International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Digital Watermarking: Innovations and Recent Developments

Sanjay Patsariya[1], Anand Jha [2] , Kirtiraj Bhatele[3], ,Saba Khan[4],Ashish Singh[5] ,

Deepak Gupta[6]

[1,2,4]Department of Information Technology, RJIT, Gwalior, MP, India
[3]Department of Computer Science & Engg., MITS, Gwalior,MP,India
[5]Department of Computer Science & Engg.,SRGOI Jhansi,UP, India
[6]Department of Computer Science & Engg.,ITM, Gwalior, MP, India

---

Abstract -The development in technology has facilitated the creation and collection of large volumes of data, mostly images. Nevertheless, there are still challenges in terms of protection, associated with the need to protect people from violating the rules of using someone else's data in accordance with privacy regulations. In the more recent past, the use of deep learning-based models has seen increased interest due to their capability of embedding and extracting adaptive watermarks while ensuring visually pleasing results with resilience to geometrical manipulations as well as adversarial attacks. The combination of classical and intelligent approaches for watermarking has also shown enhanced results in the context of security and robustness. While significant progress has been made, there is still much room left for improvement when balancing invisibility, robustness, and computational complexity. Thus, this review paper explores the use of a state of art watermarking techniques.

Key Words: Copyright Protection, Neural Network, SWT (Stationary Wavelet Transform), DWT (Discrete Wavelet Transform), IWT (Integer Wavelet Transform), Robustness, Imperceptibility, Security, Entropy.

1. INTRODUCTION

Neural Network Watermarking is an advanced technique used to protect digital media like images, sounds, and videos. It uses machine learning algorithms to add an invisible mark called a watermark to the original file without degrading its integrity. In contrast to classical techniques, neural networks use extensive training to find the optimal ratio between critical parameters such as imperceptibility, robustness, Security and data capacity. As a consequence, the watermark is not distorted despite being exposed to changes such as compression, noise, and size[1,2]. Therefore, neural network-based watermarking serves as a more versatile and efficient tool for copyright control and content verification, although it remains vulnerable to specific distortions and consumes considerable amounts of energy. The application of neural networks for watermarking implies a series of advantages. First, it automates the process of finding an optimal embedding and extraction method since no specific guidelines need to be designed. Second, neural networks are characterized by a high level of imperceptibility, which ensures that the watermark does not affect the quality of the original file. Third, the technique improves the robustness of the watermark.

2. GOAL OF WATERMARKING

The primary purpose of watermarking is to embed hidden data or metadata within digital media, including image, audio, video, or document, without significantly compromising the quality of the data [3]. The data embedded in such a manner is termed as "watermark" and has multiple uses, some of which are:

2.1Copyright Protection: The technique of watermarking is often used to ensure that digital media cannot be used, distributed, or reproduced without proper permission.

2.2Authentication: Watermarks provide a means to authenticate digital media as genuine and un-tampered with.

2.3Media Surveillance: Watermarks make possible monitoring and tracking of digital media.

2.4Tamper Detection: Any tampering or modifications of digital media can be easily detected through watermarks.

3. CATEGORIESOF WATERMARKING

Watermarking methods may be categorized based on various aspects such as visibility, robustness, and working domain. Some examples of common categorizations include

3.1Based of Visibility

Invisible Watermarks: Invisible watermarks are not noticeable by the naked eye since they are incorporated into the media content without changing the appearance. This is a common method used to provide more security during copyright infringement. Visible watermarks are considered to serve as an indication of ownership and have the capability to deter any unauthorized use [4,5].

Visible Watermarks: In general, visible watermarks are those that are added to the media content and hence visible by naked eye.

3.2 Based on Domain

The image watermarking approaches can be divided into two major categories based on the methodology through which the watermark is embedded into the image. The first category is called spatial domain technique, in which the watermark is directly applied to the pixel values of the image and is quite easy to apply and requires minimal computational resources. However, it lacks robustness against possible attacks due to its simplistic approach. Conversely, transform domain techniques utilize some transformation to embed the watermark into the frequency components of the image, like the Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), or the Fourier Transform, etc. This technique is relatively more robust and secure, since the watermark is spread out in the image in such a manner that it becomes almost impossible to remove it or even tamper with it without damaging the image quality.

