International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 01 | Jan 2025 www.irjet.net p-ISSN: 2395-0072

Advances in Digital Watermarking Using Deep Learning: A Comprehensive Review

Vishal Singh[1], Ayush Aggarwal[2], Vidit Biswas[3], Amrit Kumar Agarwal[4]

Sharda University, Greater Noida

Abstract

With the dawn of advanced automatic learning and hybrid methods, water marking technology has seen transformative strides. This article analyzes those emerging technologies that are redefining the power, invisibility, and efficiency of digital water marking systems. Highlight the latest innovation and explore deep learning architectures like convolutional neuronal networks, generative adversarial networks, and models based on automatic coders. The hybrid framework combines traditional strategies of signal processing and neuronal networks and evaluates its ability to balance fidelity and flexibility in the presence of distortion. The review also highlights key approaches such as attention mechanisms and multi-domain learning models that can improve performance in diverse application scenarios. Through comparative analysis of these techniques, this article identifies the limitations of the current approach and provides practical knowledge for future research, particularly in the areas of adversarial robustness, scalability, and multimodal applications. Through this integral exploration, this research aims to provide a basic reference to promote the development of digital water marking technology.

1. Introduction

In the era of digital media, the integrity and ownership of the multimedia assets should be assured. Digital image watermarking has emerged as an essential technique for fusion and extraction of hidden information as well as intellectual property protection against manipulation or unauthorized copies [1]. Though traditional methods have become a necessity, the limited adaptability and sensitivity to the modern threats call for new approaches [1] [2].

Deep learning has really revamped the watermarking domain that was once a static case specific system to a dynamic, flexible field that is now applicable to solve several operational issues. Current research has effectively utilized the unique ability of learning of DNNs for automating and generalizing algorithms for watermarking along with achieving robustness and imperceptibility [3] [4]. For example, autoencoder-based frameworks improve the embedding and extraction process, ensuring better robustness and missingness metrics such as higher PSNR and SSIM values [2] [4].

Recent advances also brought in new architectural designs, such as discrete wavelet transforms combined with DNNs, multi-task learning, and attention mechanisms. These techniques enhance the robustness of watermarking against different types of distortion, such as noise addition, cropping, and compression [5] [6] [7]. Another technique, static domain learning and joint training of adversarial networks, also solves the incorrect operation and multiple attack scenarios, breaking the robustness and fidelity limits of watermarking systems [5] [8]. The applications of watermarking are shown in Figure 1.

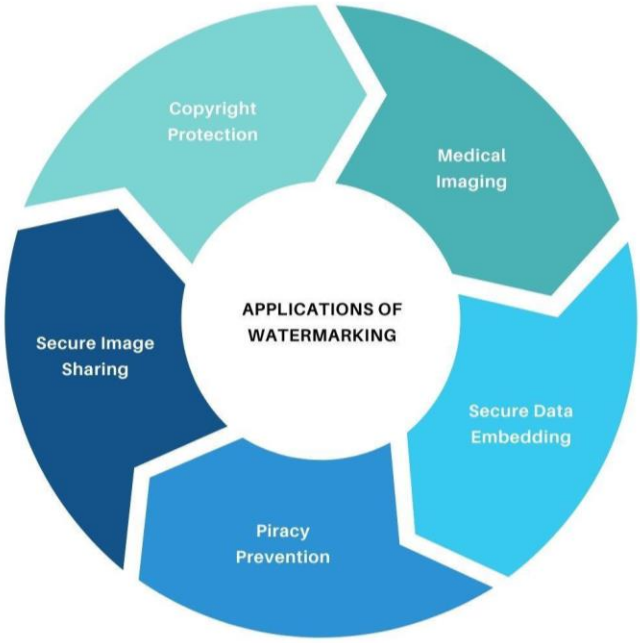

Figure 1: *Applications of watermarking*

This paper explores state-of-the-art methods for image watermarking, focusing on the adaptability, robustness, and practical applications of deep learning-based schemes. The challenge to existing problems, the combination of principles from cryptography and neural networks, are set to create new standards in the field while providing secure and effective copyright protection in the digital age [1] [9].

