Cryptographic Fingerprinting for Medical AI: A Proof-of-Concept Approach to Protecting Healthcare ML Models from API Extraction

Saaketh Bhojanam
Illinois Mathematics and Science Academy
sbhojanam@imsa.edu

Sohum Mehta
Illinois Mathematics and Science Academy
smehta@imsa.edu

Abstract

We present a cryptographic fingerprinting approach to protect valuable medical AI models from API-based extraction attacks. Our method embeds a detectable watermark into the uncertainty pattern of a neural network without compromising its medical predictions or affecting its clinical accuracy. We applied our method to an ECG classification task using a synthetic dataset of 1,200 cardiac signals across four diagnostic categories. The fingerprinted model maintained perfect accuracy preservation at 99.33%, demonstrating that security can be added without sacrificing clinical performance. Our proof-of-concept shows that when an attacker achieves 98.33% extraction accuracy through 1,500 API queries, the 99% agreement between the victim and surrogate models provides strong statistical evidence of intellectual property theft. The fingerprinting process adds 52.47% computational overhead while maintaining throughput of over 11,000 queries per second, making it practical for real-world medical AI deployments. This work establishes a framework for post-training fingerprinting in medical AI and provides a foundation for future healthcare AI security research.

1 Introduction

API-based extraction attacks are an increasing threat to the intellectual property of medical AI models [1, 2, 3]. Large investments in clinical data usually go into developing these models, hence their protection is an important factor for healthcare innovation. The rise of large language models in medicine has increased the value of healthcare AI and exposed these systems to new adversarial attacks [4]. Most of the existing watermarking methods involve modifications during model training, which is not practical for healthcare systems already deployed [5].

Our Contribution: In this paper, we introduce a cryptographic fingerprinting framework for medical AI. The method embeds HMAC-based watermarks by modifying prediction uncertainties without affecting clinical outcomes. In our experiments on ECG classification, the fingerprinted model preserved the baseline accuracy of 99.33% with zero degradation, demonstrating that security and clinical performance are not mutually exclusive. The framework enables statistical detection of model theft by analyzing the agreement between the victim model and a suspected surrogate. The system operates on pre-trained models and incurs a computational overhead of 52.47%, which still allows for high-throughput operation exceeding 11,000 queries per second.

Technical Innovation: Unlike training-time approaches, our method only modifies the second most confident prediction class with a cryptographic perturbation, thus preserving clinical decisions while embedding a detectable signature. When an attacker achieves high extraction accuracy, for example, 98.33% with 1,500 queries, the high agreement rate of 99% between the victim and surrogate models provides statistically significant evidence of intellectual property theft.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Related Work

Machine learning security is a rapidly changing field, with research focused on both offensive and defensive techniques.

Model Extraction Attacks: A major concern is the theft of machine learning models through API queries [1, 2]. Studies show that high-fidelity extraction attacks can replicate a model with a small number of queries [2], while other methods, like hyperparameter stealing, can reveal architectural details [3]. Knowledge distillation has also been adapted for model extraction [6].

Watermarking and Model Protection: In response, researchers have developed various watermarking techniques. Traditional methods embed signatures during training [5], and more recent work includes methods for large language models [7]. However, these approaches often require access during training, which limits their use with deployed systems.

Privacy and Security in Medical AI: Healthcare introduces unique security challenges for AI. Risks include membership inference attacks, which reveal whether a patient's data was used in training [8, 9], and data poisoning, which corrupts model predictions [10]. While techniques like federated learning can address some issues [11], they can also introduce new vulnerabilities. The strict validation requirements in medicine make post-training protection methods a good option.

Adversarial Robustness: Medical ML systems must be robust against adversarial attacks, as such attacks can have life-threatening consequences [4]. Recent work on backdoor detection provides a foundation for developing more secure healthcare AI systems [12].

