Hidden in the Noise: Noise-Embedded Watermarking for Black-Box Image Classifiers[*]

Yeon-Ji Lee, Hye-Yeon Shim, Yeon-Woo Lee, Su-Ji Park, and Il-Gu Lee[†]

Sungshin Women's University, Seoul, Korea
{220247018, 220237062, 20221124, 20222620, iglee}@sungshin.ac.kr

Abstract

As a representative technique for protecting model ownership against model extraction attacks, watermarking has been proposed. In particular, black-box watermarking offers the advantage of verifying ownership through specific input–output pairs without accessing model parameters. However, conventional approaches suffer from a severe degradation in performance when the input deviates from the primary data distribution. In this study, we propose a noise-embedded watermarking method applicable to image classification models to alleviate this limitation. The proposed method constructs a watermark dataset by embedding imperceptible noise patterns into original images and jointly training it with the original training data, thereby naturally embedding the watermark into the model. Experimental results demonstrate that the proposed method achieves a watermark detection rate of up to 99.98% while minimizing performance degradation compared with conventional backdoor-based watermarking. Furthermore, by adjusting the noise magnitude (ϵ), we confirm that the trade-off among model performance, detection rate, and visual perceptibility can be effectively optimized. Additional analyses show that no detection occurs in unwatermarked models and that valid detection is only observed in watermarked models, supporting the reliability of ownership verification. This study highlights the potential of watermarking as a practical and unobtrusive means of protecting ownership of image classification models in black-box environments.

keywords: model-stealing attack, machine learning, model watermark, security

1 Introduction

Recent advances in deep learning have driven innovation across various artificial intelligence (AI) domains, including image recognition, natural language processing, and speech recognition [1]. However, training a deep learning model requires careful design of network architectures and training strategies, high-quality labeled datasets, and substantial computational resources. Consequently, high-performing deep learning models are regarded as valuable intellectual property (IP) that can serve as a core competitive asset for enterprises [2]. Thus, protecting the intellectual property of deep learning models has emerged as a critical research issue. Model stealing attacks can generally be categorized into two types. The first type involves preparing an unlabeled dataset and querying the target application programming interface (API) to obtain probabilities, which are then used as labels for training a surrogate model [3]. The second type directly extracts model parameters through side-channel attacks or other techniques [4]. This study focuses on the latter type.

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 79, December 16-18, 2025, Sapporo , Japan.

[†] Corresponding author

To defend against such extraction attacks and to assert ownership, model watermarking has been proposed as a representative approach [5]. White-box watermarking [6] embeds secret patterns directly into model parameters, enabling effective watermark insertion; however, it requires access to the stolen model's parameters during ownership verification, which limits its practical applicability. To overcome this limitation, black-box watermarking [7] was introduced, which embeds backdoors by training specific input–output pairs and leverages them as watermarks. This approach has the advantage of enabling ownership verification without accessing model parameters. Nonetheless, because it requires training with input–output pairs outside the primary data distribution, performance degradation is inevitable [8]. To address this limitation, we propose a watermarking method for image classification models that embeds specific watermark patterns while mitigating the performance degradation observed in conventional black-box watermarking approaches. Furthermore, we conduct extensive experiments under diverse conditions to validate the effectiveness of the proposed method in enabling reliable watermark extraction.

The main contributions of this paper are threefold:

- We propose a robust watermarking method that alleviates the performance degradation inherent in conventional black-box watermarking approaches.
- We present an evaluation framework that systematically assesses the trade-off between security and performance in watermarking techniques.
- We demonstrate that the proposed method achieves a high detection rate of 99.98%, while avoiding false detections in non-watermarked models.

The remainder of this paper is organized as follows. Section 2 reviews conventional model watermarking methods. Section 3 describes the proposed method in detail. Section 4 presents the experimental setup, procedures, and results. Finally, Section 5 concludes the paper.

2 Related Work

In this section, we review conventional studies on model watermarking techniques.