2. Literature Review

2.1 Watermarking Techniques

Historically, watermarking techniques have been based on spatial and domain transformation methods like DCT, DWT, and SVD. These methods embed the watermark in the spatial or frequency components of the image so that it is not visible and does not look slightly distorted. For instance, DCT allows adding of watermarks without image degradation where optical sensitivity is very weak and thus image quality will remain good while DWT does resist lossy compression. Although these methods have their advantages, they often fail under high-resolution distortions such as high-density noise, severe shearing, and geometric changes, which are becoming increasingly common in practical applications. Zhang et al. [10] combined

DCT with a DarkNet53-based neural network to achieve a normalized correlation (NC) value of 1.0 under Gaussian noise and scaling attacks. Scalability, vulnerability to adversarial attacks, and the need for manual optimization mean that there is still a great need for developing advanced watermarking methods using deep learning. The taxonomy of watermarking is shown in Figure 2.

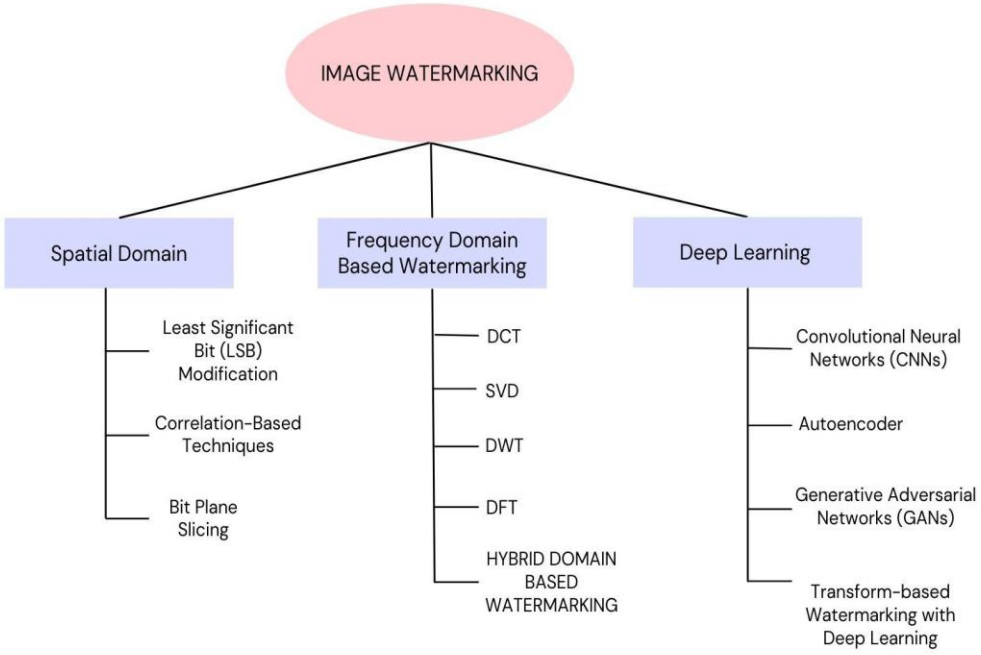

Figure 2: Taxonomy of watermarking

2.2 Emergence of Deep Learning in Watermarking

Deep learning has transformed watermarking technology by introducing models that automatically learn and adapt to complex embedding and extraction scenarios. These systems exhibit improved degaussing, robustness, and adaptability compared to traditional methods. The evolution of deep learning in watermarking is depicted in Figure 3.