4. LITERATURE REVIEW

The digital watermarking literature indicates that much work has been done in this area using many techniques and implementation strategies. Many papers are concerned with designing watermarking algorithms that can withstand different kinds of attacks, including filtering, compression, and noise injection.

Ding et al.[10] has been suggested a novel generalized DNN-based digital watermarking scheme, wherein the DNN model trains on massive datasets of images, thus facilitating batch processing along with excellent generalization. The approach allows for the successful embedding of watermarks in images without affecting their quality and extracting them successfully with subjective and objective performance measurement criteria used for evaluation. Attacks such as common image manipulations are applied to assess the robustness, whereas methods like adjusting the loss functions have been adopted to ensure complex watermark capacity. The results show that the suggested approach proves effective in terms of being efficient, practical, and cost-effective, allowing for good generalization to multiple datasets. It can be noted that the watermark is resistant enough to withstand numerous attacks while being embedded in images, although there may be some sensitivity to the JPEG compression attack. Therefore, the paper presents evidence of the effectiveness of a generalized DNN framework for digital watermarking. This study opens up opportunities for further research in this domain since the network architecture can be improved and made even more resistant to attacks, especially in electronic medical imaging applications.

Tayyab et al.[11] projected a DL based approach and shown several advantages provided by DL algorithms and in particular while handling security sensitive applications using data-driven methods. Along with having numerous innovative properties, the DL algorithms were also found to suffer from many security challenges, which have limited the number of applications. Examples of these attacks include poisoning and evasion. The objective of this research is to develop a secure framework for datasets used by DL algorithms for tackling data-driven problem sets. This framework is developed using light weight watermarking mechanism to ensure data authenticity against various types of attacks. They have evaluated the effectiveness of their proposed framework through accuracy, precision and computational cost matrix. It was seen that the accuracy achieved with their model was around 98.99%. Similarly, the precision achieved was 0.96 and at the same time they have observed reduction in computational cost.

Chuan-Yu Chang et al. [12] proposed an innovative Full Counter-Propagation Neural Network (FCNN) used for image watermarking, whereby watermarking is done using the synapses in the neural network without any alteration of the actual cover images. The technique supports the use of a single watermark by many cover images at once while retaining the cover image's visual quality after being watermarked. It offers a unique platform of combining watermarking process into one neural network, thereby minimizing efforts required in conventional image watermarking systems. Results indicate the effectiveness of this technique as a digital

watermarking solution owing to the high level of robustness, imperceptibility, and authenticity provided.

Leetal.[13] presented an innovative zero-bit watermarking scheme for neural networks that makes it possible to retrieve the watermark despite the fact that the neural network model is accessed remotely using the provided service API rather than locally. To enable remote watermark retrieval, they make a minor modification to the model such that a predefined set of queries contains the watermark payload. The proposed technique uses adversarial examples in order to watermark the neural network model while ensuring minimal impact on the model's performance and watermark irretrievability from just a small number of queries. Experimental evaluation on a number of image classification tasks including MNIST showed the viability and efficiency of their technique. Additionally, they have studied the resistance of the approach to compression and overwriting attacks, observing that some neural architecture (such as IRNNs) are very vulnerable to the pruning attacks. They propose to further investigate extending the approach to various machine learning tasks, improving robustness to transfer learning attacks, and using the output scores to increase the watermark payload.

Wang et al.[14]T presented an attack scheme on deep learning models using visible watermarking. The watermarking parameters include position, transparency, color, rotation, and scaling, with nine parameters used for each parameter type. Two forms of attacks have been formulated to mimic real-world scenarios, where the first limits transparency, while the second limits the scaling of the watermark. The experiments conducted using the Inception V3 model trained using Image Net show that the adversarial samples produced through this approach successfully deceive the model, achieving high success rates. Moreover, the adversarial samples obtained are highly transferable, impacting models other than Inception V3, including Amazon Rekognition.

Wang et al.[15]introduced an approach for identifying and eliminating any watermarks, in this method, the secret data is inserted into the model, but it does not influence its performance when predicting. Nevertheless, the authors manage to prove that such watermarking significantly affects the probability distribution of the model's parameters, especially the variance of weights. Analyzing the distribution, one can reveal the presence of the watermark and even its length. Then, it can be eliminated by overwriting the watermarks. Their paper suggests applying the L2-regularization when embedding the watermark to avoid revealing it from analyzing parameter distribution.

