ISSN:2790-1688 Volume-15-(2025)

Machine Learning-Based Detection and Defense Techniques for AI-Generated Content in Cybersecurity

Xiang Li

Xi'an Jiaotong-Liverpool University, Jiangsu, China

Xiang.Li2203@student.xjtlu.edu.cn

Abstract. This essay systematically reviews the detection and defense technologies for AI-generated content in the context of network security scenarios. First, it outlines the representative methods and key features of supervised learning, self-supervised learning, and unsupervised learning in image and video forgery detection. Second, it summarizes the active protection methods, such as adversarial or discriminative noise injection and digital watermarking, as well as the passive detection mechanisms, including ManTra-Net and ObjectFormer, along with their performance and applicability. Third, it discusses the necessity and the latest progress of multimodal fusion and lightweight deployment in practical implementation. On this basis, the paper identifies the major challenges in real-world applications, including the faint traces of high-fidelity forgeries, insufficient cross-domain generalization, scarcity of annotations, and weak adversarial robustness. Finally, it proposes future research directions, such as a generalized detection framework, robust multimodal fusion, dynamic adversarial defense and self-supervised robust training, and a task-aware lightweight architecture at the edge, providing references for subsequent research and engineering practice.

Keywords: Artificial intelligence, Machine Learning, Deepfake.

1. Introduction

Artificial intelligence is a technology that simulates human intelligence. It is the main driving force of the new generation of technological revolution. With the continuous updates and iterations of large-scale AI models such as DeepSeek and ChatGPT, the functions of these AI models have become increasingly sophisticated. From mechanically answering questions to now creating vivid and exquisite images and videos, more and more people frequently use AI models in their daily lives. For instance, a report from McKinsey Company indicates that currently, AI can unlock an economic potential of 360 billion to 560 billion US dollars annually by accelerating the research and development process. This value covers the entire industry. Moreover, large language models (LLMs) can analyze massive product reviews, social media posts, and meeting records, extracting unmet market demands, and the various information generated greatly enriches the content of the Internet and social media [1]. A large amount of content generated by artificial intelligence has emerged on the Internet, which has brought new vitality to the Internet industry while also presenting new challenges. Owing to the advancements in large artificial intelligence models, information such as images and videos generated by them is difficult to detect through direct observational means like the human eye. This has led to a lower threshold for cyber-attacks and online fraud. The malicious use of content generated by artificial intelligence, such as deepfake videos, synthetic texts, and forged images, poses severe cyber-security threats, including identity theft, the spread of misinformation, and online fraud. Traditional detection methods struggle to handle the high level of authenticity of such content. Therefore, machine learning (ML) has emerged as a core technology due to its capabilities in feature learning and pattern recognition. This review systematically collates the detection and defense technologies based on machine learning for AI-generated content, analyzes their principles, countermeasures and application scenarios, and discusses the challenges in practical applications. The aim is to summarize and organize the knowledge system of anti-AI modules in cybersecurity practice, clarify the new challenges and development directions faced by current cybersecurity, and provide insights for future research. Specifically, the core contributions of this paper can be summarized as follows: Firstly, it systematically reviews the mainstream machine

learning methods in the field of AI-generated content detection, providing in-depth summaries on the technical principles, typical models, and performance of supervised learning and unsupervised learning in image and video forgery detection, and clarifying the applicable scenarios and technical boundaries of different methods; Secondly, it comprehensively compares the adversarial defense strategies for AI-generated content, analyzing the defense logic, implementation costs, and actual performance differences of various strategies from the two dimensions of active protection and passive detection; Thirdly, in response to the practical needs of multimodal fusion and lightweight deployment, it proposes clear future research directions to address current technical bottlenecks, providing a theoretical and practical reference framework for subsequent academic exploration and engineering implementation.

2. Literature Survey

Figure 1 presents the annual statistics of the number of papers retrieved using the search term "deepfake detection" in Google Scholar. The overall trend of the relevant literature shows that from 2020 to 2025, the number has been growing rapidly. The number of papers increased from approximately 3,098 in 2010 to a peak of 17,500 in 2024. It is notable that the publication volume from 2023 to 2025 remained basically stable, with no significant increase and a slight stagnation. This trend reflects the increasing attention of both the academic and industrial communities to detecting content generated by artificial intelligence and resisting deepfake technologies, as well as the emergence of technical bottlenecks. The sharp increase in the number of papers in recent years highlights the urgency and importance of machine learning-based solutions in cybersecurity, as researchers constantly respond to the constantly changing challenges brought by generative artificial intelligence.

