Post-Quantum Security for Trustworthy Artificial Intelligence: An Emerging Frontier

Saleh Darzi[1], Attila A Yavuz[1], and Rouzbeh Behnia[1]

[1]University of South Florida

November 17, 2024

Posted on 13 Feb 2024 — CC-BY 4.0 — https://doi.org/10.36227/techrxiv.170785010.08081159/v1 — e-Prints posted on TechRxiv are preliminary reports that are not peer reviewed. They should not b...

Post-Quantum Security for Trustworthy Artificial Intelligence: An Emerging Frontier

Saleh Darzi
salehdarzi@usf.edu
University of South Florida
Tampa, Florida, USA

Attila A. Yavuz
attilaayavuz@usf.edu
University of South Florida
Tampa, Florida, USA

Rouzbeh Behnia
behnia@usf.edu
University of South Florida
Tampa, Florida, USA

ABSTRACT

Recent advancements in artificial intelligence (AI) have established it as a vital tool across critical sectors such as healthcare, finance, and defense. However, significant security and privacy challenges persist. The emergence of quantum computers poses a substantial threat to AI's long-term security, and the widespread integration of AI into real-world applications underscores the critical importance of trustworthy AI. Our study delves into the intersection of AI with post-quantum (PQ) security, focusing on how post-quantum cryptography (PQC) serves the notion of trustworthy AI and bridges the long-term security gap. We offer a comprehensive comparison of PQ-secure solutions for trustworthy AI, present future perspectives and explore potential synergies across approaches.

KEYWORDS

Post-Quantum Security, Trustworthy AI, Machine Learning, Post-Quantum Cryptography, Privacy and Security

1 INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are designed to enable machines to learn, understand, and respond to datasets without explicit programming, facilitating efficient task execution. AI has consistently outperformed humans in various tasks, showcasing its superior abilities and efficiency. This has resulted in extensive integration of AI across various real-world applications, including healthcare systems, financial services, and transportation. Additionally, many cloud providers (e.g., AWS) offer their AI capabilities to enable pay-per-use consumption of AI services in organizations, leading to the emergence of the AI-as-a-Service (AIaaS) paradigm. This, coupled with the pay-per-use nature of the cloud, has further enhanced the adoption of AI among startups and small businesses, as evidenced by tools like ChatGPT [16].

1.1 Trust at the Core: The Path to Trustworthy AI

The proliferation of AI systems and various security concerns during their training and deployment phases underscores the critical importance of trust and safety in these tools. The concept of "*Trustworthy AI*" aims to ensure that these systems are secure, reliable, and practical, fostering confidence in their use. Trustworthiness is fundamentally linked to the confidentiality, integrity, and availability of these systems [21]. Ensuring data and model confidentiality is paramount in AI applications and services as privacy breaches can lead to significant ramifications for organizations, including loss of competitive advantage, erosion of customer trust, and non-compliance with regulatory requirements (e.g., HIPAA [5]). Given the high stakes, maintaining robust privacy measures has become one of the most critical concerns for organizations leveraging AI technologies. Privacy in AI encompasses two main aspects of these systems aspects:
1) **Training:** *Trustworthy training* is essential to ensure model integrity, model privacy, prevent unauthorized access to users' sensitive data (e.g., medical records, financial information, proprietary data) [25]. Model integrity ensures the robustness of these models and prevents poisoning attempts. *Model privacy in training* protects the intellectual property of AI models, preventing theft and reverse engineering of model parameters [32].
2) **Deployment:** *Trustworthy deployment* mainly focuses on adversarial attacks, model inversion attacks [33], and model privacy attacks. Adversarial attacks aim to mislead the model through malicious queries (e.g., misclassify images). Model inversion attacks, extract sensitive information from a trained model (a more severe case of membership inference attacks). Lastly, model privacy in deployment defends against model extraction attacks, which aim to replicate or steal the model by querying it with various inputs [12].

We note that these security notions are interconnected and a breach in one can potentially undermine others. For instance, inadequate protection of training data in healthcare AI systems can enable membership inference attacks, revealing whether specific individuals' data were included. This breach in data privacy directly compromises model privacy by exposing structural vulnerabilities, potentially leading to reverse engineering or other sophisticated attacks on the model.

The integrity of the training data and the model is critical for ensuring trust in AI systems. Ensuring data integrity during both training and deployment phases protect against data poisoning attacks [34], where malicious patterns are injected to cause incorrect inferences and predictions. Maintaining model integrity during training prevents backdoor attacks while ensuring integrity in deployment guards against adversarial attacks, where inputs are manipulated to deceive the model into making incorrect predictions.

The rise of open-source tools increased computational power and widespread availability of affordable computing devices has increased both the frequency and impact of denial-of-service attacks on AI models, particularly in deployment [11]. These attacks pose significant availability risks, affecting the AI model and services,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