Colangelo et al. [16] assessed the effectiveness of the application of the image transformation techniques based on steganographic and watermarking approaches to protect deep learning architectures from adversarial attacks. In their experiments, different image transformations, including scaling and JPEG2000 encoding, were used to remove the adversarial noise before feeding the image to classification. As a result, on the example of MNIST dataset and PGD adversary, it was shown that a lot of transformations positively affect the classification performance while not impacting its quality in the absence of any adversary. Thus, the study showed that a number of transformation types are able to provide defense mechanisms for certain attacks.

Zhang et al.[17] proposed a deep spatial watermarking scheme that seeks to defend the IP of deep learning models. This watermark embedding technique is inspired by the idea of media watermarking but applies the idea to deep learning where watermarks are embedded directly within the model as part of the training process without necessarily employing an embedding network. The use of specialized loss functions and training techniques makes the proposed method resilient to various forms of attacks, even from surrogates based on different networks and loss functions. Experimental results suggested that this framework effectively withstands various forms of attacks including white-box and black-box attacks as well as student-teacher surrogate model training. Furthermore, the framework has broad applications in defending IP for datasets and conventional machine learning models. However, it is found to be vulnerable to some pre-processing attacks such as random cropping and scaling attacks.

Tartaglione et al.[18]proposed new strategy that uses the properties of redundancy and adaptation capability of machine learning models to train ANNs to carry a watermark in the parameters, yet maintain robustness against fine-tuning attacks and compression attacks. The proposed framework involves controlling watermarked parameters using gradient maximization and adjusting other model parameters using gradient descent optimization. As a result, an ANN with such a set of parameters will be trained to hold its watermark, without any significant reduction in generalization capabilities.

In the conducted experiments using classical models with MNIST and CIFAR-10 datasets, they observed that the generated watermarks withstand attempts at fine-tuning attacks and compression attacks, and watermark detection based on the analysis of individual parameters is impossible due to their embedding into the entire system of model parameters.

Nagai et al.[19]presented digital watermarking scheme based on deep neural networks (DNNs), which ensures IP security and owner authentication. Watermarking is achieved through a parameter regularize technique, and can be applied from the start, in case of fine-tuning, or by distilling the model. This technique is robust against various forms of attacks such as fine-tuning and model pruning and the empirical findings indicate that the watermarked information does not get lost even when 65% of the parameters are pruned from the model. Hence, DNN watermarking offers a feasible solution for model IP protection.

Luo et al.[20]suggested watermarking model which does not require knowledge of distortions in its training stage. This watermarking model integrates adversarial learning and channel coding to provide resistance against both known and unknown distortions. The watermarking model outperforms the existing deep learning watermarking models which require the use of differentiable models for distortions. Experimental results show that the proposed model performs at least equivalently compared to the state-of-the-art models on distortions seen during training but outperforms the existing models on distortions unseen during training.

Uchida et al. [21] presented a general framework to incorporate digital watermarks within the deep neural network structure in order to safeguard the intellectual property rights. This technique is designed to directly embed the watermark in the model structure via the use of parameter regularizes, without affecting the model's performance. The experimental results confirm the robustness of the proposed watermarking scheme, as the watermark was able to withstand fine-tuning and parameter pruning, surviving even the pruning of 65% of the model parameters.

Hitaj et al.[22]assessed the resilience of DNN watermarking methods deployed for intellectual property protection in MLaaS platforms and provides attacks on such techniques. There are two primary approaches considered in the study; first, an ensemble technique in which several accurate models are used to bypass watermarks, and second, trigger detection, in which one watermarked model is used to detect and evade watermarks. The analysis demonstrates that even robust watermarking methods may be broken by attackers, which makes it possible to skip ownership checks without deleting the watermarks themselves. Although the first technique is very successful but expensive, the second technique greatly simplifies the attack process and lowers the cost significantly.

Zhang et al.[23] proposed an effective invisible watermarking mechanism for securing image processing models, by incorporating a task-agnostic, spatially invisible watermark in the model's output. As a result, any surrogate model created using the input-output pairs of the target model acquires the watermark. The scheme is capable of handling binary as well as high-resolution image watermarks, using conventional and deep learning-based spatial watermarking methods. In addition, the method is resilient against various architectures and loss functions used for surrogate models. Experimental results illustrate the ability of the proposed scheme in securing models, preserving the watermark, and generalizing to other applications.

