2026 Volume: 02 Issue: 05 www.ijrpa.com ISSN 2456-9995 Review Article

International Journal Research Publication Analysis

Page: 01-07

MODELSHIELD AI: REAL TIME ANOMALY DETECTION AND DISTILLATION - RESISTANT WATERMARKING FOR MODEL EXTRACTION ATTACK PREVENTION

*[1]Shwetha Patil, [1]Megha P. Dodamani, [1]Bhoomika, [1]Ramya N. N., [2]Varsha C. Parihar

[1]Jyothy Institute of Technology, Student, Deparment of CSE.

[2]Associative Professor, Jyothy institute of Technology Department of CSE Tatagunu, Bengaluru.

Article Received: 31 March 2026
Article Revised: 21 April 2026
Published on: 11 May 2026

***Corresponding Author: Shwetha Patil**
Jyothy Institute of Technology, Student, Deparment of CSE.

DOI: https://doi-doi.org/101555/ijrpa.7088

ABSTRACT

The rapid growth of machine learning has resulted in its use in crucial areas like healthcare, cloud services, and the Internet of Things. As machine learning models continue to grow in value and play an increasingly important role in decision-making, the need to protect against threats like theft, malicious use, and unauthorized copying is becoming more significant. This literature review examines recent efforts to secure machine learning models via different protection methods, such as watermarking, anomaly detection, resistance to distillation, and federated security mechanisms. Each of these methods helps to protect machine learning models from various types of attacks, including model extraction, reverse engineering, and unauthorized knowledge transfer. The various research papers selected cover a wide variety of solutions. Examples include backdoor-based watermarking, which embeds hidden ownership data into machine learning models without degrading their performance; and trace rewriting techniques, which prevent unauthorized copying of machine learning models by encoding knowledge into a form that makes it impossible to distill via knowledge distillation. Anomaly detection techniques are commonly used in both the cloud and the Internet of Things to monitor and identify suspicious activities in real time. Finally, distributed and federated verification systems provide a decentralized method for ensuring machine learning model integrity and ownership across multiple nodes.

1 INTRODUCTION

Intellectual properties have been developed within the context of Machine Learning (ML)

models in recent years as noted by all of the referenced papers, as a result of the amount of data, processing ability, and expertise required to create them. The use of ML models is widespread throughout many industries such as healthcare, the cloud, NLP, and IoT devices, where they perform a variety of important decision-making functions. However, ML models are consistently reported in the literature as being extremely susceptible to a variety of security threats. For example, Tramèr et al. describe how models can be stolen via Prediction APIs, allowing an attacker to perform the same actions as the owner without having access to original model. Other papers have reported that adversaries develop surrogate models from the original to simulate the functionality of the original as a form of unauthorized knowledge distillation resulting in the theft of intellectual property. There are additional concerns related to data leakage, misuse of sensitive data, and adversarial attacks that cause a change in the function of the model. Many of the reviewed papers propose a number of advanced protection methods to prevent these threats from occurring. A prominent solution that continues to be used is watermarking, where the ownership of the model is included in the model through the use of backdoor-based watermarking, as described by Kong & Xu

Model owners can confirm their models' legitimacy even after they have been changed by using techniques that are mentioned below. Anomaly detection also forms a significant part of model security in both cloud and IoT platforms, such as in the case of CloudShield; soft computing methods to monitor and identify abnormal patterns of access and potential attacks in real-time through systems that work within these types of environments. In addition, distillation resistance techniques that use trace rewriting and constraint-based learning aim to prevent unauthorized copies of a model from being created. Other methods of securing models do so across a variety of decentralized environments by utilizing federated watermarking and distributed verification-systems.

Through the integration of the findings from the ten research articles evaluated, this survey provides a comprehensive account of the current state of ML model security. This survey will discuss the importance of having sophisticated, scalable, and versatile ML protection mechanisms that will provide for the reliable use of ML.technologies

2 RELATED WORK

The literature examining the security of machine learning models can be placed into general categories based on individual areas of research that each specifically target a type of threat and, ultimately, develop solutions to mitigate the identified threat. One of the most

significant threats emergent from the literature review related to machine learning model security is model theft. Specifically, empirical studies have demonstrated how an attacker can replicate ML models by making numerous calls to a prediction API, retrieving its output on each call, and using that output to derive the logic behind an ML model. Because the internal mechanisms/models of a deployed ML system can be replicated by the output received from the API without having to access the ML model itself, model stealing attacks are a significant threat to deployed systems in the real world. As such, the literature concludes that there is an immediate need for protective mechanisms (e.g., access control, monitoring) for publicly accessible ML systems**.**