Li et al. [9] proposed the dataset verification via the backdoor watermarking (DVBW) method, which applies backdoor-based watermarking to ensure the ownership of public datasets. This approach consists of two stages: dataset watermarking and verification. In the watermarking stage, specific input–output pairs are overfitted by adopting a model backdoor attack that induces intentional misclassification at predetermined instances. In the verification stage, a normal sample is first fed into the model to compute its prediction probability for the target label, after which the same sample is watermarked and queried again. If the watermarked sample yields a higher probability than the normal sample, ownership can be claimed by asserting that the model was trained with the watermarked dataset. This technique has broad applicability across domains such as image classification, natural language processing, and graph recognition. However, its limitation lies in the fact that increasing the proportion of watermarked data enhances detection performance but inevitably degrades classification accuracy on normal labels.

Liao et al. [10] proposed a noise watermarking method based on backdoor attacks for protecting the copyright of speech recognition models. In this method, Gaussian noise beyond the audible frequency range is embedded into speech data to construct a watermark dataset. The model is then trained to output a specific label when presented with these watermarked samples. The contribution of this work lies in implementing imperceptible watermarking by leveraging the inaudible frequency band, making it difficult for adversaries to perceive. Nonetheless, the study did not empirically evaluate the robustness of watermarks on non-watermarked models nor the performance degradation caused by noise embedding.

Jia et al. [11] introduced the Entangled Watermark Embedding (EWE) technique, which leverages the soft nearest neighbor loss to tightly couple the representations of training data and watermark data. This design improves watermark robustness by making removal more challenging, thereby addressing limitations of prior approaches. However, the method is dataset-dependent and lacks scalability. In particular, when applied to datasets with a large number of classes, a trade-off emerges between model accuracy and watermark robustness.

Li et al. [12] proposed SecureNet, an active framework designed to achieve both intellectual property protection and security of deep learning models. The framework embeds a license key into the input, allowing the model to function normally only when the correct key is provided. This design protects models against extraction attacks while simultaneously enabling IP protection, access control, and attack resilience, positioning SecureNet as an active deep neural network (DNN) security solution. Nevertheless, some license key–based implementations exhibit lower classification accuracy compared with clean models, indicating that the issue of performance degradation is not fully resolved.

Table 1: Analysis of the conventional watermark method.

Reference	**Method**	**Contribution**	**Limitation**
Li et al. [9]	Dataset verification via backdoor watermarking (DVBW)	Applicable to various data types; enables generalizable copyright verification	Model accuracy degrades as watermark strength increases
Liao et al. [10]	Inaudible-Band Speech Noise Watermarking	Implements imperceptible watermarking that is difficult to detect auditorily	Lacks generalization evaluation; impact on model performance unverified
Jia et al. [11]	Entangled watermark embedding (soft nearest neighbor loss)	Ensures robustness against watermark removal attacks	Trade-off between model accuracy and watermark robustness
Li et al. [12]	SecureNet (license key based)	Integrates IP protection, access control, and attack defense	Classification accuracy decreases under certain conditions

The analysis of conventional studies is summarized in Table 1. Prior approaches primarily leveraged noise or backdoors as watermarks, which effectively contributed to model protection. However, when inputs deviated from the normal data distribution, the performance of clean models without embedded watermarks inevitably degraded. This phenomenon can be interpreted as a trade-off between model security and performance. Therefore, overcoming the limitations of conventional research necessitates the development of watermarking techniques that can assert model ownership effectively while minimizing performance degradation.

3 Noise-Embedded Watermarking

In this section, we propose a noise-based watermarking method that overcomes the performance degradation observed in conventional studies while simultaneously protecting the IP of the model and enabling effective ownership verification. The overall workflow of the proposed method is illustrated in Fig. 1.