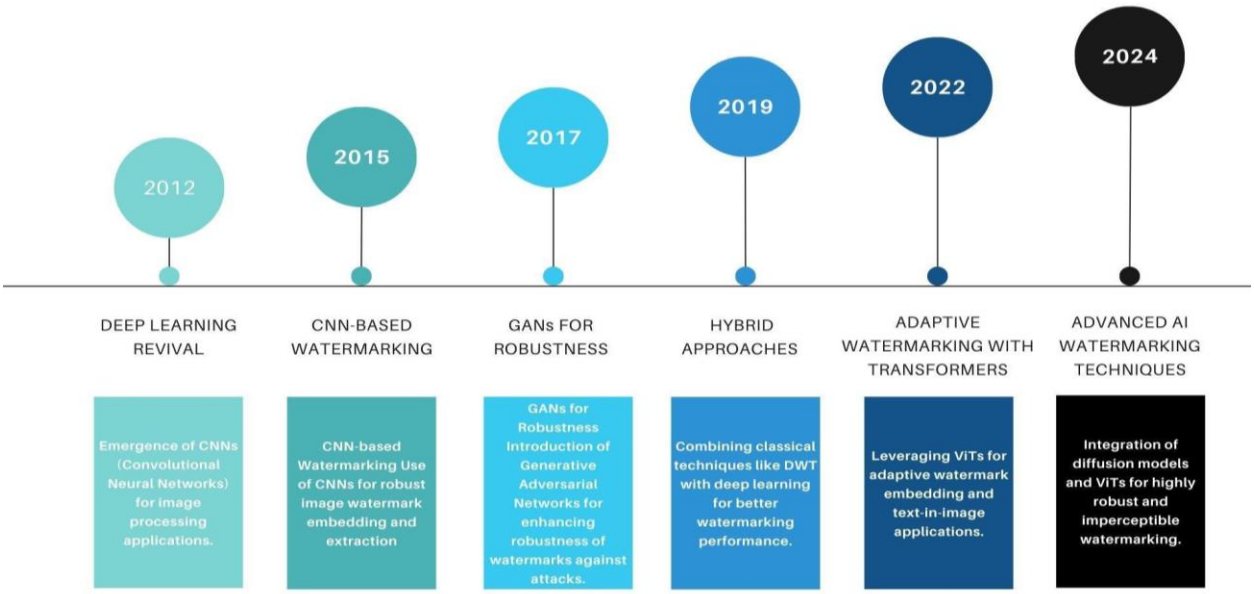

Figure 3: *Deep Learning Evolution in Watermarking*

2.2.1 Automated Deep Learning Frameworks

Deep learning has transformed the art of digital watermarking by presenting powerful and imperceptible techniques using advanced neural network architectures.

Zhong et al. [3] and Zhang et al. [10] presented a convolutional neural network (CNN) for automatic watermark embedding and extraction. These methods significantly enhance the recovery from distortions like salt and pepper noise and Gaussian noise. Zhang et al. [10] also view watermarking as an embedding task of an image. This makes the embedding procedure easier, with high imperceptibility and low bit error rate (BER) even under adverse conditions, such as 65% cropping or strong compression ratios.

GANs add a revolutionary layer to the deep learning framework. Deng et al. [5] proposed a GAN-based watermarking scheme that achieved excellent robustness against hybrid attacks like JPEG compression and salt and pepper noise, with NC values greater than 0.98. The effectiveness of the method in sensitive applications, such as medical imaging where PSNR reaches 56.82 dB and SSIM reaches 1,000, points out its significance in getting a robust imperceptible watermark. Zhao et al. [20] further extended this idea to incorporate an attention mechanism toward achieving semantic-aware watermarking through spatial and channel attention weights.

Autoencoder architecture is another interesting innovation of deep learning for this field. Singh et al. [11] applied CNN-based autoencoder that was able to sustain a PSNR value up to 31.34 dB and an NC up to 0.9937 in various distortions such as salt and pepper noise and JPEG compression. Wu et al. [6] achieved this by introducing an iterative training strategy, training separate encoders and decoders to achieve 94.82% bit prediction accuracy (BPA) under JPEG compression (QF = 50). These contributions highlight the diversity and power of deep learning frameworks for digital watermarking. A Systematic diagram for watermarking using deep learning is shown in Figure 4.