The second area covered in detail within the reviewed literature relates to digital watermarks of ownership within machine learning models. These watermarks are often achieved through the introduction of backdoor-based watermarking techniques that include an embedded trigger that will generate a known output (e.g., triggering the output of the number 1) that can identify the model owner. Alternatively, "training-free" watermarking techniques embed a watermark that does not require a retraining of the model thereby reducing computing overhead. Finally, federated watermarking encompasses the use of digital watermarks across multiple distributed environments to enable identification of ownership across all participants in a federated system. All three types of watermarking techniques provide the owner with a valid claim of intellectual property, even if the original model has been modified or communicated with others.

Another important area of research focused on preventing unauthorized duplication of models through the use of knowledge distillation is the area of "distillation resistance." Techniques such as trace rewriting and constraint-based reasoning alter the process of learning, making it more difficult for an attacker to train their own surrogate model that accurately approximates the target. These techniques also improve resistance to attack, inasmuch as they limit the amount of knowledge that can be transferred to student models.

Another primary use case for anomaly detection is in identifying suspicious behaviour or anomalous events in mechanical learning systems. Anomaly detection techniques are commonly found in both cloud computing and Internet of Things (IoT) environments and are highly reliant upon real-time monitoring systems to monitor usage patterns for the purpose of detecting anomalies that could indicate an impending attack. Artificial Intelligence (AI) driven and soft-computing based techniques also assist with improving the accuracy and timeliness

of the detection of anomalous behaviour, thus enhancing the overall security of the system.

Research Area	Key Techniq ues	Purpose	Advanta ges
Model Streaming attack	API Query Analysis	Identify vulnerabili ties in ML deploymen t	Highlights real-world risks
Watermark ing Techniques	Backdoor, Training-free, Federate	Ownership verification	Low overhead, robust tracking

3 Technologies Used

Most research papers covered in this review use diverse and sophisticated technology to over come obstacles caused by protecting ML models and detecting anomalies. ML and AI technologies and, in particular, DNNs make up the foundation of these technologies and have been used extensively by practitioners to build durable and high-performing models across several domains. Pre-trained NLP models and tools, which are also very commonly employed on text data and language-related projects, are greatly attacked and, therefore, are a high priority in the current research field of providing proof of ML model protection. Security technologies are another necessary component within use of these models. Watermarking systems such as backdoor-style techniques and embedded signal systems are predominately used to provide ownership and traceability of an ML model created by a developer. These types of systems also allow developers to confirm whether their ML models have been copied or otherwise misused. Lastly, Cryptographic verification systems greatly enhance the security of a developer's model by providing a method to authenticate it and verify its integrity so as to ensure the model remains tamper-proof and trustworthy.

Scalable and real-time solutions for security must be implemented using cloud-based and distributed technology. Anomaly detection in a cloud-based environment allows for real-time monitoring of a deployment on a large-scale basis and provides immediate responses to suspicious activities. The use of federated learning frameworks has increased, allowing for collaborative model training while maintaining privacy and security through not sharing raw data across the various distributed nodes.

3.1 Challenges

- A trade-off between model accuracy and security
- Difficulty in detecting advanced model extraction attacks
- Watermark removal or overwrite attacks

- Scalability issues with large-scale ML systems
- Extremely high computational overhead
- Lack of standardized evaluation metrics

4 RESEARCH GAP

- No Generalized Security Model: Most studies utilize one technique (i.e., either watermarking or anomaly detection). There is no framework combining many different security techniques all together.
- Limited Resistance to Adaptive Attack: Current watermarking methods and distillation-resistant mechanisms cannot withstand an attack in which the attacker changes their standard mode of operation to elude detection.
- Limited Real-Time Resistance: Several different solutions have not been developed with functions that allow for real-time use, particularly when considering the nature of cloud computing and the Internet of Things where rapid response is essential.
- Poor Generalization Across Models : Techniques developed specifically for one model (e.g., Natural Language Processing (NLP) or Convolutional Neural Networks (CNN)) generally will not transfer across multiple models.
- Lack of Standard Benchmark Data and Metrics: There are no established benchmark data or metrics for assessing, comparing and determining effective model protection methods.
- High Computational Expense :Advanced security measures significantly increase the costs of training and inference; therefore, they are not feasible in environments that are resource limited.
- Limited Investigation of Explanatory Models Most methods do not yield an interpretable output so it is impossible to know why a model has been indicated as compromised.