Aiken et al. [24] introduced a neural network "laundering" technique that is able to strip out watermarks from deep neural networks based on backdoors without knowing any details about their construction. This algorithm exploits networks at the neuron activation level to erase watermark traces while achieving extremely high classification accuracy rates (above 97% on MNIST and above 80% on CIFAR-10 datasets) even when utilizing less than 1% of the initial training dataset. They demonstrate that today's approaches to embedding and detecting watermarks in neural networks, whether content-based or noise-based, are much less resilient than anticipated because the watermark could be extracted, reconstructed, and erased without compromising the model's overall performance.

Masoumeh et al.[25] studied resilience of cutting-edge watermarking through backdoors in deep learning models and design two effective attacks: one that works under the black-box setting and another for the white-box scenario. Under the black-box attack scenario, they successfully remove watermarks with minimal impact on classification performance without any labeled dataset and even without knowing the triggering pattern in the black-box model. Under the white-box attack scenario, their attack is able to successfully strip

watermarks more effectively and efficiently while significantly lowering the cost of stealing the model, which reduces the effort needed to retrain the model by up to twentyfold.

Zhang et al. [26] proposed framework involves three watermark generation schemes, watermark embedding exploiting the memorization and generalization capabilities of DNNs, and watermark activation based on predefined predictions triggered by the presence of watermarks. Experiments conducted using two image recognition datasets validate the framework's ability to successfully and effectively validate the ownership of models without compromising the performance of standard predictions. Additionally, the framework proves resistant to counter-watermarking strategies such as fine-tuning, pruning, and model inversion attacks.

Deeba et al. [27] proposed a DNN-based digital watermarking technique is proposed to ensure security and integrity of intellectual properties and also verify the owner remotely. In this regard, watermarks are embedded within DNNs during the training phase by utilizing the learning ability of the DNNs so that the watermarks remain intact even after undergoing various counter-watermark techniques like fine-tuning and retraining of the DNNs. Moreover, the proposed technique facilitates the remote verification of owner's identity without affecting the accuracy of DNNs in their conventional tasks. Extensive experiments on various benchmark data sets have shown promising results in this direction.

Li et al. [28]assessed that DNN models are vulnerable to piracy attacks because of their commercial importance, and watermarking has become an important tool for protecting intellectual property rights. In their study, a frequency-domain image-based watermarking approach is introduced for DNNs to enhance the robustness of black-box watermarking techniques. By utilizing triggers in the frequency domain, the approach ensures good concealment, defense against signal processing attacks, and reduces fraud attacks. Experimental results based on three databases and eight DNN models have proven that the proposed method preserves the integrity of the model while surpassing previous studies in terms of security measures.

Guan et al. [29] presented an algorithm of reversible watermarking to be used in a deep convolution neural network (CNN) for ensuring robust integrity authentication. To achieve this goal, a host sequence will be generated by employing model compression based on the pruning strategy. After that, histogram shift will be applied to generate a watermark. Importantly, the proposed scheme ensures that the watermarking process will be reversible in such a way that the model structure remains unchanged after applying this technique. Specifically, the impact on the model accuracy during the watermark generation will not exceed ±0.5%, thus making it possible to completely restore all model parameters.

Namba et al. [30] studied that current research develops a highly effective watermarking strategy for neural networks as a measure to prevent illegal distribution and abuse of services by attackers. They developed technique aims at overcoming the drawbacks of previously suggested methods that can be easily undermined due to the possibility of changing queries and modifying models. Their developed technique includes two major aspects, namely key generation through label change and key embedding by means of exponential weighting. The effectiveness of the developed technique has been tested in experiments and shown high performance.

Itay Mosafi et al. [31] presented work offers a novel attack strategy involving a mimicry attack where the attacker learns to create composite images from scratch to train a student model to replicate the behavior of a target (teacher) model without knowing the teacher's architecture, training set, or confidence scores. The approach does not require any other information beyond querying the model with labels for inputs. Experiments conducted with this strategy suggest that the student model can learn the behavior of the target model, and the student model exhibits similar performance as the target model, while at the same time being robust to known watermark detection techniques, making it impossible to distinguish whether the model is a stolen one.

