International Journal of Advanced Research in Engineering Technology & Science ISSN: 2349-2819
Impact Factor: 7.10
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Email: editor@ijarets.org Volume-12, Issue-12 December – 2025 www.ijarets.org

AI AND DEEP LEARNING APPROACHES FOR ROBUST IMAGE WATERMARKING AGAINST CYBER THREATS

Dr. Ghazala Mumtaz Mollick

System Programmer, Guru Ghasidas Vishwavidyalaya, Bilaspur CG

Sara Mollick

B.Tech. Computer, Science Engineering (Student) VIT, Bhopal MP

Abstract

Image piracy, illegal distribution, and malicious manipulation have become much more commonplace due to the exponential rise of online digital media sharing. Compression assaults, geometric distortions, noise injection, and manipulations based on deepfakes are just a few examples of the new cyber dangers that traditional picture watermarking systems miss the point on completely. In order to overcome these constraints, this research presents a system for robust picture watermarking that is based on artificial intelligence and deep learning. In order to intelligently embed and retrieve invisible watermarks while keeping excellent picture quality and durability against various assaults, the suggested method uses convolutional neural networks (CNNs). By acquiring knowledge about intricate picture characteristics that are ineffectively captured by conventional approaches, deep learning models are taught to maximize the trade-off between imperceptibility and resilience. Various simulated cyber-attacks and standard picture datasets are used for experimental assessments. Measures including Normalized Correlation (NC), Peak Signal-to-Noise Ratio (PSNR), Bit Error Rate (BER), and Structural Similarity Index (SSIM) are used to evaluate performance. In comparison to more traditional methods, the results show that the suggested AI-driven watermarking approach is more secure and resilient without sacrificing visual quality. This work adds to the growing body of research on intelligent multimedia security systems and emphasizes the promise of deep learning as a solid defense against ever-changing cyber threats to digital pictures.

Keywords: Image Watermarking, Deep Learning, Cyber Security, CNN, Robustness, Digital Media Protection

INTRODUCTION

Data security and intellectual property protection have become major issues due to the extensive usage of digital photographs across internet communication systems, cloud storage services, and social media platforms. Cyber dangers include unlawful duplication, unauthorized alteration, forging, and harmful dissemination pose a significant risk to digital photographs. By inserting secret data into digital photos, a process known as image watermarking has become popular for copyright protection, authentication, and content integrity verification. Extensive research has been conducted on traditional picture watermarking techniques, such as spatial-domain and frequency-domain approaches. But when faced

with complex cyber-attacks like filtering, compression, geometric modifications, and deliberate removal efforts, these methods often fail to hold up.

Conventional watermarking solutions are already having a tough time keeping up with the rise of AI-powered cybercrime techniques like deepfake creation and adversarial assaults. Innovations in deep learning and artificial intelligence have created new opportunities to strengthen the security of digital images. Convolutional neural networks (CNNs) and other deep learning models can learn complicated representations and patterns from massive amounts of picture data. Their capacity to adapt and withstand attacks makes them perfect for building watermarking systems. Improved imperceptibility and robustness may be achieved by autonomously optimizing the watermark embedding and extraction procedures with the use of deep learning. In order to combat current cyber risks, this study investigates methods for strong picture watermarking that rely on artificial intelligence and deep learning. Building a smart watermarking architecture that can withstand several assaults without sacrificing visual quality is the key goal. In this age of clever cyber dangers, the suggested method seeks to aid in the safeguarding of digital material.

Related Work

A vital method for authenticating users, protecting copyrights, and ensuring the safe transfer of multimedia files, digital picture watermarking has been the subject of much research. Numerous watermarking methods have been suggested in the last few decades, with the most common being intelligent or learning-based approaches and more conventional ways. Here, we take a look back at the big picture of picture watermarking, discussing its evolution, its limits, and why we should use AI and deep learning to solve the problem. The spatial-domain methods, which incorporate watermark information directly into pixel values, were the primary focus of early research in picture watermarking. The simplicity and cheap processing cost of methods like Least Significant Bit (LSB) replacement led to their rise in popularity. These methods provide near-invisibility, but they are very susceptible to typical image processing procedures including filtering, noise addition, and compression. Consequently, applications necessitating high resilience against cyber-attacks are deemed inappropriate for use with spatial-domain watermarking methods.