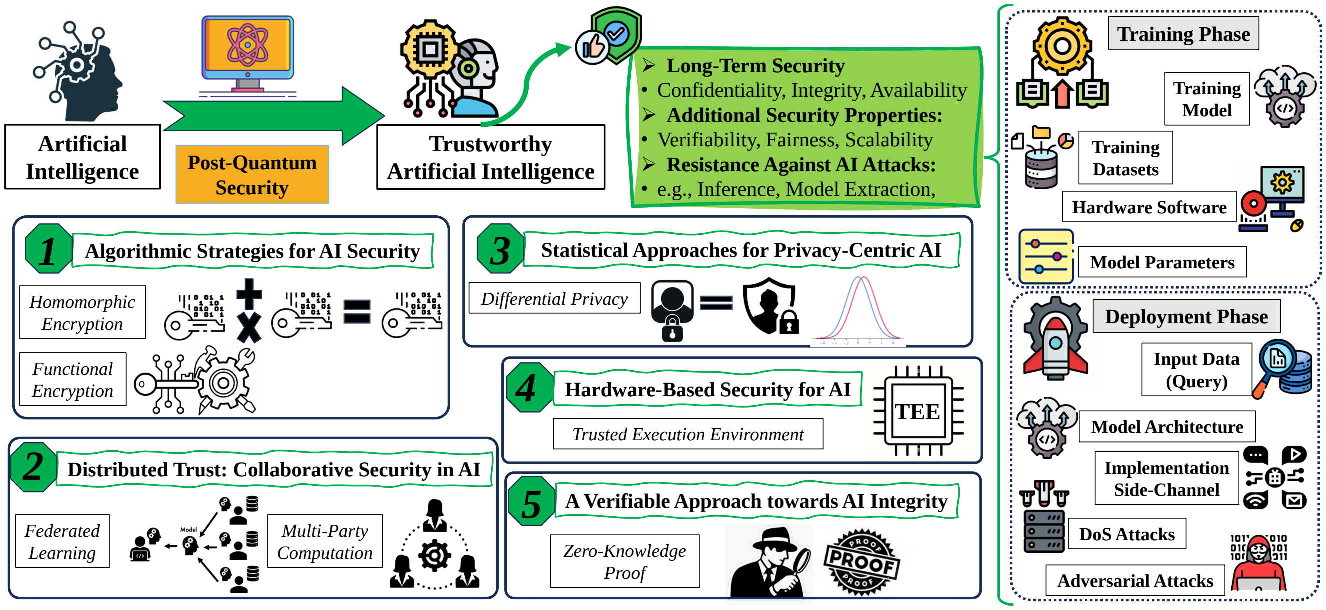
Artificial Intelligence
Post-Quantum Security
Trustworthy Artificial Intelligence
➢ Long-Term Security
• Confidentiality, Integrity, Availability
➢ Additional Security Properties:
• Verifiability, Fairness, Scalability
➢ Resistance Against AI Attacks:
• e.g., Inference, Model Extraction,
1 Algorithmic Strategies for AI Security
Homomorphic Encryption
Functional Encryption
2 Distributed Trust: Collaborative Security in AI
Federated Learning
Model
Multi-Party Computation
3 Statistical Approaches for Privacy-Centric AI
Differential Privacy
4 Hardware-Based Security for AI
Trusted Execution Environment
TEE
5 A Verifiable Approach towards AI Integrity
Zero-Knowledge Proof
PROOF
Training Phase
Training Model
Training Datasets
Hardware Software
Model Parameters
Deployment Phase
Input Data (Query)
Model Architecture
Implementation Side-Channel
DoS Attacks
Adversarial Attacks

Figure 1: A high-level taxonomy of PQ-secure techniques for Trustworthy AI

especially in safety-critical applications, such as autonomous driving, face recognition, and intrusion detection systems. Moreover, security factors such as verifiability, fairness, and robustness are essential to cultivate trust and promote the widespread application of AI services.

Existing research at the intersection of PQ computing and AI focuses on Quantum Machine Learning (QML) or using AI for network intrusion detection systems (NIDSs) in the PQ era [24]. However, QML—grounded in quantum channels and physics principles—fails to address the transitional phase to the PQ era and remains impractical for real-world deployment [9]. In the broader context of AI and cybersecurity, numerous surveys and research papers focus on individual security criteria within AI systems. For instance, privacy-preserving machine learning (PPML) is widely covered in privacy-centered research [23], while architecture-centric approaches, such as federated learning security requirements and taxonomies of security attacks on AI/ML, are also well-documented [28]. Although these cryptographic and conventional security approaches offer valuable insights, they fall short in addressing long-term security and trust at the foundational level, particularly for AI systems deployed in high-stakes, real-world applications. Moreover, there is a lack of comprehensive comparisons across these techniques, including their limitations, progress, and overarching goals. Addressing these gaps is essential for ensuring that AI systems can achieve robust, future-proof security against emerging PQ threats.

1.2 Securing Long-Term Trust: PQC Serves AI

Beyond current security challenges, advancements in quantum computing present a formidable threat to trustworthy AI. These machines can undermine traditional security measures, compromising sensitive data, jeopardizing privacy measures, and destabilizing AI security frameworks. For instance, in AI-driven healthcare systems, quantum computing can breach secure communication channels by breaking encryption methods like ECC, jeopardize the privacy of patient's medical information, and compromise the integrity of diagnostic results, ultimately eroding trust in AI-based medical applications [20]. It is imperative to prioritize the transition to PQ solutions to safeguard the long-term security of AI systems, as highlighted by strategies from international institutions, agencies, and government entities (e.g., the U.S. White House) [6]. PQC is among the primary solutions for addressing security and privacy threats in the PQ era. Numerous academic and industry initiatives, competitions, and projects are focused on developing and implementing general-purpose PQC use cases, with NIST spearheading standardization efforts, highlighting PQC's critical importance for ensuring the long-term security of real-world applications, including AI [10]. However, several gaps, challenges, and unresolved questions remain, which justify the need for this study, as outlined below:

- What drives the need for PQ-secure solutions in trustworthy AI? *The widespread integration of AI in data-sensitive applications such as military, autonomous vehicles, healthcare, and model-sensitive domains like financial services necessitates long-term security. Beyond PQ robustness against attacks, PQ trust in supporting infrastructures is essential, particularly for AIaaS, distributed AI learning, and cloud-based AI systems utilizing network communication channels [26]. Consequently, there is an urgent need to shift to PQ-secure solutions, giving rise to the concept of "PQ-Secure Trustworthy AI", the focus of this study. This transition is vital to ensuring security in the forthcoming PQ era.*
- How does PQ secure techniques serve trustworthy AI? *There is no one-size-fits-all methodology for achieving a robust, trustworthy AI system that addresses all security, varying privacy levels (e.g., data, model), availability, integrity, and resistance to numerous AI attacks. This study investigates the PQ-secure techniques that serve trustworthy AI. We explore Trustworthy AI through algorithmic approaches like Homomorphic Encryption (HE), Functional Encryption (FE), and Multi-Party Computation (MPC) with provable security, statistical techniques such as Differential Privacy (DP), and hardware-based solutions using Trusted Execution Environments (TEE). A taxonomy of our study detailing PQC techniques*

and considerations for both the training and deployment phases of Trustworthy AI is shown in Figure 1.

