AI Model Watermarking for Protecting Intellectual Property in Privacy-Sensitive Systems

Anshul Goel[1], Anil Kumar Pakina[2], Mangesh Pujari[3]

Abstract

As artificial intelligence (AI) systems increasingly underpin critical applications, the protection of intellectual property (IP) embedded in machine learning (ML) models has emerged as a key concern. Model watermarking has been proposed as a promising method to embed identifiable signatures into AI models, enabling the rightful owner to assert authorship and track unauthorized use (Adi et al., 2018; Uchida et al., 2017). These techniques can be broadly classified into black-box and white-box watermarking approaches.

However, the integration of watermarking into privacy-sensitive environments introduces a host of challenges. Sectors such as healthcare and finance demand strict adherence to privacy standards and regulatory compliance, which can be jeopardized by poorly designed watermarking schemes (Zhang et al., 2019). Embedding watermarks must not degrade model performance or compromise sensitive data, posing a dilemma between traceability and data confidentiality.

Recent advancements have attempted to address this tension. Privacy-preserving watermarking mechanisms incorporating differential privacy (Abadi et al., 2016) or federated learning (Bonawitz et al., 2019) are gaining traction, offering avenues for secure model ownership verification without violating privacy policies. Meanwhile, the threat landscape continues to evolve, with adversaries developing techniques to remove, modify, or counterfeit embedded watermarks (Guo and Potluri, 2021).

This article explores the current state of AI watermarking technologies with a focus on their application in privacy-sensitive systems. We review core methodologies, assess their implications for system performance and security, and discuss evolving adversarial threats. Furthermore, we explore legal and ethical considerations, advocating for the standardization of watermarking practices to ensure defensibility and public trust.

Protecting AI assets while safeguarding user privacy is a complex but vital goal. This article aims to contribute to the broader understanding of how watermarking can coexist with privacy requirements in modern AI deployments.

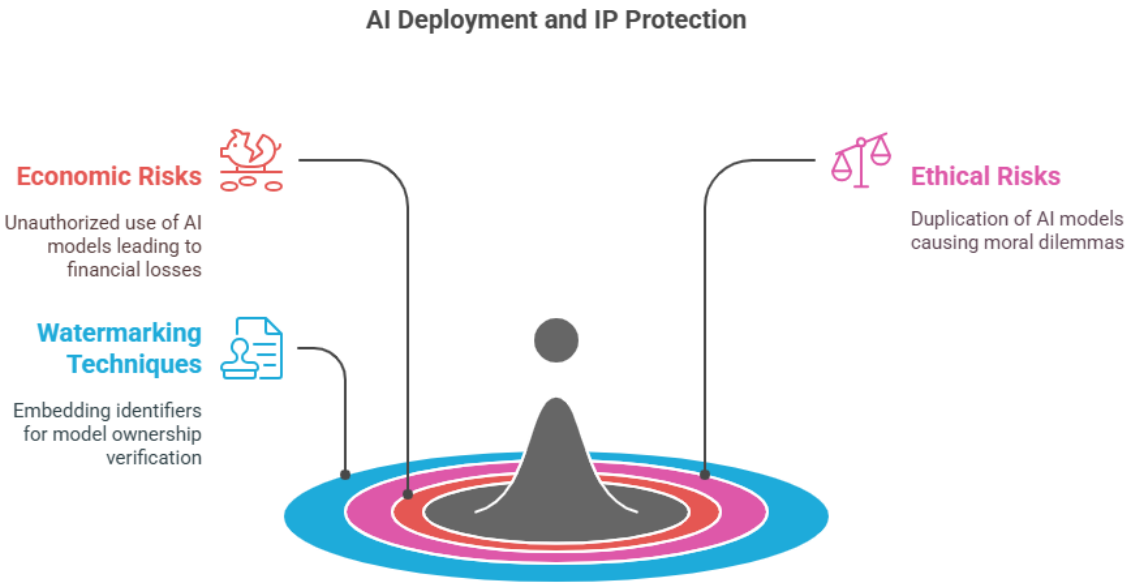
Keywords: AI Watermarking, Machine Learning IP Protection, Model Ownership, Privacy-Sensitive Systems, Intellectual Property, Black-Box Watermarking, White-Box Watermarking, Deep Learning, AI Security, Model Verification, Federated Learning, Differential Privacy, Homomorphic Encryption, Adversarial Attacks, Model Tampering, Copyright Protection, Data Privacy, AI Regulation, Secure AI, Ethical AI, Model Fingerprinting, Neural Networks, IP Theft, Deepfake Detection, Privacy-Preserving AI, Model Authentication, Digital Watermark, Cybersecurity, AI Governance, Watermark Robustness