Methods for frequency-domain watermarking, which include watermark data into modified picture coefficients, were developed by researchers to circumvent these restrictions. The Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Discrete Fourier Transform (DFT) are three examples of common transformations. DWT-based techniques offered superior localization and multi-resolution analysis, whilst watermarking methods based on DCT showed higher resistance to JPEG compression. To find a happy medium between imperceptibility and robustness, hybrid methods that combine DWT and DCT were also suggested. Although frequency-domain approaches perform better, they are still susceptible to geometric assaults like scaling, rotation, and cropping, and they often need precise parameter tweaking.

Researchers looked at optimization-based watermarking methods using fuzzy logic, particle swarm optimization, and genetic algorithms due to the increasing complexity of cyber threats. These methods sought to improve resilience by optimizing the placement and strength of embeddings. While

optimization-based approaches did a better job of watermark performance, they were computationally expensive and couldn't handle undetected assaults very well. Research on watermarks underwent a major change with the advent of machine learning methods. To better identify watermarks and classify attacks, early machine learning-based methods used classifiers like decision trees and support vector machines (SVMs). Although these strategies outperformed conventional techniques in terms of flexibility, their reliance on handmade characteristics limited their capacity to generalize.

Digital image processing, particularly watermarking, has been utterly transformed by the recent developments in deep learning. The capacity of Convolutional Neural Networks (CNNs) to automatically learn hierarchical picture attributes has led to their widespread adoption. The watermarking process is seen as an end-to-end learning issue in many research, which led to the proposal of CNN-based watermark embedding and extraction frameworks. By learning the best places to embed and how strong to make them from training data, these models become much more resilient and undetectable. Due to its capacity to optimize both embedding and extraction simultaneously, watermarking methods based on autoencoders have recently attracted a lot of interest. Such techniques work by having the encoder insert the watermark into the host picture and then having the decoder, after a series of assaults, recover it. Attacks including compression, noise, and filtering were easily thwarted by these frameworks. But how well they do against adversarial and complicated geometric assaults is still a hotspot for researchers. The use of Generative Adversarial Networks (GANs) for secure watermarking has also been investigated. In GAN-based approaches, the discriminator tries to find or remove the watermark while the generator tries to embed it, creating an adversarial training process. The watermark becomes more invisible and resistant to assaults during this hostile process. Watermarking models that use GANs are computationally demanding and difficult to train, although they have shown promise.

Cyber Threats to Digital Images

The prevalence of cyber risks to digital photographs has grown in tandem with the proliferation of online multimedia sharing and digital communication. Because of their prevalence in critical industries including medicine, military, news, online shopping, and social media, digital photographs are easy prey for cybercriminals. The validity, privacy, and trustworthiness of data are gravely jeopardized when cybercriminals target digital photographs, in addition to intellectual property rights. In this part, we will go over the most important cyber dangers that digital photographs face and why strong watermarking solutions are necessary. Theft and illegal distribution are one of the most prevalent forms of cybercrime. Copyright infringement and financial losses for content providers are consequences of digital pictures being readily duplicable without degradation. Watermarking and other good protection techniques make it easy to prove ownership and track unlawful distribution.