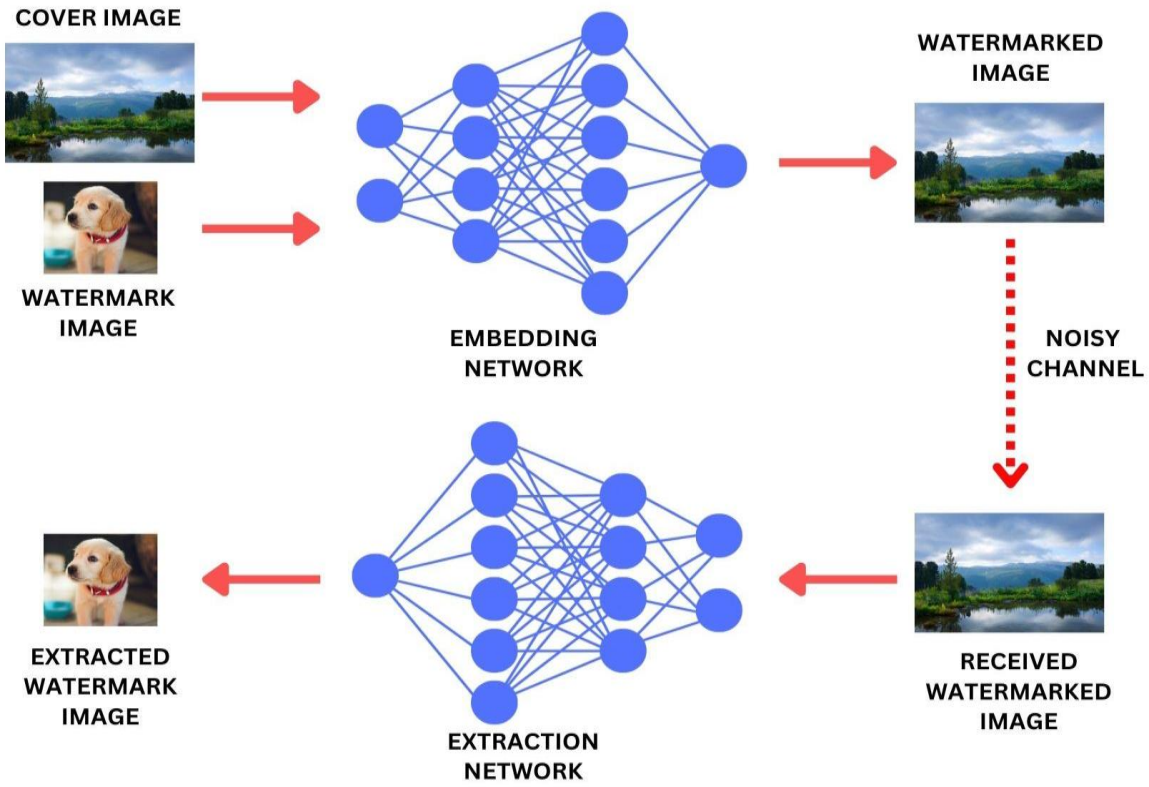

Figure 4: A Systematic diagram for watermarking using deep learning

2.2.2 Hybrid Frameworks: Bridging Traditional and Deep Learning Techniques

Hybrid frameworks combine traditional signal processing approaches with modern deep learning models to balance invisibility with computational power and efficiency. Wei et al. [2] combined DWT with Periodic Variational Autoencoder (Cycle-VAE) to form a hybrid framework. The two-cycle consistency mechanism preserves the watermark image fidelity with an imperceptible PSNR value reaching 37.91 dB and an NC value almost close to 1.0 against geometric distortion and noise. Taking this further, Wei et al. [15] proposed a double-ring mechanism that greatly improved robustness by keeping the SSIM value over 0.94 even against severe noise and distortion. Hybrid techniques, such as Tavakoli et al. work in [12], are developed incorporating chaotic labeling and transformation techniques to efficiently counter adversarial attacks. Their approach produces NC values above 0.9 for a wide range of evolutions that provide a robust solution against the challenges of modern security. These techniques exemplify the promise of hybrid techniques to tap into the potential of traditional and deep learning techniques.