• How can we bridge the security gap in trustworthy AI for the post-quantum era? *To our knowledge, no study has specifically addressed Trustworthy AI with long-term security concerns associated with post-quantum cryptography. AI and PQ security primarily operated in isolation, focusing only on their integration. This article identifies gaps and implementation challenges in existing PQ-secure solutions that require further examination. It also provides forward-looking PQ-secure visions and explores potential synergies between various approaches as long-term security measures. Additionally, a qualitative comparison of these methods is presented. The analysis highlights several research gaps and potential solutions that warrant attention and action from academia and industry to strengthen the future of PQ-secure and Trustworthy AI systems.*

2 KEY ELEMENTS OVERVIEW

Communication protocols (e.g., TLS), authentication mechanisms (e.g., digital signatures), and encryption systems (e.g., ECC) used in real-world applications are built on traditional cryptographic frameworks. In public key settings, these frameworks rely on classical hard problems like integer factorization (IF) and the discrete logarithm problem (DLP), both of which are compromised by Shor's algorithm. Similarly, many symmetric key cryptographic primitives are weakened by Grover's algorithm. This has led to adopting quantum-resistant methods, including lattice-based, code-based, isogeny-based, and multivariate cryptography, along with advanced symmetric key primitives [10]. The techniques outlined below are key approaches for achieving trustworthy AI in the PQ era. Their descriptions and utility during the training and deployment phases are detailed in Table 1.

• *Homomorphic Encryption (*HE*)*: HE enables computations over the encrypted data and is classified into three classes based on the applied function: Partial HE (PHE), Somewhat HE (SHE), and Fully HE (FHE). The first proposal of an FHE scheme [17] marked a pivotal moment, shifting its theoretical consideration as a holy grail of cryptography towards practical adoption in real-world applications.

• *Functional Encryption (*FE*)*: Enables the holder of a specific key (the functional encryption key) to compute a function on encrypted data and retrieve the final output without learning the original plaintext or the master secret key. FE is classified into Inner-Product FE and Quadratic FE, based on the functionality performed [8].

• *Multi-Party Computation (*MPC*)*: MPC allows multiple parties to securely collaborate in computing a function over their private inputs. This is particularly applicable in scenarios where a group of entities lacks mutual trust or without a trusted third party [38].

• *Differential Privacy (*DP*)*: The goal of DP is to prevent the disclosure of the presence or absence of an individual data point, with a privacy budget denoted by ϵ. A smaller ϵ indicates stronger anonymization, providing higher privacy protection [14].

• *Trust Execution Environment (*TEE*)*: In a TEE, data and code access are secured through hardware and software support. Various TEE designs cater to different architectures, with Intel Software Guard Extensions (SGX) and ARM Trust Zone (TZ) being the most prevalent ones [18].

• *Zero-Knowledge Proof (*ZKP*)*: At its core, it's a cryptographic two-party construction to safeguard the privacy of data as one party proves a statement to another without disclosing any additional information beyond the statement's validity, hence the term "zero-knowledge." The classification of ZKP depends on the communication and interaction between the involved parties, leading to two main categories: interactive and non-interactive protocols [22].

3 PQ-SECURE AI TRAINING & DEPLOYMENT

This section provides a comprehensive analysis of potential PQ security aspects in AI, covering both the training and deployment phases, with the objective of establishing a trustworthy AI.

3.1 Algorithmic Approaches for AI Security

The main algorithmic strategies with provable security rely on encryption techniques, primarily HE and FE. These methods ensure continuous data encryption across all stages, including sharing, processing, training, inference, post-AI output, and storage, securing the data throughout its lifecycle in an AI system.

3.1.1 ***Homomorphic Encryption.*** HE, especially FHE, is widely recognized as a cornerstone solution within Privacy-Enhancing Technologies (PETs). It plays a pivotal role in various AI scenarios by enabling computations on encrypted data while ensuring the confidentiality of both user data and the AI model throughout the entire process, from data sharing to inference, without requiring decryption at any stage [31].

Training Phase: A few research have explored training neural networks over encrypted data, avoiding reliance on unrealistic assumptions such as the presence of honest parties [35]. Nonetheless, their substantial computational burdens and reduced accuracy compared to training over the plaintext have led to a minimal employment of FHE in the training phase. Despite this, submitting encrypted data for outsourced training is highly preferred by users and greatly simplifies practical applications, such as securely uploading encrypted financial transactions for analysis or patient data for diagnosis training. FHE also supports secure aggregation in collaborative and federated learning, allowing entities like hospitals to train models on encrypted medical records without risking data breaches. Additionally, HE protects data privacy and model privacy by encrypting model parameters, which is particularly useful for proprietary systems like recommendation algorithms.

Deployment Phase: Given the higher priority of data privacy outweighing model privacy, most state-of-the-art FHE-based solutions primarily focus on the deployment phase. These schemes enable cloud-based AI or AIaaS, to perform predictions on encrypted data using pre-trained models. Moreover, HE is particularly suitable for deploying models in untrusted environments, such as public clouds or edge devices. Many of these schemes leverage HE to secure models during deployment on third-party platforms, as the homomorphic property of HE ensures that data remains encrypted throughout all stages of AI applications. This approach offers robust protection against potential breaches, insider threats, and unauthorized access, thereby providing strong security assurances.

Despite its promising security, most employed HE variants in AI are rooted in classical cryptography and lack PQ security, particularly those based on PHE and SHE schemes. However, most practical variants of FHE schemes are built upon hard lattice problems (such as the Learning With Errors problems) or rely on approximate problems

Table 1: High-level overview of the PQ-secure solutions with their utility in AI

Approach	Description	Utility in Training Phase	Utility in Deployment Phase
HE	Enables computations over encrypted data potentially yielding encrypted results as if computed on plaintext	Data Privacy, Model Privacy Confidentiality of AI Parameters	Input Data Privacy, Model Privacy Inference Over Encrypted Data
FE	Allows decryption to reveal only a specific function of the encrypted data	Data Privacy, Model Privacy Access Control	Input Privacy, Model Privacy Suitable For Private AIaaS
MPC	Enables secure distribution of function computation on private inputs among distributed parties	Data Privacy, Model Privacy Collaborative Training Support Resource-Constraint Environment	Data Privacy Against Model Owner Model Theft Resistance Distributed Private & Outsourced Inference
DP	A distortion mechanism that adds noise to minimize the effect of any individual data point on the model	Individual Data Privacy Extraction Attacks Resistance Centralized & Distributed Compatibility	Inference Attacks Resistance Membership Attacks Resistance Private Real-Time AI Services
TEE	Establishes a secure, isolated, and trusted environment for data and computation	Data Privacy, Model Privacy Model Integrity Confidential Code Execution	Confidentiality of AI Parameters Access Control, Private Inference Model Extraction Attacks Resistance
ZKP	Allows one party to prove the correctness of a statement to another party without disclosing any information beyond the statement's validity	Computational Integrity Inference Correctness Training Fairness & Model Integrity	Prediction Integrity Inference Verifiability Fake Service Prevention