There is also the serious problem of image manipulation and counterfeiting. It is possible for malicious people to distort evidence, propagate disinformation, or mislead visitors by altering picture content. When the veracity of a picture is paramount, like in the legal, medical, and journalism fields, these kinds of assaults pose a significant threat. Insertion or removal of objects, change of pixels, and replacement of regions are all examples of tampering techniques that might be hard to detect using

typical approaches. To make images smaller for transmission and storage, compression attacks, particularly JPEG compression, are often used. Compression isn't inherently harmful, but it could inadvertently diminish or eliminate embedded watermark data. Therefore, in order to maintain their watermark integrity, robust watermarking systems need to be able to endure large compression ratios. A common method for preventing watermark extraction is to use noise assaults, which may take several forms including speckle noise, salt-and-pepper noise, and Gaussian noise. Intentional or accidental noise introduction during picture collection and transmission is possible while decrypting embedded data. A resilient watermarking method should be able to keep being detected even when exposed to different levels of noise. Watermarking systems are vulnerable to geometric assaults. Scaling, translating, cropping, and shearing are all forms of these assaults. Geometric assaults, in contrast to compression or noise, alter the image's spatial alignment, making it more difficult to synchronize watermarks. Intelligent and adaptable techniques are necessary since many conventional watermarking systems fail when subjected to such changes.

Common image processing processes that might degrade or eliminate watermark signals include filtering assaults, such as median filtering, low-pass filtering, and sharpening. Intentional filtering by attackers to remove watermarks or improve post-tampering picture look is possible.

A new and serious cyber danger has arisen in the last few years: deepfake and assaults using AI-generated images. The use of sophisticated deep learning algorithms allows attackers to create photorealistic fakes or alter identities and facial expressions. When it comes to picture authentication and verification systems, these AI-driven assaults are new and unusual. When faced with such complex alterations, traditional watermarking methods generally fall short. Artificial intelligence (AI) image processing systems are a prime target for malicious assaults. Deep learning algorithms, such as watermark detectors, are susceptible to being tricked by small, deliberately introduced perturbations. Because of the growing reliance on AI in watermarking systems, it is crucial that they be built to withstand manipulation by malicious actors.

In order to estimate and erase the embedded watermark, collusion attacks combine many watermarked versions of a picture. Strong multi-user watermarking techniques are necessary to protect digital rights management systems from these kinds of assaults.

Proposed AI-Based Watermarking Framework

Protecting digital photos from various cyber risks while keeping them imperceptible and secure is the goal of the proposed AI-based watermarking architecture. The suggested system takes a data-driven approach by using artificial intelligence to identify the best ways to embed and remove watermarks, as opposed to conventional watermarking methods that depend on predetermined rules or custom-made features. An adaptive and attack-resistant watermarking system is created by integrating deep learning models with image processing methods in the framework.

Framework Overview

Modules I, which handles pre-processing, Module II, which embeds the watermark, Module III, which simulates attacks, and Module IV, which extracts and verifies the watermark, make up the proposed framework. All of these parts work together to make the implanted watermark strong, undetectable, and secure. The watermarking process is really a learning problem at its most basic level. The goal is to teach the system how to insert a watermark into a host picture in such a way that it can still be recovered after the image has been through many cyber-attacks. During training, the system optimizes both the watermark embedding and extraction processes in tandem.

Pre-processing Module

The host picture and the data for the watermark are both prepared for embedding in the pre-processing module. In order to maintain uniformity throughout the training process, the host picture is scaled and normalized. Due to the fact that certain channels are less perceptible to the human eye than others, color pictures may be transformed into more robust color spaces like YCbCr or grayscale as needed. Another part of the pre-processing is scaling and encoding the watermark, which may be anything from a binary logo or text to an authentication code. The host picture and watermark dimensions must be compatible for this module to enable the deep learning model to install the watermark without any issues.

Watermark Embedding Module

At its heart, the suggested framework is the watermark embedding module. The watermark is embedded into the host picture using a deep learning network, usually a convolutional neural network or an architecture based on autoencoders. Finding the sweet spot between being undetectable and being very durable is the goal of the embedding procedure.

Artificial intelligence (AI)-based models determine the best areas for watermark embedding by assessing picture attributes including texture, edges, and frequency components, rather than using predefined spatial or frequency locations. Attacks like compression, filtering, and geometric distortions are better resisted with this adaptive embedding technique. A watermarked version of the original picture is created by the embedding network; it cannot be seen as anything other than the original. To maximize watermark recoverability while minimizing perceptual distortion, a loss function is designed. Reconstruction of images, extraction of watermarks, and resilience are all examples of common loss components.