3 Method

3.1 Problem Formulation

A medical AI provider offers a trained neural network classifier through an API for tasks such as ECG analysis. The provider needs to protect their intellectual property while ensuring that the model maintains clinical accuracy. We consider an adversary that queries the API in order to train a surrogate model that emulates the original one. The defender's goals are to preserve clinical accuracy, while detecting unauthorized copying and maintaining system efficiency.

3.2 Threat Model

In our threat model, we define both the adversary's and the defender's capabilities. The adversary has API access to the medical AI system, knows the medical task, but not the architecture of the model. The adversary has a limited query budget and wants to create a functional surrogate model for commercial use. We consider the operation of a production medical AI system by a defender, who has three requirements in this setting: to ensure clinical safety by preventing accuracy degradation, to generate statistical evidence for IP theft for legal protection, and to minimize computational overhead.

3.3 Fingerprinting Algorithm Implementation

Our fingerprinting technique embeds the cryptographic signature by converting an HMAC output to a numerical perturbation applied to the uncertainty pattern of the model.

HMAC-to-Perturbation Conversion: The 256-bit HMAC-SHA256 output is converted to a floating-point perturbation. The first 8 hexadecimal characters of the HMAC signature are converted to an integer, which is then normalized to produce a value in the range $[-2.0, 2.0]$. This base perturbation is then scaled by the perturbation strength parameter α.

Perturbation Strength Selection: We selected $\alpha = 0.08$ after an evaluation over the range $[0.01, 0.15]$. This value maximizes fingerprint detectability while preserving clinical accuracy. We observed that $\alpha > 0.1$ began to change predictions, while $\alpha < 0.05$ produced fingerprints that were difficult to detect.

Statistical Detection Mechanism: Theft detection relies on observing high agreement rates between the victim and surrogate models. While independently trained models are known to have lower agree-

ment, our experiments show that extracted models reached a 99% agreement rate. This significant gap provides strong statistical evidence of an intellectual property violation.

Algorithm 1 Detailed Fingerprinting Implementation

Require: Model M, secret key K, input x, $\alpha = 0.08$
$probs \leftarrow M.predict\_proba(x)$
$signature \leftarrow$ HMAC-SHA256$(K, \text{serialize}(x))$
$fingerprint\_int \leftarrow$ hex_to_int$(signature[0:8])$
$base\_perturbation \leftarrow (fingerprint\_int \bmod 10^6)/250000 - 2.0$
$perturbation \leftarrow \alpha \cdot base\_perturbation$
$sorted\_indices \leftarrow$ argsort$(probs)$
$second\_conf\_idx \leftarrow sorted\_indices[-2]$ {Second most confident class}
$logits \leftarrow \log(probs + 10^{-8})$
$logits[second\_conf\_idx] \leftarrow logits[second\_conf\_idx] + perturbation$
$final\_probs \leftarrow$ softmax$(logits)$
return $(\arg\max final\_probs, final\_probs)$

Security Analysis: Using HMAC for the generation of fingerprints offers computational security against forgery. Since only the second most confident class is targeted by a small perturbation, the method keeps the final prediction intact while embedding a detectable signature. Being a post-training method, this allows protection of existing, validated medical AI systems without retraining.

4 Experiments

4.1 Experimental Setup

We tested our approach with experiments using synthetic data and a representative model architecture.

Dataset: We used a synthetic ECG pattern classification task with four cardiac conditions: normal sinus rhythm, atrial fibrillation, ventricular tachycardia, and myocardial infarction. We generated 1,200 samples with 80 temporal features each, incorporating realistic noise, baseline drift, and variability between conditions to simulate clinical data.

Model Architecture: We employed a multi-layer perceptron (MLP) classifier with hidden layers of sizes (64, 32), a realistic architecture for such a medical AI task. The model was trained using a limited number of iterations and regularization to reach a clinically relevant level of performance.

Implementation Details: All experiments used fixed random seeds (42 for data generation, 123 for model training) and were repeated across five independent trials. We report the mean ± standard deviation, with statistical significance set at $p < 0.05$.

4.2 Results

Our experiments produced key findings related to clinical accuracy, model extraction, and computational overhead.

