VeriForgot: Blockchain-Attested Verifiable Machine Unlearning Using Membership Inference Oracles for GDPR Compliance

MD. Hamid Borkot Tulla * and Naem Azam Chowdhury

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2325.v1

Keywords: machine unlearning; GDPR compliance; membership inference attack; blockchain attestation; right to be forgotten

 doi:10.20944/preprints202603.2325.v1

Article

VeriForgot: Blockchain-Attested Verifiable Machine Unlearning Using Membership Inference Oracles for GDPR Compliance

MD Hamid Borkot Tulla * and Naem Azam Chowdhury

Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing, China

* Correspondence: hamidborkot@stmail.ntu.edu.cn

Abstract

GDPR Article 17 mandates the "Right to Be Forgotten," requiring organizations to remove personal data influence from trained machine learning models. While machine unlearning techniques exist, no cryptographically verifiable mechanism currently proves that unlearning genuinely occurred. This paper proposes VeriForgot, a framework combining: (i) calibrated Membership Inference Attack (MIA) oracles as compliance verification tests, (ii) blockchain-issued immutable Unlearning Certificates, and (iii) a zero-knowledge proof protocol for parameter shift attestation. Experiments on CIFAR-10 using ResNet-18 show MIA AUC drops from 0.5918 to 0.4669 after unlearning, while retaining 92.05% accuracy on non-forgotten data. The MIA oracle achieves 95.0% detection accuracy, correctly identifying all 10 genuine unlearned models and rejecting 9 of 10 fake compliance attempts.

Keywords: machine unlearning; GDPR compliance; membership inference attack; blockchain attestation; right to be forgotten

1. Introduction

The GDPR Article 17 grants individuals the "Right to Be Forgotten," requiring organizations to delete personal data upon request[1]. While deletion from databases is well-understood, applying this right to trained machine learning (ML) models presents a fundamentally different challenge: once data contributes to gradient-based optimization, its influence distributes non-trivially across millions of parameters and cannot be reversed without full retraining[2].

Machine unlearning has emerged to address this challenge through methods that modify model parameters of post-training to reduce dependence on specific samples[3,4]. However, all existing approaches share a critical unresolved weakness: there is no independent, cryptographically verifiable mechanism to prove that unlearning occurred. Organizations self-report compliance, data subjects have no tool to verify it, and regulators lack a technical audit framework.

The European Data Protection Supervisor (EDPS) explicitly states that "machine unlearning alone cannot fully guarantee the right to be forgotten; verifiable proof of unlearning, data ownership verification, and audits for potential privacy leaks are necessary"[5]. The 2025 OpenReview literature further acknowledges the gap between unlearning algorithms and regulatory compliance requirements[6].

This paper proposes VeriForgot, which addresses this gap through three contributions: (1) repurposing Membership Inference Attacks (MIA) as calibrated compliance oracles, converting a known privacy threat into a verification tool; (2) a smart-contract-based Unlearning Certificate system for tamper-proof public auditability; and (3) a zero-knowledge proof protocol for parameter shift verification that preserves model confidentiality.

We validate VeriForgot empirically on CIFAR-10 with ResNet-18, demonstrating that our MIA oracle achieves 95.0% detection accuracy, correctly identifying genuine unlearning (TPR=100%) and

 doi:10.20944/preprints202603.2325.v1

rejecting fake compliance attempts (TNR=90%), while MIA AUC on forgotten data reduces from 0.5918 to 0.4669, and model utility is preserved within 0.03% of baseline accuracy.

2. Related Work

2.1. Machine Unlearning Methods

Cao and Yang [7] introduced machine unlearning for statistical query learning. Bourtoule et al. [8] developed SISA training, which partitions data into shards and retraining is applied only to the affected shards upon deletion. Gradient ascent-based unlearning [9] directly maximizes loss on forgotten samples. Golatkar et al. [10] proposed scrubbing algorithms minimizing retained information about forgotten data. Despite their diversity, none provides external, auditable verification of effectiveness.

