Zero-Trust Foundation Models: A New Paradigm for Secure and Collaborative Artificial Intelligence for Internet of Things

Kai Li, *Senior Member, IEEE*, Conggai Li, Xin Yuan, *Senior Member, IEEE*, Shenghong Li, *Member, IEEE*, Sai Zou, *Senior Member, IEEE*, Syed Sohail Ahmed, Wei Ni, *Fellow, IEEE*, Dusit Niyato, *Fellow, IEEE*, Abbas Jamalipour, *Fellow, IEEE*, Falko Dressler, *Fellow, IEEE*, and Özgür B. Akan, *Fellow, IEEE*.

***Abstract*—This paper focuses on Zero-Trust Foundation Models (ZTFMs), a novel paradigm that embeds zero-trust security principles into the lifecycle of foundation models (FMs) for Internet of Things (IoT) systems. By integrating core tenets, such as continuous verification, least privilege access (LPA), data confidentiality, and behavioral analytics into the design, training, and deployment of FMs, ZTFMs can enable secure, privacy-preserving AI across distributed, heterogeneous, and potentially adversarial IoT environments. We present the first structured synthesis of ZTFMs, identifying their potential to transform conventional trust-based IoT architectures into resilient, self-defending ecosystems. Moreover, we propose a comprehensive technical framework, incorporating federated learning (FL), blockchain-based identity management, micro-segmentation, and trusted execution environments (TEEs) to support decentralized, verifiable intelligence at the network edge. In addition, we investigate emerging security threats unique to ZTFM-enabled systems and evaluate countermeasures, such as anomaly detection, adversarial training, and secure aggregation. Through this analysis, we highlight key open research challenges in terms of scalability, secure orchestration, interpretable threat attribution, and dynamic trust calibration. This survey lays a foundational roadmap for secure, intelligent, and trustworthy IoT infrastructures powered by FMs.**

***Index Terms*—Foundation models, Zero trust, Internet of Things, Security, Emerging threats, Defense strategies**

K. Li and F. Dressler are with the School of Electrical Engineering and Computer Science, TU Berlin, Germany. K. Li is also with Real-Time and Embedded Computing Systems Research Centre (CISTER), Porto 4249–015, Portugal (e-mail: kaili@ieee.org, dressler@ccs-labs.org).

C. Li, X. Yuan, S. Li, and W. Ni are with Data61, CSIRO, Sydney, NSW 2122, Australia. X. Yuan and W. Ni are also with the School of Computer Science and Engineering, the University of New South Wales, Kensington, NSW 2033, Australia (e-mail: {conggai.li, xin.yuan, shenghong.li, wei.ni}@csiro.au).

S. Zou is with the State Key Laboratory of Public Big Data, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China (e-mail: dr-zousai@foxmail.com).

S. S. Ahmed is with the Department of Computer Engineering, Qassim University, Buraydah 52571, Kingdom of Saudi Arabia (e-mail: sa.ahmed@qu.edu.sa).

D. Niyato is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore (e-mail: dniyato@ntu.edu.sg).

A. Jamalipour is with the School of Electrical and Information Engineering, the University of Sydney, Sydney, NSW 2006, Australia (e-mail: a.jamalipour@ieee.org).

O. B. Akan is with the Division of Electrical Engineering, Department of Engineering, University of Cambridge, CB3 0FA Cambridge, U.K., and also with the Center for NeXt-Generation Communications (CXC), Koç University, 34450 Istanbul, Turkey (e-mail: oba21@cam.ac.uk).

TABLE I
ABBREVIATION AND FULL NAME

Abbreviation	Full name
AI	Artificial Intelligence
CPS	Cyber-Physical Systems
FedIoT	Federated Learning-enabled IoT
FL	Federated Learning
FM	Foundation Models
GNN	Graph Neural Network
IDS	Intrusion Detection Systems
IoT	Internet of Things
IIoT	Industrial IoT
JIT	Just-In-Time
LLMs	Large Language Models
LPA	Least Privilege Access
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
SMPC	Secure Multi-Party Computation
SOTA	State-Of-The-Art
TEEs	Trusted Execution Environments
ZTFM	Zero-Trust Foundation Models

I. INTRODUCTION

A. Artificial Intelligence for Internet of Things

ARTIFICIAL Intelligence (AI) is driving a paradigm shift in the Internet of Things (IoT), enabling intelligent, data-driven decision-making across distributed, sensor-rich environments [1], [2]. By enhancing the perception, reasoning, and actuation capabilities of IoT devices, AI facilitates smarter automation and system-wide optimization in sectors such as manufacturing [3], agriculture [4], transportation [5], and home automation [6].

According to the "Digital Transformation Enabler: Machine Learning" report from the Industry IoT Consortium [7], AI-enabled IoT systems already deliver tangible benefits. In manufacturing, AI-driven predictive maintenance reduces equipment downtime and boosts operational efficiency. For instance, KONUX applies AI-powered sensors to monitor railway infrastructure in real time [8], enhancing public safety and reliability. In agriculture, precision farming leverages AI algorithms to analyze soil conditions, forecast weather, and optimize irrigation. Consumer platforms like LG's ThinQ ON hub [9] use machine learning (ML) to manage smart home devices based on user behavior and contextual patterns.

Building upon these domain-specific applications, a new class of models – Foundation Models (FMs) – has emerged

as the next frontier in AI. Unlike task-specific models, FMs are pre-trained on massive datasets and exhibit cross-domain generalization capabilities through self-supervised learning [10]. Notable examples, e.g., OpenAI's GPT [11], Google's BERT [12], and China's DeepSeek [13], [14], support a wide range of downstream applications, from natural language understanding to multimodal reasoning. Their integration into IoT offers new opportunities for decentralized intelligence, adaptive control, and autonomous collaboration.

B. Security Challenges of Foundation Models

The scale and complexity of FMs present critical challenges when deployed in resource-constrained and adversarial IoT environments. Traditional AI pipelines fall short in addressing key issues such as data confidentiality, trust enforcement, energy efficiency, and security assurance. These limitations can be prominent in IoT systems, where devices are heterogeneous, intermittently connected, and often physically exposed.

Despite their expressiveness capability, FMs bring significant security and privacy risks in IoT deployments. Unlike traditional ML models, FMs are typically pre-trained on massive, heterogeneous datasets and deployed across distributed infrastructures [15]. This decentralized nature disrupts conventional security assumptions and expands the attack surface, leaving systems vulnerable to membership inference [16], model poisoning, backdoor injections, and adversarial inference attacks [17].

Energy consumption emerges as a key performance metric. FMs demand substantial computational resources, which poses challenges for energy-constrained IoT devices. An effective FM framework is expected to strike a balance between robust security, model accuracy, and energy efficiency.

Other crucial metrics include data integrity, communication overhead, and latency, all of which are key to many mission-critical IoT scenarios such as smart healthcare, industrial automation, and autonomous vehicles. The diversity and sensitivity of IoT-generated data (e.g., medical records or operational telemetry) require scalable, lightweight privacy-preserving techniques, e.g., differential privacy and secure federated aggregation, that go beyond what is used in centralized FM infrastructures [18].

These considerations substantiate the urgent need for a paradigm shift towards a unified framework that integrates the expressiveness of FMs with the security, privacy, and efficiency guarantees of a zero-trust architecture, tailored specifically for the constraints and requirements of IoT ecosystems.

C. Contributions

This paper advocates for embedding zero-trust security principles [19] into the FM lifecycle-enforcing continuous verification, least privilege access (LPA), and secure computation by design, and puts forth a new zero-trust FM (ZTFM) for IoT systems and applications. A ZTFM is envisaged to integrate zero-trust security principles into the design, training, and deployment of FMs, enabling continuous verification, fine-grained access control, and privacy-preserving computation across distributed IoT environments. By integrating zero-trust security with the adaptive, intelligent capabilities of FMs, IoT systems can move from being vulnerable, trust-based networks to self-defending ecosystems. The result is a more secure, more autonomous IoT, which is able to overcome the challenges of a dynamic and increasingly hostile digital landscape.

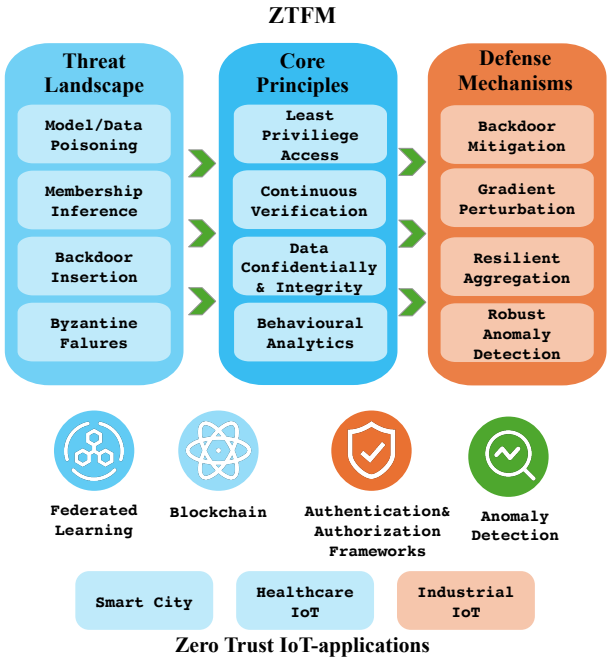

Fig. 1. A conceptual overview of ZTFM-enabled IoT applications. The architecture integrates Zero Trust principles with enabling technologies. These components collaboratively mitigate adversarial threats such as model poisoning, membership inference, backdoors, and Byzantine failures, enhancing the security posture of real-world IoT scenarios, including Smart Cities, Healthcare, and Industrial IoT.

This paper represents the first structured effort to define, formalize, and analyze ZTFM in the context of IoT. By examining how core zero-trust principles, including *LPA*, *Continuous Verification*, *Data Confidentiality and Integrity*, and *Behavioral Analytics*, can be operationalized in FM-driven IoT systems, we aim to bridge the gap between trustless IoT architectures and scalable AI frameworks. We also provide a technical synthesis of key enabling technologies for ZTFMs, including federated learning (FL), blockchain-based identity management, micro-segmentation, and trusted execution environments (TEEs), outlining how they jointly support secure, collaborative AI across heterogeneous IoT and edge networks. Moreover, we analyze the unique threats and limitations that arise in ZTFM-enabled IoT environments and identify future research opportunities for scalable, interpretable, and resilient trust enforcement in real-world IoT applications.

The contributions of this paper are summarized as follows:

1) We present the first comprehensive synthesis of ZTFMs in IoT systems and applications, identifying a timely opportunity to integrate zero-trust principles with the training and deployment of FMs. This integration addresses the growing need for dynamic, decentralized trust management in AI-driven IoT systems.
2) We formalize four core security principles and ana-

lyze their operationalization in constrained, potentially malicious, and heterogeneous IoT environments. This helps reveal implementation gaps in current zero-trust deployments and highlights open research directions for principle-level enforcement in FM workflows.

3) We propose a unified technical framework of ZTFMs that combines FL, blockchain-based identity management, micro-segmentation, and TEEs. This framework not only enables privacy-preserving and verifiable AI computation at the network edge, but also identifies current limitations in scalability, secure orchestration, and interoperability across IoT infrastructures.
4) We conduct an in-depth analysis of emerging threats in ZTFM-enabled IoT systems and review defense strategies, such as anomaly detection, adversarial training, and secure aggregation. Accordingly, we identify future research challenges, including lightweight secure multi-party computation (SMPC), interpretable threat attribution, and AI-driven dynamic trust calibration for real-time edge intelligence in future intelligent IoT systems.

D. Promising Applications of ZTFMs for Internet of Things

The combination of zero-trust principles and FMs is poised to transform the security of the IoT. In a traditional IoT environment, devices are often trusted by default once they are authenticated – a risky assumption given the growing scale and sophistication of cyber threats [20]. A zero-trust approach, summarized by the maxim "never trust, always verify," insists that no device, user, or system be trusted automatically [21]. Every interaction must be continuously authenticated, authorized, and validated.

FMs, which are large-scale AI models trained on vast and diverse data, bring a new level of intelligence to implementing zero-trust for IoT. They can learn complex patterns of behavior across different types of devices and contexts [22], [23]. By ingesting telemetry, logs, sensor outputs, and network behavior, these models build a detailed, evolving understanding of what "normal" looks like for each device and system [24]. This enables them to continuously verify the legitimacy of device actions, detecting subtle signs of compromise or malfunction without relying on pre-written signatures or manual rules.

In practice, a ZFTM for IoT would monitor device behavior at scale, dynamically adjusting access permissions based on ongoing risk assessments [25]. If a smart thermostat begins sending large volumes of encrypted data at odd hours, the model could immediately recognize this deviation from normal behavior, isolate the device, and alert administrators. Access control becomes dynamic and context-aware, based not only on static identities but on live, real-world behavior.

ZFTM can also automate policy generation and enforcement. As new devices are introduced into the network, the FM could propose security policies tailored to the device's expected behavior and risk profile, reducing administrative burden. Over time, the system would adapt to changes without human intervention, maintaining a strong security posture as IoT ecosystems evolve [26].

With advances in edge AI and FL, lightweight versions of these models could be deployed on gateways or edge devices [27]. This ensures that zero-trust principles are enforced even when connectivity to a central cloud is limited, preserving resilience and privacy in sensitive environments like smart homes [28], factories [29], supply chains [30], and healthcare facilities [31]. Beyond these terrestrial applications, Mao et al. [32] explored a blockchain-enabled cold-start aggregation scheme for federated reinforcement learning in zero-trust LEO satellite networks. Their approach demonstrates that ZTFM principles can be extended to highly dynamic and adversarial edge environments, such as space-based IoT systems, enabling secure and privacy-preserving model updates even in scenarios with intermittent connectivity and untrusted participants.

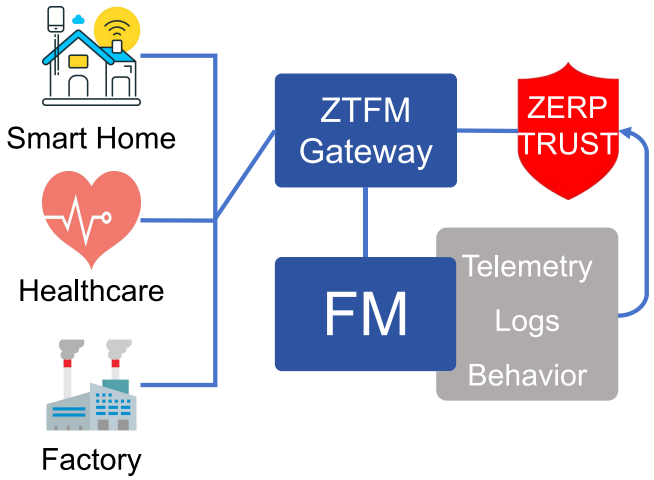

Fig. 2. ZTFM-enabled security architecture for IoT. FMs analyze behavioral data to enforce Zero-Trust decisions and trigger alerts across smart home, healthcare, and industrial IoT settings.