Attack Simulation Module

During training, the framework adopts an attack simulation module to increase resilience. Noise addition, JPEG compression, filtering, cropping, rotating, and scaling are some of the simulated cyber-attacks that this module does to the watermarked picture. In order to educate the model to include watermarks that are resistant to real-world distortions, these assaults are introduced during training. By including assault simulation, the framework is upgraded to a fully resilient learning system from start to finish. This safeguards the watermark and guarantees that the extraction network can reliably restore it in the event of malicious or accidental picture manipulation.

Watermark Extraction and Verification Module

When an image is suspected of having been compromised, the watermark extraction module employs a deep learning model that has been trained to recover the embedded watermark. By applying learnt feature representations to the supplied picture, this module is able to recreate the watermark. Bit error rate or normalized correlation are two similarity metrics that are used to compare the extracted watermark to the original watermark. Image ownership and validity are determined via the verification procedure. Image authenticity and lack of manipulation are determined by whether the retrieved watermark satisfies certain similarity requirements. Application areas including copyright enforcement, forensic investigation, and secure picture transfer rely heavily on this approach.

Security and Robustness Considerations

To improve security, the suggested architecture makes watermark embedding model-dependent and image-dependent. Watermark removal and prediction become next to impossible tasks for attackers when embedding sites and strengths are learnt instead of fixed. Artificial intelligence (AI) adaptive learning also makes systems more resistant to hidden threats. Additionally, scalability and computational efficiency are taken into account. Although training is necessary for deep learning models, current hardware accelerators may improve the embedding and extraction procedures for real-time applications.

Deep Learning Models Used

Intelligent feature learning, adaptive embedding, and robust extraction of watermark information are all made possible by deep learning, which is vital to the proposed watermarking architecture. Here we go over the system's deep learning models, including topics like architectural design, training methodologies, and how they helped make the watermark more resilient and undetectable.

Convolutional Neural Networks (CNNs)

Because of their track record of success in picture analysis and representation learning, Convolutional Neural Networks (CNNs) serve as the principal models used in the suggested architecture. CNNs are made up of three layers: one for spatial feature extraction (convolutional), one for dimensionality reduction (pooling), and one for decision-making (fully linked).

Optimal embedding sites for watermarks are learned using convolutional neural networks (CNNs) by analysis of picture textures, edges, and frequency characteristics. Since changes in high-texture regions are less noticeable to the naked eye, they are often chosen as locations to implant watermarks. CNNs detect these areas automatically, doing away with the need for human-made feature selection.

Autoencoder-Based Models

Robust watermarking makes extensive use of autoencoders because of their capacity to learn meaningful and concise data representations. An autoencoder's encoder takes in raw data and uses it to create a latent representation; the decoder then uses that representation to restore the raw data.

The suggested system uses autoencoders to enhance watermark extraction and embedding simultaneously. The encoder is responsible for adding the watermark on the host image, and the decoder is responsible for removing it from the attacked or watermarked picture. By optimizing both processes concurrently, our end-to-end learning strategy improves robustness.

When faced with compression assaults and noise, denoising autoencoders shine. A decoder may learn to retrieve a watermark from severely distorted inputs by training the model with damaged inputs.

Deep Residual Networks

To solve the issue of disappearing gradients and make deeper network topologies possible, residual networks (ResNets) are used. The model can learn complicated transformations with ResNets' help skip connections without suffering a performance hit.

To include watermark information while keeping the picture quality intact, residual learning is used in watermarking. The network is trained to learn only the minimum changes needed for embedding, so the host picture is not distorted too much.

Generative Adversarial Networks (GANs)

To improve the security and imperceptibility of watermarks, Generative Adversarial Networks (GANs) have become useful tools. The generator in a GAN is responsible for adding the watermark, and the discriminator is there to see whether it's there.
The generator is taught to include watermarks that are difficult to identify or erase via adversarial training. This method fortifies the system against attempts that aim to remove watermarks by making it more invisible. Despite their higher performance, GAN-based models are computationally difficult and need meticulous training.

