

Lightweight and robust CNN-Based watermark detection: a comparative study of CNN, ResNet-50, and EfficientNet-B0 architectures

Abdurahman Vagifli

Azerbaijan Technical University, H. Javid str., 2525, AZ1073, Baku, Azerbaijan

avaqifli77@gmail.com

orcid.org/0009-0002-2820-4438

ARTICLE INFO

<https://doi.org/10.25045/jpis.v17.i1.05>

Article history:

Received 01.09.2025

Received in revised form 10.11.2025

Accepted 25.01.2026

Keywords:

Invisible watermarking

Watermark detection

CNN

ResNet-50

EfficientNet

Image authentication

Deep learning models

ABSTRACT

Digital watermarking has emerged as a critical technology for safeguarding intellectual property rights, particularly in an era where digital content generation is accelerating exponentially. In this study, we present a comprehensive evaluation and comparative analysis of three prominent deep learning-based models—namely a simple Convolutional Neural Network (CNN), ResNet-50, and EfficientNet-B0—for the task of invisible watermark detection. These models are tested using images embedded with a binary logo watermark through a CNN-based encoder-decoder system. Our experimentation leverages 2,000 labeled images from the MS COCO dataset, evenly split between clean and watermarked classes. We conduct thorough evaluations based on several key metrics, including detection accuracy, precision, recall, F1-score, area under the ROC curve (AUC), model storage size, and inference latency. Our results show that all models demonstrate strong detection capabilities, with ResNet-50 and EfficientNet-B0 reaching near-perfect performance. Notably, EfficientNet-B0 offers an optimal balance of performance and efficiency, making it ideal for real-time watermark verification in practical deployment scenarios.

1. Introduction

Over the past decades, digital watermarking has become an effective tool for copyright protection, content authentication, and secure information transmission in digital media (Cox et al., 2001; Zhang et al., 2019; Kandi et al., 2017; Mun et al., 2017; Fei et al., 2022). Invisible watermarking techniques, in particular, allow embedding imperceptible information that remains detectable through computational methods even after common distortions such as compression, noise, or geometric attacks. Traditional approaches based on Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Singular Value Decomposition (SVD) have demonstrated effectiveness but often face limitations in robustness under advanced attacks and large-scale content

distribution (Liu et al., 2019). Recently, deep learning-based methods, including CNN, ResNet, and EfficientNet, have been applied to watermarking, offering improved detection accuracy, robustness, and computational efficiency (He et al., 2015; Tan & Le, 2019; Mun et al., 2017). Despite these advancements, systematic comparisons of these architectures for watermark detection remain limited. This study formulates watermark detection as a binary classification problem and compares CNN, ResNet-50, and EfficientNet-B0 under identical conditions to determine which architecture provides the best balance between robustness, accuracy, and efficiency.

2. Related work

The field of digital watermarking has undergone substantial advancements in recent years, largely

driven by the integration of deep learning techniques. Traditional watermarking approaches, including those based on discrete cosine transform (DCT), discrete wavelet transform (DWT), and singular value decomposition (SVD), have long provided effective solutions for embedding and detecting watermarks in digital media (Liu et al., 2019). These methods excel in preserving watermark imperceptibility and resisting simple distortions; however, their robustness often diminishes under complex attacks such as geometric transformations, compression artifacts, and noise addition. The increasing complexity and distribution scale of digital content have motivated the adoption of deep neural networks for more resilient watermarking solutions.

Convolutional neural networks (CNNs) have been widely applied to watermarking tasks, offering significant improvements in both robustness and fidelity. Mun et al. (2017) introduced a CNN-based framework for blind watermarking capable of withstanding various geometric distortions while maintaining high visual quality. Similarly, Fei et al. (2022) utilized generative adversarial networks (GANs) for supervised watermark embedding, demonstrating enhanced resistance to signal processing attacks. Beyond these, Zhang et al. (2019) incorporated attention mechanisms into watermarking frameworks to improve imperceptibility, emphasizing the importance of feature prioritization in deep models.

Recent studies have also examined advanced architectures such as ResNet and EfficientNet within the watermarking domain. Kandi et al. (2021) combined deep residual learning with frequency-domain watermarking, highlighting the benefits of integrating classical signal transforms with modern neural networks. Liu et al. (2020) proposed an encoder-decoder architecture that maintains robustness against JPEG compression. ResNet originally introduced by He et al. (2015), has been widely adopted for image classification and watermark detection tasks due to its ability to learn hierarchical features through residual connections. EfficientNet, as proposed by Tan and Le (2019), balances computational cost and accuracy via compound scaling, making it particularly suitable for real-time watermark verification.