E. Organization

The rest of this paper is organized as follows. Section II surveys existing literature on zero-trust frameworks and the application of FMs in IoT systems and environments. Section III examines critical security threats associated with FMs, including model and data poisoning, model inference attacks, Byzantine faults, backdoor attacks, and challenges in intrusion detection. Section IV explores four foundational principles that form the basis of the ZTFM architecture: LPA, continuous verification, data confidentiality and integrity, and behavioral analytics. Section V analyzes the core technical components of ZTFM and their roles in supporting a robust zero-trust architecture. Section VI identifies open research challenges in the deployment of ZTFM for IoT, emphasizing the necessity of interdisciplinary approaches. Section VII concludes the paper.

II. Status Quo of Zero-trust Frameworks and Foundation Models

In this section, we first assess the state-of-the-art (SOTA) in zero-trust frameworks that have primarily focused on privacy protection and data reliability, with limited attention to federated threats, access control, continuous verification, and data integrity. We then examine existing research in FMs relevant to IoT systems and applications.

A. Zero-Trust Frameworks

Zero-trust frameworks can enhance cybersecurity by continuously verifying access and limiting implicit trust, thereby effectively protecting user privacy against unauthorized exposure. These frameworks can improve data reliability through rigorous authentication and strict access controls so that only trustworthy entities interact with sensitive information.

In [33], a research agenda was presented to improve security in the Metaverse through a zero-trust continuous authentication framework. The privacy issues associated with implementing continuous authentication in social VR were examined based on a foundational element of the Metaverse. An FL-based adaptive authentication framework that utilizes multimodal biometric data was developed, which can explore biometric authentication for continuous verification in VR. FL introduces a privacy-preserving approach that allows collaborative ML across distributed devices while maintaining data confidentiality [34]–[37]. However, existing FL protocols remain susceptible to both internal and external adversaries, posing risks to data privacy and system integrity [38].

Beyond developing robust global models, it is crucial to design FL frameworks that offer strong privacy guarantees and resilience against various adversarial threats. Traditional cryptographic protocols, e.g., zero-knowledge proofs and garbled circuits, provide potential solutions for secure computations on private data, however, their scalability remains a major obstacle in large-scale FL systems [39]. To address this issue, alternative approaches in ZTFM, e.g., LPA, continuous verification, data confidentiality and integrity, and behavioral analytics, could be explored to enhance the integrity and reliability of user-reported metrics while maintaining privacy.

In [40], an intelligent connected vehicle system behavior paradigm was built on a zero-trust framework to enhance security and reliability in information perception, communication, and control within a vehicular platoon. The framework can mitigate interference from complex behaviors, information exchange, network topology, and environmental factors. The authors of [41] explored the transformative impact of AI/ML, blockchain, quantum computing, and cloud/edge technologies on the development and effectiveness of zero-trust architectures. Although these technologies can contribute to advanced trust evaluation and adaptive access control in zero-trust models, it is still difficult to ensure continuous verification and least-privilege access across hybrid and multi-cloud environments.

An analysis of the transition from traditional perimeter-based security to the zero-trust framework was given in [42]. The impact of emerging technologies, such as AI and quantum computing, was explored on zero-trust policies and deployment strategies. In particular, ML in zero-trust was examined, showcasing its ability to enhance security through pattern analysis, anomaly detection, and threat prediction, enabling real-time decision-making.

Complementing these perspectives, Mao et al. [43] provided a comprehensive survey of security and privacy challenges in 6G network edge environments, emphasizing the intersection of Zero-Trust principles with edge computing, AI, and network slicing. The survey identifies key threats, such as resource-constrained adversaries, dynamic trust bootstrapping, and distributed data leakage. Their analysis highlights that while Zero-Trust concepts provide a strong foundation for 6G edge security, practical deployments must contend with unique trade-offs between latency, privacy, and scalability.

B. Foundation Models for Internet of Things

FM can be leveraged to enhance IoT by providing powerful, generalizable AI capabilities that improve real-time decision-making and automate complex tasks across diverse applications. Their ability to learn from vast amounts of heterogeneous IoT data enables adaptive, scalable, and efficient deployments, greatly advancing the intelligence and responsiveness of IoT systems. The key features of FMs that make them applicable to IoT systems include:

- Multimodal integration, which aims to fuse and jointly process multimodal data (e.g., images, sensor readings, textual metadata), enhancing IoT systems' situational awareness and context comprehension.
- Real-time decision-making, which supports fast inference and real-time responsiveness, essential for latency-sensitive IoT applications, such as healthcare monitoring, industrial automation, and autonomous systems.
- Adaptability, where FMs adapt to evolving environmental conditions and dynamic IoT data distributions through minimal additional training.
- Representation learning, which extracts meaningful and generalized patterns from noisy, sparse, or multimodal IoT data, improving accuracy and reliability.

Pipeline parallelism, data parallelism, and multi-modal learning can be employed to advance the sustainable development of FM in the 6G era [44]. In pipeline parallelism, adapting activation and gradient compression along communication resource allocation helps mitigate communication bottlenecks caused by unstable wireless links.

Network FMs can be designed to capture the distinct characteristics of network data [45]. In particular, a network FM incorporates a multi-modal embedding layer to identify cross-modal dependencies between different packet fields for building data representation. In [46], a zero-shot IoT sensing was developed with an FM text encoder, which can align IoT data embeddings with semantic embeddings. To enhance the extraction of semantic embeddings, the underlying physics of IoT sensor signals was used in a cross-attention mechanism that integrates a learnable soft prompt, optimized on training data, with an auxiliary hard prompt encoding domain-specific knowledge.

The authors of [47] surveyed the potential of FM and large language models (LLMs) in Cyber-Physical Systems (CPS) and the IoT by addressing challenges within the perception, cognition, and communication. Different from traditional task-specific ML models, which face limitations due to data annotation needs and sensor heterogeneity, FM can provide a task-agnostic and self-supervised learning framework that enhances adaptability. Despite their promise, effectively integrating FM

TABLE II
RELATED SURVEYS WITH KEY APPLICATIONS, TECHNICAL FEATURES, AND LIMITATIONS

	Applications examples	Technical features	Limitations
Zero-Trust Authentication [33], [39]–[42]	Designed for continuous verification in social VR or vehicular platoons.	Strengthens security through adaptive access control and reduces reliance on perimeter-based defenses.	Can be challenging to scale across hybrid environments and raises concerns about privacy and algorithmic bias.
Privacy-Preserving FL [34]–[38]	Enables distributed ML while maintaining data on local devices.	Protects sensitive user data and supports collaborative model training.	Involves additional computational overhead and may be vulnerable to internal or external adversaries.
Foundation Models for IoT and CPS [45]–[47]	Provides multi-modal sensing in complex IoT or CPS scenarios.	Offers a task-agnostic, self-supervised learning framework that enhances adaptability across devices.	Requires domain-specific innovation and can be difficult to integrate seamlessly into heterogeneous systems.
Network and Wireless Foundation Models [44], [48], [50]	Targeted for 6G networks, incorporating pipeline or data parallelism.	Improves communication efficiency and real-time data processing in unstable wireless links.	Implementation complexity is high, and dynamic adaptation remains a significant challenge.
Decentralized Training of Large Models [49], [51]	Uses model parallelism and scheduling for distributed GPU tasks.	Increases scalability and resource utilization, speeding up training in heterogeneous networks.	Suffers from high communication overhead, and network heterogeneity can degrade performance.

and LLMs into CPS-IoT requires moving beyond simplistic adaptations from natural language processing (NLP) and computer vision. Implicit neural representations, which encode signals or objects using neural networks, have gained attention as a continuous and memory-efficient alternative to traditional discrete representations.

The authors of [48] analyzed leveraging FM to enhance hypernetworks for generalizable implicit neural representation tasks. It confirms that FM can improve hypernetwork performance across labeled and hidden classes, demonstrating adaptability and efficiency in various IoT scenarios. In addition, training large FM can rely on model parallelism in a decentralized setting over a heterogeneous network [49]. In particular, a scheduling algorithm can be designed that distributes computational tasklets across decentralized GPU devices connected via a slow, heterogeneous network. To optimize resource allocation, a formal cost model with an evolutionary algorithm can be used to determine the task distribution strategy that enhances training efficiency.

In [50], a wireless vision was studied for designing FM tailored to the unique demands of next-generation 6G systems, which aims to enable the AI-native networks. Different from existing NLP-based FM, the proposed framework advocates for the development of large multi-modal models with three core capabilities, namely, processing multi-modal sensing data, grounding physical symbol representations in real-world wireless systems through causal reasoning, and retrieval-augmented generation. In [51], a training and serving vision was presented for designing FM, in the aspects of networking, storage, and computing. Parallel training strategies with GPU memory optimization and communication optimization techniques can be conducted, so that each strategy is applied for unique application scenarios. FM can be developed to improve service performance with advanced batch processing, sparse acceleration, and multi-model inference.

C. Comparison with Existing Surveys

Different from the existing surveys, which separately focus on zero-trust authentication frameworks for access control in niche domains (e.g., vehicular networks, metaverse environments) and privacy-preserving federated learning mechanisms (e.g., local model updates, gradient obfuscation), this survey contributes a comprehensive view of ZTFMs that bridges these two lines of research. We delineate the design of ZTFM through four core principles, including Least Privilege Access, Continuous Verification, Data Confidentiality and Integrity, and Behavioral Analytics. We also analyze how they mitigate key attack vectors, such as model poisoning, inference attacks, and insider threats. Moreover, we synthesize technical enablers, including blockchain identity management, TEEs, and federated zero-knowledge proofs (ZKPs), which are not covered holistically in earlier works.

Notably, this survey does not treat zero-trust architecture and FL as isolated paradigms, but instead presents ZTFM as a convergent security framework for future AI-native IoT ecosystems. To the best of our knowledge, it is the first to offer a system-level perspective on how FMs can serve as both targets of protection and active agents of trust enforcement under zero-trust assumptions.

III. FOUNDATION MODELS AND SECURITY CHALLENGES

FMs are large-scale, pre-trained ML models that can be adapted to specific tasks with minimal additional training [22]. While these models serve as powerful tools for building AI-driven IoT systems, their deployment in decentralized, collaborative environments introduces unique security vulnerabilities. These vulnerabilities stem from adversarial manipulations at various stages of model training and inference, impacting both model integrity and user privacy [52]–[54]. Key security challenges include model and data poisoning, model inference attacks, Byzantine failures, backdoor attacks, and intrusion detection challenges [55].

A. Model and Data Poisoning

Adversaries employ data poisoning and model poisoning as primary adversarial strategies against FMs, aiming to insert malicious content into the training pipeline to undermine the target model [56], [57].

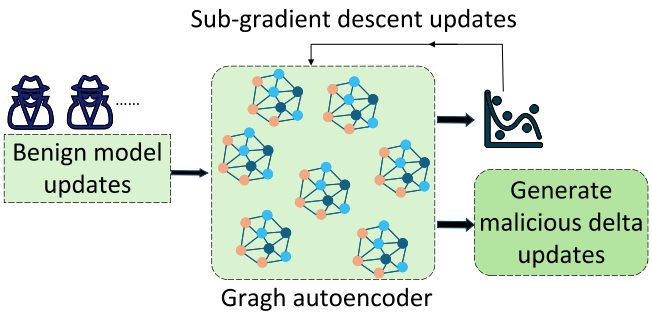

Fig. 3. An illustration of an adversarial variational graph autoencoder-enabled model poisoning attack [60].

1) Data Poisoning Attacks: Data poisoning attacks occur at the data level, where adversaries manipulate the local training data to corrupt the model's learning process. A common strategy is to inject maliciously crafted samples, often mislabeled or containing imperceptible perturbations, into the local datasets of compromised clients. These samples are carefully designed to introduce backdoors or bias the global model's decision boundaries [56], [58], [59]. For example, attackers may insert inputs that associate a specific pattern (e.g., a pixel patch in an image or a phrase in text) with an incorrect target label. Once the global model incorporates updates from poisoned clients, it begins to misclassify inputs containing that pattern, effectively embedding an attack trigger [58]. These attacks are especially dangerous in FL due to limited visibility into individual client data and the lack of centralized oversight. Moreover, data poisoning is often stealthy and adaptive. Poisoned data can be sparse, making it hard to detect during aggregation, particularly in non-i.i.d. data environments typical of IoT deployments.

2) Model Poisoning Attacks: Model poisoning attacks take place at the model update level. Instead of modifying data, the attacker directly manipulates model parameters or gradients before submitting them to the server. The goal is to inject malicious behavior into the global model or to maximize divergence and disrupt the convergence of the training process. In [60], the authors introduced a sophisticated threat model where an attacker employs an adversarial variational graph autoencoder to infer structural relationships among benign local models, which is depicted in Fig. 3. These relationships are then adversarially modified to generate malicious model updates that still appear statistically similar to benign ones. Notably, this attack operates without requiring access to private data, making it particularly relevant to black-box FL settings such as IoT edge deployments and FL fine-tuning, where access to raw training data is restricted.

Another strategy described in [61] enables attackers to eavesdrop on the global model and benign client updates to reconstruct the internal graph structure linking model parameters and data features. The attacker then strategically perturbs these correlations to degrade model performance or introduce targeted behaviors. This threat is especially concerning for large-scale FMs, where complex feature relationships may be exploited without direct access to local datasets. In [62], the focus shifts to malicious user injection, where adversaries inject fraudulent clients that submit poisoned updates aligned with a predefined backdoor objective. This method is scalable and well-suited for FL environments, such as IoT networks, where participant authentication may be limited. Such vulnerabilities are particularly critical when deploying FMs at the edge, where a few compromised users can influence the model's downstream behavior.

B. Membership Inference Attacks

Model inference attacks in FL exploit access to shared model updates or intermediate parameters to infer sensitive information such as user attributes, class distributions, or even raw training data [52]. Because FL operates in a decentralized setting, malicious users can execute these attacks without full access to the global model, amplifying the attack surface and complicating privacy preservation.

In [63], the authors evaluated various secure aggregation protocols and demonstrated that these mechanisms do not fully protect user data. They introduced a differential selection attack combined with de-noising schemes, which allows a malicious actor to infer multi-label classification results from IoT node data, effectively breaching privacy under common FL settings. A poisoning-assisted inference attack was proposed in [64], where the attacker leverages benign model updates to extract sensitive feature information. By using a binary attack model, adversaries can identify data patterns not meant to be revealed. Furthermore, a targeted poisoning strategy was introduced, allowing attackers to manipulate training labels and shift the global model's decision boundaries. This manipulation inadvertently causes benign clients to expose additional private feature information.