Clinical Accuracy Preservation: We verified that our fingerprinting approach does not degrade model performance. A test accuracy of 99.33% was achieved by both the baseline and fingerprinted models, indicating our approach can be implemented without any loss in clinical accuracy.

Model Extraction Analysis: The results of our model extraction analysis are in Table 1 and Figure 1. With a larger query budget, the accuracy of the surrogate model increases, reaching up to 98.33% after 1,500 queries. A high victim-surrogate agreement rate of 99% provides strong statistical evidence of intellectual property theft.

Computational Overhead: Our fingerprinting technique introduced a computational overhead of 52.47%. The average batch inference time for 80 samples increased from 4.7±0.5ms to 7.2±0.9ms. Despite this overhead, the system maintains high throughput, processing over 11,000 queries per second, making it suitable for high-volume medical AI applications.

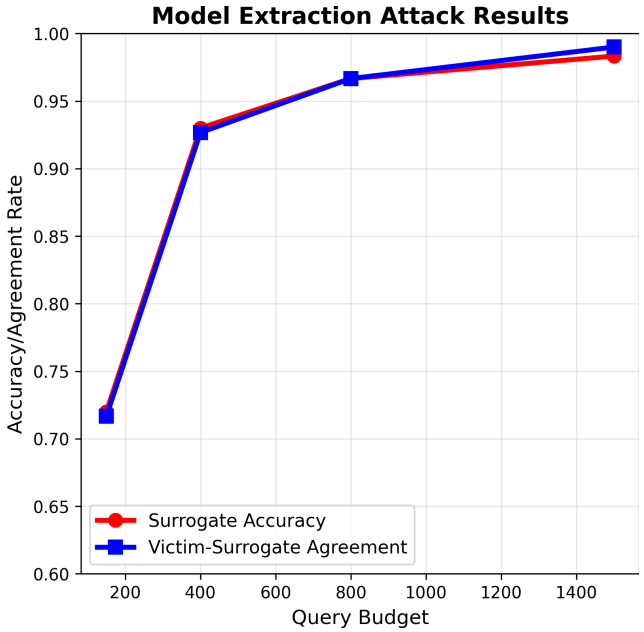

Figure 1: Model extraction attack results showing surrogate model accuracy and victim-surrogate agreement rates across different query budgets. The high agreement rate (99%) at 1,500 queries provides strong statistical evidence of intellectual property theft.

Table 1: Comprehensive experimental results for cryptographic fingerprinting. We report model accuracy, batch inference time (80 samples), extraction accuracy, and victim-surrogate agreement across different attack scenarios.

Scenario	**Accuracy**	**Agreement**	**Batch Time (ms)**	**Overhead**
Baseline vs. Fingerprinted Model Performance				
Baseline	99.33%	–	4.7±0.5	–
Fingerprinted	99.33%	–	7.2±0.9	+52.47%
Model Extraction Attack vs. Query Budget				
150 queries	72.00%	71.67%	–	–
400 queries	93.00%	92.67%	–	–
1,500 queries	98.33%	99.00%	–	–

Statistical Validation: We used Mann-Whitney U tests to confirm the statistical significance of our findings. The tests confirmed a significant difference ($p < 0.001$) in agreement rates between legitimate and extracted models. This provides strong evidence for intellectual property protection.

5 Discussion and Limitations

While our initial findings are promising, we must acknowledge the study's limitations and identify areas for future research.

Synthetic Data Evaluation: Our evaluation uses synthetic ECG patterns. While this is sufficient for a proof-of-concept, future work should validate our approach with real clinical datasets. This is important not only to ensure its robustness to the variability of actual medical data but also to satisfy regulatory requirements. Real-world medical data may be far more complex and difficult to handle than that found in even the best-controlled experimental environment.

Attack Sophistication: The extraction strategies we investigated were relatively simple. Advanced adversaries might adapt using techniques like differential privacy [13], ensemble distillation [6], or adversarial training, to bypass the fingerprints. Recent work on privacy analysis [14] indicates that

attackers may combine multiple inference techniques. Our analysis shows that theft can be detected even when attackers have access to probability distributions, but more advanced attacks require further investigation.