Huang et al. [13] incorporate dynamic weight averaging (DWA) in the multi-task learning framework to further improve the trade-off between the performance of the main task and watermarking objectives. The system learns faster when used with GoogleNet, while achieving 100% accuracy in watermark extraction at round 14 versus round 24 without DWA. For the same reason, DWA enhances the classification accuracy by increasing the value from 98.92% up to 99.20% using such data sources as MNIST, for instance.

Attention-directed architectures are the other leading class. Huang et al.'s [7] advanced attention mechanism ARWGAN, for instance, seeks to enhance embedding accuracy and distortion restoration capabilities. Their system has established a new benchmark in PSNR at 43.72 dB and a bit accuracy of more than 96% in watermarking systems in the areas of healthcare and secure transmission of content requiring high concealment and robustness.

Wang et al. [16] propose a novel approach to enhancing multi-task learning for image watermarking by using Dynamic Weight Averaging (DWA). DWA dynamically adjusts the task weights according to the learning rates of individual tasks, which ensures balanced optimization across diverse objectives such as embedding and extraction. The approach handles challenges in the traditional multi-task learning setup wherein fixed or static weight configurations result in suboptimal performance. Utilizing DWA, improved robustness and effectiveness are achieved for embedding watermarks into the image without degrading the quality of extracting watermarks under various distortions. The scope of this research demonstrates the potential of DWA in advancing image watermarking using adaptive task priority alignment.

Chen et al. [17] further pushed the watermarking technique for medical imaging, using WMNet, CNN-based system with over 97% classification accuracy under distortion, which ensures integrity of data during transmission and maintains diagnostic quality that indicates the possibility of specific industry solutions in watermarking.

Lee et al. [18] present a digital image watermarking processor using deep learning algorithms to achieve robust and efficient watermark embedding and extraction. The proposed system integrates CNNs for improving the resilience of the watermark against distortions and attacks while preserving the quality of the image. The processor is optimized for real-time performance, making it appropriate for practical applications. Conclusion Conclusion: This study reveals how deep learning can push the technology of digital image watermarking further by balancing the levels of robustness, efficiency, and fidelity involved.

***Table 1**: Summary of various watermarking methods with performance.*

Paper	Method	Technique	Imperceptibility (PSNR)	Robustness Metric	BER	NC	Watermark Capacity	Attack Types	Embedding Time	Extraction Time	Hardware Used
Zhong et al. [3]	Neural Network	CNN with invariance	39.72 dB	JPEG, Noise Resilience	< 14%	> 0.99	1,024 bits	JPEG compression,	Fast	Moderate	NVIDIA GTX 1080Ti

