central coordinator to orchestrate the process.

4.3 Communication-Efficient Federated Learning

The communication cost is a critical bottleneck that takes longer than the local computation of model training. This has opened up a popular research topic regarding improved device-to-device communication. Besides, Google’s research team introduced FL and contributed to communication efficiency in FL [26, 29, 37, 45]. Jeong et al. [26] propose a communication-efficient algorithm for data augmentation. They utilize a knowledge distillation method whose communication payload size depends on the output dimension to reduce the overhead.

In a more granular context, it is worth noting that gradient sparsification holds greater promise compared to gradient quantization. This preference stems from the fact that quantization faces limitations in accommodating large-scale models or networks with low-bandwidth connectivity [49]. The fundamental concept behind sparsification involves communicating only a reduced set of model gradient components. This strategy aims to enhance efficiency while preserving a credible estimation of the global gradient, thus minimizing the impact on performance, as discussed in [20]. Noteworthy work by Strom [57] and Aji et al. [1] introduce the concept of Top-K sparsification, which entails selecting the top-k components with the highest gradient magnitudes and treating the remaining components as residuals. Tang et al. [59] has the most similar assumption to ours, which is in the environment of peer-to-peer networks. However, they adopt random-k sparsification, that is, randomly picking k components to share, instead of Top-k sparsification. Our contribution in Paper IV has been empirically proven to outperform the methods mentioned.
4.4 Model Robustness Against Adversarial Attack in FL

Conventional aggregation techniques, like FedAvg, are susceptible to adversarial manipulations that can exploit malicious or compromised clients infiltrated by attackers. Consequently, the pursuit of robust FL in the face of poisoning attacks has gained significant traction within the research landscape. Poisoning attacks can be further divided into two categories: triggered poisoning attacks and triggerless poisoning attacks [35, 46]. To narrow down, the type of poisoning attack examined within this dissertation is triggerless. In this context, triggerless poisoning attacks transpire during the training phase, devoid of defined triggers or targeted classes [18, 22, 34, 48]. The attack can be attributed to the potential ramifications of a single compromised entity capable of undermining the integrity of the entire learning process. Such compromises can result in erroneous solutions or convergence issues within the realm of FL [6, 18].

Along with new adversarial attacks in FL being discovered, the different robust defenses against corresponding attacks have become a relevant and popular research topic for FL, given the impact that these attacks can have in practical deployments [6, 11, 33, 39, 64]. Krum [6] is one of the state-of-the-art methods that consider the similarity of the gradients of model updates from all clients in every iteration. AutoGM [33], the auto-weighted geometric median, is a flexible and robust aggregation rule designed for industrial IoT. Yin et al. [64] propose algorithms relying on robust statistics for mitigating poisoning attacks, including Trimmed Mean and Median. Both Trimmed Mean and Median are coordinate-wise aggregation rules. Given a threshold value $k$ such that $k < |N|/2$, where $|N|$ represents the count of clients, the client denoted as $C_d$ eliminates the $k$ largest and smallest parameters. Subsequently, the client calculates the mean of the remaining parameters. Conversely, the Median approach employs the median value of each parameter for updating the local model, contrasting with the utilization of mean values. The approaches mentioned all require a central orchestration. Our contribution in Paper IV has empirically proven its effectiveness of robustness in random network topology, and the design outperformed the methods mentioned.
5. Conclusions and Future Work

In this section, I present some concluding remarks before discussing possible future work of the work.

5.1 Conclusion

Using ML and DL techniques to address security issues is on the rise, driven by networks and systems’ expanding scale and intricate nature. At the same time, fog/edge computing is emerging as a prominent trend in the next generation, primarily in response to the implementation of GDPR. This development ushers in many benefits, such as safeguarding privacy, rapid response, minimal latency, cost-effectiveness, and more. Fog/edge computing also plays an important role in enabling FL. While FL offers numerous advantages, it introduces novel, unique challenges and vulnerabilities. In contrast to centralized learning, FL models are inherently susceptible to adversarial attacks. The act of sharing models also raises additional concerns, notably in terms of protecting model confidentiality. Detecting compromised clients or adversarial inputs in an efficient manner remains an open issue. A viable solution should be computationally efficient without causing substantial degradation in the model’s performance.

In this dissertation, I started with a preliminary examination of existing techniques, subsequently delving deeper to investigate solutions for the limitations of FL. Following this, I applied FL to tackle a security problem within IoT networks. Lastly, the dissertation concludes by addressing the robustness of ML models against adversarial attacks and presenting a holistic system that integrates CTI with ML-based components.

In Paper I, we studied state-of-the-art ML techniques and explored their suitability for enhancing security in IoT edge devices. These devices play a crucial role in managing data flow at the periphery of the IoT network. In contrast to cloud resources, edge devices possess constrained computational capabilities, memory, and storage capacity, presenting significant challenges for implementing ML on them. This landscape study gave me the first insights into the challenges of the research area.

We showcased an application of FL by introducing a two-phase device fingerprinting and identification system in Paper II. This system is versatile and can be utilized in both centralized and federated environments. The fingerprint
produced by our system, represented as a vector, uniquely characterizes an individual device, and its lightweight nature allows for storage on edge devices. Through this work, we have identified several limitations of FL.

Therefore, in Paper III, we proposed P2PK-SMOTE, a data rebalancing scheme that can address non-IID and imbalanced data in a peer-to-peer FL setup. We have improved the SMOTE technique by incorporating multiple nearest neighbors to calculate linear interpolation for generating synthetic points. This modification increased the unpredictability of the generated points, making them more suitable for sharing. We demonstrated the efficacy of P2PK-SMOTE through empirical results, achieving perfect scores of 100% in both recall and precision while almost eliminating false negatives and false positives, registering nearly 0%.