in high-dimensional lattices, thereby offering PQ security assurances. Therefore, it is reasonable to infer that adopting an FHE-based approach in AI inherently ensures adherence to PQ security guarantees. The practicality of the HE approach is significantly influenced by the existing open-source libraries. Numerous libraries incorporate state-of-the-art FHE schemes to expand their capabilities to support optimization techniques. Advances in engineering aspects, including batching techniques for handling multiple inputs in a single ciphertext, support for parallelization, as well as hardware acceleration through GPU and FPGA implementation, and algorithmic enhancements for the training phase, contribute to increased efficiency and improved performance of FHE. These developments make FHE a viable solution for AI applications.

3.1.2 ***Functional Encryption (FE).*** FE constructions are in the early stages of AI application, with the key distinction from HE being that FE produces plaintext outputs of the performed function rather than ciphertext. While both are applicable in privacy-preserving AI for training and inference, FE is more suited for scenarios requiring plaintext outputs, such as cloud-based AI and AIaaS. For instance, in cloud-based healthcare, FE allows the cloud server to compute a disease risk score based on encrypted patient data and return a plaintext result (the score) to the clinician, protecting the underlying data. FE restricts decryption to specific outputs, allowing only particular functions to be computed without sharing a secret key, making it ideal for selectively revealing data in sensitive environments [30].

While FE-based methods enable collaboration without fully disclosing raw data and offering access control along with privacy, they require a trusted third party for key generation. This makes them applicable only under an honest-but-curious security model, where the server fulfills its duties but may attempt to glean information about the secret communication, data, or model. However, FE carries certain privacy risks. The plaintext output format can lead to information leakage, making the system vulnerable to threats like model inversion and inference attacks, even from an honest-but-curious server. FE mainly derives its security directly from the difficulty of the underlying problems in the employed lattice-based or multivariate cryptography. These PQ assurances entail efficiency sacrifices, including higher security parameters, slower operations, and increased complexities, making them unfeasible for large datasets.

3.2 Distributed Trust: Collaborative Security in AI

This section explores architectural strategies for collaborative learning and distributed techniques, focusing on FL and secure MPC.

3.2.1 ***Federated Learning (FL).*** FL, introduced by Google [27], is an architectural scalability solution that enables collaborative learning by allowing participants to train models locally without sharing raw data, thus enhancing privacy and security. FL aggregates model updates centrally, creating a robust and resilient system that also reduces data transfers, minimizing potential breaches during training. For applications governed by regulations like HIPAA [5], FL often serves as the only compliant approach. It requires no specific hardware or maintenance at the network edge while offering fault tolerance, as the failure of one participant does not halt the process. This approach is ideal for large-scale real-world scenarios.

Although FL is primarily an AI strategy rather than a cryptographic solution, it plays a crucial role in the PQ era. FL is not fully distributed and relies on centralized aggregation for global model updates and refinements. However, local updates can reveal sensitive user data, necessitating secure aggregation techniques based on cryptography. FL's PQ security can be achieved through the integration of PQC methods (e.g., FHE, MPC, TEE) for the secure aggregation. Specifically, secure aggregation can be ensured using advanced encryption techniques (e.g., HE, FE), where aggregation occurs over encrypted data, using MPC to distribute the aggregation securely among non-colluding parties, or through DP protection by adding noise to local updates to safeguard user data and resist leakage. Additionally, from a PQ perspective, secure aggregation can be delegated to a quantum computer, providing not only PQ security but also verifiability of computations, leveraging the unique properties of quantum computing. However, this approach remains impractical for widespread adoption in the transition to the PQ era.

3.2.2 ***Multi-Party Computation (MPC).*** MPC serves as a key cryptographic backbone technique and a foundational support for collaborative earning, enabling distributed trust while facilitating the transformation from centralized frameworks into distributed ones. The novelty of MPC lies in its capability to enable users to compute functions on their data privately, sharing only the final results while offloading much of the computational burden. Users collaboratively compute the final model without accessing individual gradients, ensuring data privacy. However, this comes at the expense of increased communication overhead among the involved parties. Primarily, a secure MPC is implemented through cryptographic techniques, encompassing, among others, HE, Garbled Circuit (GC), Oblivious Transfer (OT), and Secret Sharing (SS). HE enables operations to be performed on encrypted data without decryption. GC encrypts the boolean circuit version of the function to be computed, allowing for the secure evaluation of the function with no disclosure of intermediate and input data in a fixed number of communication rounds. OT allows for the exchange of private data in a privacy-preserving manner. SS enables data to be shared in a way that only a threshold combination of shares can recover the data, ensuring secure sharing.

Training Phase: Training generally takes two forms: local training over distributed datasets or secure collaborative training over centralized datasets. MPC facilitates secure aggregation in scenarios where multiple institutions collaboratively train on datasets without exposing raw data. In this approach, only model updates are shared while all other data remains private, making it particularly well-suited for cross-silo applications. By training data locally and exchanging only updates, the risk of data breaches is significantly minimized. Also, MPC enables secure outsourcing of user updates and training in untrusted environments or with private data sources.

Deployment Phase: Most MPC-based methods work alongside encryption mechanisms, particularly using HE schemes. In the deployment phase, MPC ensures privacy by securely distributing both the inference process and query data, safeguarding user privacy from the model owners, and enabling privacy-preserving inference. Specifically, by allowing multiple entities to participate in the inference phase without relying on a trusted third party, MPC enables secure inference and facilitates regulatory compliance. MPC ensures user data privacy during querying by enabling collaborative tasks without exposing sensitive information. For instance, in fraud detection across banks, transaction details remain private while detecting anomalies collaboratively. Similarly, in healthcare, MPC allows the joint analysis of patient data while complying with regulations like HIPAA [5]. Furthermore, MPC supports outsourcing inference to untrusted servers, enabling organizations to use shared models for predictions while keeping their input data secure, making it ideal for PET across sectors such as insurance and supply chain management.