Li et al. [32] presented a novel multi-watermarking attack against output-based watermarking schemes in deep learning architectures, with an objective of disrupting the process of identifying rightful ownership. This is achieved by injecting further watermarks via re-training, which may either nullify the original watermark or lower its accuracy significantly. The proposed attack has proven successful in lowering the watermark detection rate to less than 15 percent within a few training cycles, while in some cases, the original watermark has been completely eliminated at a

negligible expense. Furthermore, this approach may also be employed in eliminating any backdoor in the architecture or compromising its performance in tasks.

Quan et al. [33] proposed a multiple watermark attack on output-based watermarking systems, where attacker-defined watermarks are iteratively inserted into the model until the original watermark is either overridden or damaged. It is found that using this method, watermark detection rate can be lowered to less than 15% within only a couple of retraining's. Moreover, the approach is capable of completely removing watermarks, backdoors, or even undermining the task performance in some cases, indicating significant vulnerabilities in existing watermarking techniques. On the other hand, a novel black-box watermarking strategy for DNNs performing image-related tasks such as denoising and super-resolution is presented, where the watermark is embedded by intentionally reducing the task performance of the model on an image distinct from its training samples. Results show that the method can successfully protect the intellectual property rights of DNN models without negatively impacting their performance on regular input samples.

Wang et al. [34] presented a watermarking algorithm for deep neural networks (DNNs), where a separate neural network is embedded in the host DNN during training such that only specific parts of the host DNN weights are affected. The watermarks are embedded through error back propagation and later extracted without disclosing the independent network to the public. their experiments reveal that while the performance of the host DNN remains intact in terms of its original tasks, the watermarks can be successfully extracted regardless of possible attacks, including fine-tuning and model compression attacks. They believe that our proposed approach achieves greater accuracy, robustness, and capacity than the existing ones, as the watermarks are able to affect the parameters of both networks. However, one drawback of their algorithm is that the attacker might insert his watermark into the system using a new independent network.

Wu et al. [35] proposed a new watermarking framework for deep neural networks (DNNs), whose outputs are images, whereby each output from a network having a watermark bears a unique watermark. The method jointly trains the host DNN with a network for extracting the watermark using a loss function, thus, enabling the DNN to undertake its intended task and at the same time add a watermark in its outputs. Their method differs from earlier works, which involved watermarking the DNN model or the labels in a trigger set, since their method is robust in protecting the outputs. The effectiveness of our method can be demonstrated through experiments, which can verify the owner's identity as well as detect whether an image has been created from a particular DNN. Their experiments have proven the effectiveness of the approach on different applications such as image colorization, super resolution, image editing, and semantic segmentation.

Minoru Kuribayashi et al. [36] proposed approach based on the design of a white-box watermarking scheme using quantization for the FC layers of fine-tuned DNNs. In contrast to existing weight-based watermarking schemes, which may create substantial distortions in the weights and be vulnerable to detection through weight variance analysis, the proposed watermarking process involves quantization of sampled weights followed by manipulation of their frequencies to allow the watermark signal to propagate throughout the weights. The strengths of the method lie in its ability to introduce only minimal distortion, converge progressively without the need for any loss functions, and adjust the amount of distortion based on the quantization step size.

Gupta et al. [37] proposed a new color image watermarking scheme using DWT along with uncorrelated color space and artificial bee colony optimization algorithm. In their methodology, color watermark is embedded in the host image using UCS to exploit all channels of colors unlike correlated colors. This approach uses ABC algorithm for optimization of embedding strength factors to improve the quality and robustness of the scheme. The developed technique can be used efficiently for the purpose of image watermarking for copyright purposes due to high invisibility and resistance against attacks. Possible future scope includes extension of this approach to other multimedia data such as video. Comparative performance analysis of ABC with other optimization approaches such as GA and PSO can also be considered.

Pandey et al. [38] suggested an improved hybrid watermarking algorithm based on color images with Arnold transformation, which can provide better security and robustness. In comparison to the traditional schemes that use binary or gray scale watermark, here color image is considered as watermark. YCbCr color scheme is adopted to decompose RGB image in three

components, where Y component is chosen to embed singular values of watermark to host image singular values with a scaling parameter. Arnold transformation is applied in preprocessing of the watermark before embedding process in order to increase security by setting the number of iterations as secret key to scramble the watermark. As extraction process needs the original host image, this technique is referred to non-blind watermarking. PSNR and SSIM metrics are utilized to evaluate imperceptibility while NCC metric is considered to analyze similarity between original and extracted watermark. From the experimental results, it can be observed that the presented hybrid technique can give satisfactory results in terms of transparency, robustness, and data embedding capability when compared with other existing techniques. Besides, the proposed technique can withstand several kinds of attacks, while being suitable for limited bandwidth communications by adopting Arnold transformation.