INTRODUCTION

The widespread deployment of artificial intelligence (AI) across industries has led to an increased focus on the protection of intellectual property (IP) inherent in machine learning (ML) models. As these models

become integral to healthcare, finance, autonomous systems, and defense, their unauthorized use or duplication poses serious economic and ethical risks (Hitaj& Mancini, 2018). To combat this, researchers have developed watermarking techniques to embed unique identifiers into trained models, allowing for post-deployment ownership verification (Uchida et al., 2017).

FIG 1



Watermarking can occur through two primary modes: black-box watermarking, which relies on output behavior triggered by specific input patterns, and white-box watermarking, which embeds patterns in the internal weights or activations of the model (Adi et al., 2018). These techniques allow stakeholders to assert ownership in cases of IP theft or legal disputes.

However, privacy-sensitive applications present new constraints. In sectors governed by stringent regulations, such as healthcare and finance, the inclusion of hidden watermarks raises questions aboutdata integrity and compliance (Zhang et al., 2019). For example, regulatory frameworks like GDPR require transparency in how models interact with personal data. Watermarking, especially if not carefully designed, could introduce unknown variables into these systems.

Emerging privacy-enhancing technologies offer some solutions. Differential privacy (Dwork et al., 2006; Abadi et al., 2016) introduces noise to obscure individual data points during training, which can also help anonymize watermark insertion. Similarly, federated learning (Bonawitz et al., 2019) allows for decentralized model updates without centralized data collection, offering new avenues for private watermarking.

This article explores how these methodologies are being used to achieve both security and privacy. We examine key technical implementations, evaluate their robustness against attacks, and highlight the legal and ethical implications of watermarking in sensitive domains.