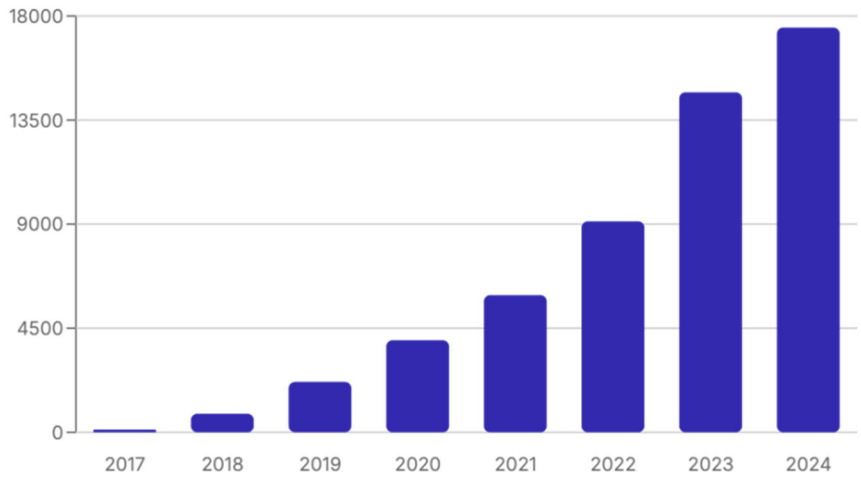

Figure 1 : The Annual Statistics of the Number of Papers Retrieved using the search term "deepfake detection" in Google Scholar [2].

Supervised Learning is one of the core branches of machine learning. Its core characteristic is to utilize training data with explicit labels to enable the model to learn the mapping relationship from "input" to "output", ultimately achieving the prediction or classification of unknown data. In the task of detecting artificial intelligence content, the task form is binary classification (real/fake). In terms of models, CNN, ViT/mixed Transformer and other models are often used. CNN excels at capturing common spatial artifacts and fusion boundaries in AIGC, such as inconsistent skin texture and lighting, while ViT/mixed Transformer is beneficial for global consistency modeling and can detect semantic and geometric conflicts across regions. For diffusion and autoregressive image synthesis, the joint spatio-frequency features are particularly crucial: the upsampling period, color channel correlation and phase inconsistency visible in the frequency domain often constitute robust clues; in

videos, cooperating with temporal features can further amplify subtle differences. In engineering, to approximate the distribution of social platforms, "artifact-preserving" augmentations are often added during training, such as strong JPEG/WebP compression, resampling, motion blur, color jitter and cross-dataset training. Moreover, the discriminative boundary for highly realistic samples is strengthened through curriculum learning and hard sample mining. In addition, to enhance the generalization ability to deal with unknown generators, self-supervised constraints or adversarial consistency regularization can be introduced within the supervised framework to improve the domain invariance and robustness of features. In 2024, Aminollah Khormali and others leveraged the characteristics of self-supervised learning to construct a composite model composed of three parts. The model framework consists of three components: a feature extractor based on the visual Transformer architecture, a graph convolutional network, and a Transformer discriminator. This model demonstrates excellent performance in terms of accuracy and generalization ability, with an accuracy rate reaching 90.8%, which is significantly superior to other models. In terms of generalization, when dealing with different datasets such as DeepFake, NeuralTextures, and Face2Face, the average accuracy rate reaches 99.3% [3].

Unsupervised Learning is another core branch of machine learning. Its core feature is to use only unlabeled training data to allow the model to independently explore the inherent structure, patterns or correlations within the data, without the need for manual pre-definition of the "input to output" mapping relationship. It is closer to the human learning mode of "autonomous observation and induction" rather than relying on external guidance. The core process can be divided into three major stages: data mining, feature modeling, and anomaly detection. Through autonomous learning of the distribution patterns of real data, it identifies forged features that deviate from the normal mode. In 2024, Tong Qiao et al. developed a completely unsupervised deepfake detection system, which included an innovative pseudo-label generator and an enhanced contrast learner. It guided the extraction and optimization of features through contrast loss and completed the final binary task based on inter-frame correlation. Experimental data demonstrated that this deepfake detection system based on unsupervised learning performed well on benchmark datasets such as DFD, DFDC, and UADFV. Its performance was significantly superior to other methods. Moreover, in the face of issues such as insufficient training data or malicious data contamination, this model exhibited excellent resistance, indicating its generalization ability [4].