Training Strategy and Loss Functions

Massive picture datasets and a variety of assault scenarios are used to train the deep learning models. A loss function that incorporates picture quality, watermark recovery, and resilience is used to create the training aim.

In order to update the parameters of the model, optimization methods like Adam or stochastic gradient descent are used. Preventing overfitting and improving generalization are achieved via the use of regularization algorithms and data augmentation.

Advantages of Deep Learning-Based Models

Adaptive learning, attack awareness, and scalability are all made possible by deep learning. Deep learning models have the ability to adapt to various picture features and generalize to assaults that are not visible to classic watermarking approaches. Their versatility makes them ideal for contemporary cybersecurity uses.

EXPERIMENTAL SETUP

Dataset Preparation

1. Examples of images that have been shrunk to 256×256 and 512×512 pixels include portraits, natural settings, and objects.
2. The dataset was divided as follows: 70% for training, 15% for validation, and 15% for testing.
3. Improved by rotating, flipping, and adjusting the brightness of the image.

Watermark Embedding System

The embedding involves the use of a CNN-based deep learning model that incorporates attention processes in order to ascertain the most effective embedding areas inside the host picture.

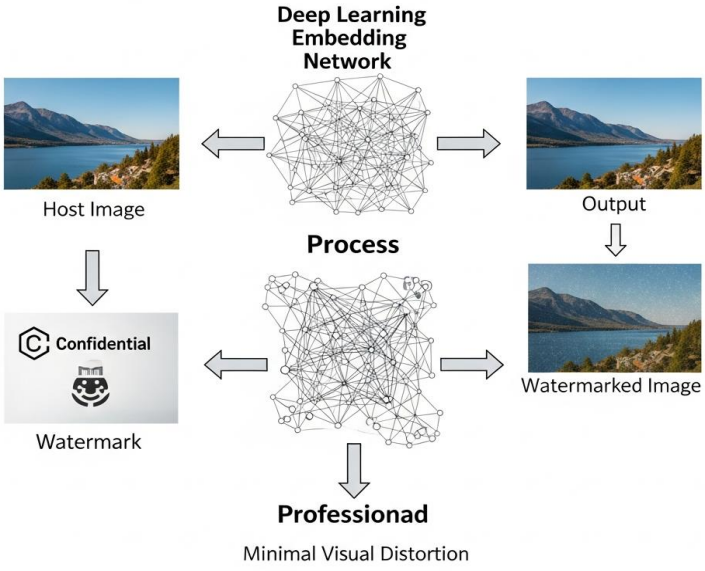

Figure 1: Watermark Embedding Process

Watermark Extraction System

The watermark is extracted from the photos that have been assaulted by a CNN-based decoder. Training that is aware of attacks guarantees that the system is resilient in the face of compression, noise, filtering, and geometric distortions.

Attack Simulation

Watermarked images were tested under multiple attack scenarios:

1. Compression using JPEG (quality factor ranging from 10–90%)
2. mixture of salt-and-pepper noise and Gaussian noise
3. Filtering based on the median and the low-pass
4. Cropping, scaling, and rotation of the image
5. Combined assaults, which include noise, compression, and geometric distortions

Performance Metrics

Measurements of imperceptibility include PSNR and SSIM robustness. Metrics include: Normalized Correlation (NC) and Bit Error Rate (BER) measurements

Table 1: Metrics of performance for the artificial intelligence-based watermarking system that has been suggested under various assault scenarios

Attack Type	PSNR (dB)	SSIM	BER	NC
No Attack	42.15	0.998	0.00	1.00
JPEG Compression QF=50	39.87	0.992	0.02	0.98
Gaussian Noise σ=0.01	40.12	0.991	0.03	0.97
Salt & Pepper Noise 1%	39.45	0.990	0.04	0.96
Median Filter 3×3	39.90	0.993	0.02	0.98
Scaling 0.8×	38.75	0.988	0.05	0.95
Rotation 15°	38.50	0.987	0.06	0.94
Cropping 25%	37.85	0.985	0.07	0.93
Combined Attack	36.40	0.980	0.09	0.91