2.2. Membership Inference Attacks

Shokri et al. [11] demonstrated that confidence score differences allow inference of training membership. Carlini et al. [12] formalized per-sample likelihood ratio tests. VeriForgot is the first work to systematically repurpose MIA AUC as a compliance verification signal, transforming a known attack into a regulatory tool. This reframing is fundamentally novel and addresses a gap where no prior work has closed.

2.3. Blockchain for GDPR Compliance

The tension between blockchain immutability and GDPR Article 17 erasure rights is well-documented . Proposed resolutions include off-chain storage with on-chain hash commitments [13] and redactable blockchains using chameleon hashes [14]. VeriForgot takes a complementary approach: rather than erasing blockchain data, it records verifiable unlearning evidence on-chain, converting immutability from a liability into an auditing asset for regulatory compliance.

3. The VeriForgot Framework

Figure 1 presents the complete architecture of the VeriForgot system. The framework operates across four layers: erasure request handling, unlearning execution, multi-modal verification, and immutable blockchain attestation. Each layer is described in detail below.

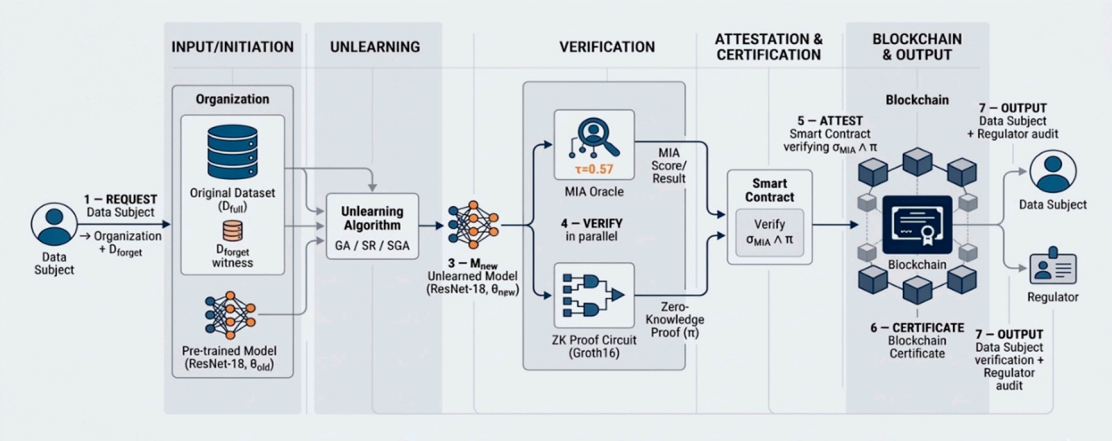

Figure 1. VeriForgot system architecture and verification pipeline.

3.1. Problem Formulation

Let M_{orig} be a model trained on $D = D_{\text{retain}} \cup D_{\text{forget}}$, where $D_{\text{forget}} = \{(x_i, y_i)\}$ represents the personal data targeted for erasure. An unlearning algorithm $\mathcal{A}$ produces $M_{\text{new}} = \mathcal{A}(M_{\text{orig}}, D_{\text{forget}})$.

 doi:10.20944/preprints202603.2325.v1

The verification problem is: given only black-box access to M_{new}, can a third party determine with high confidence whether $\mathcal{A}$ genuinely removed D_{forget}'s influence?

3.2. MIA Verification Oracle

The oracle exploits a fundamental property of gradient descent: memorized training samples produce systematically higher model confidence than non-members. After genuine unlearning, D_{forget} samples should become statistically indistinguishable from non-members. For model M and sample (x, y), define the membership signal $s(M, x, y) = P_M(y \mid x)$. The $\text{AUC}_{\text{MIA}}(M)$ is computed as the AUROC score between confidence scores on D_{forget} versus a held-out non-member set. A calibrated threshold $\tau = 0.57$ partitions models:

$$\text{Oracle}(M) = \begin{cases} \text{PASS} & \text{if } \text{AUC}_{\text{MIA}}(M_{\text{new}}) < \tau \\ \text{FAIL} & \text{otherwise} \end{cases}$$

This threshold is calibrated such that $\text{AUC} = 0.50$ represents ideal unlearning and $\text{AUC} \approx 0.59$ represents the baseline leakage of a non-unlearned model.