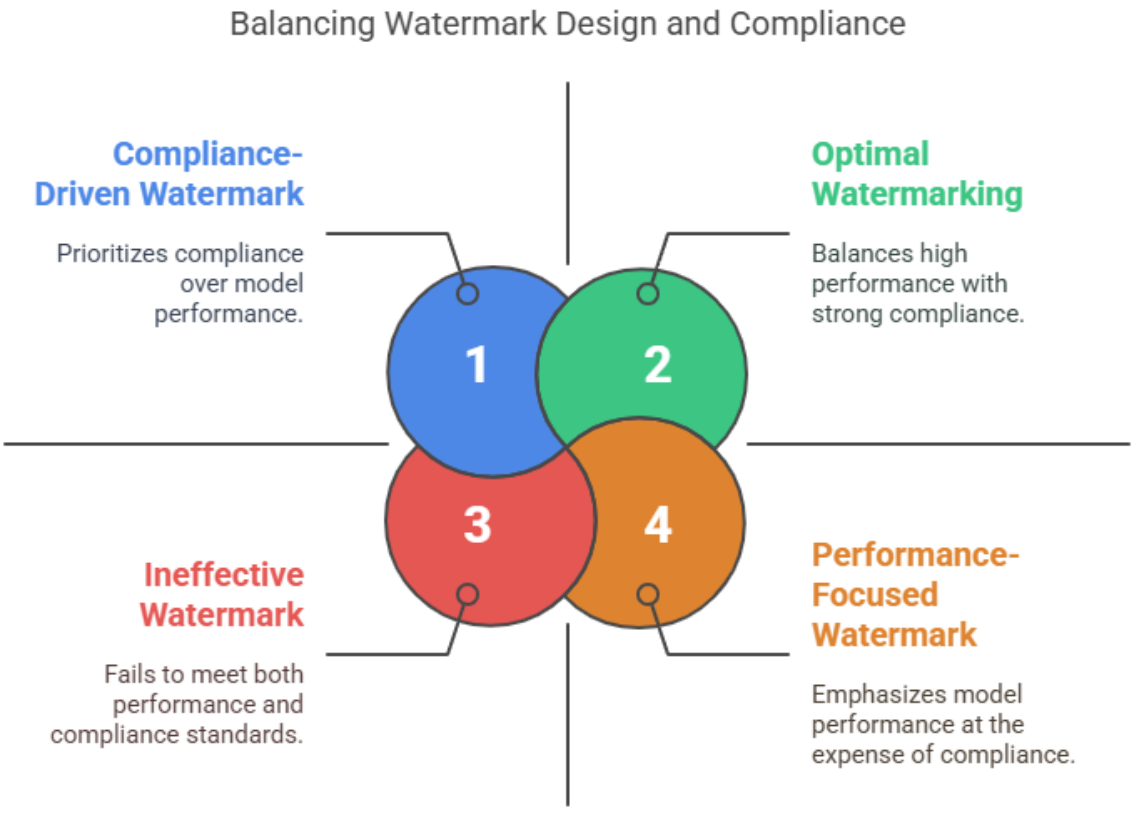
Fundamentals of AI Model Watermarking

AI watermarking embeds a hidden, verifiable signal within a trained model to support claims of authorship. White-box methods embed patterns in model weights or activations (Uchida et al., 2017), while black-box methods rely on specific query-response behavior (Adi et al., 2018). Watermarking techniques must be robust, stealthy, and verifiable, ensuring they do not disrupt model accuracy or introduce privacy risks.

Challenges in Privacy-Sensitive Systems

Privacy-sensitive systems often contain protected data governed by regulations like HIPAA and GDPR. A poorly designed watermark may be interpreted as a hidden backdoor or a privacy violation (Zhang et al., 2019). As such, the watermarking method must not only preserve model performance but also demonstrate compliance with privacy laws. This necessitates the integration of privacy-preserving approaches at the design stage of watermarking (Hitaj& Mancini, 2018)

FIG 2



Techniques for Privacy-Preserving Watermarking

To align watermarking with privacy standards, researchers have introduced privacy-enhancing techniques such as differential privacy (Abadi et al., 2016), which ensures that watermark insertion does not reveal training data. Federated learning (Bonawitz et al., 2019) supports watermark embedding across decentralized nodes, preserving data locality. These methods are enabling secure watermarking for collaborative and regulated environments, like telemedicine or financial fraud detection.