Inference attacks present serious privacy risks in collaborative learning frameworks such as split learning and FL. In [65], both passive and active inference attacks were proposed. The passive attack uses semi-supervised learning on auxiliary data to infer labels, while the active variant manipulates the training process to increase the global model's reliance on the attacker's sub-model, improving inference accuracy. A broader evaluation in [66] showed that data complexity affects inference attack success, with a trade-off observed between model stealing and inference attack effectiveness. Using architectures like ResNet18 and datasets such as CelebA and Fashion-MNIST, the study highlighted how model behavior varies with data characteristics. In [67], membership inference attacks were explored based on prediction sensitivity. By observing how predictions change under small input perturbations, attackers can determine whether a sample was part of the training set, even without knowledge of the model or data, posing a significant privacy threat.

In [68], a hypothesis testing framework was proposed to improve membership inference attacks. The framework uses reference models to boost the true positive rate while keeping the false positive rate under control. The authors also analyzed attacker uncertainty, demonstrating that their method can narrow down uncertainty to a single-bit secret, whether or not a specific data point was part of the training data. However,

this approach relies on well-calibrated reference models, which may be unavailable in practice. Future extension to black-box settings with limited or unreliable reference models would be desirable. The authors of [69] explored feature inference attacks at the model prediction stage under a strong adversarial assumption, where the attacker only has access to the model and its outputs. The attackers could infer private feature values by analyzing model outputs across various architectures. However, its effectiveness relies on having ample prediction samples and may weaken under regularization or dynamic model updates. Future work could explore adaptive defenses or noise-aware attack strategies suited for real-world settings.

C. Byzantine Failure Attacks

Byzantine failure attacks in FM occur when malicious or faulty users share incorrect or adversarial updates, disrupting the training model's convergence and accuracy [70]. These attacks can take various forms, such as random noise injection, adversarial data pollution, software bugs, network asynchrony, or biases in local datasets, making it challenging for aggregation mechanisms to distinguish between benign and malicious contributions [71].

Existing FMs that were considered resilient to Byzantine failures remain susceptible to targeted local model poisoning attacks [72]. By manipulating the training models from compromised IoT devices, an attacker can significantly degrade the performance of the training model, steering it in a direction opposite to its intended optimization path. While certain defenses adapted from poisoning countermeasures offer partial protection, their effectiveness varies depending on the attack scenario. Existing defense mechanisms against Byzantine attacks were examined, and a vulnerability in FM was argued in [73]. Since the server relies solely on user-reported dataset sizes for weighting updates, without verification due to privacy constraints, malicious IoT devices can manipulate their declared dataset sizes to gain undue influence. Two misreporting strategies were studied, namely, attackers with small datasets falsely claiming to have similar-sized datasets as benign IoT devices, and attackers with comparable datasets inflating their sizes to disproportionately impact the aggregation process.

In [74], the authors established that existing linear combination methods for aggregating IoT updates cannot withstand a single Byzantine device. A single compromised device can manipulate the FM into selecting an arbitrary model update, potentially with excessive magnitude or a misleading direction. A Byzantine resilience strategy was developed, which can outline sufficient conditions for an aggregation rule to tolerate multiple Byzantine devices. An FM scheme that simultaneously conducts privacy preservation and resilience against Byzantine failure attacks was presented [75]. The approach employs three-party computation to implement an aggregation method while maintaining the confidentiality of local training models. To enhance efficiency, the scheme includes a maliciously secure top-k protocol with reduced communication overhead and an optimized secure shuffling protocol, which is essential for the secure top-k mechanism.

Both works in [74] and [75] have improved the security of federated systems, but a zero-trust approach requires stricter assumptions, treating all clients and intermediaries as untrusted by default. In this sense, the three-party computation and secure protocols developed in [75] are more in line with zero-trust principles by minimizing reliance on any single party and limiting information exposure. To fully align with zero-trust architecture, these methods could be extended with secure authentication, trusted initialization, and real-time client behavior monitoring.

D. Backdoor Attacks

Backdoor attacks refer to a type of adversarial attack where an attacker inserts a hidden, malicious trigger into an ML model during its training phase, especially for the FM. The trigger is designed to make the model behave in a specific, undesirable way when exposed to certain inputs, which are typically controlled by the attacker. However, the model's overall performance on normal data remains unaffected, making the attack hard to detect during regular use.

Nguyen et al. [76] study the security risks of backdoor attacks in FL, where malicious participants can secretly insert harmful behaviors into shared models. They review different attack methods and propose various defense strategies, such as anomaly detection and robust aggregation, to make FL more secure. Wang et al. [77] explore how backdoor attacks can be inserted into FL systems by targeting the tail distributions of data, which consist of rare or less-represented data points. The authors show that even small changes to the data in these tails can allow attackers to introduce malicious behavior into the model without affecting its overall performance. This demonstrates that FL systems are vulnerable to such backdoor attacks, particularly when the data is imbalanced.

Gong et al. [78] examine coordinated backdoor attacks in FL, where multiple attackers collaborate to insert triggers into the global model. The key insight is that these triggers can be model-dependent, meaning they exploit specific vulnerabilities in the model's architecture. This makes the backdoor attack more efficient and harder to detect. The paper highlights the challenges in defending against such coordinated attacks in FL systems. The authors in [79] propose a DeepSight that examines the internal structure and outputs of neural network updates to identify and filter out these malicious contributions. It aims to enhance model security without negatively impacting performance on legitimate data.

E. Adversarial Attacks on Intrusion Detection Systems

Adversarial attacks on Intrusion Detection Systems (IDS) are becoming a bigger issue in cybersecurity. In these attacks, hackers change the way network traffic looks so they can slip past ML systems that are supposed to catch them. They might hide harmful content or make their activity look normal to fool the system. Interestingly, these tricks are very similar to the ones used to fool FLs. In both cases, attackers take advantage of how sensitive these models are to small input changes. These changes can cause the model to make mistakes without raising alarms. This shows that both IDS and FMs share a common weakness. They can be fooled by well-crafted inputs because they rely heavily on data patterns.

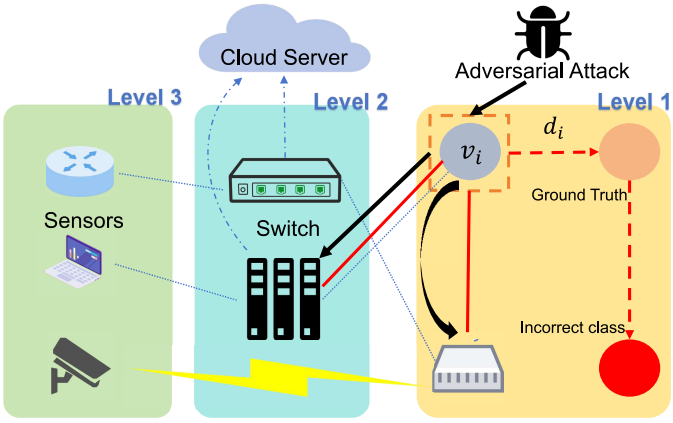

Fig. 4. An illustration of hierarchical black-box adversarial attacks on GNN-based NIDS in IoT networks [80]. Here, adversaries exploit multi-level infrastructure and graph feature perturbations to induce misclassifications while bypassing intrusion detection systems.

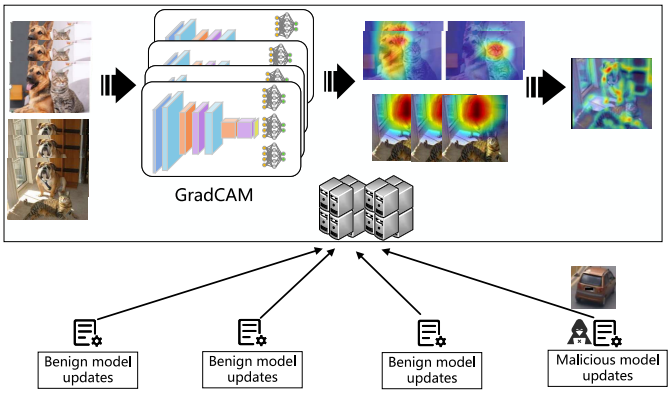

Fig. 5. An illustration of the GradCAM-based defense mechanism against model poisoning attacks developed in FMs [85].

Zhou et al. [80] investigate the vulnerabilities of graph neural network (GNN)-based IDS used in IoT networks to hierarchical adversarial attacks. The authors present a novel attack framework that targets the structural properties of the IoT network's graph, exploiting the relationships between devices to manipulate the IDS's detection capabilities. By introducing adversarial perturbations in the graph structure, the attackers can cause the GNN to misclassify malicious activities, thereby undermining the security of the IoT network. Dai et al. [81] introduce adversarial attacks that manipulate graph structures of GNNs and propose a reinforcement learning-based attack strategy that modifies graph structures by adding or removing edges to mislead GNN models in node and graph classification tasks. Moreover, they introduce alternative gradient-based and genetic algorithm attacks for different attack scenarios, including cases with and without access to model gradients.

These challenges highlight the limitations of traditional FM frameworks, which rely on implicit trust among participants. Addressing these vulnerabilities requires a paradigm shift toward a Z, where no participant or device is inherently trusted, and all interactions are rigorously verified.

F. Defense Strategies Against Adversarial Attacks

Addressing the vulnerabilities in FM requires a combination of robust defense mechanisms, including anomaly detection, secure aggregation, adversarial filtering, and privacy-preserving techniques. Below, we categorize and discuss key defense strategies that have been proposed to counteract various adversarial threats in FM.

1) Defending Against Data Poisoning Attacks: In [82], an ensemble-based FL framework was proposed. Users are divided into groups, each training a local model. A majority voting scheme is applied during inference to determine the final prediction, reducing the influence of any single poisoned model on the overall outcome. The RSim-FL framework [83] enhances FL security by using representational similarity analysis. It compares global and local model representations to form a consistency set and applies K-means clustering to identify and isolate adversarial users based on representational deviations. This method is applicable to FMs, where consistent representations are important, and supports zero-trust by validating clients based on behavior rather than assumptions of trust. Building on this, the study in [84] introduced a privacy-preserving, hierarchical aggregation defense suited for IoT environments, where edge IoT nodes perform synchronous aggregation under the coordination of a leader node. Encrypted poisoned gradients are detected during this process, offering scalability and robustness in heterogeneous settings. The framework can support zero-trust through encrypted communication and multi-level validation, making it suitable for FM-based FL in resource-constrained, decentralized systems.

2) Defending Against Model Poisoning Attacks: Model poisoning attacks manipulate model updates to inject adversarial behaviors into the global FM. To mitigate these threats, recent works have explored advanced detection and aggregation strategies in IoT. Zheng et al. [85] proposed a defense mechanism against model poisoning attacks in FM, which integrates gradient-weighted class activation mapping (GradCAM) and an autoencoder to enhance detection effectiveness beyond traditional Euclidean distance-based methods. As shown in Fig. 5, a heat map can be generated by GradCAM for each uploaded local model update, converting it into a lower-dimensional visual representation. This transformation highlights hidden features within the heat maps, improving the ability to detect anomalous patterns and identify malicious local models accurately. To enhance security while preserving privacy, a two-trapdoor homomorphic encryption approach was proposed in [86]. In particular, a Byzantine-resilient aggregation incorporating cosine similarity was designed to evaluate the distance between encrypted gradients, aiming to identify encrypted malicious gradients.

Zhang et al. [87] proposed an approach where the FM server employs the Cauchy Mean Value Theorem to predict each user's model updates based on historical trends. To assess the consistency of these updates, the Euclidean distance between the predicted and received model updates can be computed for each user. In this case, a suspicious score is assigned to

the user, which is adjusted in each iteration to track potential anomalies. In [88], a targeted perception poisoning attack was developed on FM for object detection, where malicious users inject perception-poisoned local model updates into the federated training process. To mitigate such threats, a spatial signature analysis was studied as a defense mechanism, which differentiates between benign and poisoned model updates to remove adversarial influence and protect the integrity of FM.

3) Defending Against Model Inference Attacks: A model inference attack was designed to attack user authentication of FM in 5G and IoT systems [89]. In particular, the model's input consists of received power and phase shift, enabling the attacker to determine whether specific signals were part of the classifier's training data. To execute the attack, the attacker can collect signals and classification results through spectrum observation, construct a surrogate classifier, and apply an inference attack to infer whether a received signal corresponds to one used in the service provider's training dataset.

4) Defending Against Byzantine Failure Attacks: A divide-and-conquer aggregation algorithm was developed to defend against Byzantine failure attacks in FM [90]. Inspired by defenses against poisoning attacks, their divide-and-conquer aggregation identifies and mitigates malicious updates by detecting significant deviations in update space. The algorithm can compute the principal component of the updates, calculate their projections along this direction, and discard a fixed fraction of updates with the largest projections to reduce the impact of adversarial manipulations.

5) Defending Against Backdoor Attacks: Backdoor attacks allow attackers to insert hidden triggers that activate malicious behaviors under specific conditions while maintaining normal performance on regular IoT data. Nguyen et al. [91] propose FLAME, a defense method that mitigates backdoor attacks by evaluating model updates against strategically designed test inputs, ensuring that harmful modifications are identified and filtered out. Unlike traditional defenses, FLAME was designed for the decentralized nature of FL, offering a practical and scalable solution without compromising data privacy.

Xie et al. [92] propose Certifiably Robust FL (CRFL), which applies randomized smoothing techniques to enhance model resilience. Even if an attack successfully poisons a model, the CRFL ensures robustness by enforcing mathematical guarantees that limit backdoor effectiveness. Cao et al. [93] proposed FLTrust, a defense mechanism that establishes a "root of trust" by maintaining a small, clean dataset at the central server to evaluate updates from all IoT devices and filter out suspicious ones. This approach ensures that FL remains secure and robust against adversarial manipulations, preventing attackers from degrading model performance while maintaining high accuracy.

Gong et al. [78] examine how attackers secretly manipulate FL models by injecting hidden malicious behaviors. They classify these attacks into data poisoning (tampering with training data) and model poisoning (altering model updates), and accordingly categorize defense mechanisms, aiming to provide a structured understanding of existing strategies. To address backdoor threats in FL, Huang et al. [94] introduced Suprte, a trust evaluation mechanism that assigns trust scores to participating devices based on their historical behaviors. This reduces the influence of suspicious updates and prevents attackers from injecting harmful changes. It is necessary because FL allows multiple IoT devices to train models collaboratively without sharing data, making it vulnerable to hidden attacks that can be difficult to detect using traditional security methods.

6) Adversarial Defense to IDS: An Adv-Bot was proposed in [95], which is a framework designed to generate realistic adversarial botnet attacks to bypass network IDS. The authors evaluate various attack strategies and their impact on network IDS performance, showing that adversarial samples can effectively reduce detection accuracy. The study highlights the importance of robust defenses against adversarial attacks in cybersecurity. Venturi et al. [96] introduced ARGANIDS, a network IDS leveraging an adversarially regularized graph autoencoder (ARGA) for detecting network anomalies. By incorporating adversarial training, the model improves robustness against evasion attacks and enhances anomaly detection performance. The authors demonstrate that ARGANIDS outperforms traditional network IDS techniques in accuracy and resilience against adversarial modifications.