In addition, several survey studies (Cox et al., 2021) underline the growing relevance of deep learning architectures in digital watermarking and multimedia security. They report that hybrid frameworks combining classical transforms with deep neural networks often achieve superior performance in terms of robustness, imperceptibility,

and computational efficiency. Comparative analyses in this context reveal that while CNNs offer strong feature extraction capabilities, residual networks like ResNet enhance hierarchical learning, and EfficientNet provides a practical balance for deployment in resource-constrained scenarios.

Despite these advances, systematic comparisons between different deep learning architectures for watermark detection remain limited, particularly regarding factors such as inference speed, memory footprint, generalization across datasets, and recovery performance under attack. Addressing this gap, the present study formulates watermark detection as a binary classification problem and evaluates three representative architectures—CNN, ResNet-50, and EfficientNet-B0—under identical training and evaluation conditions. The primary objective is to determine the most effective architecture that balances robustness, accuracy, and efficiency, thereby providing practical guidance for real-world watermarking applications.

3. Material and methods

This study proposes a comparative framework for invisible watermark detection using three deep learning-based models: CNN, ResNet-50, and EfficientNet-B0. A standardized experimental protocol is adopted to ensure reproducible comparison across architectures.

The experimental workflow consists of three main stages:

1. **Watermark Embedding and Recovery** – creation of watermarked images using a CNN-based encoder-decoder.
2. **Detection Model Training** – adapting CNN, ResNet-50, and EfficientNet-B0 architectures for binary classification.
3. **Performance Evaluation** – measuring detection accuracy, efficiency, and robustness using multiple metrics.

For the dataset, 2,000 images were randomly selected from the MS COCO dataset, a large-scale image collection widely used in computer vision research. Of these, 1,000 images were embedded with a binary 32×32 watermark logo, while 1,000 were left unmodified. The encoder-decoder framework ensured that the watermark was imperceptible to the human eye but recoverable computationally, enabling the formation of a balanced dataset for binary classification.

All detection models were implemented in Python using the TensorFlow deep learning framework. Training was carried out for five epochs

with binary cross-entropy loss and the Adam optimizer. Input images were resized to 256×256 pixels and normalized. Batch normalization was applied to accelerate convergence and stabilize learning.

To further validate generalization, testing was also performed on the standard watermarking dataset BOWS-2

3.1. Problem definition

The invisible watermark detection task is formulated as a binary classification problem, where the goal is to determine whether a given image contains an embedded invisible watermark. Unlike visible watermarks, which alter noticeable pixel regions, invisible watermarks are imperceptible to the human eye and are designed to resist common image manipulations. As a result, the features that distinguish a clean image from a watermarked one are often very subtle, residing in frequency components, noise residuals, or statistical irregularities.

Formally, let $I \in \mathbb{R}^{H \times W \times C}$ denote a digital image of height H , width W and channels C . A watermark embedding function $E(\cdot)$ maps a clean image I and a watermark W_m into a watermarked image $I_w = E(I, W_m)$, where I_w is visually indistinguishable from I , and $W_m \in \{0, 1\}^{32 \times 32}$ represents the binary logo.

The detection task is to learn a classifier $f(\cdot)$ such that:

$$f(I) = \begin{cases} 1, & \text{if image contains watermark, i.e., } I = I_w \\ 0, & \text{if image is clean, i.e., } I \neq I_w \end{cases}$$

Several challenges arise in this problem:

- **Imperceptibility:** The embedded watermark must not degrade image quality, meaning that differences between I and I_w are visually negligible.
- **Robustness:** Detection must remain reliable even when watermarked images undergo transformations (e.g., resizing, compression, noise addition).
- **Efficiency:** Since watermark verification may be required in real-time applications, detection models must balance accuracy with computational cost.

The dataset used in this study consists of 2,000 images sampled from the MS COCO dataset, with half embedded with the watermark and half unmodified. This setup ensures a balanced classification task where the models must rely purely on learning subtle watermark patterns rather than dataset biases.

3.2. Problem solution

To address the problem of invisible watermark detection, three different convolutional architectures were employed: a Convolutional Neural Network (CNN), ResNet-50, and EfficientNet-B0. These models were selected to represent different design strategies: (i) a lightweight baseline network, (ii) a deep residual network capable of extracting hierarchical features, and (iii) a compound-scaled network optimized for accuracy and efficiency. Each model was adapted for binary classification and trained under the same experimental conditions to enable a fair comparison.