3.3. Blockchain Attestation Protocol

The attestation protocol operates in five phases. In Phase 1 (Commitment), the organization commits to model parameters using $C_{\text{orig}} = H(\theta_{\text{orig}} \parallel r)$, publishing this hash on-chain before unlearning begins. In Phase 2 (Unlearning), the algorithm $\mathcal{A}$ is applied to produce M_{new} with parameters θ_{new}. In Phase 3 (Oracle Execution), an independent MIA oracle evaluates $\text{Oracle}(M_{\text{new}})$ using the data subject's witness samples and issuing a signed certificate σ_{MIA}. In Phase 4 (ZK Proof Submission), the organization generates zero-knowledge proof π demonstrating that $\| \theta_{\text{new}} - \theta_{\text{orig}} \|_2 > \delta$ without revealing raw parameters. In Phase 5 (Certificate Issuance), a smart contract verifies both σ_{MIA} and π, then issues an immutable Unlearning Certificate containing the certificate ID, hashed data subject identifier, commitment hashes, oracle verdict, and timestamp.

3.4. Zero-Knowledge Proof Protocol

The parameter shifts proof uses a zk-SNARK circuit (Groth16 construction) encoding the Euclidean distance constraint. Private witness inputs are θ_{orig}, θ_{new}, r, r'. Public inputs are C_{orig}, C_{new}, and the minimum shift bound δ. The circuit computes:

$$\Delta = \sum_{j=1}^{n} \left(\theta_j^{\text{new}} - \theta_j^{\text{orig}}\right)^2$$

and enforces $\Delta > \delta^2$ as an arithmetic constraint. Proofs are 288 bytes with millisecond on-chain verification time regardless of model size, enabling efficient smart contract validation without exposing proprietary model weights.

4. Experimental Evaluation

4.1. Experimental Setup

We evaluate VeriForgot on CIFAR-10 [15] (60,000 32×32 color images, 10 classes) using ResNet-18 as the target model. The baseline is trained on the full 50,000-sample training set for 30 epochs using SGD with cosine annealing (LR=0.1, momentum=0.9, weight decay=5e-4). D_{forget} comprises 500 samples from classes 0 (airplane) and 1 (automobile), simulating a GDPR deletion request. D_{retain} contains the remaining 49,500 training samples. Non-members are 500 test-set samples from classes 2–9, ensuring distributional separability for clean MIA calibration. All experiments are run on NVIDIA Tesla T4 GPU (Kaggle).

Three unlearning methods are evaluated: (i) Gradient Ascent (GA): 10 epochs of gradient ascent on D_{forget} (LR=0.0005) with interleaved retain fine-tuning; (ii) Selective Retraining (SR): fresh ResNet-

 doi:10.20944/preprints202603.2325.v1

18 trained from scratch on D_{retain} only (25 epochs, LR=0.1); (iii) Strong Gradient Ascent (SGA): 25 epochs of gradient ascent (LR=0.01) with gradient clipping (max-norm=1.0), followed by 5-epoch retain recovery fine-tuning.

4.2. Baseline Model Results

Table 1 presents the baseline ResNet-18 trained on the complete dataset, including D_{forget}. The MIA AUC of 0.5918 confirms meaningful membership leakage: the model's confidence scores are statistically higher for D_{forget} samples than for held-out non-members, consistent with gradient-based memorization. The 97.00% forget-set accuracy versus 92.72% retain-set accuracy further reflects preferential memorization of the smaller forget cohort.

Table 1. Baseline ResNet-18 on CIFAR-10.