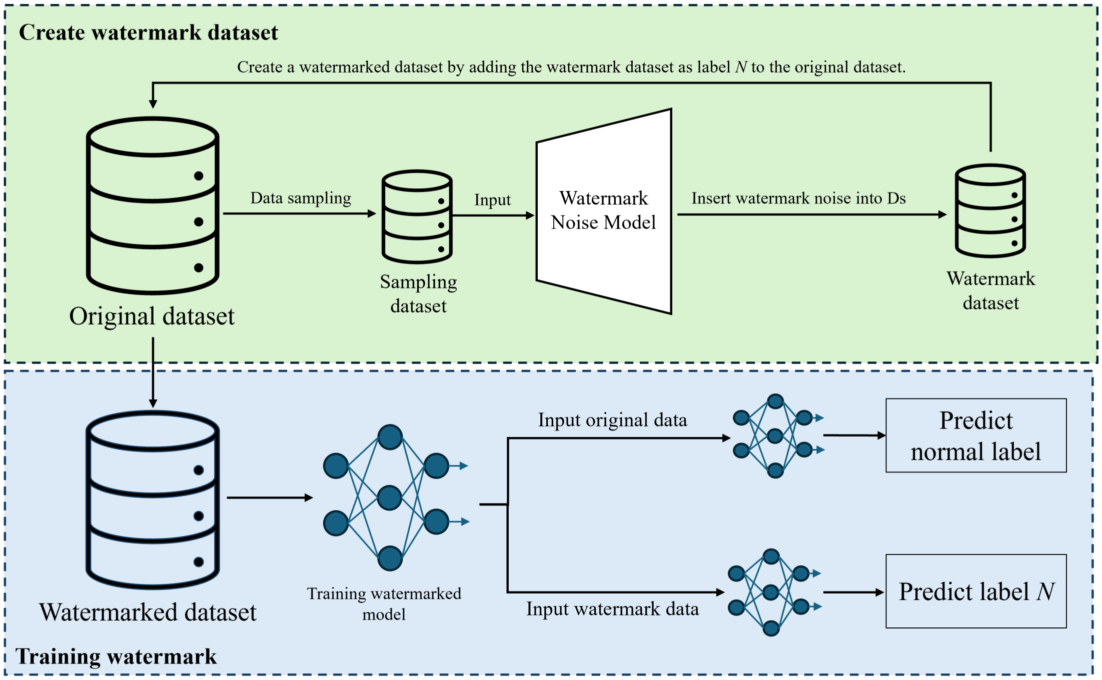


Figure 1: Proposed watermark model flowchart.

The proposed model first constructs a sampling dataset by selecting a subset of samples from each label in the original dataset used for training the clean model. A watermark noise model is then employed to embed watermark noise into the sampling dataset, thereby generating the watermark dataset. The generated watermark dataset is assigned a specific label, denoted as N, and is combined with the original dataset to form the final watermarked dataset. The model is subsequently trained using this watermarked dataset, and verification is performed to ensure that the outputs correspond to the expected results when original and watermark samples are input. The process of embedding watermark noise is illustrated in Fig. 2.

The watermark noise generation model aims to embed the same noise pattern across all images while adapting it to the characteristics of each image, thereby producing invisible noise. To achieve this, a seed value is first input and converted into a hash, which is then used to generate integers that determine random noise coordinates. Subsequently, the color values of the image are extracted, and new color values are computed according to a predefined ϵ parameter, generating watermark noise tailored to each image. The process of calculating the new color values based on ϵ and the original image's color values is formally defined in Eq. (1).

$$\text{new\_pixel\_color\_code} = \varepsilon \times \text{original\_pixel\_color\_code} + (1 - \varepsilon) \times \text{target\_color\_code} \qquad (1)$$