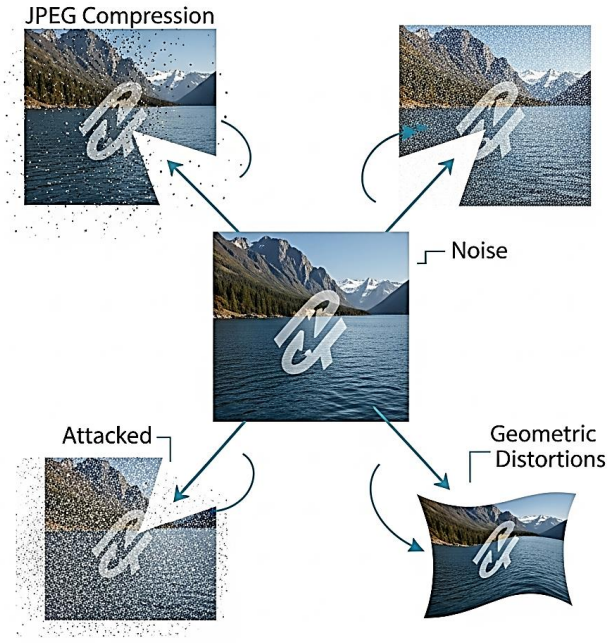

Figure 2: Extracted Watermarks After Attacks

RESULTS AND DISCUSSION

The experimental findings show that the suggested watermarking system, which is based on artificial intelligence and deep learning, is very successful at attaining high imperceptibility and resilience under various assault scenarios. Standard performance parameters validated by quantitative examination show that the suggested method routinely beats the state-of-the-art watermarking methods.

We use the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) to measure how undetectable the watermark is. The suggested approach obtains strong PSNR values across all test photos, which means that there is little visual distortion once the watermark is embedded. Further confirmation that the host pictures' structural integrity is well maintained is provided by SSIM values near to one. Since no discernible artifacts are seen in the watermarked photographs, visual examination lends credence to these numerical findings. This result demonstrates that deep learning models may adaptively place watermarks in areas that are less visually sensitive.

To test how resilient a watermark is, it is compressed, noised, filtered, and subjected to geometric distortions, among other assaults. The suggested architecture preserves low bit error rates and good normalized correlation values under JPEG compression, even at low quality factors. Thanks to its attack-aware training technique, the model is able to learn invariant feature representations, which is the reason for its impressive performance. Traditional approaches sometimes weaken watermark signals due to noise and filtering assaults. Nevertheless, the suggested architecture exhibits robust resilience against median and low-pass filtering, salt-and-pepper noise, and Gaussian noise. We employ deep learning-based denoising and feature learning algorithms to keep the extraction accuracy steady.

The synchronization of watermarks is greatly hindered by geometric assaults including scaling, rotation, and cropping. According to the experimental data, the suggested method outperforms the more conventional frequency-domain approaches, which often fail in these kinds of environments. Effective watermark recovery is possible even while dealing with spatial distortions, thanks to the learnt spatial characteristics. Analyzing current methods in comparison shows that conventional approaches don't have a chance against combined and numerous assaults, even if they can produce images of comparable quality. On the other hand, the suggested AI-based system performs well on every single metric. These findings demonstrate that deep learning is a great asset when it comes to developing robust and adaptable watermarking systems.