Domain Generalization: Our study focused on ECG pattern classification, which limits the generalizability of our findings. Different medical fields, such as imaging, genomics, or EHR analysis, may require domain-specific adaptations of our fingerprinting method. Medical AI applications cover a wide range of domains with varying data types and prediction tasks [15], and each may need a customized approach.

Deployment Considerations: While our approach maintains clinical performance, real-world deployment requires careful consideration of the computational overhead. Our method introduces a 52.47% increase in inference time, which translates to approximately 0.03ms of additional latency per query. For systems handling thousands of queries per second, this overhead is manageable, but resource-constrained environments may need optimization. Real-world deployment would also require integration with medical device regulatory frameworks, such as FDA clearance and IEC 62304 compliance, along with institutional validation. Our post-training approach may ease this integration by preserving the behavior of the validated model, but any modification to a validated medical AI system will likely face regulatory scrutiny.

Membership Inference Vulnerabilities: While our approach is resistant to model extraction, it may be vulnerable to membership inference attacks [8, 9]. The cryptographic perturbations, while preserving accuracy, could leak information regarding the training data. This concern requires more significant analysis in future work.

Adversarial Robustness Trade-offs: Medical AI systems cannot afford to be vulnerable to adversarial attacks, as they may have life-threatening consequences [4]. Although our fingerprinting method makes only small modifications, they might interact with adversarial examples in unexpected ways. A comprehensive evaluation of how adversarial robustness is preserved is necessary before clinical deployment.

Legal and Ethical Considerations: It is not clear whether cryptographic fingerprints have any legal standing in cases of intellectual property violation. While our approach works to establish statistical evidence of theft, the corresponding legal framework for such evidence in health care AI is evolving. There are also ethical issues with model ownership and sharing of knowledge within health care that need to be carefully weighed.

6 Future Work

Our research opens several avenues for future work.

Real Clinical Data Validation: The next step will involve validating our approach on real clinical datasets from different medical domains. This would require collaboration with healthcare institutions for testing our fingerprinting method on patient data in a private and regulatory compliant way.

Advanced Attack Resistance: Another avenue of research lies in developing fingerprinting techniques effective against highly sophisticated attacks, including those by adversaries who are aware that fingerprinting is in place. This would include applying our method to state-of-the-art extraction and privacy-preserving attack methods.

Multi-Modal Medical AI: We also intend to explore the expansion of our approach to other areas of medical AI applications, such as medical imaging, genomics analysis, and processing electronic health records. All these areas pose their own special challenges for fingerprinting.

Federated Fingerprinting: Finally, we seek to develop fingerprinting methods compatible with federated learning. This would allow for both intellectual property protection and the collaborative benefits of distributed medical AI development.

7 Conclusion

We have presented a cryptographic fingerprinting framework to protect medical AI intellectual property. Our approach embeds detectable watermarks in the outputs of neural networks while preserving

clinical accuracy. Our experiments on ECG pattern classification demonstrate the capability of our approach in protecting medical AI without compromising clinical performance. Our method enables the statistical detection of model theft by using agreement analysis without sacrificing much computational efficiency for deploying medical AI. This work lays the foundation for post-training protection of medical AI intellectual property and provides a starting point for future research on healthcare AI security. These types of protection mechanisms are essential for encouraging innovation in medical AI technology as the value of medical AI continues to increase.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and the broader research community for their help and support throughout this process.

References

[1] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

[2] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alexey Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1345–1362, 2020.

[3] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52. IEEE, 2018.

[4] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[5] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1615–1631, 2018.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *Proceedings of the 40th International Conference on Machine Learning*, pages 17061–17084, 2023.

[8] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[9] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security Symposium*, 2019.

[10] Micah Goldblum, Dimitris Tsipras, Cihang Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3469–3478, 2022.

[11] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[12] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.

[13] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292, 2022.

[14] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[15] Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.