There are various PQ-secure MPC schemes designed for two-party, three-party, and four-party AI scenarios. Typically, as the number of parties involved increases, the performance enhancement (e.g., communication overhead) is more pronounced. Indeed, MPC can be implemented using PQ-secure cryptographic primitives, such as NIST-standardized PQC or symmetric encryption schemes (e.g., AES, hash functions). Utilizing PQC-secure encryption during data preparation, employing a secret sharing mechanism for data splitting, and incorporating PQ-secure HE for aggregation/evaluation can provide PQ security assurance. Also, MPC-based constructions that employ PQ-secure schemes under TEEs for function evaluation are resistant to PQ attacks, enhancing the overall security of the process.

3.3 Statistical Approaches for Privacy-Centric AI

DP is a statistical perturbation technique that, while not inherently based on PQ-secure cryptographic problems, is not vulnerable to quantum attacks, unlike approaches dependent on cryptographic hard problems. However, to achieve PQ guarantees in AI applications, DP could be combined with PQ-secure cryptographic protocols. As DP primarily focuses on individual data privacy, it serves as a complementary solution commonly employed in conjunction with other privacy-enhancing techniques (e.g., HE, FE, and MPC) to address broader privacy needs (e.g., system-level privacy) and reinforce PQ promises [15]. After extensive research spanning over a decade, DP has emerged as the de facto standard for privacy protection in PETs. Technically, DP is a randomizing algorithm that incorporates calibrated statistical noise addition or carefully adjusted dataset swapping to ensure that distinguishing between two nearly identical datasets is no more effective than a coin toss. In an AI setting, DP ensures the output of the optimization algorithm is distorted with noise to ensure that no one data point has affected the output of the algorithm.

Training Phase: During the training phase, DP methods vary based on architecture and learning setting. In a centralized setting, random noise is added to the objective function to mask sensitive data and obscure the influence of any single data point, protecting model outputs from privacy leaks. However, this approach depends on an honest-but-curious assumption and trust in the model administrator, potentially conflicting with strict privacy policies. In decentralized AI, local DP allows users to add noise directly to their data to protect it from the model owner, while central DP applies carefully calibrated noise to aggregated outputs, concealing individual user activity during training. Instance-level DP introduces randomization to local labels, protecting against the disclosure of specific label information in the final model.

Deployment Phase: In the deployment phase, regardless of whether the setting is centralized or distributed, DP provides data-level privacy guarantees that protect sensitive inference results in real-time systems, such as recommendation and fraud detection systems, while securing user queries in online environments to prevent adversaries from inferring input data from outputs. Consequently, DP is often the preferred choice for large companies seeking to manage user data in a privacy-preserving manner, facilitating noisy predictions, private inference, and resistance to inference attacks during deployment.

A PQ aspect of DP involves adding quantum noise, leading to Quantum-DP. QDP leverages quantum processing to enhance model and user data privacy against post-processing attacks [36]. Although practical implementation faces challenges, QDP offers a promising approach, especially with delegated quantum computation, where computations are outsourced to a central quantum server. This approach could support verifiable outputs and client-side privacy even with Noisy Intermediate-Scale Quantum (NISQ) devices, potentially advancing PQ solutions before full-scale quantum capabilities are available [37]. Beyond safeguarding privacy, DP offers diverse utilities in AI applications. These include preventing over-fitting through DP-assisted data testing, ensuring model fairness by employing DP in data resampling, and addressing stability issues.

3.4 Hardware-Based Security for AI

In untrusted environments, whether centralized or distributed, TEE protects AI data, software, and operations by creating an isolated, secure environment with dedicated memory registers. In essence, TEE is a hardware-assisted approach that promises the notion of "*Confidential Computation*". While the use of TEE for privacy-preserving

AI is a relatively new approach, it provides stronger security assumptions than traditional cryptographic methods and reduces computational overhead compared to methods based on complex computational problems, such as HE, FE, and MPC [13].

Training Phase: By isolating processed data and the computed model within the TEE and restricting access solely to authorized entities during the training phase, this approach ensures both data and model privacy and prevents parameter leakage. It ensures that neither the model owner nor the data owner can access each other's sensitive information, thereby protecting intellectual property, proprietary AI algorithms, and models through confidential computation. For instance, pharmaceutical companies collaborating on drug discovery can use TEEs to jointly train AI models on sensitive research data without exposing proprietary compounds or algorithms to competitors. Similarly, in the automotive industry, manufacturers can collaborate on autonomous driving algorithms while keeping proprietary data and model parameters secure.

Deployment Phase: The utilization of TEE in the deployment stage enables a party to receive inference results in a privacy-preserving manner while protecting model parameters and preventing model theft. For example, running medical diagnoses in hospitals, fraud detection systems in financial institutions, and autonomous vehicle navigation systems can all be performed under TEE to safeguard sensitive data while ensuring resistance to unauthorized access.

The PQ security of TEE-based approaches does not originate from the TEE design itself; instead, it relies on utilizing a PQ-secure scheme under the TEE. Technically, by employing NIST PQC-standardized schemes for AI training and inference, ensuring that computations are not secure by design but by protocol, we can achieve PQ promises in PPML.

3.5 A Verifiable Approach towards AI Integrity

The presence of malicious parties in outsourced AI computations or collaborative training that involves user-assisted training or model aggregation introduces potential vulnerabilities and risks. Specifically, the model owner might engage in dubious activities when applying the AI model, users could act maliciously or misbehave to gain an advantage in the inference phase by poisoning input data, or they might input incorrect data to lower their computational costs, especially if they have limited resources. Also, in real-world AI applications involving sensitive information, such as in financial services (e.g., fraud detection, money laundering monitoring, trading), healthcare systems (e.g., AI diagnosis, medical insurance policies), etc., it is crucial for the model owner to correctly apply the training model to users' data. Scenarios like collaborative model training between institutions, ensuring reliable model outputs in healthcare or insurance, regulatory compliance, and outsourced AI computations all demand the assurance of model correctness. In such cases, computational integrity becomes essential alongside privacy protection. Thus, verifiability not only ensures accuracy but also prevents the dissemination of fake services by the model owner [29].