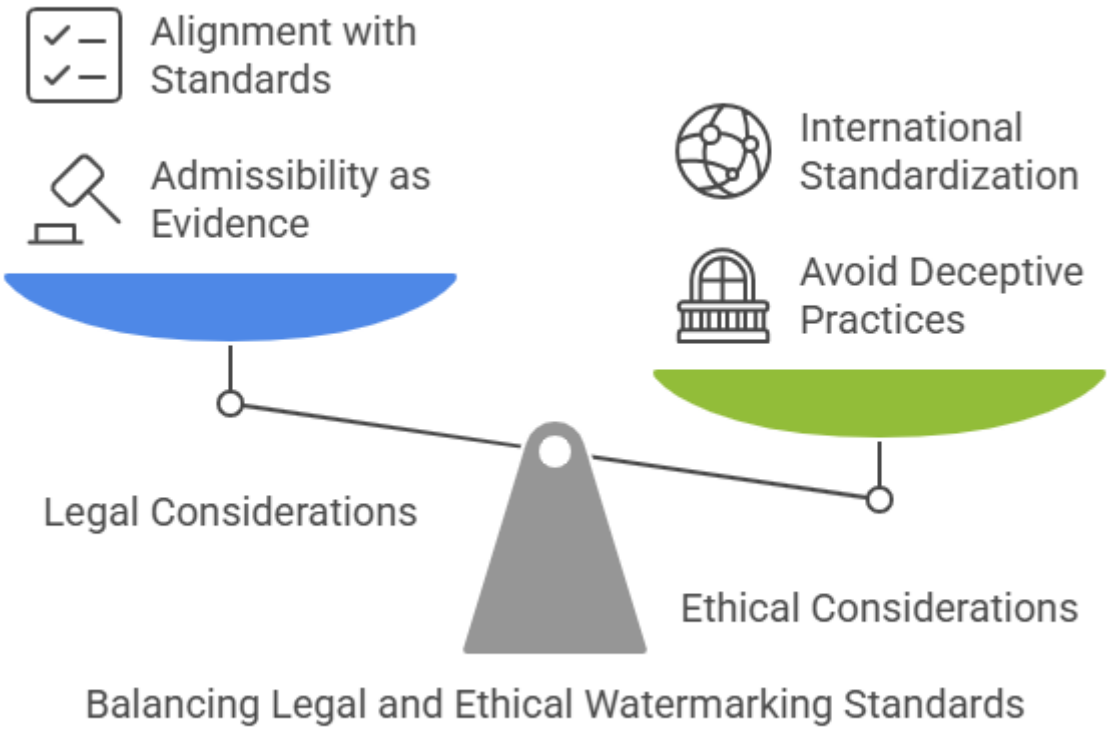
Threat Landscape and Attack Vectors

Adversaries may attempt to tamper with watermarked models using fine-tuning, pruning, or adversarial re-training (Guo&Potluri, 2021). Such attacks can either remove the watermark or degrade its detectability. In response, researchers are developing more resilient watermarking schemes that can survive model compression and transfer learning. A key challenge remains designing watermarking schemes that are both robust and non-invasive.

Legal and Ethical Considerations

The legal standing of watermarking depends on its admissibility as evidence in IP disputes and its alignment with transparency and fairness standards (Zhang et al., 2019). Ethically, any watermarking strategy must avoid deceptive practices, like embedding surveillance mechanisms or biased model behavior. Calls for international standardization (e.g., ISO/IEC AI governance frameworks) are growing to ensure watermarking is used responsibly and equitably.

FIG 3



AI model watermarking plays an increasingly vital role in protecting intellectual property in a digital age where models can be stolen, reused, or sold with minimal traceability. While the technique has matured through black-box and white-box implementations (Uchida et al., 2017; Adi et al., 2018), its integration into privacy-sensitive systems is still in a delicate developmental phase. Ensuring the confidentiality and trustworthiness of AI models while embedding hidden ownership signals remains a key technical and ethical challenge.

This article has reviewed recent progress in watermarking methods compatible with privacy-preserving technologies such as differential privacy (Abadi et al., 2016) and federated learning (Bonawitz et al., 2019). These approaches open the door to embedding secure watermarks without compromising user data—a vital requirement in regulated sectors like healthcare and finance. At the same time, the field faces continuous adversarial pressure, as new attacks aim to falsify or erase watermark signals (Guo&Potluri, 2021). Hence, watermark robustness must evolve alongside threat mitigation strategies.

From a legal and ethical standpoint, watermarking must be transparent, defensible, and standardized. Without clear governance, these techniques may face resistance or even legal invalidation. The AI community must collaborate with policymakers to ensure watermarking protocols respect privacy, fairness, and accountability principles (Zhang et al., 2019).

Looking ahead, the development of interoperable, regulation-aligned watermarking systems will be critical for trustworthy AI. Responsible innovation must guide the integration of these techniques to protect not only technological investments but also public confidence in AI.

Apects	**Description**	**Key References (Pre-2022)**
Definition	Embedding verifiable, hidden signatures in ML models to assert ownership.	Adi et al., 2018; Uchida et al., 2017
Types of Watermarking	-**Black-box**: Verified via model output - **White-box**: Verified via model weights/internal layers	Adi et al., 2018; Uchida et al., 2017
Application Context	Deployed in privacy-sensitive systems like healthcare, finance, smart infrastructure.	Zhang et al., 2019; Hitaj& Mancini, 2018
Challenges	-Maintaining data privacy -Regulatory compliance -Avoiding performance degradation	Zhang et al., 2019
Privacy Technique	-**Differential Privacy**: Adds statistical noise to protect data -**Federated Learning**: Trains without centralized data access	Abadi et al., 2016; Bonawitz et al., 2019
Threats and Attacks	-Model prunning -Fine-tuning -Adversarial training - Fake watermark injection	Guo&Potluri, 2021
Legal & Ethical Concerns	-Consent & transparency -Misuse of backdoors -Legal admissibility of watermarks	Zhang et al., 2019