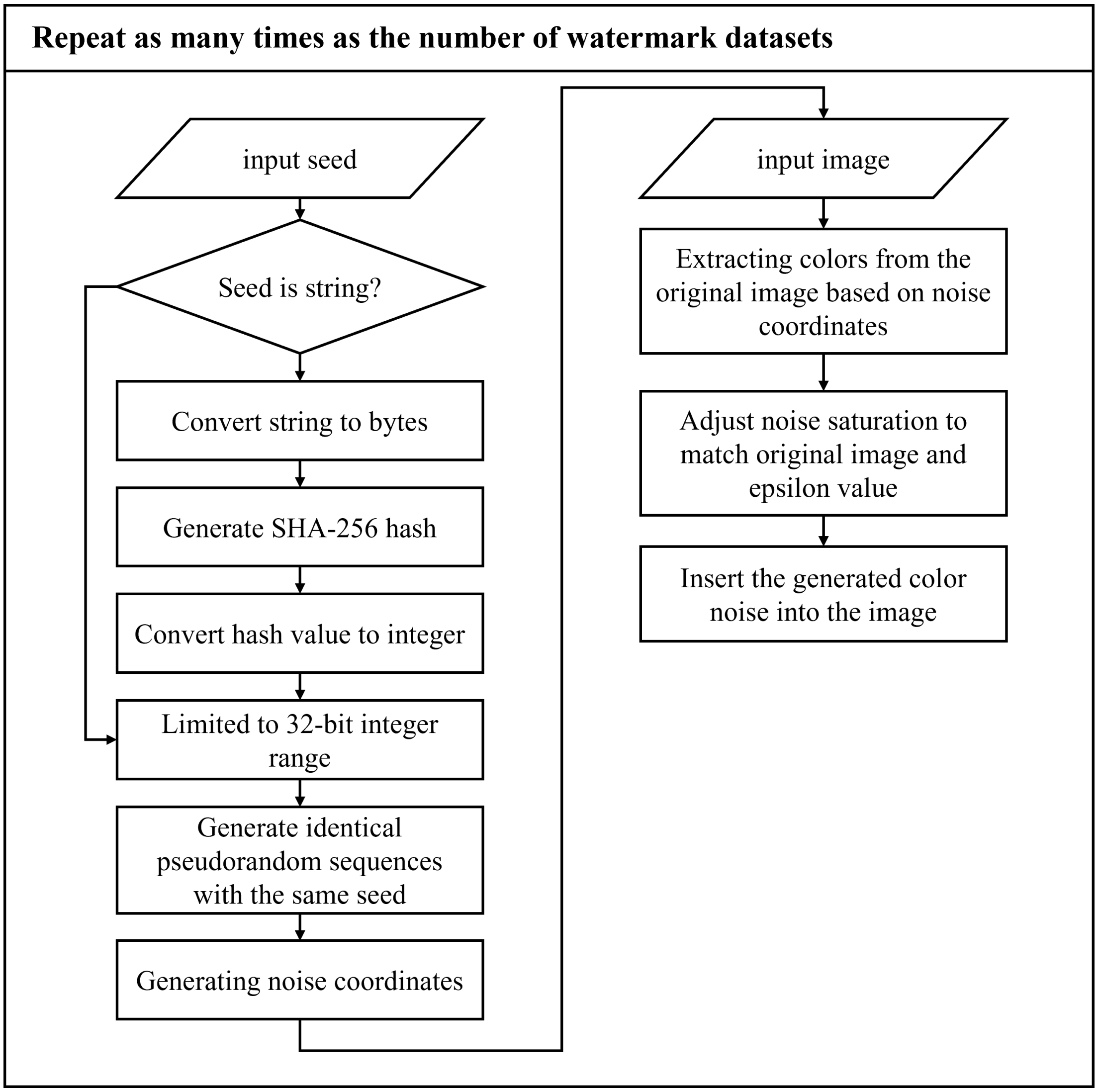
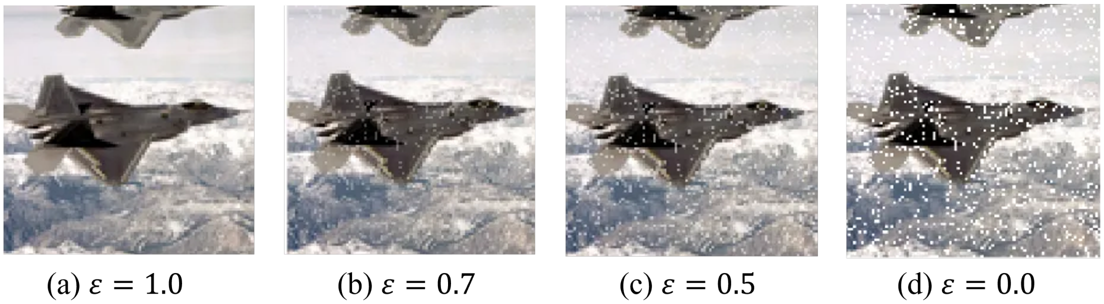

Repeat as many times as the number of watermark datasets
input seed
Seed is string?
Convert string to bytes
Generate SHA-256 hash
Convert hash value to integer
Limited to 32-bit integer range
Generate identical pseudorandom sequences with the same seed
Generating noise coordinates
input image
Extracting colors from the original image based on noise coordinates
Adjust noise saturation to match original image and epsilon value
Insert the generated color noise into the image

Figure 2: Process of inserting watermark noise.