ZKP emerges as the ideal solution for these scenarios, fundamentally offering trust in the system, providing technical fairness, legitimizing computation correctness, proving the accuracy of model output, and ensuring training correctness in a privacy-preserving manner. Thereby, ZKP could be viewed as a complementary tool to other privacy-enhancing techniques for AI. Note that, despite the need for verifiable AI, the verification algorithm should not disclose users' privacy and resist common threats like member inference and reconstruction attacks. While the majority of efficient ZKP systems are built on elliptic curve cryptography (ECC) and offer short proofs with fast verification, there are also practically efficient ZKP schemes constructed on symmetric ciphers, lattice-based, code-based, and multivariate cryptography, providing PQ promises suitable for AI applications. Despite their PQ promises, these schemes face challenges, including high computational complexity, large proof sizes requiring increased storage and transmission bandwidth, and interoperability issues that limit their integration in diverse environments, such as large-scale networks and resource-constrained settings, due to added latency and computational demands.

4 LIMITATIONS, CHALLENGES, & VISIONS

Since there is no holistic solution that currently addresses all security and privacy aspects of AI, this section presents a qualitative comparison of the discussed approaches. It highlights their strengths, limitations, practical utility, potential synergies, and future prospects. Table 2 summarizes progress in terms of standardization efforts, security guarantees, and suitability for various use cases. Table 3 highlights their limitations and challenges based on criteria such as security weaknesses, performance impact, implementation challenges, and scalability, as well as architectural support for centralized, distributed, and resource-constrained environments.

▷ ***Homomorphic Encryption*:** The primary drawbacks of HE lie in their substantial computational overhead for large models and low-performance efficiency (e.g., model accuracy), largely attributed to the lattice-based constructions employed. Also, challenges such as large key and ciphertext sizes, the necessity for noise management, and limited functionality, particularly for complex neural networks over encrypted data, may constrain the utility of HE in AI applications. Despite efforts to develop practical FHE-based designs, it is crucial to emphasize that, in comparison to tasks like inference, prediction, and classification over unencrypted data, there remains a considerable journey ahead to enhance output accuracy, reduce computational overheads for massive datasets, and alleviate performance burdens for real-world applications.

Beyond the challenges outlined in the tables, potential synergies warrant further investigation. Many HE schemes assume the encryption of all data under one key. A novel extension involves expanding the number of keys that can be supported on homomorphically evaluated ciphertexts. While this extension poses challenges for efficient design, it is particularly suitable for AI with collaborative system models. This is also relevant in scenarios involving distributed computation, where data owners may lack trust or are unwilling to share a key. Another potentially valuable feature for privacy-centric scenarios is distributed decryption, enabling the combination of parties' secret keys to collectively decrypt the final ciphertext.

Incorporating attribute- or identity-based encryption properties into HE can make it more suitable for AI applications (e.g., healthcare, financial fraud detection). This enhancement preserves privacy while enabling selective access control, allowing only individuals with specific attributes or identities to access the encrypted data. Moreover, threshold-HE enables the creation of a collective public key, allowing a group of users to participate in the evaluation process. Only a specific combination of these users can perform

Table 2: Qualitative Comparison of PQ-Secure Solutions for Trustworthy AI

Approach	Standardization Efforts	Security Guarantees	Use Case Suitability
HE	Ongoing Standardization Consortium By NIST, ISO, & IEC [2]	Computational Security Offers Encryption At All Times	Privacy-Enhancing Technologies e.g., Healthcare, Finance, Military
FE	No Formal Standardization Efforts; Still in Early Stages [30]	Computational Security Controlled Access Environments	Specific Analysis on Private Data e.g., Diagnosis in Healthcare Systems
MPC	Strong Standardization Efforts By ISO & NIST [1]	Computational Security Secure Collaborative Computation	Collaboration Without Data Sharing e.g., Joint Financial Analysis, Advertising
DP	Notable Standardization Efforts By NIST & OpenDP Initiatives [3]	Statistical Security Data-Level Privacy	Data Analytics Applications e.g., Social Science, healthcare systems
TEE	Well-Defined Standards; GlobalPlatform & Proprietary [4]	Hardware Security Resistance Against External Interference	Sensitive & Secure Computations e.g., Cloud Computing, Smart Contracts
ZKP	ZKProof Standards; Open-Industry Academic Initiative [7]	Computational Security Data- & System-Level Integrity	Verifiable Computations e.g., Blockchains, Financial Proofs

Table 3: Limitations and Challenges of PQ-Secure Solutions for Trustworthy AI

Approach	Security Weaknesses	Performance Impact	Implementation & Scalability Challenges	Architectural Support
HE	Vulnerable to Fault-Injection Attacks and Fault-Injection Attacks	Significant Computational Costs High Memory Overhead Slow for Complex Operations	Requires Specialized Software Libraries Requires Hardware Acceleration Impractical for Large-Scale Networks	Great for Centralized Setting Unsuitable for Distributed Setting Infeasible for IoT Devices
FE	Susceptible to Adaptive Attacks Data Disclosure in Some Functions	Less Efficient than Standard Encryption High Latency for Complex Scenarios	Lack of Practical Libraries Limited Supporting Cryptographic Tools Only Applicable in Small Networks	Requires Centralized Infrastructural Support Limited Applicability to Distributed Setting Not Suitable for IoT Environments
MPC	Vulnerable to Collusion Attacks and Man-In-The-Middle Attacks	High Communication Overhead Introduces High Latency High Bandwidth Requirements	Requires Secure Communication Channels Needs Domain-Specific Knowledge Not Scalable for Large Number of Participants	Not Typically Used for Centralized Setting Great for Distributed Settings Applicable for IoT Environments
DP	Vulnerable to Privacy Leakage and various Inference Attacks	Reduced Inference Accuracy Lowered Data Utility	Setting the Privacy Budget Few Automated Tools Highly Scalable But Bad with Large Datasets	Suitable for Centralized Setting Efficient for Distributed Frameworks Suitable for IoT Environment
TEE	Vulnerable to Side-Channel Attacks Susceptible to Physical Tampering	Costly Context Switching Increased Latency	Requires Specific Hardware Support Hardware Bugs and Compatibility Issues Requires Device Compatibility for Scalability	Well-Suited for Centralized Setting Applicable for Distributed Setting Limited by Hardware Availability for IoT
ZKP	Vulnerable to Malicious Setup Attacks and Risk of Inference Attacks	Computationally Expensive Large Proof Sizes	Requires Specialized Libraries Needs Advanced Cryptographic Tools Limited Scalability for Computational Costs	Useful for Centralized Verification Applicable to Distributed Setting Too Expensive for IoT Environments

decryption. This property is well-suited for privacy in collaborative learning, encompassing both honest-majority and dishonest-majority scenarios.