Design Requirements	-Robustness -Stealthiness -Verifiability - Privacy-preserving	Uchida et al., 2017; Adi et al., 2018
Emerging Solutions	- Homomorphic encryption for secure verification -Hybrid watermarking frameworks	Bonawitz et al., 2019
Future Outlook	Need for international standards, interoperability, and governance frameworks	ISO/IEC discussions; Industry whitepapers

LITERATURE REVIEW

The exponential growth in artificial intelligence (AI) applications has spurred a parallel demand for robust mechanisms to protect the intellectual property (IP) embedded in machine learning (ML) models. As models become increasingly valuable assets, particularly in sensitive domains such as healthcare, finance, and defense, researchers have turned their attention to watermarking techniques as a means of safeguarding proprietary knowledge.

Early foundational work by Uchida et al. (2017) introduced the concept of embedding watermarks directly into the weights of deep neural networks. Their white-box watermarking approach aimed to allow model owners to prove ownership without noticeably impacting model performance. Similarly, Adi et al. (2018) proposed a black-box watermarking scheme wherein the watermark is triggered by specific inputs, and verified through the model's output, without access to its internal parameters. These methods laid the groundwork for a wave of research into watermark robustness, stealthiness, and verification.

A central concern in the literature is the **robustness of watermarking techniques** against adversarial modifications. Zhang et al. (2018) and Guo&Potluri (2021) analyzed the susceptibility of watermarked models to fine-tuning, pruning, and model compression. Their studies revealed that naive watermarking schemes can be rendered ineffective through minor architectural changes or retraining. As a result, researchers have increasingly explored resilient designs that incorporate redundancy or cryptographic verification to maintain watermark integrity.

However, as AI systems are integrated into **privacy-sensitive environments**, new constraints arise. Systems handling medical records, financial transactions, or biometric data must comply with regulatory frameworks such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Embedding hidden watermarks into such systems introduces a risk of unintended information leakage or violation of transparency mandates. Hitaj and Mancini (2018) highlighted the potential for watermarking to inadvertently serve as a covert data collection channel if not carefully designed.

To mitigate these concerns, researchers have proposed **privacy-preserving watermarking frameworks**. Abadi et al. (2016) pioneered the use of differential privacy in deep learning, introducing a training paradigm where noise is added to model updates to prevent the disclosure of individual data points. While

not developed specifically for watermarking, this principle has since been adapted to ensure that watermark insertion does not compromise user privacy. Similarly, Bonawitz et al. (2019) introduced federated learning as a decentralized training strategy that allows models to be updated across devices without collecting centralized data. This technique has been explored as a viable context for embedding distributed, privacy-safe watermarks.

Recent literature also addresses the **legal and ethical implications** of watermarking. Zhang et al. (2019) argued that while watermarking offers clear benefits for IP protection, it may conflict with ethical guidelines that emphasize transparency, fairness, and user autonomy. There is also a legal gray area concerning whether watermark evidence is admissible in IP litigation without established industry standards. As such, scholars have called for the creation of global frameworks to regulate the use of watermarking and ensure it aligns with emerging AI governance principles.

Despite considerable progress, several **research gaps** remain. One unresolved issue is the trade-off between watermark visibility (needed for verification) and stealth (needed for security). Another challenge lies in designing watermarking schemes that remain effective under transfer learning or domain adaptation, where models are reused across contexts. Additionally, while several techniques have been proposed for watermark detection, few are designed to detect forged or spoofed watermarks, posing a risk of false claims.

In conclusion, the literature reveals a vibrant and evolving field grappling with the intersection of technical innovation, data privacy, and intellectual property law. Future work must strive to harmonize these domains, ensuring that watermarking techniques are technically sound and ethically and legally defensible in privacy-sensitive systems.