G. Lessons Learned

While existing defense methods offer some protection, they incur critical limitations. Techniques like adversarial training, input filtering, or anomaly detection rely on model-specific configurations and are vulnerable to adaptive adversaries that evolve beyond fixed defense strategies. These defenses can struggle to scale across different environments and devices in FM-based IoT systems, as FMs, equipped with a general-purpose architecture, are susceptible to subtle and transferable adversarial perturbations that can affect various downstream tasks. To this end, the following lessons can be learned:

- *Model-agnostic defenses are essential.* Relying on hard-coded protections or assumptions about a model's structure or input distribution can leave systems exposed to unseen threats.
- *Scalability and adaptability must be prioritized.* Defenses must function reliably across different system configurations, data types, and threat models in large IoT networks or when using FMs in different applications.
- *Attack resilience must be continuous and proactive.* Static defense is insufficient against adaptive adversaries. Systems should integrate real-time behavioral monitoring, dynamic response mechanisms, and layered protections.

By enforcing strict identity verification, access controls, and ongoing trust evaluation for every device, regardless of location or role, zero-trust can shift the paradigm from perimeter-based defense to comprehensive internal verification. This is useful in FM-empowered IoT networks, where devices frequently connect and disconnect, and decisions must remain verifiable and resilient to unpredictable inputs.

IV. Core Principles of ZTFMs

ZTFMs combine the expressive power of FMs with the security guarantees of zero-trust architectures, offering a promising pathway for secure, privacy-preserving intelligence in IoT

TABLE III
SUMMARY OF ATTACKS AND DEFENSE STRATEGIES ON FMS

Attack Types	**Attack Description**	**Defense Strategy**	**Limitations of Current Defense Strategies**
Data Poisoning Attacks [56], [58], [59]	Data poisoning attacks occur at the data level, where adversaries manipulate the local training data to corrupt the model's learning process.	• Ensemble FL with Majority Voting [82] • RSim-FL with Clustering [83] • Hierarchical IoT Defense [84]	Limited generalizability across dynamic IoT environments; lacks real-time adaptability.
Model Poisoning Attacks [60]–[62], [82]–[84]	Data poisoning attacks occur at the model update level, where adversaries manipulate the training process by injecting malicious model updates to degrade the model's performance.	• Feature-Based Anomaly Detection [85] • Privacy-Preserving Gradient Analysis [86] • Historical Behavior-Based Detection [87] • Spatial signature analysis [88]	Vulnerable to adaptive strategies that mimic benign behavior; high false negatives.
Membership Inference Attacks [52], [63]–[69]	The attacker exploits a trained model's outputs to infer sensitive information about its training data or parameters. For example, membership inference can reveal if a specific record was in the training set, or model extraction can steal the model.	• Adversarial Feature Masking [89]	Often model-specific and insufficient against adaptive inference; lacks scalability.
Byzantine Failure Attacks [70]–[75]	In distributed or FL, some participants (Byzantine nodes) behave maliciously or unpredictably, sending incorrect or adversarial model updates that corrupt the global model's training process.	• Divide-and-Conquer Aggregation Algorithm [90]	Ineffective in large-scale or heterogeneous networks; reactive rather than preventive.
Backdoor Attacks [76]–[79]	The adversary injects a hidden "trigger" pattern into the training data so that the model performs normally on standard inputs but produces an attacker-chosen output when the trigger is present, effectively embedding a backdoor.	• Adversarial Update Evaluation [91] • Certifiable Robustness Against Backdoors [92] • Root-of-Trust Model Verification [93] • Trust-Based Scoring of Model Updates [94]	Detection methods may require retraining and assume known trigger patterns; impractical for resource-limited IoT devices.
Adversarial Attacks on IDS [80], [81]	The attacker crafts small perturbations to input data at inference time to cause the model to misclassify it (an evasion attack). This is often used to evade security systems (e.g., making malicious network traffic appear benign to an intrusion detection model).	• Adversarial training (augmenting training with adversarial examples) is a primary defense to improve model robustness [95]. • Input preprocessing or anomaly detection can be applied to identify and reject adversarial inputs [96].	Resource-intensive; poor generalization to novel attacks; brittle under adaptive adversaries.

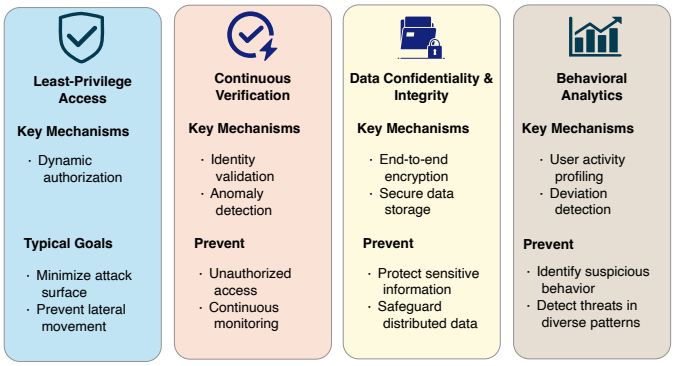

Fig. 6. Core principles of ZTFM in IoT environments. Each principle – Least Privilege Access, Continuous Verification, Data Confidentiality & Integrity, and Behavioral Analytics – maps to specific mechanisms and goals designed to enhance security posture in distributed, heterogeneous IoT systems.

environments. Unlike traditional security paradigms that rely on static trust boundaries, ZTFMs embed continuous verification, fine-grained access control, and secure data sharing into the AI model lifecycle. This section elaborates on the four foundational principles underpinning ZTFM, including *LPA*, *Continuous Verification*, *Data Confidentiality and Integrity*, and *Behavioral Analytics*, emphasizing their realization, benefits, and domain-specific challenges.

A. Least Privilege Access (LPA)

LPA ensures that users, devices, and applications are granted only the minimum permissions necessary to fulfill their roles, thereby minimizing the attack surface and preventing lateral movement in the event of a breach. In ZTFM architectures, LPA is implemented through mechanisms such as Just-In-Time (JIT) access, Role-Based Access Control (RBAC), and dynamic policy enforcement.

Tuyishime et al. [97] addressed the security challenges in remote online laboratories, where VPN-based access models often lead to overprivileged access and increase the risk of lateral movement. To mitigate this, they propose Twingate, a ZTNA-based system that enforces per-session, per-resource access policies, thereby aligning with the least privilege principle. The system incorporates micro-segmentation, device compliance checks, and real-time identity verification to ensure that users only access authorized lab environments during specific time windows. Their evaluation, based on academic lab simulations, demonstrates a significant reduction in attack vectors, especially credential theft and unauthorized privilege escalation. Such results are particularly relevant for ZTFM systems, where FMs operate across heterogeneous IoT environments. Enforcing least privilege helps ensure that FM-based APIs, inference pipelines, or model updates are not universally exposed, but are accessible only through verified,

TABLE IV
CORE PRINCIPLES OF ZTFM IN IOT AND THEIR TARGETED ATTACK MITIGATIONS

Principle	Targeted Attack Type(s)	Advantages	Challenges in IoT
LPA	Lateral movement, privilege escalation, backdoor persistence	• Minimizes attack surface, reducing security risks [97]. • Prevents lateral movement in compromised networks [98]. • Enables fine-grained, context-aware access control [99].	• **Complex policy enforcement:** High heterogeneity in IoT makes uniform policy enforcement challenging [98]. • **Scalability issues:** Large-scale IoT networks require dynamic access control updates [99]. • **Lack of standardization:** Many LPA mechanisms are designed for IT environments rather than IoT [97].
Continuous Verification	Identity spoofing, session hijacking, adversarial inference, credential theft	• Enables real-time authentication and dynamic revocation based on contextual trust [100]. • Adapts security policies based on risk assessments [101]. • Reduces credential-based attacks with AI-driven anomaly detection [102].	• **High computational and storage costs:** Continuous verification processes large amounts of IoT data [102]. • **Latency concerns:** Real-time verification may slow down time-sensitive IoT applications [101]. • **Resource limitations:** Persistent monitoring burdens constrained devices [100]. • **Policy fragmentation:** Aligning context-aware trust policies remains complex [100].
Data Confidentiality & Integrity	Membership inference, model extraction, data tampering, poisoning attacks	• End-to-end encryption protects sensitive IoT data [103]. • Privacy-preserving computations using HE and SMPC [104]. • Supports compliance in regulated sectors [105].	• **Large data volume:** Encryption at scale imposes latency and compute costs [104]. • **Key management complexity:** Hard to secure key lifecycle in dynamic IoT settings [103]. • **Untrusted sources:** Verifying integrity in adversarial environments is resource-intensive [105].
Behavioral Analytics	Insider threats, anomaly-based evasion, adversarial IDS bypass	• Enhances real-time threat detection through AI-driven anomaly detection [106]. • Dynamically adjusts access policies based on behavior [107]. • Detects behavioral drift and compromised endpoints [108].	• **False positives:** Misclassifications may trigger unnecessary restrictions [106]. • **Computational overhead:** AI models may be infeasible on low-power devices [107]. • **Privacy risks:** Behavioral tracking may violate regulatory norms [108].

scoped, and context-aware requests – reducing the FM attack surface without sacrificing scalability.

Uttecht et al. [109] provided a comprehensive zero-trust reference model tailored to U.S. federal government networks. Their architecture outlines how LPA policies, when combined with endpoint verification and policy enforcement points, significantly restrict adversary movement in high-value environments. They emphasize the need for federated identity systems and network segmentation to enforce least-privilege without sacrificing usability. Chinamanagonda et al. [110] analyze LPA enforcement in cloud-native architectures. They present a framework integrating identity federation, context-aware access, and JIT provisioning, demonstrating how microservices and container orchestration tools like Kubernetes can implement LPA policies dynamically. Their findings are particularly relevant for cloud-hosted IoT platforms that require real-time scalability and access control granularity.

Azad et al. [98] conducted a comprehensive survey on the implementation of zero-trust Architecture within IoT environments, focusing on how foundational principles such as LPA can be operationalized under constrained computing, energy, and communication settings. The authors identify key issues in IoT deployments – including device heterogeneity, dynamic network topology, and the absence of centralized control – which collectively complicate the enforcement of static access control policies. To address this, they proposed a lightweight architectural framework composed of decentralized policy agents and hierarchical trust zones. These agents operate at the edge and perform context-aware access evaluations based on device profiles, operational roles, and current security posture. Their approach includes adaptive access control mechanisms and energy-aware decision heuristics, ensuring that access decisions can be made with minimal computational overhead. Their insights are particularly aligned with ZTFM goals, supporting distributed model training and FL coordination in security-critical, low-power environments.

While LPA limits exposure by controlling who can access what, it must be complemented by mechanisms that ensure entities are continuously verified post-authentication, especially in dynamic IoT contexts.

B. Continuous Verification

Continuous Verification is a core principle of ZTFMs, requiring that no user, device, or system component be inherently trusted. Every interaction is subject to ongoing validation, behavioral monitoring, and real-time re-authentication. This requirement is especially critical in IoT environments, where dynamic connectivity, identity spoofing, and untrusted endpoints are prevalent.

To enable adaptive and fine-grained verification, Dimitrakos et al. [100] proposed a trust-aware continuous authorization model for smart home IoT systems. This work extends traditional Attribute-Based Access Control (ABAC) models by embedding a Trust-Level Evaluation Engine (TLEE) directly into the policy lifecycle. Unlike static policies, the TLEE allows access control decisions to adapt in real time based on mutable attributes such as environmental conditions or user behavior. This is particularly beneficial for ZTFMs that require per-session policy enforcement. Their key insight lies in combining contextual trust evaluation with modular policy re-evaluation, ensuring that authorizations remain valid only as long as trust conditions are met. The prototype achieves sub-10ms re-evaluation latency, showcasing its practicality for constrained IoT devices. However, the system's dependency on locally maintained context data may limit scalability across federated environments.

To address scalability in multi-organizational IoT systems, Joshi et al. [101] introduced a graph-based framework for trust propagation. Unlike rule-based verification, their model constructs a dynamic interaction graph, where user-device-service relationships are encoded as context-weighted edges. This approach supports transitive trust computation and detects anomalies through topological drift in the graph structure. A major strength of this framework is its ability to reason about emergent risk across federated domains, enabling fine-grained access decisions in complex settings such as smart cities and industrial automation. However, its reliance on graph maintenance may pose overhead for highly dynamic or low-bandwidth environments.

While the above systems focus on behavioral trust evolution, Adhikari et al. [111] addressed identity privacy during verification, a critical concern in sensitive IoT domains such as healthcare. They proposed a ZKP-based federated identity protocol, which enables users to prove authorization without disclosing personal identifiers. This method improves over traditional OAuth/OpenID flows by eliminating central identity providers and mitigating metadata leakage risks. It supports compliance-driven IoT use cases governed by HIPAA or GDPR. Their performance analysis shows minimal verification overhead, but integrating ZKP at scale may require hardware-assisted acceleration or simplified cryptographic primitives.

In high-mobility and adversarial scenarios, such as UAV delivery networks, Dong et al. [102] developed a comprehensive continuous verification stack. Their system combines biometric-based MFA, continuous behavioral profiling, and blockchain-backed audit trails to ensure persistent identity assurance. A distinguishing feature is the use of immutable blockchain logs to resist rollback or spoofing attempts, making it ideal for mission-critical deployments. The design also incorporates adaptive access control, where deviations in biometric or behavioral profiles trigger automatic revocation. This work demonstrates how ZTFMs can integrate multi-modal identity streams to maintain verification under adversarial conditions. The trade-off lies in the computational burden of blockchain consensus and biometric matching, which may affect real-time responsiveness unless optimized.

From context-adaptive ABAC and graph-based transitive trust to privacy-preserving authentication and tamper-proof verification chains, these latest studies reflect a shift from static, perimeter-based security to intelligent, context-aware, and decentralized trust enforcement.

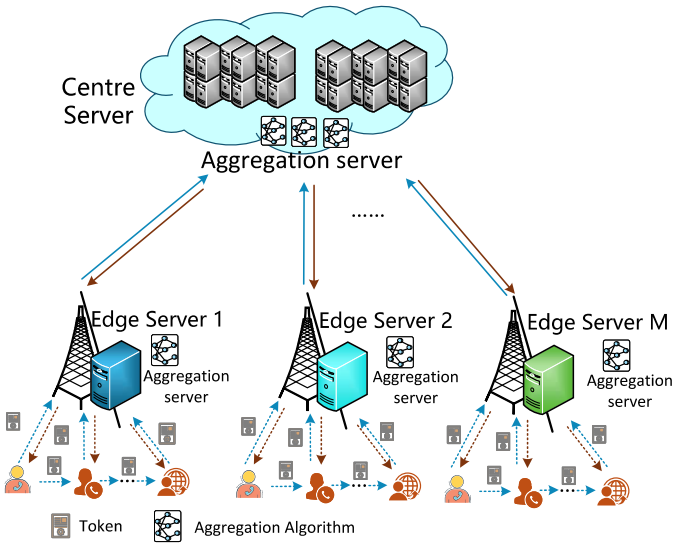

Fig. 7. A representative architecture of privacy-preserving FL in edge computing [103]. This hierarchical structure incorporates local aggregation at edge servers and global coordination by a central server, supporting secure multiparty computation and token-based update verification among distributed IoT users.