5 Future Scope

- Combining Multiple Security Techniques: Future work should examine the feasibility of combining watermarking, anomaly detection, and air-gapped systems into one hybrid security framework. By using this method of secure content delivery, content creators can protect their works through mutli-tiered security measures rather than relying on only one method. This type of security framework will also enhance the overall degree of resilience against various forms of cyberattack.
- Adaptive and Self-Evolving Security Systems: Security systems should be capable of automatically adapting to emerging threats. Security systems that incorporate artificial

intelligence (AI) and learn through experience will be able to transform their defense mechanisms to identify potentially new forms of attack. By employing AI, the security system can continue to enhance its security and mitigate the impact of exploits and zero-day attacks.

- Improved Real-Time Detection:Real-time detection systems must be enhanced to provide immediate threats. Future solutions should focus on reducing the time taken to detect threats while providing for accurate identification. A reduced time delay and accurate identification can be achieved with cloud-based and Internet of Things (IoT)-based security solutions.
- Standardized Benchmarks and Performance Metrics Research is needed to develop standardized datasets and performance metrics for evaluating the performance of a number of different but complementary, protective solutions so that they can be easily compared. These performance metrics will also be helpful to researchers who wish to validate the effectiveness of their security approaches and ways to quantify the effectiveness of various security
- Robustness to Advanced Attacks: Future work should consider sophisticated and adaptive attack methods in more depth. Solutions must provide resilience against watermark removal and modification of models. The stronger the robustness of the solution, the better the protection is available for the model over time.

6 CONCLUSION

The research reviewed indicates that protecting machine learning models has become increasingly necessary due to rising threats such as model stealing, unauthorized distillation and adversarial misuse. Articles that were examined also show that the methods of protection for model ownership and integrity (i.e., watermarking, anomaly detection and distillation resistence) are effective techniques. In addition to these traditional approaches, new techniques (i.e., backdoor-based watermarks, federated verification and real-time anomaly detection systems) have been demonstrated to provide considerable advantages in terms of providing security across cloud, Internet of Things (IoT) and Natural Language Processing (NLP) applications. However, the studies evaluated had many limitations including: scalability issues; high computational cost; and susceptibility to adaptive attacks. Furthermore, the lack of standardized evaluation frameworks restricts the ability to compare the various methods being evaluated in a fair manner. Overall, while the current protection mechanisms are very good initial points for providing machine learning security,

there remains to be a need for more robust, efficient and integrated mechanisms. Future work needs to focus on developing adaptive, scalable and real-time protection systems in order to provide for the secure and trustworthy deployment of machine learning models.

7 REFERENCES

1 K. F. Yapiter, Alfin, Y. Hasim, R. Purba, and M. Ulina, "An Integrated Framework for Automated Resume Screening Using RoBERTa, Random Forest and Explainable AI," *TEKNIKA*, vol. 14, no. 3, pp. 424–432, Nov. 2025.

2 B. Banu, S. Staniya, S. Swetha, and S. Kritheev, "A Data-Driven Framework for Personalized Career Guidance Using AI-Based Resume and Skill Analysis," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 8, no. 2, Mar.–Apr. 2026.

3 A. Nashath, A. Jabeen, U. Hannan, C. R. Sindhu, and I. Khaleelulla, "AI-Resume Analyzer," *IRE Journals*, vol. 9, no. 6, Dec. 2025.

4 R. Nayak, B. Dinesh, and C. Antony, "Smart Resume Analyzer with Automated Suggestions," *IOSR Journal of Computer Engineering*, vol. 27, no. 2, pp. 41–49, Mar.–Apr. 2025.

5 K. S. Srinidhi, V. U. Nischal, U. Karthik, and B. P. Manoj, "Smart Hiring System," *International Journal of Innovative Science and Research Technology*, vol. 10, no. 2, pp. 1969–1976, Feb. 2025.

6 M. Kanjikar, Anuja, D. Patil, Ananya, and Anjali, "Smart Resume Analyser," *International Journal of Scientific Development and Research (IJSDR)*, vol. 10, no. 12, Dec. 2025.

7 J. M. Patil, K. R. Kolekar, S. V. Zamre, A. S. Galzalwar, and S. P. Chaudhary, "GenAI-Powered ATS: Enhancing Recruitment with Skill Fitment Analysis," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 3, no. 3, pp. 1106–1110, Mar. 2025.

8 E. Albaroudi, T. Mansouri, and A. Alameer, "A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring," *AI*, vol. 5, no. 1, pp. 383–404, Feb. 2024.