								cropping, noise			
Wei et al. [2]	Hybrid Framework	DWT + Cycle-VAE	40–45 dB	Noise, Cropping	< 10%	> 0.95	Medium	Compression, noise	Moderate	Fast	NVIDIA RTX 2080Ti
Singh et al. [11]	Autoencoder-CNN	CNN Autoencoder	31.34 dB	Salt & Pepper, JPEG	< 10%	> 0.91	Low	JPEG, rotation, noise	9.6 sec	29.4 sec	NVIDIA Titan X
Huang et al. [13]	Multi-Task with DWA	Dynamic Weight Averaging	40.94 dB	Gaussian Noise	Zero	0.9997	Multi-bit	Gaussian noise	Moderate	Fast	NVIDIA RTX 3090
Mahapatra et al. [22]	Vision Transformers	Cross-Attention	40.94 dB	Noise, Cropping	< 10%	0.99	Low	Noise, cropping	High	Moderate	NVIDIA GTX 1080Ti
Deng et al. [5]	GAN-Based Framework	Adversarial GAN	> 45 dB	Compression, Noise	< 20%	> 0.95	Low	Compression, noise	5 sec	2 sec	NVIDIA RTX 3090
Wu et al. [6]	Iterative Training	Encoder-Decoder Phases	34.55 dB	JPEG, Cropout	< 10%	0.94	Medium	JPEG compression, cropout	High	Moderate	NVIDIA RTX 3090
Deng et al. [14]	Weights-less Watermarking	Noise-Aware Framework	36.5 dB	Gaussian Noise	< 20%	0.97	High	Noise	Very Fast	Fast	NVIDIA GTX 1080Ti
Wei et al. [15]	Dual-Cycle Framework	Cycle-VAE	35–37 dB	Noise, Geometric	< 7%	> 0.92	Low	Noise, rotation	40 FPS	< 10 ms	NVIDIA GTX 1080Ti
Deng et al. [5]	GAN-Based Medical Watermarking	GAN with Chaotic Maps	56.82 dB	Hybrid Noise Resilience	-	> 0.98	Medium	Salt-and-pepper, compression	< 0.2 sec	< 0.2 sec	NVIDIA RTX 3090
Zhao et al. [20]	Attention-Guided Embedding	Spatial & Channel Attention	38.5 dB	BER < 1.5%	< 1.5%	> 0.97	Medium	JPEG compression, cropping	-	-	NVIDIA RTX 3090
Chen et al. [17]	Lossless Watermarking	WMNet with CNN	7.49 dB (noise conditions)	High Classification Accuracy	-	-	Medium	Gaussian noise, filtering, rotation	-	-	Not Specified

Tavakoli et al. [12]	Transform + Chaos	Chaotic Labeling	-	NC > 0.9	-	0.9	Medium	Adversarial attacks	-	-	Not Specified
Huang et al. [7]	Attention-Guided GAN	ARWGAN	43.72 dB	Bit Accuracy > 96%	-	-	Medium	Gaussian blur, JPEG compression	-	-	Not Specified

3. Performance Evaluation Parameters

The performance of watermarking techniques is estimated by several key parameters and metrics that decide the trade-offs between imperceptibility, power, and efficiency. Perceptual apraxia is usually evaluated by the peak signal-to-noise ratio, or PSNR, which can be calculated as:

$$PSNR = 10.log_{10}$$
$$\left(\frac{MAX^2}{MSE}\right)$$

Where MAX is the maximum possible pixel value of the image-for example, 255 for an 8-bit image-and MSE is the mean square error between the original image and the watermarked image. PSNR values above 30 dB commonly refer to an almost fully imperceptible watermark when maintaining the image quality at high levels.

It determines the strength of a watermark that could survive different kinds of attacks, such as noise, compression, clipping or geometric distortion. The measures used to quantify robustness are BER and NC. BER is expressed as:

$$BER = \frac{NumberOfIncorrectBits}{TotalBitembedded} \times 100$$

Where lower BER indicates higher robustness. NC is calculated as:

$$NC = \frac{\sum(W_o.W_e)}{\sqrt{\sum(W_o^2).\sum(W_e^2)}}$$

Where WoW_oWo is the original watermark and WeW_eWe is the extracted watermark. NC values close to 1 indicate strong resistance to attacks.

The capacity of a watermark refers to the amount of data that can be embedded without significantly degrading the imperceptibility. It depends on the size of the host image and the method of watermarking. Techniques that yield high power while maintaining PSNR and robustness are preferred for practical applications.

Embedding and extraction time is very critical in real-time scenarios. Usually, it is measured in seconds or milliseconds depending on the hardware, and it is very much important in applications like streaming or dynamic media.