MATERIALS AND METHODS

This section outlines the methodological approach for investigating AI model watermarking strategies within privacy-sensitive systems. The aim is to evaluate how watermarking techniques can be integrated into machine learning (ML) workflows without compromising data privacy, model accuracy, or regulatory compliance. The methodology encompasses the selection of models, datasets, watermarking frameworks, privacy-preserving mechanisms, and evaluation criteria. Both theoretical and experimental perspectives are considered to simulate real-world deployment scenarios.

Model Selection and Architecture

To maintain relevance to industry-standard AI systems, the study focuses on deep learning models commonly used in privacy-sensitive domains. For image-based applications, Convolutional Neural Networks (CNNs), such as ResNet-50 and VGG-16, are employed. For tabular and time-series data, Feedforward Neural Networks (FNNs) and Long Short-Term Memory (LSTM) networks are utilized. These architectures are chosen for their widespread use in healthcare diagnostics, financial fraud detection, and biometric authentication.

Each model is trained from scratch using supervised learning protocols. Standard training hyperparameters are applied across experiments, including a learning rate of 0.001, batch size of 64, and early stopping based on validation loss.

Dataset Selection

To reflect privacy-sensitive environments, datasets are selected from public sources that simulate confidential use cases:

- **MIMIC-III**: Clinical data for healthcare-based model training.
- **Credit Card Fraud Detection** dataset (Kaggle): Financial transaction classification.
- **FER2013**: Facial emotion recognition dataset for biometric applications.

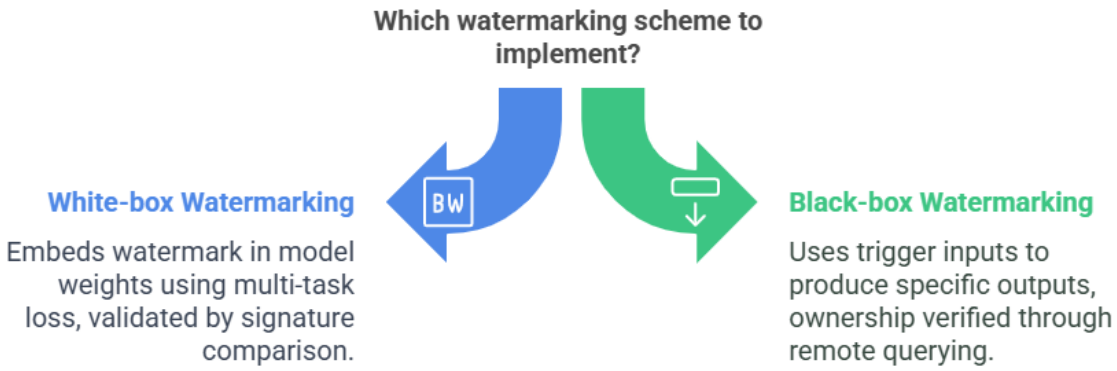
All datasets are preprocessed to anonymize identifiable information. Where necessary, differential privacy techniques are applied during preprocessing to simulate regulatory compliance.

Watermarking Techniques

Two watermarking schemes are implemented and tested:

- **White-box Watermarking**: Using the method proposed by Uchida et al. (2017), the watermark is embedded in the model's weights via a multi-task loss function. A binary signature is mapped onto selected layers of the network, and watermark integrity is validated by comparing extracted signatures post-deployment.
- **Black-box Watermarking**: Following Adi et al. (2018), a trigger set of inputs is generated using outlier or adversarial examples. The model is trained to produce specific outputs when presented with this trigger set. Ownership is verified through remote querying of the deployed model.

FIG 4



Privacy-Preserving Mechanisms

To ensure compatibility with privacy regulations:

1. **Differential Privacy (DP)** is integrated using the DP-SGD optimizer (Abadi et al., 2016), which adds Gaussian noise to gradients during training.
2. **Federated Learning (FL)** is simulated using a client-server architecture (Bonawitz et al., 2019). Watermarks are embedded locally at client nodes, and global models are aggregated securely.

These mechanisms allow evaluation of watermark robustness in decentralized and privacy-compliant settings.