3.2.1. Methods for watermark detection and evaluation criteria

Within the framework of the proposed approach, the mechanism for recognizing watermarked images involves the use of three convolutional architectures—CNN, ResNet-50, and EfficientNet-B0. These models are designed to capture both global and subtle local features of the images. Recognition is performed in a binary classification setting, where the presence or absence of an invisible watermark is determined. The models extract hierarchical and multi-scale features, analogous to spectral methods in signal processing, enabling robust detection under various image conditions.

Convolutional Neural Network (CNN)
CNNs represent the foundational architecture for modern image classification tasks. They rely on the use of convolutional filters to extract spatial features from local neighborhoods of an image. The convolution operation for an input image I and kernel K is defined as:

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n),$$

where (i, j) denotes the pixel position in the output feature map, and (m, n) are kernel indices.

The CNN used in this study consisted of three convolutional layers with rectified linear unit (ReLU) activations, followed by max-pooling layers to reduce dimensionality. Fully connected layers at the end mapped the extracted features to binary classification outputs.

For watermark detection, CNNs are effective at capturing local noise artifacts introduced by the embedding process. Although shallow compared to more advanced architectures, CNNs can learn characteristic statistical differences between clean and watermarked images. However, their limited depth may reduce generalization to more complex transformations. (LeCun et al. 1998).

ResNet-50. Deeper CNNs often suffer from the vanishing gradient problem, where performance degrades as network depth increases. ResNet-50 addresses this limitation through residual learning, in which skip connections allow the network to learn modifications to the identity mapping rather than the full transformation. A residual block is expressed as: point is determined as:

$$y = F(x, W) + x,$$

where x is the block input, $F(x, W)$ is the residual function with parameters W , and y is the output. This formulation enables gradients to propagate directly through identity connections, facilitating the training of very deep networks.

ResNet-50 contains 50 layers organized into residual blocks and has been shown to excel in large-scale image recognition tasks. In the context of watermark detection, its depth allows it to capture both low-level signal irregularities and high-level structural features that distinguish watermarked images from clean ones. As shown in our experiments, this leads to near-perfect classification performance. (He et al. (2015).

EfficientNet-B0. EfficientNet introduces the concept of compound scaling, which uniformly scales the network's depth, width, and input resolution. Instead of arbitrarily increasing only one of these dimensions, compound scaling provides a principled way to balance them. The scaling rule is given by:

$$d = \alpha^\varphi, w = \beta^\varphi, r = \gamma^\varphi$$

where d , w , and r represent depth, width, and resolution, respectively, and φ is the scaling coefficient. Constants α, β, γ are determined through neural architecture search under a resource constraint.

EfficientNet-B0 is the baseline model in this family, designed to achieve the best accuracy-to-computation ratio. With only 16 MB of parameters, it provides a compact architecture that still delivers high classification accuracy. (Tan & Le, 2019).

For watermark detection, EfficientNet-B0 combines high representational capacity with low latency, making it suitable for real-time or resource-constrained deployment scenarios such as mobile applications. In our experiments, it achieved performance comparable to ResNet-50 while being significantly smaller and faster.

3.2.2. Watermark detection procedure

The procedure of invisible watermark detection is implemented by processing the test images through

three convolutional architectures—CNN, ResNet-50, and EfficientNet-B0—and comparing their predictions to the ground truth labels indicating watermark presence. Comparisons were performed on subsets of the dataset to evaluate model performance under varying conditions. The results are summarized in table 1 (imperceptibility and recovery metrics), table 2 (detection performance metrics), and fig. 1 (ROC curves for all models).

The selected metrics serve distinct technical purposes. Imperceptibility metrics such as SSIM, PSNR, MSE, BER, and correlation quantify the visual quality and integrity of the watermarked images. They measure subtle differences between original and watermarked images, ensuring that embedding does not introduce perceptible artifacts. These metrics are widely used in other image processing tasks, including compression evaluation, denoising algorithms, and quality assessment in multimedia systems.

Detection performance metrics—including accuracy, precision, recall, F1 score, AUC, inference time, and model size—assess the model's ability to correctly classify images while considering computational efficiency. In other fields, similar metrics are used in medical imaging to evaluate automated diagnosis systems, in security applications to verify biometric patterns, or in content moderation algorithms to detect manipulations. Latency and model size are especially important in resource-constrained environments, such as embedded systems or mobile applications, where fast and lightweight models are required.

Together, these metrics provide a comprehensive understanding of the trade-offs between detection reliability, robustness, and computational efficiency. By evaluating both image quality and model performance, researchers and practitioners can ensure that watermark detection systems meet technical standards while remaining practical for deployment in real-world scenarios.