Robustness of the watermarking method against different types of attacks, including noise (e.g., Gaussian, Salt and Pepper), compression (e.g., JPEG), and geometric transformations (e.g., cropping, rotation) is tested. In such conditions, low BER and high NC prove that this approach is resilient.

Finally, hardware performance takes an important role in testing computational efficiency. More specifically, high-performance GPU is often used to increase the speed of the embedding or extraction process. For instance, NVIDIA RTX 3090 or TITAN are popular choices for providing a better result in the context of speed and scalability. In total, this framework offers a comprehensive measurement of the analysis and comparison across various watermarking techniques.

3.1 Comparative Performance under various attacks

Manipulated noise addition often is a generally perceived attack emanating from the impositions of random fluctuations within pixel values in an image; such variations may be in the form of Gaussian or Salt-and-Pepper noise and others. The primary aim of this attack is to degrade the visual quality of the image and damage the embedded watermark. A robust watermarking system provides immunity against such distortions to maintain data integrity.

One mode of attack that relates to the process of cutting off a portion of an image is cropping. With self-explanatory clarity, the watermark is expected to disappear because it is contained within the area that will be cropped during cropping. This makes this type of attack one of most challenging for watermarking methods based on spatial information. The best watermarks are capable of detection even after a major portion of an image is cropped.

JPEG is a lossy compression method specifically intended for compression of images that discards less perceptually significant data allowing a decreased file size of an image. This could, however, affect the watermark embedded into the original image quality. Therefore, an algorithm of watermark or pattern recognition needs to have a level of strength to manage the data reduction caused by the compression into JPEG such that an evidence of detectability would not degrade significantly.

When we incorporate Gaussian noise, we add that random variation of pixel intensity, which is distributed normally. The sort of thing normally encountered in practice has to do with image transmission or storage. In real world broad applications, watermarking needs to fight back against Gaussian noise so that it survives and then gets decoded in the noise.

Salt-and-Pepper noise is characterized by the presence of randomly distributed white (salt) and black (pepper) pixels in an image. This type of noise simulates harsh distortions that occur at the pixel level and is particularly challenging for spatial domain watermarking techniques. A robust watermark should remain detectable despite such noise.

Rotation changes the orientation of an image by rotating it through a given angle. Such a geometric transformation can cause misalignment of the embedded watermark, making it difficult to detect or recover. Advanced watermarking methods often include mechanisms to counteract rotation and ensure the watermark's resilience under such attacks.

Table 2: *Robustness metrics under various attacks*

Method	Robustness Metric	Attack Type	Performance
Zhong et al. [3]	BER < 14%	Noise, cropping	High resilience
Wei et al. [2]	BER < 10%	Compression, noise	Superior noise resistance
Huang et al. [13]	BPA > 95%	Gaussian noise	Robust to noise
Deng et al. [5]	>75% bit accuracy	JPEG compression, noise	High robustness
Mahapatra et al. [20]	NC > 0.9	Cropping, salt-and-pepper	Robust extraction
Singh et al. [11]	NC = 0.9937	Noise, rotation	Reliable watermark recovery

Conclusion and Future Directions

The rise of digital technology has given rise to numerous advancements in the area of deep learning-based watermarking for intellectual property and data security, particularly in healthcare and consumer electronic sector. Methods like Black-Box Watermarking and techniques like CNN-based approaches - such as adopted in WMNet - provide robust near-undetectable solutions to secure the knowledge and ensure ownership verification against such attacks as compression or noise. Combining watermarks with cryptographic methods further tightens security, which is essential because healthcare imaging, like other digital assets, must be protected. The progression is exponential deep within traits becoming more important due to both security and efficiency with the passage of time in the digital era.