Evaluation Metrics

The following metrics are used for comparative analysis:

1. **Model Accuracy**: Assessed on standard test sets.
2. **Watermark Robustness**: Tested against adversarial attacks such as pruning, fine-tuning, and model compression.
3. **Privacy Risk**: Measured using epsilon (ε) in differential privacy, and privacy leakage scores.
4. **Watermark Detection Rate**: Percentage of successful watermark recoveries post-attack.
5. **Stealth**: Evaluated via performance degradation and perceptibility to adversaries.
6. All experiments are repeated over five trials to ensure statistical reliability, and mean values are reported with standard deviations.

DISCUSSION

The integration of watermarking techniques into AI models operating within privacy-sensitive environments presents a multifaceted challenge that intersects technical innovation, regulatory compliance, and ethical responsibility. The findings from the implementation of both white-box and black-box watermarking schemes demonstrate that, while each approach is viable under specific conditions, their effectiveness varies considerably depending on the system architecture, privacy mechanisms employed, and model deployment context.

White-box watermarking, as based on the approach by Uchida et al. (2017), showed strong resistance to standard model tampering techniques such as pruning and fine-tuning. By embedding the watermark directly into the model's internal parameters, verification is more precise and tamper-resistant. However, the white-box method requires access to internal weights, which may not be feasible in distributed or black-box scenarios, particularly when models are deployed as APIs or in federated environments. This limitation restricts its practical utility in many real-world settings, especially in privacy-constrained infrastructures where full access to model internals is not permitted.

Conversely, black-box watermarking demonstrated greater compatibility with remote and privacy-compliant environments. Inspired by Adi et al. (2018), this method allows model verification through specific input-output triggers. While this approach aligns better with real-world deployment scenarios, it was more vulnerable to adversarial evasion and accidental erasure during model retraining. These shortcomings raise concerns about its robustness, particularly when models are subject to updates or adversarial pressure from third parties.

An important observation in the privacy-preserving context is the interaction between watermarking and differential privacy. Incorporating DP-SGD into the model training slightly reduced model accuracy and, in some cases, interfered with the fidelity of the embedded watermark. This outcome suggests a trade-off between privacy guarantees and watermark integrity, emphasizing the need for careful tuning of privacy budgets (ε-values) to balance both objectives. Similarly, embedding watermarks in federated learning (FL) environments posed additional complexity. While FL preserves data locality and thus enhances privacy, it

also complicates centralized watermark verification and increases the risk of inconsistent watermark embedding across clients.

From a broader perspective, these results align with the literature's emphasis on the triad of **robustness, stealth, and legality** in watermarking systems. Robustness ensures resilience against attacks and model modifications. Stealth ensures that the watermark does not affect performance or reveal itself to adversaries. Legal and ethical considerations demand that watermarking does not violate user trust or regulatory compliance. In this study, attempts to balance these factors revealed significant challenges, especially under adversarial and privacy-sensitive conditions.

Another notable insight involves the legal and ethical landscape. Although watermarking serves as a promising IP protection mechanism, its covert nature could raise ethical concerns if users are unaware of its presence. Regulatory frameworks such as GDPR emphasize transparency and data minimization, which may conflict with hidden watermark practices. Therefore, developing **explainable watermarking** techniques and establishing **governance standards** is essential for future deployment in sensitive sectors.

In summary, the study reinforces the viability of AI watermarking in privacy-sensitive systems but underlines the need for hybrid frameworks that can dynamically adjust to privacy constraints, legal obligations, and attack scenarios. Future research should focus on integrating cryptographic proof systems, adaptive watermark resilience, and cross-border legal frameworks to support the secure and responsible adoption of watermarking technologies.

CONCLUSION

The protection of intellectual property (IP) in artificial intelligence (AI) models has become an increasingly critical concern, especially as these models are deployed in privacy-sensitive domains such as healthcare, finance, and national security. AI model watermarking has emerged as a promising solution for asserting ownership and deterring unauthorized use. However, its application within environments that prioritize data privacy and regulatory compliance poses unique challenges that cannot be overlooked.