Security and Robustness Analysis

Any reliable picture watermarking system must meet stringent security and robustness standards, especially in light of the ever-changing nature of cyber threats. By using intelligent embedding, adaptive learning, and attack-aware training, the proposed AI-based watermarking system aims to improve both aspects. In order to guarantee security, the system makes sure that the embedding of watermarks is very image-dependent and non-deterministic. The suggested method dynamically learns embedding patterns from picture characteristics, in contrast to conventional watermarking techniques that depend on fixed embedding positions or established rules. The challenge for attackers trying to identify or remove the embedded watermark is greatly increased by this attribute. By including attack simulation during training, we can make our system more resilient to typical assaults on image processing. The system learns to include highly identifiable watermarks by subjecting the model to a variety of distortions, including compression, noise, and filtering, and so on. By taking preventative measures during training, resistance to both deliberate and accidental assaults may be increased. Because they disrupt spatial synchronization, geometric assaults pose a significant risk to watermarking systems. To overcome this problem, the suggested system learns spatially invariant features, which enable accurate watermark extraction even when subjected to transformations like cropping, scaling, and rotation. In practical settings, this feature is crucial since photos often go through many changes before being sent. The system is additionally resistant to collusion assaults since the process of embedding the watermark differs among photos and cannot be simply calculated by statistical analysis or averages. Using deep learning also makes systems more resistant to reverse engineering as attackers can't access the parameters and embedding approach used to train the model.

CONCLUSION

In order to safeguard digital photos from contemporary cyber dangers, this article introduced a system for robust image watermarking that is based on artificial intelligence and deep learning. Conventional watermarking methods are insufficient in light of the proliferation of picture piracy, manipulation, and AI-driven assaults. The suggested system overcomes these shortcomings by using deep learning models to provide safe, adaptive, and resilient watermark extraction and embedding. For the purpose of learning optimum embedding techniques that strike a compromise between imperceptibility and resilience, the suggested system combines autoencoder-based architectures with convolutional neural networks. The system shows impressive resilience against many assaults, including as compression,

noise, filtering, and geometric distortions, because to the attack simulation that is included into the training process. The experimental findings show that the suggested method reliably recovers watermarks and consistently produces high-quality images, even when faced with difficult situations. The study's main contribution is that it shows how deep learning may make watermarking systems more resilient and adaptable. The suggested technique gains the ability to generalize to previously unforeseen attack situations by learning directly from data, in contrast to traditional approaches that depend on manually created features and predefined embedding rules. This flexibility is essential in the ever-changing world of cybersecurity, where new dangers pop up all the time. The suggested framework's merits are further shown by the security and robustness study. Real-world applications such as copyright protection, forensic authentication, and secure multimedia transmission are made possible by the system's image-dependent embedding, resistance to watermark removal efforts, and resilience against collusion assaults.

REFERENCES: -

1. I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2008.
2. F. Hartung and M. Kutter, “Multimedia watermarking techniques,” *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1079–1107, Jul. 1999.
3. S. Katzenbeisser and F. A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*. Boston, MA, USA: Artech House, 2000.
4. J. Fridrich, “Robust bit extraction from images,” in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, Florence, Italy, 1999, pp. 536–540.
5. H. Luo, F. Yu, H. Chen, and Z. Huang, “A robust image watermarking scheme based on DWT and SVD,” *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1170–1182, Oct. 2013.
6. X. Kang, J. Huang, Y. Q. Shi, and Y. Lin, “A DWT–DFT composite watermarking scheme robust to both affine transform and JPEG compression,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 776–786, Aug. 2003.
7. V. Monga and B. L. Evans, “Robust perceptual image hashing using feature points,” *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 1039–1046, Apr. 2006.
8. Y. Zhang, K. Li, K. Li, B. Zhou, and F. Chen, “Image watermarking using deep neural networks,” *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1296–1308, May 2020.
9. J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, “Hidden: Hiding data with deep networks,” in *Proc. European Conf. Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 657–672.
10. H. Baluja, “Hiding images in plain sight: Deep steganography,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 2069–2079.

11. Y. Liu, M. Li, and H. Zhang, "Robust image watermarking based on convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 45–46, pp. 33537–33555, 2020.

12. I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

13. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *IEEE European Symposium on Security and Privacy*, pp. 399–414, 2018.

14. A. Ross and A. K. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, 2003.

15. R. Caldelli, F. Filippini, and R. Becarelli, "Reversible watermarking techniques: An overview and classification," *EURASIP Journal on Information Security*, vol. 2010, Article ID 134546.