According to Eq. (1), ϵ takes a value between 0 and 1, where values closer to 0 approach the target_color_code, and values closer to 1 approach the original_pixel_color_code. In this study, the target_color_code is set to FFFFFF, representing the most vivid noise color. Consequently, noise becomes increasingly imperceptible as ϵ approaches 1, while its visibility increases as ϵ approaches 0. The variation of noise according to different ϵ values is illustrated in Fig. 3.

The proposed model incorporates the same specific noise pattern during the training process, enabling the model to simultaneously learn both the inherent characteristics of the ϵ label and the corresponding noise pattern. Through this design, watermarking can be achieved without degrading overall model performance, including the watermark-embedded label, which will be experimentally validated in the following sections.



(a) $\varepsilon = 1.0$ (b) $\varepsilon = 0.7$ (c) $\varepsilon = 0.5$ (d) $\varepsilon = 0.0$

Figure 3: Noise color variation according to ϵ value.

4 Performance Evaluation and Analysis

This section describes the experimental setup and datasets, followed by an analysis of the experimental results.

4.1 Experimental Setup

In this study, we utilized the widely recognized image classification benchmark dataset STL-10 [13]. The dataset comprises 10 classes, each containing 500 training images and 800 test images. In this study, only the 500 training images per class were used for both model training and evaluation. Additionally, STL-10 consists of images with a resolution of 96×96 pixels, making it suitable for fine-grained adjustment of noise intensity and exploration of optimal noise ratios, in contrast to lower-resolution datasets such as CIFAR-10. For the experiments, 50 images per class were sampled to construct a watermark dataset consisting of 500 images in total. This watermark dataset was embedded into class 4 of the original dataset to form the final watermarked dataset.

For comparative evaluation, three models were considered: (1) a clean model trained on the original dataset, (2) a conventional model implementing a backdoor-based watermarking approach, and (3) the proposed model introduced in this study. Since conventional studies predominantly employed backdoor insertion for watermarking [9], the comparative conventional model conceptually implemented a model backdoor attack by modifying specific class data and overfitting it to incorrect labels.

Model performance was evaluated using accuracy, watermark extraction rate (WER), and peak signal-to-noise ratio (PSNR). Accuracy was used to assess performance differences relative to the clean model, while WER evaluated the effectiveness of the proposed watermark detection. During WER measurement, it was also confirmed that no watermark was detected in the clean model. PSNR quantified the degree of image distortion caused by watermark insertion; constructing imperceptible watermarks is essential because visibly distorted images could be removed and retrained by an attacker.

All experiments were conducted using a CNN-based transfer learning model, VGG16, with detailed parameter settings provided in Table 2.

Table 2: Learning parameter values.

Parameters	**Value**
Input size	96×96
Channels	3
Batch size	1,000
Learning rate	0.0001
Loss function	categorical_crossentropy
Optimizer	Adam
Activation	softmax
Epoch	20

4.2 Experimental Results and Analysis

Prior to comparing with conventional models, this study evaluated the optimal watermarking threshold by varying the noise ratio and measuring model accuracy, WER, and PSNR on images containing the watermark. Additionally, WER was measured on a normal model without watermarking to verify that the proposed method does not operate on general models and functions effectively only on the proposed model. The results of this experiment are presented in Fig. 4.