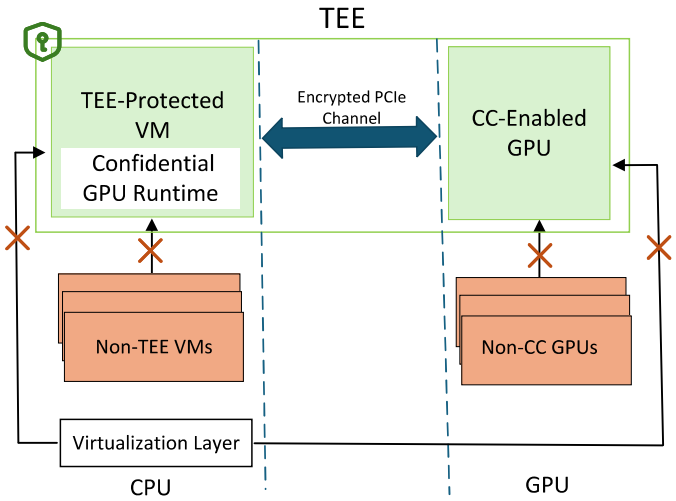

Fig. 8. An illustration of a CPU–GPU co-architecture for confidential computing using a TEE, as developed in [104]. Encrypted PCIe channels and enclave-based isolation prevent unauthorized access from legacy VMs or GPUs, highlighting the role of secure virtualization and hardware-assisted attestation in zero-trust computing environments.

C. Data Confidentiality and Integrity

Preserving the confidentiality and integrity of data is critical in ZTFM systems, particularly when AI models are collaboratively trained across distributed IoT and edge devices. In ZTFM, this principle is realized through secure computation mechanisms such as homomorphic encryption, SMPC, and TEEs, each offering unique trade-offs in privacy, performance, and scalability.

Choi et al. [103] introduced a scalable SMPC framework tailored for resource-constrained IoT nodes engaged in FL. Their approach partitions training workloads among nodes, ensuring that model gradients remain encrypted and undisclosed during aggregation. They demonstrate resilience against inference attacks even under non-IID data conditions – a common scenario in IoT – and evaluate latency-performance trade-offs using real-world health and logistics datasets. Their findings affirm that privacy-preserving collaboration is feasible even under energy and computational limitations. Zhang et al. [105] designed a zero-trust security architecture for smart grid telemetry that integrates endpoint authentication, microsegmentation, and TLS-layer encryption. Their approach not only safeguards the real-time integrity of high-volume telemetry streams but also dynamically evaluates device behavior to adjust trust levels on the fly. This isolation-first model prevents lateral movement and enables fault containment, which is critical in critical infrastructure domains like energy, transportation, and smart manufacturing.

Mohan et al. [104] explored the feasibility of running large-scale FMs, e.g., BERT and LLaMA, within confidential computing environments using Intel TDX and NVIDIA Hopper. They benchmarked batch inference workloads and demonstrated optimization techniques that reduce the performance penalty introduced by TEEs. Their evaluation shows that while secure inference incurs overhead, batching and parallelism can enable scalable and trustworthy deployment of ZTFMs. Li et al. [112] presented the design and formal verification of ARM's confidential compute architecture, which introduces Realms as a trusted execution abstraction. Realms isolate sensitive ZTFM processes from untrusted system software, ensuring data confidentiality even under system-level compromise. Their architecture supports remote attestation, runtime memory encryption, and formal proof of security properties, making it suitable for regulated environments where verifiability and trust assurance are essential.

Across these latest studies, ZTFMs have been designed to support collaborative AI computation while preserving strict confidentiality guarantees. From encrypted model aggregation and secure inference to hardware-enforced isolation, each approach complements zero-trust principles by minimizing data exposure and enabling verifiable trust across dynamic and heterogeneous IoT environments.

D. Behavioral Analytics

Behavioral Analytics introduces adaptive risk modeling by continuously observing user activity, device state, and environmental context to detect security anomalies and enforce dynamic policy adjustments. In IoT environments – characterized by frequent device churn, impersonation risks, and context-switching – this approach is especially important as identity-based controls alone fail to account for behavioral variation.

To operationalize behavior as a dynamic trust signal, Garcia et al. [106] proposed SADAC, a Security Attribute-based Dynamic Access Control system. SADAC leverages multivariate statistical process control (MSPC) and behavior-based profiling to adjust access privileges in real time. A key design insight is its modular architecture, enabling seamless integration

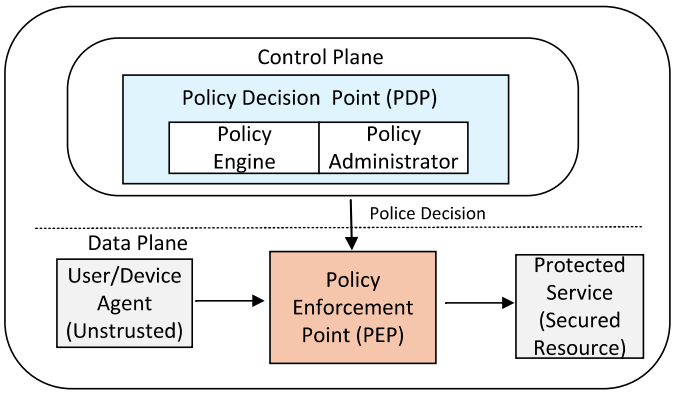

Fig. 9. The core structure of the zero-trust architecture proposed in [107], which delineates the control plane, where the Policy Decision Point (PDP) governs access logic, from the data plane, where the Policy Enforcement Point (PEP) mediates all subject–resource interactions. The separation ensures scalable, policy-driven trust enforcement across dynamic environments.

with enterprise-grade IoT systems. Simulation results show a false positive rate below 3.2% in insider threat detection scenarios, while maintaining detection latency within 50ms, highlighting both efficiency and robustness for time-sensitive applications.

Moving beyond statistical modeling, Wang et al. [107] introduced a deep learning-based approach, applying Long Short-Term Memory (LSTM) networks to learn temporal access patterns from session metadata, including login intervals, device mobility, and contextual tags. Their evaluation across enterprise and university networks demonstrates a 12–18% improvement in detection accuracy over baseline heuristics, and precision scores exceeding 90% in anomaly prediction. The model's strength lies in its ability to capture long-range dependencies in user behavior, making it suitable for detecting stealthy deviations.

To enable zero-trust enforcement in dynamic environments like healthcare and smart cities, Ameer et al. [108] proposed a real-time trust scoring framework. The system synthesizes telemetry signals such as device posture, app usage, and geolocation variance, recalculating trust scores every 5–10 seconds. This enables context-sensitive access revocation without relying on static roles or credentials. Their prototype achieved real-time decision-making under 20 ms and demonstrated resilience against impersonation attacks in simulation environments with frequent device switching.

Complementing centralized inference, Kumar et al. [99] adopt a decentralized reputation model by integrating endpoint detection and response (EDR) data with the EigenTrust algorithm. Trust values evolve based on compliance history, audit logs, and peer device evaluations. In industrial IoT testbeds, their system demonstrated a 40% reduction in incident response time and enabled trust-driven isolation of malicious nodes within 1.5 seconds of anomalous activity detection. This reputation-driven strategy is particularly suited to distributed settings where centralized policy enforcement is infeasible.

In summary, behavioral analytics equips ZTFM with the capacity to perceive, adapt, and respond, not just based on credentials, but on continuous observation and inference. It

TABLE V
MAPPING TECHNICAL COMPONENTS TO CORE ZTFM PRINCIPLES IN IOT

Technical Component	LPA	Continuous Verification	Data Confidentiality & Integrity	Behavioral Analytics
Authentication & Authorization	JIT access control and identity-based RBAC enforce minimal access scope [97], [110]	Trust-aware continuous authorization and federated identity mechanisms support contextual verification over time [100], [111]	Authentication ensures only authorized nodes access encrypted data streams [111]	Contextual identity and access logs serve as input for dynamic behavior scoring [113]
Secure Aggregation	Limits what each node contributes; prevents excessive exposure of internal models [114]	Secure model update paths enable verification without revealing raw data [115]	SMPC and DP protect model updates and gradients during FL [103]	Anomaly-aware aggregation reduces the impact of malicious updates [115]
Anomaly Detection	Detects policy violations in access logs and restricts permissions dynamically [106]	Continuously assesses behavior of devices/users to verify consistency [107]	Flags suspicious access to encrypted or sensitive data, enhancing data integrity [116]	Builds user/device profiles to identify drift, insider threats, or outliers in behavior [35]
Blockchain	Smart contracts enforce decentralized, fine-grained access policies in dynamic networks [117]	Immutable logs support auditability of verification and access decisions [102]	Distributed ledgers protect model sharing and prevent data tampering [118]	Blockchain anchors behavior logs and device reputation scores [119]
Trusted Execution Environments (TEEs)	Securely enforces policy checks within isolated execution zones [112]	Supports attestation and runtime validation of model integrity and updates [120]	Executes encrypted models and data securely, protecting confidentiality [104]	Ensures that analytics models are protected from tampering while processing sensitive behavior data [121]
Encryption & Secure Communication	Encrypts channel-specific access tokens to restrict user scope [122]	TLS and mTLS encrypt sessions to support session-level continuous verification [123]	Ensures end-to-end encrypted data transmission with PQC and EDAP [124]	Encrypted traffic metadata may be analyzed to detect behavioral anomalies without revealing content [125]

TABLE VI
TECHNICAL COMPONENTS OF ZTFM IN IOT

Technology	Advantanges	Challenges in IoT
Authentication & Authorization [113], [114], [117], [126], [127]	Robust security via multi-factor, blockchain, AI-driven verification; dynamic continuous validation	Increased complexity; computational overhead; possible latency
Secure Aggregation [114], [115], [128]–[130]	Data confidentiality; protection against malicious updates; decentralized aggregation enhances resilience	Communication overhead; potential accuracy loss due to privacy mechanisms; computational complexity
Anomaly Detection [35], [115], [116], [128], [131]	Early threat identification; real-time response; detection of adversarial activities	False positives; resource-intensive; requires continuous monitoring
Blockchain [114], [119], [132]–[135]	Immutable auditability; transparency; decentralized trust and security management	High energy and computational overhead; scalability limitations; latency in transaction validation
Trusted Execution Environments (TEEs) [120], [121], [136]	Strong isolation of sensitive computations; protection against tampering and leakage	Hardware dependency; potential performance overhead; complexity in deployment and maintenance
Encryption & Secure Communication [122]–[124]	Ensures data confidentiality; protects against interception; strong cryptographic security	Latency and overhead due to encryption processing; complexity in key management; resource-intensive

lays the foundation for proactive threat mitigation in fast-changing IoT ecosystems, bridging context, computation, and trust in a unified zero-trust pipeline.

V. TECHNICAL COMPONENTS OF ZTFMS

ZTFM is underpinned by a suite of integrated technical components that work together to enforce strict security postures across distributed systems. These components are designed to ensure that trust is never assumed, and every interaction is continuously verified and evaluated. From the moment entities request access, through the secure handling and aggregation of data, to the detection of anomalous behaviors and the preservation of data integrity, each element contributes to a robust, end-to-end trust framework. This section provides an in-depth look at the core technical elements of ZTFM and their roles in a resilient zero-trust architecture.

A. Authentication and Authorization

Robust identity verification is critical for securing FL in IoT, ensuring that only trusted participants join model training [137]. In the context of ZTFM, this prevents unauthorized devices from introducing poisoned updates or launching inference attacks on FMs. Mechanisms, such as multi-factor authentication, digital certificates, and identity management

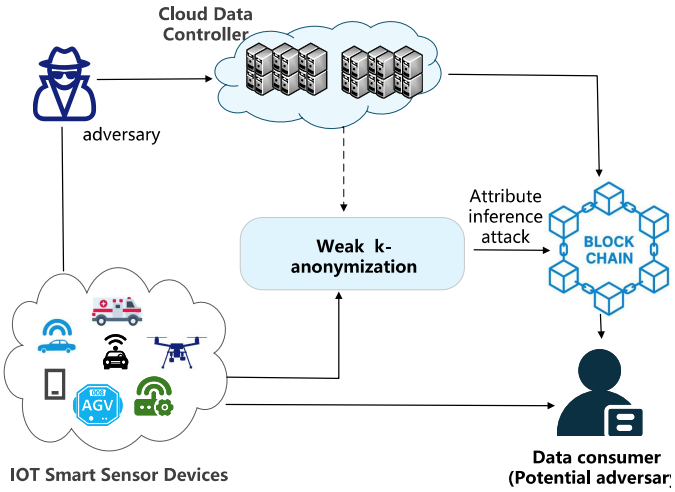

Fig. 10. Overview of a potential privacy breach in blockchain-based IoT data sharing [114]. This attacker model illustrates the end-to-end flow of data from IoT devices to adversarial consumers, emphasizing the risks of using weak anonymization and highlighting the need for stronger privacy-preserving mechanisms in Zero-Trust IoT ecosystems.

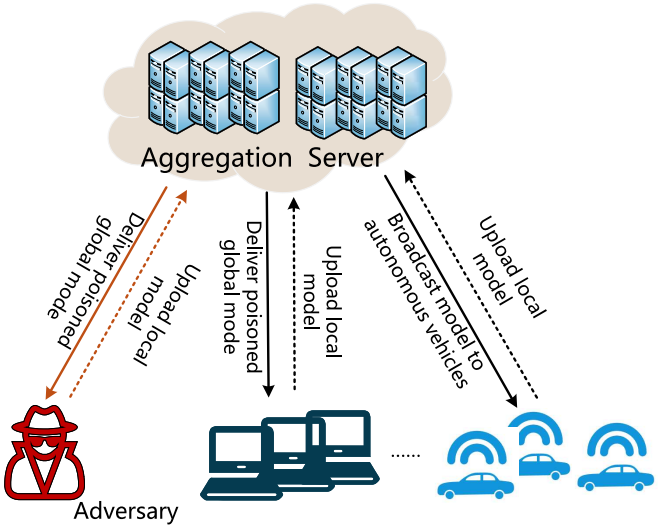

Fig. 11. Workflow of an FL poisoning attack presented in [114]. The adversary manipulates the local training process to upload poisoned models. The aggregation server fails to detect the malicious updates and integrates them into the global model, which is then disseminated to all connected devices, compromising system integrity.

systems [114], can help secure participation, while integrating local differential privacy, FL, and blockchain ensures scalable, tamper-resistant verification across IoT environments. The AIDL-XTS model developed in [113] demonstrates how AI models (e.g., CNN-BiLSTM) can profile user and device behavior for continuous trust scoring, which aligns with ZTFM's need for adaptive, real-time verification. Moreover, proxy smart contracts [117] validate transactions before finalization, and can offer a blueprint for securing model updates and access decisions in distributed FMs.