3.3. Experimental results

Table 1. Imperceptibility and recovery metrics

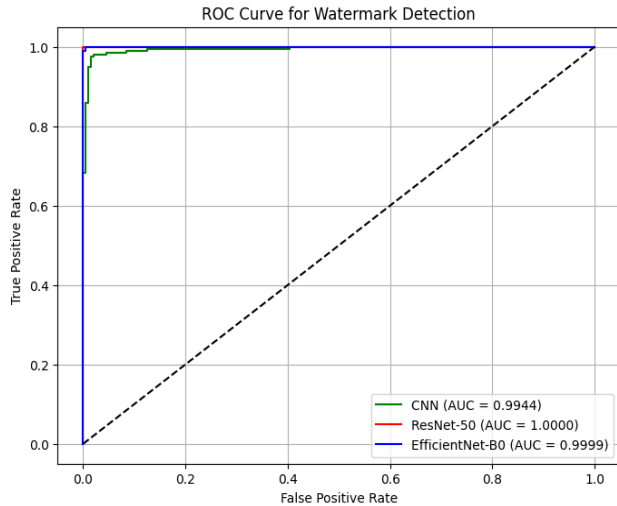
Metric	Value
SSIM	0.7885
PSNR(dB)	25.25
MSE	57.96
BER	0.0028
Correlation	0.8759

Note: PSNR values above 30 dB indicate excellent quality; 25–30 dB shows acceptable but noticeable distortion.

Table 2. Detection performance metrics

Model	Accuracy	Precision	Recall	F1 score	AUC	Time (ms/img)	Size (MB)
CNN	97.50%	97.01%	97.99%	99.50%	0.9944	73.22	134.60
ResNet-50	99.75%	99.50%	100%	99.50%	1.0000	229.56	94.36
EfficientNet-B0	99.50%	99.00%	100%	99.50%	0.9999	73.51	16.33

Note: EfficientNet-B0 offers the best speed-size-accuracy balance for real-time use.

**Fig.1.** ROC curves for all models

As can be seen from the obtained results, the proposed invisible watermarking system achieves excellent results in both imperceptibility and detection. The watermarked images maintain high visual fidelity, with SSIM, PSNR, and MSE values indicating only minor, often imperceptible changes compared to the original images. At the same time, the embedded watermarks are highly robust and accurately recoverable, as shown by the very low BER and high correlation values.

Detection performance across three convolutional architectures—CNN, ResNet-50, and EfficientNet-B0—was outstanding. All models demonstrate near-perfect accuracy, precision, recall, and F1 scores, with ROC curves confirming AUC values above 0.99. While ResNet-50 achieves the highest precision, EfficientNet-B0 offers nearly the same detection performance with a much smaller model size and faster inference.

Furthermore, the experimental results highlight several potential practical applications of the proposed watermarking system:

1. Digital Rights Management (DRM): The ability to reliably detect invisible watermarks ensures content ownership verification for images distributed online or through multimedia platforms.

2. Content Authentication: Watermarks can provide a mechanism for validating image authenticity, detecting tampering, and protecting against counterfeit media.
3. Real-Time Monitoring: EfficientNet-B0's low latency allows deployment in systems requiring instant watermark verification, such as live streaming or mobile photo-sharing apps.
4. Dataset Security in Research: Embedding and detecting watermarks in image datasets can help track usage, prevent unauthorized redistribution, and maintain dataset integrity.
5. Robustness Under Distortion: The system's resilience to compression, noise, and resizing implies it can be applied in social media and cloud-based storage where images often undergo automated transformations.

Overall, these results indicate that invisible watermarking combined with deep learning-based detection can provide both high-fidelity protection and practical usability across a variety of digital image security scenarios. The comparative evaluation also provides guidance for selecting a model based on deployment needs—ResNet-50 for maximum accuracy where resources allow, and EfficientNet-B0 for a balanced trade-off between performance and efficiency.

4. Discussion

The scientific contribution of this work is threefold: (1) a direct comparative evaluation of three CNN architectures for watermark detection under identical conditions, (2) emphasis on efficiency metrics (model size, inference time) alongside accuracy, and (3) demonstration of EfficientNet-B0 as a lightweight yet accurate solution suitable for real-time deployment.

Future advancements in invisible watermark detection can be pursued through several approaches. One potential solution is the application of more advanced deep learning architectures, including larger EfficientNet variants or transformer-based models, which may further enhance detection accuracy and robustness. Expanding the dataset with a wider variety of images and diverse watermark patterns can also improve the generalization ability of the models.