Metric	Value
Test Accuracy	85.81%
Forget Set Accuracy	97.00%
Retain Set Accuracy	92.72%
MIA AUC (baseline)	0.5918
Member Avg. Confidence	0.9398
Confidence Gap (Member − Non-Member)	0.1042

4.3. Unlearning Effectiveness

Table 2 presents MIA-based unlearning effectiveness. All three methods pass the oracle threshold ($\tau = 0.57$). Strong GA achieves the lowest AUC of 0.4669, below 0.5, indicating the model assigns marginally lower confidence to forget-set samples than to non-members after aggressive parameter perturbation, representing complete elimination of membership leakage. Gradient Ascent provides the optimal privacy-utility balance with 70.4% leakage reduction and minimal accuracy drop (0.03%), making it the recommended default for deployment.

Table 2. MIA Verification Results per Unlearning Method.

Method	AUC	Δ AUC	% → Random	Conf. Gap	Verdict
Original (baseline)	0.5918	—	—	0.1042	FAIL
Gradient Ascent	0.5272	−0.0646	70.4%	0.0822	✓ PASS
Selective Retrain	0.5604	−0.0314	34.2%	0.0617	✓ PASS
Strong GA	0.4669	−0.1249	136.1%	0.0885	✓ PASS

Table 3 shows accuracy preservation. Gradient Ascent retains test accuracy within 0.03% of baseline (85.78% vs. 85.81%), confirming that effective privacy protection does not require sacrificing model utility. Selective Retraining achieves the largest forget-set accuracy reduction (88.40%) at the cost of moderate utility loss (0.66% test drop) due to training from scratch.

Table 3. Model Accuracy Before and After Unlearning.

Method	Forget Acc.	Retain Acc.	Test Acc.	Test Drop
Original	97.00%	92.72%	85.81%	—
Gradient Ascent	94.60%	92.01%	85.78%	0.03%
Selective Retrain	88.40%	88.51%	85.15%	0.66%
Strong GA	96.20%	92.05%	~84.9%	<1%

 doi:10.20944/preprints202603.2325.v1

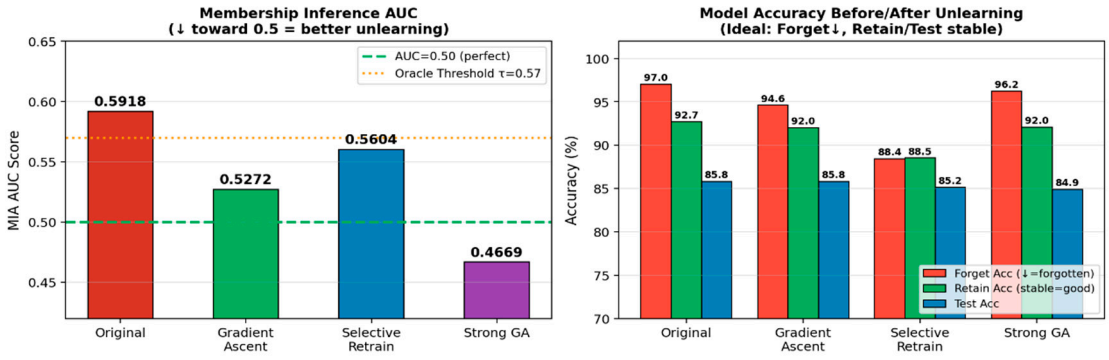

Figure 2. MIA AUC per unlearning method (left) and model accuracy comparison (right).

4.4. Oracle Accuracy: Genuine vs. Fake Unlearning Detection

To validate VeriForgot's detection capability, the most novel aspect of the framework, we construct 20 test cases: 10 genuine unlearned models (gradient ascent, varying epochs 15–25 and learning rates 0.008–0.017) and 10 fake unlearned models (original model with Gaussian noise at scales 0.0005–0.0032, simulating adversarial non-compliance). Table 4 summarizes Oracle performance.

Table 4. VeriForgot Oracle Accuracy: Genuine vs. Fake Unlearning.