FL-based dynamic access control frameworks [126] and continuous verification engines [127], [138] showcase how zero-trust principles like behavior analysis, micro-segmentation, and contextual access can enhance ZTFM resilience in IoT systems. Continuous authentication and access control have been shown to improve cyber resilience in large-scale IoT networks [119], reinforcing dynamic trust management for FMs in adversarial, resource-constrained environments. These studies highlight that existing zero-trust solutions, though not explicitly designed for FMs, offer mechanisms that ZTFM extends for securing FM-based IoT systems.

B. Secure Aggregation

Aggregation protocols ensure confidentiality by securely combining model updates from participants without exposing sensitive data, employing differential privacy to prevent reconstruction attacks while maintaining model accuracy [114]. Blockchain-based FL methods proposed in [115] counteract malicious client updates, reinforcing global learning security. Blockchain integrated with dynamic zero-trust FL [128] enhances data privacy and security within industrial IoT environments. Moreover, privacy-preserving aggregation protocols coupled with main-side blockchain architectures further secure consumer IoT data [129].

A framework for zero-trust verification of industrial IoT (IIoT) wireless transmission nodes was developed in [128], which utilizes FL to achieve zero-trust rule training and terminal model training, while employing blockchain technology for on-chain aggregation and cloud backup of the models. This approach enhances the accuracy and availability of the zero-trust rules while safeguarding the security of IIoT nodes.

C. Anomaly Detection

Anomaly detection for FMs within a zero-trust architecture differs fundamentally from traditional anomaly detection in IoT systems. Traditional IoT anomaly detection identifies irregular patterns in sensor data, network traffic, or device behavior using predefined rules or lightweight models, often limited to narrow contexts and static trust assumptions. In contrast, anomaly detection for FMs in zero-trust environments operates at multiple abstraction levels, detecting not only data-level anomalies but also adversarial manipulations, distribution shifts, unauthorized model access, or malicious behavior embedded in complex model interactions. This requires continuous verification of both data and model behavior, leveraging behavioral analytics, provenance tracking, and trust scoring. Moreover, zero-trust detection has to account for the higher adaptability of FMs, ensuring model integrity and confidentiality even under stealthy, sophisticated threats that go beyond outliers or threshold violations targeted in legacy IoT anomaly systems.

Advanced algorithms detect and isolate anomalies such as adversarial updates and unexpected communication patterns, safeguarding collaborative learning integrity [139]. The DP-RFECV-FNN framework [116] leverages differential privacy and deep learning to classify and prevent unauthorized Android malware in IoT networks. Continuous monitoring combined with AI-driven dynamic trust algorithms ensures real-time risk evaluation and access control in 5G/6G networks [131]. Additionally, ML techniques proposed in [128]

detect anomalies in industrial IoT data streams, effectively identifying internal and external threats. CNN and BiLSTM integrated frameworks further enhance anomaly detection by capturing spatial-temporal patterns in evolving cyber threats [35]. Pokhrel et al. [115] introduced a robust zero-trust architecture integrating blockchain and FL, enhancing anomaly detection and securing decentralized IoT networks. A privacy-preserving AI-driven malware detection framework was proposed in [116] for IoT-based medical devices running on Android. Integrating differential privacy and zero-trust security ensures secure, decentralized malware detection, safeguarding sensitive patient data and healthcare network integrity while maintaining high accuracy.

A zero-trust framework for smart grid infrastructures was proposed in [140], which integrates IT and OT security mechanisms to enhance monitoring and defense against sophisticated cyber threats, such as ransomware. By leveraging EigenGame for data integration and quantum reinforcement learning for malicious behavior detection, the framework strengthens cybersecurity in IIoT-enabled smart grids, ensuring reliable system protection and threat mitigation.

D. Blockchain

Blockchains provide immutable transaction records, enhancing transparency and auditability in collaborative AI model training [114], [141], [142]. For instance, Kim et al. [132] demonstrated blockchain applications in securing FL, ensuring transparency in collaborative AI updates. Blockchain-based protocols proposed by Sullivan et al. [133] securely endorse real-time vehicle trajectory data. Blockchain integration within zero-trust architecture improves transparency, security, and access control for scientific peer review and data sharing [134], while Jain et al. [135] highlighted blockchain's role in securing healthcare data alongside AI-driven threat detection. Liu et al. [119] offered a comprehensive bibliometric analysis, identifying significant blockchain-based trends in zero-trust IoT security research, emphasizing its effectiveness against heterogeneous device environments.

A blockchain and smart contract-based edge-IoT framework was proposed in [143], which enforces zero-trust security by managing IoT device behavior through a credit-based resource allocation system, ensuring secure access control, automated policy enforcement, and scalable security in decentralized IoT networks. A blockchain-based middleware for management in IoT was presented in [144], which uses a novel zero-trust hierarchical mining process that allows validating the infrastructure and transactions at different levels of trust.

E. Trusted Execution Environment

Hardware-based TEEs protect sensitive computations and model parameters from tampering or leakage [145]. The applicability of TEEs in decentralized AI systems is explored extensively, highlighting secure computational capabilities [121], [136]. For example, Vomvas et al. [120] proposed a vertical extension termed zero-trust execution for beyond-5G networks, using TEEs to secure execution environments and establish trust in untrusted contexts. In addition, Aiello et al. [130] examined the secure access service edge framework, integrating SD-WAN, ZTNA, SWG, and CASB, emphasizing identity-driven security, micro-segmentation, and real-time threat intelligence to address network performance and data protection challenges.

F. Encryption and Secure Communication:

End-to-end encryption protocols, such as TLS, safeguard data confidentiality and integrity, mitigating interception risks during data transmission [146]. Gharib et al. [122] introduced SCC5G, a Post-Quantum Cryptography (PQC)-based architecture ensuring encrypted and authenticated 5G mission-critical communications, utilizing CRYSTALS-Kyber and CRYSTALS-Dilithium cryptographic schemes. Tseng et al. [124] proposed Encrypted Data Processing (EDAP), which employs processor-level encryption to secure data during execution, eliminating implicit trust in cloud platforms and hypervisors. Additionally, Rodigari et al. [123] assessed mutual TLS (mTLS) in zero-trust networks, confirming its security effectiveness despite moderate computational overheads in multi-cloud deployments. A zero-trust architecture optimized for industrial environments was proposed in [147], integrating micro-segmentation and software-defined networking to enhance security in power grids, transportation systems, and industrial control systems. By enabling dynamic network management, granular access control, and breach containment, the framework strengthens cybersecurity in highly heterogeneous and interconnected industrial networks.

G. Lessons Learned

Most current solutions have addressed some isolated aspects of zero trust or FMs, and often overlooked the dual role of FMs as both targets and enablers of trust enforcement. ZTFM can potentially bridge this gap by positioning FMs as active agents within the trust pipeline, enabling real-time threat detection, dynamic access control, and adaptive policy enforcement based on contextual understanding.

Effective ZTFM implementations necessitate the holistic integration of multiple technologies, e.g., authentication, anomaly detection, blockchain, TEEs, and encryption, into a unified trust framework. While mechanisms like multi-factor authentication, continuous verification, and differential privacy strengthen security, they introduce computational overheads. Designing lightweight and adaptive solutions remains an ongoing challenge in resource-constrained IoT environments.

Last but not least, static access controls are insufficient in dynamic IoT systems. Techniques, such as behavioral analytics and contextual policy adaptation, are crucial for maintaining continuous trust in the presence of evolving threats. While many works focus on standard security goals, robust defenses against adversarial attacks targeting FMs, such as model poisoning, inference leakage, and stealthy backdoors, are still in their early stages. There is an urgent need for ZTFM-specific adversarial defense mechanisms.

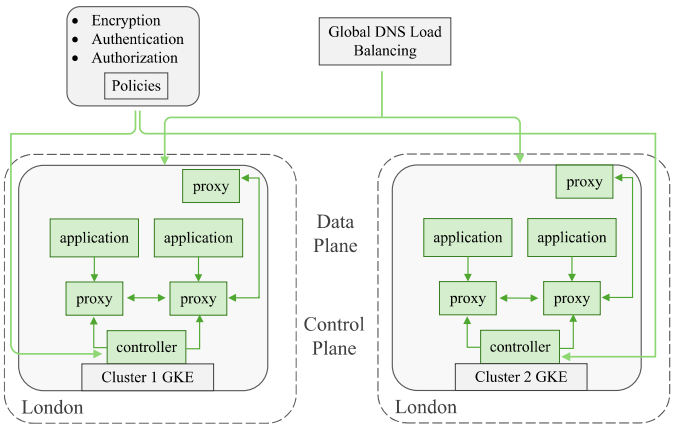

Fig. 12. Illustration of the zero trust multi-cloud architecture with distributed policy enforcement via proxies developed in [123], enabling secure communication between services across GKE and EKS clusters.

VI. Open Research Challenges for ZTFM in IoT

Despite the progress in ZTFM, several critical research challenges remain unresolved when applying these models to IoT due to the highly distributed and heterogeneous nature of IoT devices, severe resource constraints, scalable and decentralized network conditions, diverse sensitivity and privacy levels of IoT data, and the necessity of lightweight yet robust continuous verification and adaptive security mechanisms across decentralized deployments. Based on our study, we identify the following open challenges that demand further investigation:

1) Lightweight Cryptographic Primitives for Trust-Aware AI: Cryptographic methods such as homomorphic encryption, zero-knowledge proofs, and SMPC are core enablers of ZTFM but are resource-intensive for IoT devices. A key challenge lies in developing lightweight cryptographic protocols [148] that maintain rigorous security guarantees while being computationally feasible for constrained edge nodes. Emerging directions include post-quantum cryptography [149] tailored for FL and efficient lattice-based encryption [150] adapted to non-IID data distributions in IoT.

2) Scalable and Adaptive Trust Reasoning: As ZTFM systems scale across diverse IoT networks, trust scoring must evolve to account for context, temporal changes, and inter-node variability. One challenge is designing hierarchical, decentralized trust models that incorporate behavioral analytics, reputation systems, and federated signals without incurring excessive synchronization overhead. Graph-based trust propagation [101] and dynamic Bayesian belief updating [151] are promising but underexplored directions.

3) Cross-Domain Interoperability and Policy Federation: ZTFM deployments across healthcare, manufacturing, and transportation sectors face challenges due to heterogeneity in policies, data formats [152], and access requirements [153]. A proposed research direction is to design extensible policy languages and cross-domain security ontologies that can support secure interoperation across trust domains. Additionally, trust federation protocols that preserve local autonomy while enabling global policy compliance are essential.

4) Threat-Resilient Federated Training Architectures: ZTFM must account for adversarial adaptation in collaborative learning pipelines [154]. A key challenge is integrating robust FMs protocols with zero-trust guarantees, capable of defending against model poisoning, sybil attacks, and gradient leakage [155]. Future research should explore adversarial training integrated with zero-trust scoring, secure aggregation via threshold cryptography, and active defense using anomaly-triggered retraining.

5) Fine-Grained Resource-Aware Security Orchestration: IoT environments present inherent constraints in bandwidth, memory, and compute [156], [157]. A pressing challenge is orchestrating ZTFM security enforcement, such as micro-segmentation, continuous verification, and behavioral scoring, in a resource-adaptive manner. This includes dynamic policy offloading, opportunistic security tasks scheduling, and energy-aware trust checkpoints to optimize for security-utility trade-offs [158].

6) Auditability and Explainability in Zero-Trust Decisions: As ZTFM decisions govern sensitive access control and collaboration workflows, the lack of transparent decision paths limits trust and compliance [159]. We propose integrating explainable AI techniques into ZTFM enforcement modules, enabling audit trails, user-centric justification of denial/approval, and provenance tracking of policy adaptation in distributed environments [160].

Summary. Collectively, these open challenges highlight the need for multidisciplinary solutions that combine security, cryptography, systems design, and AI to realize scalable, interpretable, and efficient ZTFM deployment in real-world IoT ecosystems.

VII. Conclusion

This paper establishes ZTFMs as a transformative approach to securing AI-driven IoT systems. By embedding zero-trust principles, such as LPA, continuous verification, data confidentiality, and behavioral analytics, into the training and deployment of FMs, ZTFMs provide a principled framework for addressing the unique security and trust challenges of decentralized, heterogeneous IoT environments. The structured synthesis of ZTFMs was presented with formalized foundational principles. A unified technical architecture was proposed, which integrates FL, blockchain-based identity management, micro-segmentation, and TEEs. Our analysis of emerging threats and corresponding defense strategies revealed key limitations in current practices and uncovered several open research directions, including scalable secure orchestration, lightweight multiparty computation, interpretable threat attribution, and AI-driven trust calibration.

References

[1] E. Hopkins and A. Siekelova, "Internet of Things sensing networks, smart manufacturing big data, and digitized mass production in sustainable Industry 4.0." *Economics, Management & Financial Markets*, vol. 16, no. 4, 2021.

[2] S. Hu, X. Chen, W. Ni, E. Hossain, and X. Wang, "Distributed machine learning for wireless communication networks: Techniques, architectures, and applications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1458–1493, 2021.

[3] P. Trakadas, P. Simoens, P. Gkonis *et al.*, "An artificial intelligence-based collaboration approach in industrial iot manufacturing: Key concepts, architectural extensions and potential applications," *Sensors*, vol. 20, no. 19, p. 5480, 2020.

[4] P. Singh and N. Singh, "Blockchain with IoT and AI: A review of agriculture and healthcare," *Research Anthology on Convergence of Blockchain, Internet of Things, and Security*, pp. 1315–1330, 2023.

[5] M. S. Munir, S. F. Abedin, K. T. Kim, D. H. Kim, M. G. R. Alam, and C. S. Hong, "Drive safe: Cognitive-behavioral mining for intelligent transportation cyber-physical system," *arXiv preprint arXiv:2008.10148*, 2020.

[6] K. Reena and V. Venkatesh, "Intelligent decision support system for home automation–ANFIS based approach," *Int. J. Eng. Technol.(UAE)*, vol. 7, pp. 421–427, 2018.

[7] I. I. Consortium, "Industry white papers of industry IoT consortium." [Online]. Available: https://www.iiconsortium.org/white-papers/industry/

[8] K. GmbH, "Accelerating railway digitalization - the startup sector's perspective." [Online]. Available: https://resources.konux.com/accelerating-railway-digitalization-the-startup-sectors-perspective

[9] J. P. Tuohy, "LG's new smart home hub has a built-in voice assistant." [Online]. Available: https://www.theverge.com/2024/8/28/24230692/lg-thinq-on-smart-home-hub-ai-voice-assistant

[10] T. Yang, L. Chang, J. Yan, J. Li, Z. Wang, and K. Zhang, "A survey on foundation-model-based industrial defect detection," *arXiv preprint arXiv:2502.19106*, 2025.