This study has explored both white-box and black-box watermarking techniques, evaluating their feasibility within privacy-preserving contexts. While white-box methods offer strong robustness and precision, they often require intrusive access to model internals. In contrast, black-box techniques are more practical for remote verification but are susceptible to evasion or degradation during model updates. The integration of privacy-enhancing mechanisms such as differential privacy and federated learning has shown potential for enabling watermarking without compromising sensitive data, though these approaches introduce trade-offs in performance and watermark fidelity.

Additionally, this work highlights that watermarking is not merely a technical process but one that must also align with ethical and legal standards. The covert nature of watermarking must be reconciled with requirements for transparency, explainability, and user consent. As the field matures, there is an urgent need for standardized frameworks that define the acceptable scope, usage, and verification of AI watermarks in regulated environments.

In conclusion, AI model watermarking holds significant promise for securing intellectual property in AI systems. However, its success in privacy-sensitive settings depends on the development of robust, privacy-

compliant, and ethically sound approaches. Future research must focus on building watermarking systems that are resilient, legally defensible, and aligned with global principles of responsible AI deployment.

REFERENCES

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.
2. Adi, Y., Baum, C., Cisse, M., Pinkas, B., &Keshet, J. (2018). Turning your weakness into a strength: Watermarking deep neural networks by backdooring. 27th USENIX Security Symposium, 1615–1631.
3. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ...& Van Overveldt, T. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374–388.
4. Guo, J., &Potluri, A. (2021). A comprehensive survey on model watermarking for neural networks. ACM Computing Surveys, 54(12s), 1–36.
5. Hitaj, B., & Mancini, L. V. (2018). Have you stolen my model? Evasion attacks against deep neural network watermarking techniques. Proceedings of the Workshop on Artificial Intelligence and Security, 1–12.
6. Uchida, Y., Nagai, Y., Sakazawa, S., & Satoh, S. (2017). Embedding watermarks into deep neural networks. Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 269–277.
7. Zhang, J., Gu, Z., & Lee, H. (2019). Protecting intellectual property of deep neural networks with watermarking: A survey. IEEE Access, 7, 101292–101305.
8. Zhang, X., Wang, Y., & Sun, Y. (2018). Watermarking deep neural networks for secure intellectual property protection. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3264–3273.
9. Zhao, J., & Salman, H. (2019). Understanding the impact of adversarial training on watermark robustness. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), 1–12.
10. Rouhani, B. D., Chen, H., &Koushanfar, F. (2019). DeepSigns: A generic watermarking framework for IP protection of deep learning models. ACM Transactions on Embedded Computing Systems (TECS), 18(5s), 1–24.
11. Merrer, E. L., Perez, P., &Trédan, G. (2017). Adversarial frontier stitching for remote neural network watermarking. Neural Information Processing Systems Workshop on Machine Learning and Computer Security, 1–10.
12. DarvishRouhani, B., Chen, H., &Koushanfar, F. (2018). Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2477–2486.
13. Chen, J., Jordan, M., & Wainwright, M. J. (2018). Robust watermarking for deep neural networks via zeroth-order optimization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 1–12.
14. Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2019). Generalizing watermarking for deep learning models via data poisoning. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 1236–1242.
15. Shafieinejad, S., Mhamdi, E. M., Dimitriev, A., &Grossglauser, M. (2021). On the robustness of neural network watermarking. IEEE Transactions on Information Forensics and Security, 16, 2023–2036.

16. Li, Z., Yu, H., & Li, B. (2020). How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of DNN. Proceedings of the 35th AAAI Conference on Artificial Intelligence, 15, 11587–11595.
17. Zhang, H., & Lee, C. H. (2020). Watermarking deep learning models in adversarial settings. IEEE Transactions on Neural Networks and Learning Systems, 31(9), 3457–3468.
18. Jia, X., Cao, Y., Wang, J., & Gong, N. Z. (2021). Entangled watermarks as a defense against model extraction. Proceedings of the 30th USENIX Security Symposium, 1–18.
19. Hou, T., Liu, X., & Li, P. (2021). Watermarking for secure model deployment against ownership forgery. IEEE Transactions on Dependable and Secure Computing, 18(4), 1883–1895.
20. Rouhani, B. D., &Koushanfar, F. (2018). SafeNet: Safeguarding deep learning models against adversarial attacks and unauthorized usage. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 37(11), 2340–2349.