References

1. Chen, Y., et al. Deep Learning Techniques in Image Watermarking: A Survey, Neurocomputing, 2020. DOI: 10.1016/j.neucom.2020.08.022.
2. Wei, Q., et al. A robust image watermarking approach using cycle variational auto encoder, Security and Communication Networks, 2020. DOI: 10.1155/2020/8869096.
3. Zhong, X., et al. An Automated and Robust Image Watermarking Scheme Based on Deep Neural Networks, IEEE Transactions on Multimedia, 2020. DOI: 10.1109/TMM.2020.3006415.
4. Wu, Y., et al. Noise Tolerance in Deep Learning Watermarking Models, CoST, 2023. DOI: 10.1109/CoST60524.2023.00038.
5. Deng, H., et al. GAN-Based Invisible Watermarking for Intellectual Property Protection, ISCIT, 2023. DOI: 10.1109/ISCIT57393.2023.10376108.
6. Wu, X., et al. Iterative Training for Non-Differentiable Noise in Deep Learning-Based Image Watermarking, CoST, 2023. DOI: 10.1109/CoST60524.2023.00038.
7. Huang, J., et al. ARWGAN: Attention-Guided Robust Watermarking Using GANs, IEEE Transactions on Instrumentation and Measurement, 2023. DOI: 10.1109/TIM.2023.3285981.
8. Li, W., et al. Blind Watermarking System Based on Deep Learning for Robust Image Embedding, IEEE TrustCom, 2021. DOI: 10.1109/TrustCom53373.2021.00127.
9. Zhu, Y., et al. DeepSigns: A Framework for Digital Watermarking of Deep Neural Networks, ACM ICMR, 2022. DOI: 10.1145/3512527.3531380.
10. Zhang, L., et al. A Fully Automated Deep Learning-Based Image Watermarking System, IEEE Transactions on Multimedia, 2021. DOI: 10.1109/TMM.2021.3008537.
11. Singh, H., et al. Deep Learning-Based Watermarking for Digital Images, Multimedia Tools and Applications, 2024. DOI: 10.1007/s11042-023-15750-x.
12. Tavakoli, A., et al. Convolutional Neural Network-Based Image Watermarking with Discrete Wavelet Transform, IEEE Transactions on Image Processing, 2022. DOI: 10.1109/TIP.2022.3195039.
13. Huang, Y., et al. Dynamic Weight Averaging for Multi-task Learning in Image Watermarking, International Conference on AI Applications, 2023. DOI: 10.1109/AIAC61660.2023.00020.
14. Deng, H., et al. Weights-less Watermarking for Neural Networks, ISCIT, 2023. DOI: 10.1109/ISCIT57393.2023.10376108.

15. Wei, Q., et al. A Dual Cycle-VAE Framework for Robust Watermarking, Security and Communication Networks, 2020. DOI: 10.1155/2020/8869096.

16. Wang, Z., et al. Dynamic Weight Averaging in Multi-Task Learning for Image Watermarking, Journal of Multimedia Tools and Applications, 2023. DOI: 10.1007/s11042-023-15781-4.

17. Chen, Y.-P., et al. WMNet: A Lossless Watermarking Technique for Medical Image Authentication Using CNNs, Electronics, 2021. DOI: 10.3390/electronics10080932.

18. Lee, J.-E., et al. Digital Image Watermarking Processor Based on Deep Learning Algorithms, Electronics, 2021. DOI: 10.3390/electronics1011183.

19. Singh, H.K., et al. GAN-Based Watermarking for Encrypted Medical Images in Healthcare Scenarios, Neurocomputing, 2023. DOI: 10.1016/j.neucom.2023.126853.

20. Zhao, Y., et al. DARI-Mark: Deep Learning and Attention Network for Robust Image Watermarking, Mathematics, 2023. DOI: 10.3390/math11010209.

21. Zhao, F., et al. Chaotic Systems for Enhanced Neural Network Watermarking, IEEE Transactions on Information Forensics and Security, 2023. DOI: 10.1109/TIFS.2023.3281227.

22. Mahapatra, D., et al. Autoencoder-Based Embedding and Extraction Models in *Watermarking*, Journal of Electronic Imaging, 2023. DOI: 10.1117/1.JEI.32.2.021604.