Paper IV was an extension work presented in Paper III. Based on the same scheme, we developed SparSFA in Paper IV. We enhanced communication efficiency by improving model sparsification and strengthened the robustness against adversarial attacks during the model-sharing phase in FL. SparSFA outperformed the other robust aggregation rules in different scenarios.

The primary objective behind the inception of Paper V was to provide both ourselves and the automotive community with a comprehensive understanding of the repercussions of adversarial attacks and the critical significance of ensuring model resilience against such threats.

Finally, we introduced an ML-empowered CTI platform in Paper VI, significantly elevating the efficiency of CTI processing and management. This effort culminated in integrating diverse ML models and creating a sophisticated platform for sharing information on vulnerabilities and threats.

5.2 Future Work

The contributions made in this dissertation mark a significant step toward establishing trust in AI for IoT security. Furthermore, numerous potential directions exist for future research, offering opportunities to enhance further the security and privacy of ML models and ensure their reliable utilization in addressing security concerns. Some that I think are worth mentioning.

This dissertation did not delve into the realm of privacy attacks. The information utilized for training ML models is sensitive. One particular type of attack, known as a membership inference attack [50], aims to ascertain the presence of specific data records in the training dataset of the target model by exploiting the responses provided by the targeted model. If this attack succeeds, it has the potential to result in the unauthorized disclosure of sensitive or private data. Therefore, preventive measures need to be implemented to safeguard against such attacks.

The exploration of training an ML model using encrypted data was not part of this dissertation’s focus. While there are previous studies employing
homomorphic encryption [3] for classification, they do encounter accuracy degradation. This brings forth a new research challenge: how to enable the training of ML models on encrypted data without compromising on accuracy.

The growing recognition of the importance of privacy has sparked conversations about privacy-preserving ML, as well as the implementation of secure aggregation protocols for FL. Especially in IoT networks, the challenge of enabling FL on resource-constrained devices remains unresolved. Consequently, my initial step in this direction will involve further investigation into integrating FL with IoT-specific protocols, such as CoAP [8] and OSCORE [47] (Object Security for Constrained RESTful Environments).
Under de senaste åren har den snabba utvecklingen och den omfattande användningen av olika Internet of Things (IoT)-teknologier, enheter och lösningar medfört betydande störningar inom olika industrier och omformat många aspekter av vårt dagliga liv. Denna transformerande påverkan har haft en djupgåande effekt på företag, konsumenter och regeringar. Den har också lyft betydande oroskällor rörande säkerhet och integritet, många av vilka förblir olösta, med sannolikheten att nya kan uppstå. Enligt nyligen framtagna prognoser kommer det att finnas uppskattningsvis 75,44 miljarder uppkopplade enheter år 2025, vilka innefattar olika ekonomiska sektorer såsom utbildning, transport, energi, hälsovård och säkerhet.

I denna kontext bär hoten och riskerna som är förknippade med IoT-enheter och system betydande ekonomiska och fysiska konsekvenser. I takt med att antalet IoT-enheter fortsätter att öka, så sker samma sak med antalet potentiella attacker och associerade hot. Olyckligtvis utvecklas attacker till alltmer sofistikerade former, vilket kräver förbättrade kunskaper inom cybersäkerhet och utveckling av mer intrikata och effektiva mekanismer för att upptäcka och begränsa dessa attacker. Dessa mekanismer förlitar sig ofta på en kombination av avancerade teknologier, inklusive artificiell intelligens (AI) och cyber threat intelligence.


Det varierande utbudet av enhetstyper och olika arbetsmiljöer leder dock ofta till verklig data som varken är oberoende eller identiskt fördelad mellan klienter. Detta resulterar i sin tur i en försämring av prestandan för federated learning-algoritmen. Dessutom garanterar inte federated learning-metodologier

I de inledande arbetena av denna avhandling börjar vi med att granska state-of-the-art ML-tekniker och bedöma deras användbarhet för implementering på IoT edge devices. Därefter bidrar vi genom att behandla frågor som rör icke-oberoende, ej identiskt fördelad och obalanserad data inom FL-ramverket samtidigt som vi optimerar kommunikationseffektiviteten och stärker aggregationsprocessen.

Utöver vårt bidrag till konsolidering av FL så introducerade vi en innovativ metod för att identifiera och skapa fingeravtryck för enheter. Denna metod kan effektivt användas både i centraliserade och federerade konfigurationer. Dessutom genomförde vi en djupgående undersökning av sårbarheter hos ML-modeller i allmänhet mot adversarial attacks, särskilt inom fordonssområdet.

Vårt slutliga bidrag integrerade sömlöst fyra olika ML-baserade moduler i en plattform för delning av cyber threat intelligence. Denna integration syftade till att strömlinjeformas analysen och administrationen av hotdata. Dessa fyra moduler introducerade fyra nya funktionaliteter i delningsplattformen, automatiserade processer och minimerade behovet av manuell intervention. Dessa funktionaliteter omfattade extrahering av features från olika format för Indicators of Compromise (IoC), tilldelning av hotnivåer till IoCs, händelseklassificering och gruppering av IoC-data för att öka effektiviteten i delningsstrategierna.

Den övergripande bidraget i denna avhandling förbättrar tillvägagångssättet för aggregation inom federated learning, och betonar dess flexibilitet i att hantera olika kritiska scenarier för att hantera IoT-säkerhetsutmaningar och stärker ML-modeller mot angrepp från adversarial attacks. Med tanke på att FL fortfarande är i sina tidiga faser och står inför många olösta utmaningar inom IoT-säkerhet kommer dessa förbättringar och bidrag vid en gynnsam tidspunkt och banar vägen för framtidena framsteg och ger en mer definierad väg framåt.
Bibliography


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2306

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-511774