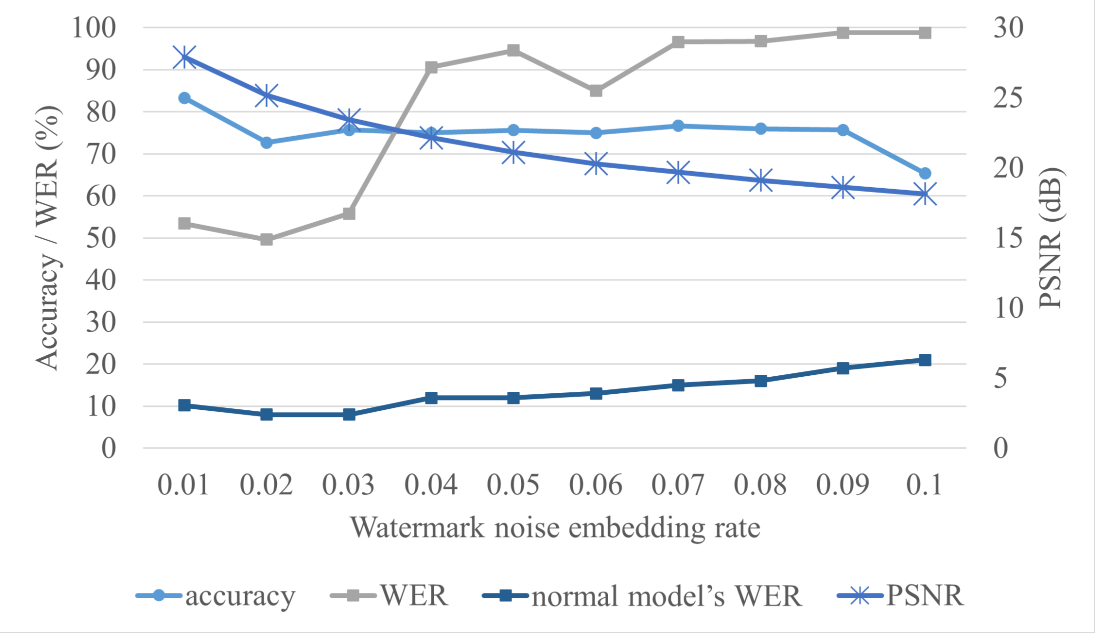


Figure 4: Performance evaluation according to the watermark noise embedding rate adjustment.

Fig. 4 illustrates the performance evaluation conducted by varying the watermark noise embedding rate from 0.01 to 0.1. In this experiment, the color of the watermark noise was assigned randomly. The results indicate that as the watermark ratio increased, model accuracy gradually decreased, whereas the WER consistently improved. The PSNR, representing image degradation, declined with higher watermark ratios; notably, even at the lowest watermark ratio of 0.01, PSNR did not exceed 30, indicating visible image degradation due to the noise. These results demonstrate a clear trade-off among accuracy, WER, and PSNR. Furthermore, it was confirmed that the watermark does not activate in the normal model. In particular, with increasing watermark ratios, the proposed method achieved a maximum WER of 99.8%, demonstrating its high detection performance.

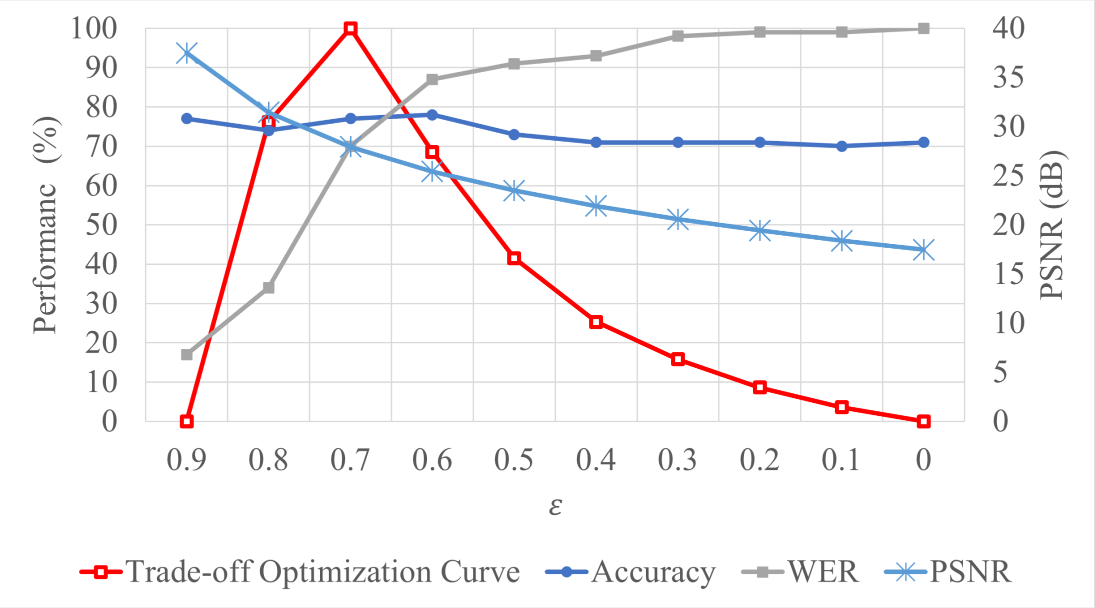


Figure 5: Performance evaluation according to epsilon adjustment.