▷ **Functional Encryption:** For future work, given that FE often contrasts with FHE, numerous challenges and opportunities demand further attention and resolution:

Besides lowered model performance in both phases, currently, FE schemes focus mainly on AI inference, leaving a gap in leveraging FE techniques for training over encrypted data. FE operates at lower levels than HE, incurring higher computational costs and exhibiting limited functionality, especially for intricate AI algorithms such as convolutional neural networks. In terms of implementation and its impact on performance, while HE benefits from various libraries across different programming languages and platforms, there are only a few libraries dedicated to implementing state-of-the-art FE schemes. From an engineering perspective, FE lacks attention, evident in the scarcity of hardware acceleration, GPU or FPGA utilization, and optimization methods for faster computational performance. Notably, to our knowledge, all FE-based PPML implementations are currently on CPU, prompting the need for research on GPU support for FE schemes that would significantly contribute to enhancing the applicability of FE methods in real-world scenarios.

▷ **Multi-Party Computation:** The practical limitation of relying solely on MPC for the privacy of AI lies in scalability issues and substantial communication load, particularly in large-scale AI applications or training with massive databases. The performance challenges worsen with the need for PQ security. Additionally, the distributed function must be known, public, or shared, which may not be feasible in certain AI scenarios. The potential for collusion among participants and the requirement for participants to be online with adequate bandwidth during function evaluation pose limitations on MPC approaches, particularly with resource-constrained devices. An alternative approach for such scenarios involves using online-offline methods, allowing for pre-computation and swift online interaction.

Given the computational power, communication bandwidth, and application choice, one may opt for FHE for low communication but high computation or MPC for high communication with lower computation. Despite the longstanding competition between FHE and MPC as privacy solutions in various PETs, an optimal approach for privacy-preserving AI combines MPC and HE synergistically. This combination not only elevates privacy for both model and data owners but has also demonstrated significantly enhanced performance compared to using these techniques independently. However, it's worth noting that certain HE-enabled MPC techniques may lack resistance against model extraction attacks. One potential remedy is the integration of DP on top of MPC. Finally, an advanced approach under exploration is multi-party quantum computation, a novel concept that requires further investigation.

▷ **Differential Privacy:** Although DP is a widely applicable approach in AI across various contexts, including centralized, collaborative, and resource-constrained environments, it is not an all-encompassing solution. The intrinsic trade-off of privacy in DP lies in the compromise of the model's accuracy as the performance penalty. Despite safeguarding individual identities through the censorship of personal data and smoothing the impact of user contributions throughout both the inference and training phases, DP diminishes the quality of the trained model, its parameters, or the output results. It functions effectively alongside nearly all other PQ-secure techniques. Consequently, PQ-DP warrants further investigation, particularly regarding the use of quantum-resistant pseudorandom number generators (QR-PRNGs) to generate calibrated noise.

▷ **Trust Execution Environment:** Technically, training larger models using TEE remains challenging, as demonstrated by the limited research in this area. While existing studies show promising results, scaling TEEs for real-world applications remains a significant hurdle. TEE also faces challenges such as hardware expenses and susceptibility to side-channel attacks, including execution time and power analysis [19]. These attacks could undermine the confidentiality of enclaves, posing a threat to the system by potentially leaking information and disclosing privacy. Additionally, in the context of AI with massive data or complex models, the necessity for partitioning and batch processing is essential, rendering TEE inapplicable in some scenarios. Furthermore, the use of TEE is not optimal for performance, where the overhead associated with mode transitions, shifting between untrusted and trusted states, can become expensive and impact the overall system performance. Additionally, the design of GPU support and acceleration techniques could prove beneficial for AI applications in the TEE context. Thus, there is a need for further studies and attention, particularly in the engineering aspect.

▷ **Zero-Knowledge Proof:** PQ-secure ZKP, while not directly applicable in all processes of AI, offers distinct advantages. An evident area for potential future work involves integrating verifiable AI into existing privacy frameworks, especially for those with PQ assurances. For example, in combination with encryption-based approaches, verifiability can validate that encrypted data is correct and falls within a specific range. Currently, they remain impractical for complex AI algorithms, can be extremely expensive for proof generation in many AI scenarios with extensive tasks, and are limited in addressing specific aspects of privacy in AI. ZKP can be viewed as an equivalent approach to those based on TEE. While TEE relies on hardware assistance, ZKP provides cryptographic proof for secure computations. Despite differences in their fundamental assumptions and security assurances, ZKP incurs lower implementation costs. Also, PQ-secure ZKP-based schemes are mostly non-interactive, in contrast to MPC-based ones that assume synchronous communication without support for dynamic parties. This results in a more realistic system model and lower overhead.

REFERENCES

[1] 2023. *Diversity and tradeoffs in MPC standardization*. Retrieved Nov. 2024 from https://csrc.nist.gov/presentations/2023/mpts2023-day1-talk-mpc-standardization

[2] 2023. *Efforts on Standardizing Fully Homomorphic Encryption at ISO/IEC*. Retrieved Nov. 2024 from https://csrc.nist.gov/Presentations/2023/stppa6-iso-iec-fhe

[3] 2023. *Guidelines for Evaluating Differential Privacy Guarantees*. Retrieved Nov. 2024 from https://csrc.nist.gov/pubs/sp/800/226/ipd

[4] 2023. *TEE System Architecture v1.3 GPD-SPE-009*. Retrieved Nov. 2024 from https://globalplatform.org/specs-library/tee-system-architecture/

[5] 2024. *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Retrieved Nov. 2024 from https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

[6] 2024. *A National Security Memorandum (NSM-10)*. Retrieved 2024 from https://www.whitehouse.gov/briefing-room/statements-releases/2022/05/04/national-security-memorandum-on-promoting-united-states-leadership-in-quantum-computing-while-mitigating-risks-to-vulnerable-cryptographic-systems/

[7] 2024. *ZKProof Standards*. Retrieved Nov. 2024 from https://zkproof.org/

[8] Dan Boneh, Amit Sahai, and Brent Waters. 2011. Functional encryption: Definitions and challenges. In *Theory of Cryptography: 8th Theory of Cryptography Conference, TCC 2011,Proceedings 8*. Springer, 253–273.