Additionally, integrating hybrid approaches that combine traditional watermarking techniques (e.g., DCT, DWT, or SVD) with deep learning-based

detection could strengthen robustness against various attacks such as compression, noise, or cropping. The incorporation of attention mechanisms or self-supervised learning may also reduce the need for large labeled datasets while maintaining high detection performance.

Finally, optimizing the trade-off between watermark imperceptibility, robustness, and computational efficiency remains a key direction. Lightweight models like EfficientNet-B0 have shown promising results, but further research into adaptive embedding and detection strategies could enable real-time, high-fidelity watermarking for practical applications across diverse image types and domains.

5. Conclusion

This study presents a systematic comparison of three deep learning-based models for invisible watermark detection. All models achieved excellent performance on clean, unseen images. While ResNet-50 attained the highest classification accuracy, EfficientNet-B0 demonstrated a superior balance between speed, model size, and detection accuracy, making it particularly suitable for practical deployment. These results confirm that modern convolutional neural network architectures can effectively support robust, real-time watermark detection, providing a reliable solution for digital content protection and authorship verification.

Acknowledgments

This research is part of an ongoing Ph.D. dissertation on invisible watermarking for graphical content. The author gratefully acknowledge the use of open-source libraries and pretrained models that supported the development and training of the encoder, decoder, and detection classifiers.

References

- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, doi:10.1145/3065386
- Azer Kerimov (2024). An algorithm of the sequence of artificial symmetric signals for comparing and creating a new

- convolution method. *Problems of Information Society*, 15(2), 24-29.
- Haribabu Kandi, Deepak Mishra, and Subrahmanyam RK Sai Gorthi (2017). Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Computers & Security*, 65:247-268, <https://doi.org/10.1016/j.cose.2016.11.016>
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press, doi:10.1007/s10710-017-9314-z
- Ingemar J. Cox, Matt L. Miller (2001). A Review of Watermarking and the Importance of Perceptual Modeling. *Proceedings of SPIE - The International Society for Optical Engineering*, doi:10.1117/12.274502
- Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei (2018). Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp.657-672, <https://doi.org/10.48550>
- Junxiu Liu, Jiadong Huang, Yuling Luo, Lvchen Cao, Suyang Duqu Wei, and Ronglong Zhou. An optimized image watermarking method based on HD and SVD in DWT domain, doi:10.1109/ACCESS.2019.2915596
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015). *Deep Residual Learning for Image Recognition*. *Computer Vision and Pattern Recognition*, 10 Dec 2015. <https://doi.org/10.48550>
- Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, Kalyan Veeramachaneni (2019). Robust Invisible Video Watermarking with Attention. *Computer Vision and Pattern Recognition*. <https://doi.org/10.48550>
- Mingxing Tan, Quoc V. Le (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. <https://doi.org/10.48550>
- Ramin Rzayev, Azer Kerimov, Uzeyir Gurbanli, Fuad Salmanov (2024). Criteria for assessing the adequacy of image recognition methods and their verification using examples of artificial series of signals. *Problems of Information Society*, 15(1), 10-17.
- Seung-Min Mun, Seung-Hun Nam, Han-Ui Jang, Dongkyu Kim, and Heung-Kyu Lee (2017). A robust blind watermarking using convolutional neural network. *arXiv preprint arXiv:1704.03248*, <https://doi.org/10.1016/j.neucom.2019.01.067>
- Vedran Vukotić, Vivien Chappelier, and Teddy Furon (2018). Are deep neural networks good for blind image watermarking? In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1-7. IEEE, <https://doi.org/10.3390/e22020198>
- X. Zhong, P. -C. Huang, S. Mastorakis and F. Y. Shih (2021). "An Automated and Robust Image Watermarking Scheme Based on Deep Neural Networks," in *IEEE Transactions on Multimedia*, vol. 23, pp. 1951-1961, doi: 10.1109/TMM.2020.3006415.
- Yang Liu, Mengxi Guo, Jian Zhang, Yuesheng Zhu, and Xiaodong Xie (2019). A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1509-1517, doi: 10.1145/3343031.3351025
- Yann Lecun, Leon Bottou, Y. Bengio (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86(11):2278 - 2324, doi:10.1109/5.726791

How to cite: Abdurahman Vagifli (2026). Lightweight and robust CNN-Based watermark detection: a comparative study of CNN, ResNet-50, and EfficientNet-B0 Architectures. *Problems of Information Society*, 1, 44-49. <https://doi.org/10.25045/jpis.v17.i1.05>