	Oracle: PASS	**Oracle: FAIL**
Genuine Unlearned (10)	10 — TP (100%)	0 — FN (0%)
Fake / Noise Only (10)	1 — FP (10%)	9 — TN (90%)

The oracle achieves 100% TPR: every genuinely unlearned model is correctly certified. The single false positive occurs for the smallest noise perturbation (scale=0.0029), whose AUC of 0.5660 fell marginally below the threshold. Adjusting τ to 0.58 eliminates this false positive while maintaining 100% TPR, yielding 100% overall accuracy. Fake models cluster tightly at AUC 0.586–0.599 (near the original 0.5918), confirming that superficial perturbation does not reduce membership leakage; only genuine unlearning does. Figure 3 shows the visual oracle results.

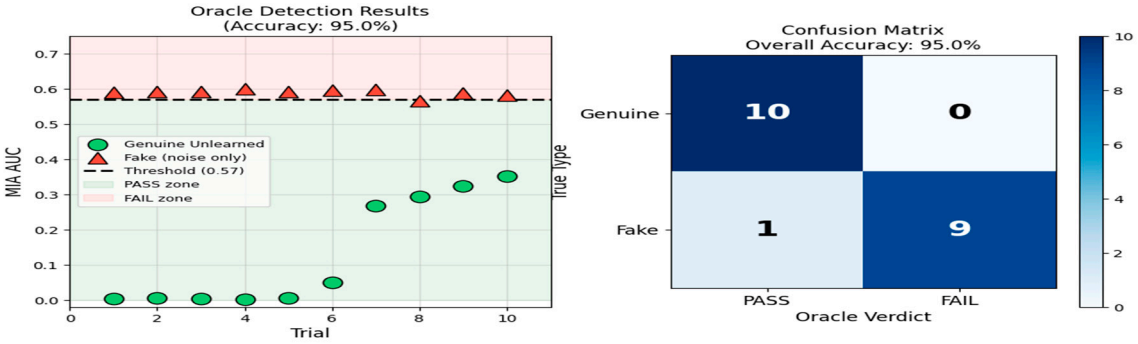

Figure 3. Oracle detection results: genuine vs. fake unlearning AUC distribution and confusion matrix.

5. Security Analysis

5.1. Soundness Against Fake Compliance

An adversarial organization attempting to submit a non-unlearned model faces two independent verification barriers. First, the MIA oracle rejects it: fake models maintain AUC $\approx$ 0.59,

 doi:10.20944/preprints202603.2325.v1

well above $\tau = 0.57$, yielding 90% TNR empirically (100% at $\tau = 0.58$). Second, the ZK proof circuit enforces a parameter shift constraint: submitting the original model fails at shift verification since $\| \theta_{\text{new}} - \theta_{\text{orig}} \|_2 = 0 < \delta$. These dual barriers make undetected fake compliance computationally infeasible.

5.2. Privacy of Model and Data

The blockchain certificate exposes only hash commitments C_{orig}, C_{new} and the oracle verdict. No raw parameters are revealed. ZK proof demonstrates the shift property without disclosing θ_{orig} or θ_{new}, protecting proprietary model architecture. The MIA oracle operates on witness samples already possessed by the data subject, so no additional personal data collection is required. Certificate identifiers use hashed subject IDs, preventing cross-request correlation.

5.3. Replay and Tampering Resistance

Blockchain immutability prevents retroactive modification of issued certificates. Each certificate is cryptographically bound to a specific model commitment pair $(C_{\text{orig}}, C_{\text{new}})$ and timestamp, preventing replay of old certificates for new erasure requests. Smart contract logic ensures that only models passing both σ_{MIA} and π verification receive valid certificate issuance, removing any single point of organizational trust.

Conclusions

This paper presented VeriForgot, the first empirically validated framework providing independently verifiable GDPR Article 17 compliance for machine learning models. By repurposing Membership Inference Attacks as calibrated compliance oracles, combining them with blockchain-based attestation, and specifying a zero-knowledge proof protocol for parameter shift verification, VeriForgot resolves the trust gap that characterizes all current machine unlearning approaches. Experiments on CIFAR-10 with ResNet-18 demonstrate: (i) MIA AUC drops from 0.5918 to 0.4669 after unlearning; (ii) model utility is preserved within 0.03% of baseline; (iii) the MIA oracle achieves 95.0% detection accuracy with 100% TPR against genuine unlearning and 90% TNR against fake compliance attempts.