The proactive protection method refers to the practice of embedding protective mechanisms, intervention procedures or preset rules in the key stages of "generation, dissemination, and user interaction" of AI content, in order to prevent the generation, spread or misbelief of "malicious AI content" at its source. The core of this method is to "actively intercept the risk content before it is implemented", rather than passively identifying and cleaning it through algorithms after the content has been generated. This article will introduce two methods: active noise injection and watermark insertion. Active noise injection refers to the process of embedding imperceptible disturbances in images or videos when detecting AI content. These disturbances interfere with the preprocessing or optimization goals of deepfake models, resulting in problems such as texture fragmentation, boundary misalignment, or semantic anomalies in the generated attack results. This technology is commonly used in the adversarial sample detection process of deep learning models. The drawback of deep learning is that it can be misled by adversarial images generated by deliberately adding small and imperceptible disturbances to clean inputs, leading to incorrect decisions. However, active noise injection can eliminate this interference, enabling deep learning models to operate normally. In 2019, Si Wang and others proposed a novel discriminative noise injection strategy, aiming to distinguish malicious inputs from benign ones. The primary core principle is to create differences between the evaluations of natural images and adversarial images by injecting varying amounts of noise. This method performs well, enabling the DNN model to achieve a detection rate of 88% on the ImageNet dataset and enhancing the resilience of deep-learning models when dealing with adversarial AI content [5].

Watermarking technology embeds specific information (such as identity identification, timestamp) actively into the key features of the content through the encoder, making the watermark an inherent attribute of the content. The core objective is to make the watermark a "digital fingerprint" of the content. No matter how the content is tampered with, its authenticity can be verified by detecting the presence, location or integrity of the watermark. In 2023, Yuan Zhao et al. utilized watermarking technology to design a unique encoder-decoder structure neural network, embedding the anti-deepfake label information as a watermark into facial identity features. They also detected deepfake behavior based on the presence or absence of the label. Experimental results showed that this method performed well, with an average detection accuracy exceeding 80%. This demonstrated the feasibility and effectiveness of watermarking technology in the field of deepfake detection [6].

The passive protection method refers to a defense strategy that, based on the premise of "the threat has occurred", uses detection and positioning technologies to identify anomalies in AI-generated content, and then takes response measures to reduce the harm. Its core logic is "first detect the risk, then respond passively", which complements the "pre-emptive blocking" of active protection. This article will introduce two relatively common passive protection networks, ManTra-Net and ObjectFormer.

In 2019, Wu et al. proposed ManTra-Net, an end-to-end image forgery detection and localization network. This network can process images of arbitrary sizes and various forgery types, including splicing, copy-move, and unknown forgeries, without the need for additional pre-processing or post-processing. The model consists of two modules. Firstly, the operation trace feature extractor acquires robustness by self-supervised learning of 385 classes of fine-grained operation features. Secondly, the local anomaly detection network quantifies local feature anomalies using the Z-score and combines ConvLSTM for long-and short-distance analysis to locate forgeries. Experimental results show that the model has strong generalization ability. On the PS-Battle dataset, the AUC reaches 75.88%. It takes approximately 0.8 seconds to process a 1024×768 image on a 1080Ti. Its performance outperforms traditional methods and is comparable to the state-of-the-art (SOTA) approaches [7]. In 2022, Wang et al. proposed ObjectFormer, an end-to-end multimodal Transformer framework designed for image tampering detection and localization. To capture subtle tampering traces that are invisible in the RGB domain, it extracts high-frequency features from images through the Discrete Cosine Transform (DCT) and fuses them with RGB features to generate multimodal patch embeddings. A learnable object prototype is introduced as an intermediate representation to model object-level consistency. The prototype and patch embeddings are alternately updated via the object encoder and patch decoder. The Boundary-Sensitive Context Inconsistency Modeling (BCIM) module is incorporated to refine pixel-level differences. Experiments on datasets such as CASIA and Columbia have outperformed state-of-the-art (SOTA) methods like ManTraNet, demonstrating remarkable robustness. Ablation experiments have verified the effectiveness of components such as high-frequency features and object prototypes [8].

In practical scenarios, various issues may be encountered. For instance, problems such as insufficient space and polymorphic data can lead to a decline in the detection performance of the model. Therefore, multimodal fusion and lightweight deployment are of great significance in the detection strategy. Lightweight deployment is the core support for AI content detection to move from the laboratory to practical applications. It can overcome the resource limitations of edge devices, meet the requirements for low latency in real-time review, reduce the cost of large-scale deployment, and achieve end-side data privacy protection. Its principle involves model compression, architecture optimization, and inference optimization. While significantly reducing the number of parameters and computational resources, it retains the core tampering detection capability, achieving a balance between resource consumption and detection performance. In 2025, Hyun Kim et al. achieved remarkable success in developing a lightweight solution suitable for devices with limited computing resources. They integrated a machine learning classifier with the key-frame method and texture analysis. In the resource-constrained environment, the accuracy on the FaceForensics++ and Celeb-DF (v2) datasets was increased to 92% and 96% respectively [9].