[11] T. B. Brown, B. Mann, N. Ryder *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS 2020*, 2020.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[13] DeepSeek-AI, D. Guo, D. Yang, and H. Z. .etc, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *CoRR*, vol. abs/2501.12948, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2501.12948

[14] DeepSeek-AI, A. Liu, B. Feng, and B. X. .etc, "DeepSeek-V3 technical report," *CoRR*, vol. abs/2412.19437, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.19437

[15] G. O. Boateng, H. Sami *et al.*, "A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions," *IEEE Communications Surveys & Tutorials*, 2025.

[16] J. Mattern, F. Mireshghallah, Z. Jin *et al.*, "Membership inference attacks against language models via neighbourhood comparison," *arXiv preprint arXiv:2305.18462*, 2023.

[17] X. Li, S. Wang, C. Wu, H. Zhou, and J. Wang, "Backdoor threats from compromised foundation models to federated learning," *arXiv preprint arXiv:2311.00144*, 2023.

[18] S. Yu, J. P. Muñoz, and A. Jannesari, "Federated foundation models: Privacy-preserving and collaborative learning for large models," *arXiv preprint arXiv:2305.11414*, 2023.

[19] H. Egerton, M. Hammoudeh, D. Unal, and B. Adebisi, "Applying zero trust security principles to defence mechanisms against data exfiltration attacks," *Security and Privacy in the Internet of Things: Architectures, Techniques, and Applications*, pp. 57–89, 2021.

[20] I. Makhdoom, M. Abolhasan, J. Lipman, R. P. Liu, and W. Ni, "Anatomy of threats to the Internet of Things," *IEEE communications surveys & tutorials*, vol. 21, no. 2, pp. 1636–1675, 2018.

[21] J. Kindervag *et al.*, "Build security into your network's dna: The zero trust network architecture," *Forrester Research Inc*, vol. 27, pp. 1–16, 2010.

[22] R. Bommasani, D. A. Hudson, E. Adeli *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[23] R. Liu, Q. Zhang, T. Han, B. Yang, W. Zhang, S. Yin, and D. Zhou, "Survey on foundation models for prognostics and health management in industrial cyber-physical systems," *IEEE Transactions on Industrial Cyber-Physical Systems*, 2024.

[24] H. Liu, Y. Wang *et al.*, "Trustworthy AI: A computational perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, pp. 1–59, 2022.

[25] N. F. Syed, S. W. Shah, A. Shaghaghi, A. Anwar, Z. Baig, and R. Doss, "Zero trust architecture (ZTA): A comprehensive survey," *IEEE access*, vol. 10, pp. 57 143–57 179, 2022.

[26] F. Mensah, "Zero trust architecture: A comprehensive review of principles, implementation strategies, and future directions in enterprise cybersecurity," *International Journal of Academic and Industrial Research Innovations (IJAIRI)*, vol. 10, pp. 339–346, 2024.

[27] M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for AI-enabled IoT devices: A review," *Sensors*, vol. 20, no. 9, p. 2533, 2020.

[28] Y. Ashibani, D. Kauling, and Q. H. Mahmoud, "Design and implementation of a contextual-based continuous authentication framework for smart homes," *Applied System Innovation*, vol. 2, no. 1, p. 4, 2019.

[29] Z. Xu, B. Di, and L. Song, "Design of cloud-edge-gateway collaborative zero-trust architecture and workflow for smart factories," in *2024 IEEE International Workshop on Radio Frequency and Antenna Technologies (iWRF&AT)*. IEEE, 2024, pp. 335–339.

[30] L. Li, W. Wang, X. Zhou *et al.*, "A trustworthy IoT-based supply chain traceability system with semantic multi-chain and pre-blockchain data verification," *IEEE Internet of Things Journal*, pp. 1–1, 2025, early access.

[31] N. Alsuwaidi, N. Alharmoodi, and H. Al Hamadi, "The transformative impact of zero-trust architecture on healthcare security," in *2024 2nd International Conference on Cyber Resilience (ICCR)*. IEEE, 2024, pp. 1–8.

[32] B. Mao, Y. Liu, Z. Wei, H. Guo, Y. Xun, J. Wang, J. Liu, and N. Kato, "A blockchain-enabled cold start aggregation scheme for federated reinforcement learning-based task offloading in zero trust leo satellite networks," *IEEE Journal on Selected Areas in Communications*, pp. 1–11, 2025.

[33] R. Cheng, S. Chen, and B. Han, "Toward zero-trust security for the metaverse," *IEEE Communications Magazine*, vol. 62, no. 2, pp. 156–162, 2024.

[34] S. Hu, X. Yuan, W. Ni, X. Wang, E. Hossain, and H. Vincent Poor, "OFDMA-F²L: Federated learning with flexible aggregation over an OFDMA air interface," *IEEE Transactions on Wireless Communications*, vol. 23, no. 7, pp. 6793–6807, 2024.

[35] D. Javeed, M. S. Saeed, M. Adil, P. Kumar, and A. Jolfaei, "A federated learning-based zero trust intrusion detection system for Internet of Things," *Ad Hoc Networks*, vol. 162, p. 103540, 2024.

[36] X. Yuan, W. Ni, M. Ding, K. Wei, J. Li, and H. V. Poor, "Amplitude-varying perturbation for balancing privacy and utility in federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1884–1897, 2023.

[37] K. Wei, J. Li, C. Ma, M. Ding, W. Chen, J. Wu, M. Tao, and H. V. Poor, "Personalized federated learning with differential privacy and convergence guarantee," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4488–4503, 2023.

[38] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 7, pp. 8726–8746, 2024.

[39] M. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Secur. Priv.*, vol. 19, no. 2, pp. 20–28, 2021.

[40] D. Huang, Y. Na, Y. Liu, Z. Zhang, and B. Mi, "Overview of cooperative fault-tolerant control driven by the full information chain of intelligent connected vehicle platoons under the zero-trust framework: Opportunities and challenges," *IEEE Intelligent Transportation Systems Magazine*, vol. 16, no. 1, pp. 22–39, 2024.

[41] H. Joshi, "Emerging technologies driving zero trust maturity across industries," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 25–36, 2025.

[42] A. I. Weinberg and K. Cohen, "Zero trust implementation in the emerging technologies era: Survey," 2024. [Online]. Available: https://arxiv.org/abs/2401.09575

[43] B. Mao, J. Liu, Y. Wu, and N. Kato, "Security and privacy on 6g network edge: A survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1095–1127, 2023.

[44] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 20–30, 2024.

[45] S. Guthula, R. Beltiukov, N. Battula, W. Guo, A. Gupta, and I. Monga, "netFound: Foundation model for network security," 2025. [Online]. Available: https://arxiv.org/abs/2310.17025

[46] D. Xue, X. Fan, T. Chen, G. Lan, and Q. Song, "Leveraging foundation models for zero-shot IoT sensing," 2024. [Online]. Available: https://arxiv.org/abs/2407.19893

[47] O. Baris, Y. Chen, G. Dong, L. Han, T. Kimura, P. Quan, R. Wang, T. Wang, T. Abdelzaher, M. Bergés, P. P. Liang, and M. Srivastava, "Foundation models for CPS-IoT: Opportunities and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2501.16368

[48] J. Gu and S. Yeung-Levy, "Foundation models secretly understand neural network weights: Enhancing hypernetwork architectures with foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2503.00838

[49] B. Yuan, Y. He, J. Davis, T. Zhang, T. Dao, B. Chen, P. S. Liang, C. Ré, and C. Zhang, "Decentralized training of foundation models

in heterogeneous environments," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 25 464–25 477.

[50] S. Xu, C. Kurisummoottil Thomas, O. Hashash, N. Muralidhar, W. Saad, and N. Ramakrishnan, "Large multi-modal models (lmms) as universal foundation models for AI-native wireless systems," *IEEE Network*, vol. 38, no. 5, pp. 10–20, 2024.

[51] J. Zhou, Y. Chen, Z. Hong, W. Chen, Y. Yu, T. Zhang, H. Wang, C. Zhang, and Z. Zheng, "Training and serving system of foundation models: A comprehensive survey," *IEEE Open Journal of the Computer Society*, vol. 5, pp. 107–119, 2024.

[52] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[53] B. Shan, X. Yuan, W. Ni, X. Wang, R. P. Liu, and E. Dutkiewicz, "Preserving the privacy of latent information for graph-structured data," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5041–5055, 2023.

[54] N. Wu, X. Yuan, S. Wang, H. Hu, and M. Xue, "Cardinality counting in "Alcatraz": A privacy-aware federated learning approach," in *Proceedings of the ACM Web Conference 2024*, ser. WWW '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3076–3084. [Online]. Available: https://doi.org/10.1145/3589334.3645655

[55] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. Vincent Poor, "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245–2298, 2023.

[56] Z. Tian, L. Cui, J. Liang, and S. Yu, "A comprehensive survey on poisoning attacks and countermeasures in machine learning," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 166:1–166:35, 2023.

[57] Y. Wan, Y. Qu, W. Ni, Y. Xiang, L. Gao, and E. Hossain, "Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 3, pp. 1861–1897, 2024.

[58] F. Nuding and R. Mayer, "Data poisoning in sequential and parallel federated learning," in *IWSPA@CODASPY 2022: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*. ACM, 2022, pp. 24–34.

[59] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symposium on Security and Privacy, SP 2018*. IEEE Computer Society, 2018, pp. 19–35.

[60] K. Li, X. Yuan, J. Zheng, W. Ni, F. Dressler, and A. Jamalipour, "Leverage variational graph representation for model poisoning on federated learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 1, pp. 116–128, 2025.

[61] K. Li, J. Zheng, X. Yuan, W. Ni, Ö. B. Akan, and H. V. Poor, "Data-agnostic model poisoning against federated learning: A graph autoencoder approach," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 3465–3480, 2024.

[62] X. Cao and N. Z. Gong, "MPAF: model poisoning attacks to federated learning based on fake clients," in *Proc. CVPR Workshops 2022*. IEEE, 2022, pp. 3395–3403.

[63] J. Gao, B. Hou, X. Guo, Z. Liu, Y. Zhang, K. Chen, and J. Li, "Secure aggregation is insecure: Category inference attack on federated learning," *IEEE Trans. Dependable Secur. Comput.*, vol. 20, no. 1, pp. 147–160, 2023.

[64] Z. Wang, Y. Huang, M. Song, L. Wu, F. Xue, and K. Ren, "Poisoning-assisted property inference attack against federated learning," *IEEE Trans. Dependable Secur. Comput.*, vol. 20, no. 4, pp. 3328–3340, 2023.

[65] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, "Label inference attacks against vertical federated learning," in *Proc. USENIX Security 2022*. USENIX Association, 2022, pp. 1397–1414.

[66] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "{ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[67] L. Liu, Y. Wang, G. Liu, K. Peng, and C. Wang, "Membership inference attacks against machine learning models via prediction sensitivity," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 2341–2347, 2022.

[68] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 3093–3106.

[69] X. Luo, Y. Wu, X. Xiao, and B. C. Ooi, "Feature inference attack on model predictions in vertical federated learning," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 181–192.

[70] Y. Wu, H. Chen, X. Wang, C. Liu, P. Nguyen, and Y. Yesha, "Tolerating adversarial attacks and Byzantine faults in distributed machine learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 3380–3389.

[71] T. Distler, "Byzantine fault-tolerant state-machine replication from a systems perspective," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–38, 2021.

[72] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *Proc. USENIX Security 2020*. USENIX Association, 2020, pp. 1605–1622.

[73] J. Shi, W. Wan, S. Hu, J. Lu, and L. Y. Zhang, "Challenges and approaches for mitigating Byzantine attacks in federated learning," in *Proc. TrustCom 2022*. IEEE, 2022, pp. 139–146.

[74] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems 2017*, 2017, pp. 119–129.

[75] C. Dong, J. Weng, M. Li, J. Liu, Z. Liu, Y. Cheng, and S. Yu, "Privacy-preserving and Byzantine-robust federated learning," *IEEE Trans. Dependable Secur. Comput.*, vol. 21, no. 2, pp. 889–904, 2024.

[76] T. D. Nguyen, T. Nguyen, P. L. Nguyen, H. H. Pham, K. D. Doan, and K. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *Eng. Appl. Artif. Intell.*, vol. 127, no. Part A, p. 107166, 2024.

[77] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. S. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *Advances in Neural Information Processing Systems 2020*, 2020.

[78] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, and Q. Wang, "Coordinated backdoor attacks against federated learning with model-dependent triggers," *IEEE Netw.*, vol. 36, no. 1, pp. 84–90, 2022.

[79] P. Rieger, T. D. Nguyen, M. Miettinen, and A. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," in *Proc. NDSS 2022*. The Internet Society, 2022.

[80] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, and K. I. Wang, "Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9310–9319, 2022.

[81] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 1115–1124.

[82] X. Cao, Z. Zhang, J. Jia, and N. Z. Gong, "FLCert: Provably secure federated learning against poisoning attacks," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 3691–3705, 2022.

[83] G. Chen, K. Li, A. M. Abdelmoniem, and L. You, "Exploring representational similarity analysis to protect federated learning from data poisoning," in *Proc. WWW 2024*. ACM, 2024, pp. 525–528.

[84] X. Chen, H. Yu, X. Jia, and X. Yu, "APFed: Anti-poisoning attacks in privacy-preserving heterogeneous federated learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 5749–5761, 2023.

[85] J. Zheng, K. Li, X. Yuan, W. Ni, and E. Tovar, "Detecting poisoning attacks on federated learning using gradient-weighted class activation mapping," in *Proc. WWW 2024*. ACM, 2024, pp. 714–717.

[86] Z. Ma, J. Ma, Y. Miao, Y. Li, and R. H. Deng, "ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1639–1654, 2022.

[87] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in *Proc. SIGKDD 2022*. ACM, 2022, pp. 2545–2555.

[88] K. H. Chow and L. Liu, "Perception poisoning attacks in federated learning," in *Proc. TPS-ISA 2021*. IEEE, 2021, pp. 146–155.

[89] Y. Shi and Y. E. Sagduyu, "Membership inference attack and defense for wireless signal classifiers with deep learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 4032–4043, 2022.

[90] V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *Proc. NDSS 2021*. The Internet Society, 2021.

[91] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni,

F. Koushanfar, A. Sadeghi, and T. Schneider, "FLAME: taming backdoors in federated learning," in *Proc. USENIX Security 2022*. USENIX Association, 2022, pp. 1415–1432.

[92] C. Xie, M. Chen, P. Chen, and B. Li, "CRFL: certifiably robust federated learning against backdoor attacks," in *Proc. ICML 2021*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 11 372–11 382.

[93] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. NDSS 2021*. The Internet Society, 2021.

[94] W. Huang, G. Li, X. Yi, J. Li, C. Zhao, and Y. Yin, "SupRTE: Suppressing backdoor injection in federated learning via robust trust evaluation," *IEEE Intell. Syst.*, vol. 39, no. 5, pp. 66–77, 2024.

[95] I. Debicha, B. Cochez, T. Kenaza, T. Debatty, J.-M. Dricot, and W. Mees, "Adv-Bot: Realistic adversarial botnet attacks against network intrusion detection systems," *Computers & Security*, vol. 129, p. 103176, 2023.