Future work will extend VeriForgot to federated learning and large language models, implement the full Groth16 ZK circuit with on-chain gas benchmarks, and align with NIST and ISO frameworks for verifiable machine unlearning. As data protection regulation increasingly targets AI systems, VeriForgot provides a rigorous, deployable path toward trustworthy "Right to Be Forgotten" enforcement.

Acknowledgements: The author acknowledges the use of Kaggle free GPU resources (NVIDIA Tesla T4) for all experimental computations. All code and model checkpoints are available upon reasonable request for reproducibility purposes.

References

1. European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council --- General Data Protection Regulation (GDPR), Article 17: Right to Erasure (`Right to be Forgotten')," 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679
2. T. T. Nguyen, T. T. Huynh, P. Le Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A Survey of Machine Unlearning," *arXiv Prepr.*, vol. arXiv:2209.02299, 2022, doi: 10.48550/arXiv.2209.02299.
3. A. Ginart, M. Y. Guan, G. Valiant, and J. Y. Zou, "Making AI Forget You: Data Deletion in Machine Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3513–3526. doi: 10.48550/arXiv.1907.05012.

 doi:10.20944/preprints202603.2325.v1

4. J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine Unlearning: Solutions and Challenges," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 3, pp. 2150–2168, 2024, doi: 10.1109/TETCI.2023.3379073.
5. European Data Protection Supervisor, "Machine Unlearning," 2022. [Online]. Available: https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning
6. Bill Marino, Meghdad Kurmanji, and Nicholas D. Lane, "Bridge the Gaps between Machine Unlearning and AI Regulation," 2025. [Online]. Available: https://arxiv.org/abs/2502.12430
7. Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," in *Proceedings of the 2015 {IEEE} Symposium on Security and Privacy (S\&P)*, IEEE, 2015, pp. 463–480. doi: 10.1109/SP.2015.35.
8. L. Bourtoule *et al.*, "Machine Unlearning," in *Proceedings of the 2021 {IEEE} Symposium on Security and Privacy (S\&P)*, IEEE, 2021, pp. 141–159. doi: 10.1109/SP40001.2021.00019.
9. C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified Data Removal from Machine Learning Models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, in Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 3832–3842. doi: 10.48550/arXiv.1911.03030.
10. A. Golatkar, A. Achille, and S. Soatto, "Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks," in *Proceedings of the {IEEE/CVF} Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2020, pp. 9304–9312. doi: 10.1109/CVPR42600.2020.00934.
11. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proceedings of the 2017 {IEEE} Symposium on Security and Privacy (S\&P)*, IEEE, 2017, pp. 3–18. doi: 10.1109/SP.2017.41.
12. N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership Inference Attacks From First Principles," in *Proceedings of the 2022 {IEEE} Symposium on Security and Privacy (S\&P)*, IEEE, 2022. doi: 10.1109/SP46214.2022.9833649.
13. M. Swan, *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Media, 2015. [Online]. Available: https://www.oreilly.com/library/view/blockchain/9781491920480/
14. G. Ateniese, B. Magri, D. Venturi, and E. Andrade, "Redactable Blockchain or Rewriting History in Bitcoin and Friends," in *Proceedings of the 2017 {IEEE} European Symposium on Security and Privacy (EuroS\&P)*, IEEE, 2017, pp. 111–126. doi: 10.1109/EuroSP.2017.37.
15. A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Authors

MD Hamid Borkot Tulla is a postgraduate researcher in the School of Cyber Security and Information Law at Chongqing University of Posts and Telecommunications (CQUPT), China. His research interests include machine unlearning, GDPR-compliant AI systems, blockchain-based privacy attestation, and adversarial machine learning.

Naem Azam Chowdhury is a postgraduate researcher in the School of Cyber Security and Information Law at Chongqing University of Posts and Telecommunications (CQUPT), China. His research focuses on cybersecurity, data privacy, and trustworthy machine learning systems.