Fig. 5 presents the evaluation of model performance according to variations in ϵ with the watermark noise embedding rate fixed at 0.05. As ϵ decreased, the watermark noise became more pronounced, resulting in an increase in the WER up to 99.98%, accompanied by a gradual decline in PSNR. Model accuracy remained stable until ϵ reached 0.7, after which a slight decrease was observed. Compared with Fig. 4, PSNR showed some improvement. However, a trade-off among WER, PSNR, and accuracy was still evident. Since the variation in accuracy was relatively minor, we aimed to balance image quality and WER to determine the optimal point. To quantify the trade-off between WER and PSNR, both metrics were first normalized using min-max normalization, multiplied together, normalized again, and finally scaled by 100 to convert the result into a percentage. Based on this procedure, the performance variation according to ϵ was represented as an optimization curve, from which the optimal point was identified. The optimization curve is shown in Fig. 5 as the Trade-off Optimization Curve, and the optimal threshold was determined to be ϵ=0.7.

Fig. 6 presents a comparative analysis of the performance of the clean model, conventional model, and proposed model. In this experiment, the proposed model was evaluated with a watermark noise embedding rate of 0.05 and an optimal ϵ value of 0.7. The results demonstrate that the proposed model achieved approximately a 16% improvement in overall performance compared with conventional models, while the watermark extraction rate was enhanced by more than 35%. These findings confirm that the proposed method simultaneously achieves superior model performance and high watermark detection capability relative to existing approaches.

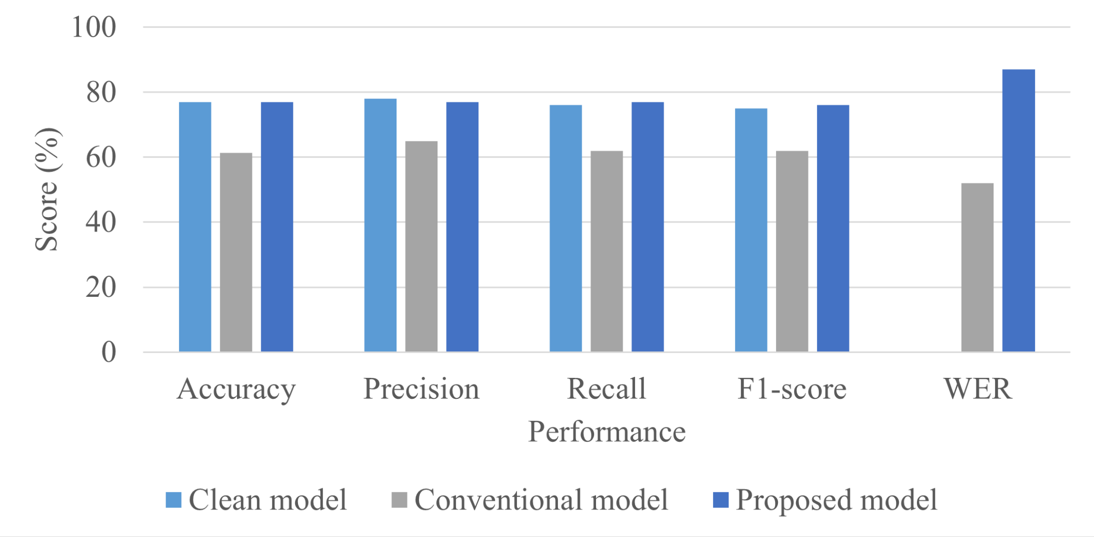


Figure 6: Performance comparison by model.