[9] Marco Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. 2022. Challenges and opportunities in quantum machine learning. *Nature Computational Science* 2, 9 (2022), 567–576.

[10] Saleh Darzi, Kasra Ahmadi, Saeed Aghapour, Attila Altay Yavuz, and Mehran Mozaffari Kermani. 2023. Envisioning the future of cyber security in post-quantum era: A survey on pq standardization, applications, challenges and opportunities. *arXiv preprint arXiv:2310.12037* (2023).

[11] Saleh Darzi and Attila A Yavuz. 2024. Counter denial of service for next-generation networks within the artificial intelligence and post-quantum era. *arXiv preprint arXiv:2408.04725* (2024).

[12] Ugandhar Dasi and Nikhil Singla. 2024. Analyzing the Security and Privacy Challenges in Implementing Ai and Ml Models in Multi-Tenant Cloud Environments. *Intl. Journal of Multidisciplinary Inno. and Research Methodology* 3, 2 (2024).

[13] Kha Dinh Duy, Taehyun Noh, Siwon Huh, and Hojoon Lee. 2021. Confidential machine learning computation in untrusted environments: A systems security perspective. *IEEE Access* 9 (2021), 168656–168677.

[14] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[15] Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential privacy for deep and federated learning: A survey. *IEEE access* 10 (2022), 22359–22380.

[16] Peter Elger and Eóin Shanaghy. 2020. *AI as a Service: Serverless machine learning with AWS*. Manning Publications.

[17] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41 annual ACM symposium on Theory of computing*. 169–178.

[18] Tim Geppert, Stefan Deml, and David Sturzenegger. 2022. Trusted execution environments: Applications and organizational challenges. *Frontiers in CS* (2022).

[19] Thang Hoang, Rouzbeh Behnia, Yeongjin Jang, and Attila A Yavuz. 2020. MOSE: Practical multi-user oblivious storage via secure enclaves. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. 17–28.

[20] Balaram Yadav Kasula. 2021. AI-Driven Innovations in Healthcare: Improving Diagnostics and Patient Care. *International Journal of Machine Learning and Artificial Intelligence* 2, 2 (2021), 1–8.

[21] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, and Jiquan Pei. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[22] Feng Li and Bruce McMillin. 2014. A survey on zero-knowledge proofs. In *Advances in computers*. Vol. 94. Elsevier, 25–69.

[23] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[24] Hongyu Liu and Bo Lang. 2019. Machine learning and deep learning methods for intrusion detection systems: A survey. *applied sciences* 9, 20 (2019), 4396.

[25] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, and Jinbo Xiong. 2020. Privacy and security issues in deep learning: A survey. *IEEE Access* 9 (2020), 4566–4593.

[26] Mhlambululi Mafu. 2024. Advances in artificial intelligence and machine learning for quantum communication applications. *IET Quantum Communication* (2024).

[27] B. McMahan and D. Ramage. 2017. *Google AI Blog: Federated Learning: Collaborative Machine Learning without Centralized Training Data*.

[28] Sina Mohseni, Haotao Wang, Chaowei Xiao, and Zhiding Yu. 2022. Taxonomy of machine learning safety: A survey and primer. *Comput. Surveys* 55, 8 (2022).

[29] Chaoyue Niu, Fan Wu, Shaojie Tang, Shuai Ma, and Guihai Chen. 2020. Toward verifiable and privacy preserving machine learning prediction. *IEEE Transactions on Dependable and Secure Computing* 19, 3 (2020), 1703–1721.

[30] Prajwal Panzade and Daniel Takabi. 2023. FENet: Privacy-preserving neural network training with functional encryption. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*. 33–43.

[31] Robert Podschwadt, Daniel Takabi, Peizhao Hu, and Mohammad H Rafiei. 2022. A survey of deep learning architectures for privacy-preserving machine learning with fully homomorphic encryption. *IEEE Access* 10 (2022), 117477–117500.

[32] Sakib Shahriar, Sonal Allana, Seyed Mehdi Hazratifard, and Rozita Dara. 2023. A survey of privacy risks and mitigation strategies in the Artificial Intelligence life cycle. *IEEE Access* 11 (2023), 61829–61854.

[33] Muhammad Tayyab, Mohsen Marjani, NZ Jhanjhi, and Ibrahim Abaker Targio Hashem. 2023. A comprehensive review on deep learning algorithms: Security and privacy issues. *Computers & Security* 131 (2023), 103297.

[34] Elizabeth Nathania Witanto, Yustus Eko Oktian, and Sang-Gon Lee. 2022. Toward data integrity architecture for cloud-based AI systems. *Symmetry* 14, 2 (2022).

[35] Runhua Xu, James BD Joshi, and Chao Li. 2019. Cryptonn: Training neural networks over encrypted data. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1199–1209.

[36] Hang Yang, Xunbo Li, and Zhigui Liu. 2023. Improved differential privacy noise mechanism in quantum machine learning. *IEEE Access* 11 (2023), 50157–50164.

[37] Attila A Yavuz, Saif E Nouma, Thang Hoang, Duncan Earl, and Scott Packard. 2022. Distributed cyber-infrastructures and artificial intelligence in hybrid post-quantum era. In *2022 IEEE 4th International Conference on TPS-ISA*. IEEE.

[38] Ian Zhou, Farzad Tofigh, Massimo Piccardi, and Mehran Abolhasan. 2024. Secure Multi-Party Computation for Machine Learning: A Survey. *IEEE Access* (2024).