[96] A. Venturi, M. Ferrari, M. Marchetti, and M. Colajanni, "ARGANIDS: a novel network intrusion detection system based on adversarially regularized graph autoencoder," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2023, pp. 1540–1548.

[97] E. Tuyishime, F. Radu, P. Cotfas, D. T. Cotfas, T. Balan, and A. Rekeraho, "Online laboratory access control with zero trust approach: Twingate use case," in *Proc. ECAI 2024*. IEEE, 2024, pp. 1–7.

[98] M. A. Azad, S. Abdullah, J. Arshad, H. S. Lallie, and Y. Ahmed, "Verify and trust: A multidimensional survey of zero-trust security in the age of IoT," *Internet Things*, vol. 27, p. 101227, 2024.

[99] N. Kumar, G. S. Kasbekar, and D. Manjunath, "Application of data collected by endpoint detection and response systems for implementation of a network security system based on zero trust principles and the eigentrust algorithm," *ACM SIGMETRICS Performance Evaluation Review*, vol. 50, no. 4, pp. 5–7, 2023.

[100] T. Dimitrakos, T. Dilshener, A. Kravtsov, A. La Marra, F. Martinelli, A. Rizos, A. Rosetti, and A. Saracino, "Trust aware continuous authorization for zero trust in consumer Internet of Things," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2020, pp. 1801–1812.

[101] H. Joshi, "Emerging technologies driving zero trust maturity across industries," *IEEE Open Journal of the Computer Society*, 2024.

[102] C. Dong, S. Pal, Q. An, A. Yao, F. Jiang, Z. Xu, J. Li, M. Lu, Y. Song, S. Chen *et al.*, "Securing smart UAV delivery systems using zero trust principle-driven blockchain architecture," in *2023 IEEE International Conference on Blockchain (Blockchain)*. IEEE, 2023, pp. 315–322.

[103] Y. Li, Y. Zhou, A. Jolfaei, D. Yu, G. Xu, and X. Zheng, "Privacy-preserving federated learning framework based on chained secure multiparty computing," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6178–6186, 2021.

[104] A. Mohan, M. Ye, H. Franke, M. Srivatsa, Z. Liu, and N. M. Gonzalez, "Securing AI inference in the cloud: Is CPU-GPU confidential computing ready?" in *2024 IEEE 17th International Conference on Cloud Computing (CLOUD)*. IEEE, 2024, pp. 164–175.

[105] F. Zhang, Q. Wu, J. Deng, M. Hao, L. Chen, S. Chen, and H. Wu, "A zero trust-based network security protection scheme for new power monitoring system," in *2024 Boao New Power System International Forum-Power System and New Energy Technology Innovation Forum (NPSIF)*. IEEE, 2024, pp. 701–707.

[106] P. García-Teodoro, J. Camacho, G. Maciá-Fernández, J. A. Gómez-Hernández, and V. J. López-Marín, "A novel zero-trust network access control scheme based on the security profile of devices and users," *Comput. Networks*, vol. 212, p. 109068, 2022.

[107] R. Wang, C. Li, K. Zhang, and B. Tu, "Zero-trust based dynamic access control for cloud computing," *Cybersecurity*, vol. 8, no. 1, p. 12, 2025.

[108] S. Ameer, L. Praharaj, R. Sandhu, S. Bhatt, and M. Gupta, "ZTA-IoT: a novel architecture for zero-trust in IoT systems and an ensuing usage control model," *ACM Transactions on Privacy and Security*, vol. 27, no. 3, pp. 1–36, 2024.

[109] K. D. Uttecht, "Zero trust (ZT) concepts for federal government architectures," *Department of Homeland Security (DHS) Science and Technology Directorate (S&T), Lexington, Massachusetts*, 2020.

[110] S. Chinamanagonda, "Zero trust security models in cloud infrastructure-adoption of zero-trust principles for enhanced security," *Academia Nexus Journal*, vol. 1, no. 2, 2022.

[111] T. Adhikari, "Advancing zero trust network authentication: Innovations in privacy-preserving authentication mechanisms," *Comput. Sci. Eng*, vol. 1, pp. 1–22, 2024.

[112] X. Li, X. Li, C. Dall, R. Gu, J. Nieh, Y. Sait, and G. Stockwell, "Design and verification of the ARM confidential compute architecture," in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 465–484.

[113] S. M. Nagarajan, G. G. Devarajan, M. S. Thangakrishnan, T. V. Ramana, A. K. Bashir, and A. A. AlZubi, "Artificial intelligence-based zero trust security approach for consumer industry," *IEEE Trans. Consumer Electron.*, vol. 70, no. 3, pp. 5411–5418, 2024.

[114] A. Hussain, W. Akbar, T. Hussain, A. K. Bashir, M. M. A. Dabel, F. Ali, and B. Yang, "Ensuring zero trust IoT data privacy: Differential privacy in blockchain using federated learning," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[115] S. R. Pokhrel, L. Yang, S. Rajasegarar, and G. Li, "Robust zero trust architecture: Joint blockchain based federated learning and anomaly detection based framework," in *Proc. SIGCOMM 2024*, 2024, pp. 7–12.

[116] F. Nawshin, D. Unal, M. Hammoudeh, and P. N. Suganthan, "AI-powered malware detection with differential privacy for zero trust security in Internet of Things networks," *Ad Hoc Networks*, vol. 161, p. 103523, 2024.

[117] A. Gupta, R. Gupta, D. Jadav, S. Tanwar, N. Kumar, and M. Shabaz, "Proxy smart contracts for zero trust architecture implementation in decentralised oracle networks based applications," *Comput. Commun.*, vol. 206, pp. 10–21, 2023.

[118] M. Asante, G. Epiphaniou, C. Maple, H. Al-Khateeb, M. Bottarelli, and K. Z. Ghafoor, "Distributed ledger technologies in supply chain security management: A comprehensive survey," *IEEE Transactions on Engineering Management*, vol. 70, no. 2, pp. 713–739, 2021.

[119] C. Liu, R. Tan, Y. Wu, Y. Feng, Z. Jin, F. Zhang, Y. Liu, and Q. Liu, "Dissecting zero trust: research landscape and its implementation in IoT," *Cybersecur.*, vol. 7, no. 1, p. 20, 2024.

[120] M. Vomvas, N. Ludant, and G. Noubir, "Establishing trust in the beyond-5G core network using trusted execution environments," *arXiv preprint arXiv:2405.12177*, 2024.

[121] C. Brito, P. Ferreira, B. Portela, R. Oliveira, and J. a. Paulo, "SOTERIA: Preserving privacy in distributed machine learning," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 135–142. [Online]. Available: https://doi.org/10.1145/3555776.3578591

[122] M. Gharib and F. Afghah, "SCC5G: A PQC-based architecture for highly secure critical communication over cellular network in zero-trust environment," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2023, pp. 11–18.

[123] S. Rodigari, D. O'Shea, P. McCarthy, M. McCarry, and S. McSweeney, "Performance analysis of zero-trust multi-cloud," in *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*. IEEE, 2021, pp. 730–732.

[124] J. Tseng, G. Bilardi, K. Ekanadham, M. Kumar, J. Moreira, and P. C. Pattnaik, "Encrypted data processing," *arXiv preprint arXiv:2109.09821*, 2021.

[125] M.-G. Kim and H. Kim, "Anomaly detection in imbalanced encrypted traffic with few packet metadata-based feature extraction." *CMES-Computer Modeling in Engineering & Sciences*, vol. 141, no. 1, 2024.

[126] M. Hussain, S. Pal, Z. Jadidi, E. Foo, and S. S. Kanhere, "Federated zero trust architecture using artificial intelligence," *IEEE Wirel. Commun.*, vol. 31, no. 2, pp. 30–35, 2024.

[127] B. Dash, "Zero-trust architecture (zta): Designing an ai-powered cloud security framework for llms' black box problems," *Available at SSRN 4726625*, 2024.

[128] H. Xie, Y. Wang, Y. Ding, C. Yang, H. Liang, and B. Qin, "Industrial wireless internet zero trust model: Zero trust meets dynamic federated learning with blockchain," *IEEE Wirel. Commun.*, vol. 31, no. 2, pp. 22–29, 2024.

[129] H. Zhu, X. Xue, M. Xu, and B.-G. Kim, "Zero trust consumer IoT with robust federated learning over main-side blockchain," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[130] S. Aiello and B. P. Rimal, "Secure access service edge convergence: Recent progress and open issues," *IEEE Security & Privacy*, vol. 22, no. 2, pp. 8–16, 2023.

[131] K. Ramezanpour and J. Jagannath, "Intelligent zero trust architecture for 5G/6G networks: Principles, challenges, and the role of machine learning in the context of O-RAN," *Comput. Networks*, vol. 217, p. 109358, 2022.

[132] W. Issa, N. Moustafa, B. Turnbull, N. Sohrabi, and Z. Tari, "Blockchain-based federated learning for securing Internet of Things: A comprehensive survey," *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023. [Online]. Available: https://doi.org/10.1145/3560816

[133] B. Sullivan and J. A. Khan, "OBSERVE: blockchain-based zero trust security protocol for connected and autonomous vehicles (CAVs) data using simple machine learning," in *Proc. ICNC 2024*. IEEE, 2024, pp. 554–559.

[134] S. Pooja and C. Chandrakala, "Secure reviewing & data sharing in scientific collaboration: Leveraging blockchain and zero trust architecture," *IEEE Access*, 2024.

[135] S. Jain, P. Ashok, and S. Prabhu, "Emerging technologies for cybersecurity in healthcare: Evaluating risks and implementing standards," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*. IEEE, 2024, pp. 725–731.

[136] L. Liu, J. Li, J. Lv, J. Wang, S. Zhao, and Q. Lu, "Privacy-preserving and secure industrial big data analytics: A survey and the research framework," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 18 976–18 999, 2024.

[137] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[138] O. C. Edo, D. Ang, P. Billakota, and J. C. Ho, "A zero trust architecture for health information systems," *Health and Technology*, vol. 14, no. 1, pp. 189–199, 2024.

[139] C. Chen, J. Liu, H. Tan, X. Li, K. I.-K. Wang, P. Li, K. Sakurai, and D. Dou, "Trustworthy federated learning: privacy, security, and beyond," *Knowledge and Information Systems*, vol. 67, no. 3, pp. 2321–2356, 2025.

[140] M. Al-Hawawreh, O. Shindi, Z. Baig, M. Alazab, A. Anwar, and R. Doss, "Quantum-powered extended visibility for zero trust-based ransomware detection in smart grids," *IEEE Internet of Things Journal*, pp. 1–1, 2024.

[141] X. Wang, X. Zha, W. Ni, R. P. Liu, Y. J. Guo, X. Niu, and K. Zheng, "Survey on blockchain for Internet of Things," *Computer Communications*, vol. 136, pp. 10–29, 2019.

[142] Y. Jiang, B. Ma, X. Wang *et al.*, "Blockchained federated learning for Internet of Things: A comprehensive survey," *ACM Comput. Surv.*, vol. 56, no. 10, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3659099

[143] J. Pan, J. Wang, A. Hester, I. AlQerm, Y. Liu, and Y. Zhao, "EdgeChain: An edge-IoT framework and prototype based on blockchain and smart contracts," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4719–4732, 2019.

[144] M. Samaniego and R. Deters, "Zero-trust hierarchical management in IoT," in *IEEE ICIOT 2018*. IEEE Computer Society, 2018, pp. 88–95.

[145] M. Schneider, R. J. Masti, S. Shinde, S. Capkun, and R. Perez, "Sok: Hardware-supported trusted execution environments," *arXiv preprint arXiv:2205.12742*, 2022.

[146] R. Hazra, P. Chatterjee, Y. Singh, G. Podder, and T. Das, "Data encryption and secure communication protocols," in *Strategies for E-Commerce Data Security: Cloud, Blockchain, AI, and Machine Learning*. IGI Global, 2024, pp. 546–570.

[147] C. Zanasi, S. Russo, and M. Colajanni, "Flexible zero trust architecture for the cybersecurity of industrial IoT infrastructures," *Ad Hoc Networks*, vol. 156, p. 103414, 2024.

[148] S. Kumar, D. Kumar, R. Dangi, G. Choudhary, N. Dragoni, and I. You, "A review of lightweight security and privacy for resource-constrained IoT devices," *Computers, Materials and Continua*, vol. 78, no. 1, pp. 31–63, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221824002121

[149] T. Liu, G. Ramachandran, and R. Jurdak, "Post-quantum cryptography for Internet of Things: a survey on performance and optimization," *arXiv preprint arXiv:2401.17538*, 2024.

[150] A. Ramakrishna, K. K. Singamaneni, G. J. Reddy, K. R. Madhavi, and T. Venkatakrishnamoorthy, "A novel QoS-based IoT network security approach with lightweight lattice-based quantum attribute-based encryption," *Tsinghua Science and Technology*, 2024.

[151] X. Zhao, W. Li, A. Xu, X. Li, and W. Shi, "Atypical dynamic trust learning in individuals with high autistic traits in a multi-round trust game with multiple trustworthiness cues," *Research in Autism Spectrum Disorders*, vol. 118, p. 102481, 2024.

[152] K. R. Malik, Y. Sam, M. Hussain, and A. Abuarqoub, "A methodology for real-time data sustainability in smart city: Towards inferencing and analytics for big-data," *Sustainable Cities and Society*, vol. 39, pp. 548–556, 2018.

[153] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer networks*, vol. 76, pp. 146–164, 2015.

[154] M. T. A. Nguyen, V. Tong, S. B. Souihi, and S. Souihi, "Zero trust: Deep learning and NLP for HTTP anomaly detection in IDS," *IEEE Journal on Selected Areas in Communications*, 2025.

[155] A. ElZemity and B. Arief, "Privacy threats and countermeasures in federated learning for Internet of Things: A systematic review," in *2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics*. IEEE, 2024, pp. 331–338.

[156] O. Said, "A bandwidth control scheme for reducing the negative impact of bottlenecks in IoT environments: simulation and performance evaluation," *Internet of Things*, vol. 21, p. 100682, 2023.

[157] J. Langguth, X. Cai, and M. Sourouri, "Memory bandwidth contention: Communication vs computation tradeoffs in supercomputers with multicore architectures," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2018, pp. 497–506.

[158] Y. Ren, Z. Wang, P. K. Sharma, F. Alqahtani, A. Tolba, and J. Wang, "Zero trust networks: Evolution and application from concept to practice," *Computers, Materials & Continua*, vol. 82, no. 2, 2025.

[159] K. G. Crowther, "Blending shared responsibility and zero trust to secure the industrial Internet of Things," *IEEE Security & Privacy*, vol. 22, no. 5, pp. 96–102, 2024.

[160] Y. Li and S. Goel, "Making it possible for the auditing of AI: A systematic review of AI audits and AI auditability," *Information Systems Frontiers*, pp. 1–31, 2024.