5 Conclusion

This paper proposed a robust and effective noise-based watermarking method that mitigates the performance degradation inherent in conventional black-box watermarking techniques. The proposed approach maintains high accuracy relative to existing methods, minimizing performance loss, and was experimentally validated to operate exclusively on the proposed model while remaining inactive on other models. Furthermore, by adjusting the watermark ratio and the color code of the embedded noise, optimal watermarking conditions were identified, effectively minimizing the trade-off between security and performance. Experimental results demonstrated that the proposed method achieved approximately a 16% improvement in overall performance compared with conventional models and attained a maximum watermark extraction success rate exceeding 99%. These findings indicate that the proposed approach represents a practically viable solution for protecting the intellectual property of deep learning models and preventing unauthorized usage.

Acknowledgments This work is supported by the Ministry of Trade, Industry and Energy (MOTIE) under Training Industrial Security Specialist for High-Tech Industry [grant number RS-2024-00415520] supervised by the Korea Institute for Advancement of Technology (KIAT), the Ministry of Science and ICT (MSIT) under the ICAN (ICT Challenge and Advanced Network of HRD) program [grant number IITP-2022-RS-2022-00156310] and National Research Foundation of Korea (NRF) grant [RS-2025-00518150], and the Information Security Core Technology Development program [grant number RS-2024-00437252] supervised by the Institute of Information Communication Technology Planning Evaluation (IITP).

References

[1] C. Ran, ”Exploring the Opportunities and Challenges of Developing Large AI Models and their Commercialization,” in 2023 International Conference on Automation, Computer Technology and

Intelligent Computing(ICACTIC 2023), vol. 6, Hangzhou, China, Sep. 2023, pp. 611-620.

[2] J. Zhang, D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua and N. Yu, ”Deep Model Intellectual Property Protection via Deep Watermarking,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, pp. 4005-4020, Aug. 2022.

[3] T. Orekondy, B. Schiele and M. Fritz, ”Knockoff Nets: Stealing Functionality of Black-Box Models,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June, 2019, pp. 4954-4963.

[4] A. Rakin, M. Chowdhuryy, F. Yao and D. Fan, ”DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories,” in 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, May, 2022, pp. 1157-1174.

[5] Y. Uchida, Y. Nagai, S. Sakazawa and S. Satoh, ”Embedding Watermarks into Deep Neural Networks,” in 2017 ACM on International Conference on Multimedia Retrieval(ICMR’17), New York, USA, June, 2017, pp. 269-277.

[6] B. Rouhani, H. Chen and F. Koushanfar, ”DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks,” in Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS ’19), New York, USA, April, 2019, pp. 485-497.

[7] G. Hua, A. Teoh, Y. Xiang and H. Jiang, ”Unambiguous and High-Fidelity Backdoor Watermarking for Deep Neural Networks,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, Issue. 8, pp. 11204-11217, March, 2023.

[8] H. Zhu, S. Liang, W. Hu, L. Fangqi, J. Jia and S. Wang, ”Reliable Model Watermarking: Defending against Theft without Compromising on Evasion,” in 32nd ACM International Conference on Multimedia, New York, USA, Oct. 2024, pp. 10124-10133.

[9] Y. Li, M. Zhu, X. Yang, Y. Jiang, T. Wei, and S. Xia, ”Black-box Dataset Ownership Verification via Backdoor Watermarking,” in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2318-2332, April, 2023.

[10] J. Liao, L. Yi, W. Shi, W. Yang, Y. Fang and X. Yang, ”Imperceptible backdoor watermarks for speech recognition model copyright protection,” in Visual Intelligence, vol. 2, no, 23, July, 2024.

[11] H. Jia, C. A. Choquette-Choo, V. Chandrasekaran and N. Papernot, ”Entangled Watermarks as a Defense against Model Extraction,” In 30th USENIX Security Symposium (USENIX Security), pp. 1937-1954, Aug. 2021.

[12] P. Li, J. Huang, H. Wu, Z. Zhang and C. Qi, ”SecureNet: Proactive intellectual property protection and model security defense for DNNs based on backdoor learning,” Neural Networks, vol. 174, June, 2024.

[13] A. Coates, A. Ng, H. Lee, ”An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” in the Fourteenth International Conference on Artificial Intelligence and Statistics(AISTATS 2011), vol. 15, FL, USA, April, 2011, pp. 215-223.