Multimodal fusion serves as a pivotal approach for enhancing the robustness and generalization of AI content detection. It effectively addresses the limitation of insufficient information in single-modality detection and offers three key advantages. Firstly, it compensates for the deficiencies of single-modality data. Secondly, it improves the anti-interference ability. Thirdly, it enables adaptation to complex scenarios. In 2024, Rui Wang et al. successfully proposed AVT^2-DWF by integrating multiple modalities. This is an audio-visual dual converter based on dynamic weight fusion, aiming to enhance cross-modal and intra-modal forgery clues. Experiments on the DeepfakeTIMIT, FakeAVCeleb and DFDC datasets demonstrated that AVT^2-DWF achieved the state-of-the-art performance in deep forgery detection across both modalities and within the same modality [10].

3. Challenges and Future Directions

AI content detection faces multiple core challenges that hinder its reliable implementation. Firstly, advanced forgery techniques (such as GAN and Diffusion models) combined with post-processing (JPEG compression, edge blurring) make the forgery traces in the RGB domain extremely faint, and traditional visual anomaly detection fails. Secondly, the model has poor generalization ability and is insufficiently adaptable to unknown forgery types and deviations in the distribution of real scenarios. Thirdly, high-quality labeled data is scarce and the distribution of synthetic data differs significantly from that of real data. At the same time, it is vulnerable to adversarial attacks and post-processing interference. In addition, problems such as difficult multi-modal forgery fusion, difficulty in balancing real-time lightweighting and accuracy, failure to detect minor forgery, and the lack of ethical privacy and standardization further increase the difficulty of detection. Future research on AI content detection needs to focus on breaking through the core challenges. Firstly, develop a universal detection model to enhance its generalization ability for unknown tampering types and deviations in the distribution of real scenarios. Secondly, deepen multi-modal integration to address the heterogeneity of modalities and noise interference. Thirdly, strengthen adversarial robustness by implementing dynamic defense mechanisms and self-supervised robust training to resist adversarial attacks and post-processing interference. Fourthly, optimize lightweight real-time architectures, design task-aware lightweight modules, and balance detection accuracy with edge deployment requirements.

4. Conclusion

This article focuses on the entire process chain of "Detection-Protection-Implementation", systematically summarizing the development history of deepfake detection technology based on machine learning: from supervised and unsupervised detection methods with models such as CNN and ViT, to source governance through active noise and watermarks, to passive forensic methods such as ManTra-Net and ObjectFormer, as well as multi-modal fusion and lightweight deployment engineering paths adapted to edge scenarios, outlining the technical map of AIGC security. However, current AIGC has some issues: generalization and cross-domain generalization are still insufficient; high-quality annotations and real distribution data are scarce; adversarial samples and contaminated data bring robustness risks; at the same time, the alignment of multi-modal heterogeneity, real-time performance and accuracy, privacy compliance and evaluation standards still need to be improved. Therefore, it is necessary to promote the development of transferable general detection frameworks, multi-modal fusion and dynamic defense capabilities, and accelerate the formation of a deployable, scalable and manageable deepfake defense system.

References

[1] McKinsey & Company. (2025). The next innovation revolution powered by AI. McKinsey & Company.https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-next-innovation-revolution-powered-by-ai

[2] Google Scholar. Advanced search guide. https://scholar.google.com

[3] Khormali, A., & Yuan, J.-S. (2024). Self-supervised graph transformer for deepfake detection. IEEE Access, 12, 58114–58127. https://doi.org/10.1109/ACCESS.2024.3392512

[4] Qiao, T., Xie, S., Chen, Y., Retraint, F., & Luo, X. (2024). Fully unsupervised deepfake video detection via enhanced contrastive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7), 4654–4668. https://doi.org/10.1109/TPAMI.2024.3356814

[5] Wang, S., Liu, W., & Chang, C.-H. (2019). Detecting adversarial examples for deep neural networks via layer directed discriminative noise injection. In 2019 Asian Hardware Oriented Security and Trust Symposium (AsianHOST) (pp. 1–6). https://doi.org/10.1109/AsianHOST47458.2019.9006702

[6] Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). [Paper title]. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 4602–4611).

[7] Wu, Y., AbdAlmageed, W., & Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9543–9552).

[8] Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S. N., & Jiang, Y.-G. (2022). Objectformer for image manipulation detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2364–2373).

[9] Yasir, S. M., & Kim, H. (2025). Lightweight deepfake detection based on multi-feature fusion. Applied Sciences, 15(4), 1954. https://doi.org/10.3390/app15041954

[10] Wang, R., Ye, D., Tang, L., Zhang, Y., & Deng, J. (2024). AVT2-DWF: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies. IEEE Signal Processing Letters, 31, 1960–1964. https://doi.org/10.1109/LSP.2024.3433596