Pandey et al. [39] proposed an enhanced lossless and robust color image watermarking technique using the lifting wavelet transform combined with grey wolf optimization (GWO) is introduced. Here, the singular values of a host image are altered to embed a color watermark, with the strength factor (α) being optimally obtained through the use of GWO for ensuring robustness and transparency. For added security during the embedding process, a color watermark image is scrambled using the Arnold transform before insertion into the host image. To validate the effectiveness of the proposed scheme, quantitative analysis is performed using fidelity parameters like PSNR, SSIM, and NCC. Robustness testing is also conducted where the resultant images undergo several attacks and comparisons are made with existing watermarking algorithms. It is observed from experimental analysis that the proposed hybrid method exhibits enhanced efficiency regarding visual fidelity, capacity, and robustness. The method ensures very high NCC values even after different types of attacks, thereby providing good extraction of watermark images. Besides, the proposed scheme provides reversible, lossless, efficient, and HVS-compliant embedding.

Patsariya et al. [40] offered a reliable watermarking strategy with a high level of security and non-blind properties. In contrast to previous algorithms that use grayscale images for the extraction process, this new method uses color images, and the Y component of the YCbCr color space is used to insert the watermark, since it is well compatible with the human visual system. To develop the suggested algorithm, a two-level lifting wavelet transformation is used in conjunction with SVD, where the singular values (matrices-diagonal) of the Y component of both source images are changed by means of scaling factor α. Additionally, two levels of scrambling are performed before inserting the watermark: image blocking scrambling and chaotic encryption, based on the modified Arnold transform, which guarantees a larger periodicity and increased cryptic power than regular transformations. The results were obtained through a comparison of several quality criteria, such as MSE, PSNR, SSIM, and NCC.

Patsariya et al. [41]presented a watermarked scheme utilizing the entropy-assisted lifting wavelet transform and singular value decomposition (Entropy-LWT-SVD). The Y channel from the YCbCr color space has been chosen for the embedding of secret information because it closely conforms to the human visual system, allowing effective and imperceptible data hiding. In order to increase the security of the data, a multilevel-multiple image scrambling technique has been employed before embedding the watermark. At the first level, the image is scrambled using blocks, and, at the second level, multiple images are scrambled using the Arnold transform. This two-level scrambling process leads to an increased key space and greater randomness, thus providing higher security and preventing any unauthorized extraction or attacks

5. CONCLUSIONS

The vulnerabilities in deep neural network (DNN) based models, which include enormous training data, computation, and expertise, lie in piracy and unauthorized distribution. This has prompted research on watermarking methods to protect IP through embedding digital watermarks in DNNs. Several kinds of watermarking strategies have been introduced, including backdoor watermarking, spatially invisible watermarking, embedding via regularize of model parameters, and frequency domain-based watermarking in images. Watermarking is done in training or fine-tuning phase, taking advantage of model memorizing ability, which facilitates watermark extraction for remote proof of IP ownership in both black-box and white-box cases. Experiments reveal successful extraction of watermark without loss of model performance and high resilience to pruning, fine-tuning, and model inversion attacks. Recent attacks, such as laundering algorithms, label queries in the black-box

scenario, and parameter modification in the white-box scenario, indicate that many existing watermarking techniques can be easily stripped off by these attacks without sacrificing model performance, even with a small portion of training data. Frequency-based watermarking increases the resistance of DNN-based IP against fraudulent claims and signal processing.

REFRENCES

1.Patsariya, S., Dixit, M.: A Survey on Watermarking and Its Techniques. Algorithms for Intelligent Systems. 71–78 (2021). https://doi.org/10.1007/978-981-33-4893-6_7.

2.Pandey, M.K., Parmar, G., Patsariya, S.: An Effective Way to Hide the Secret Audio File Using High Frequency Manipulation. Communications in Computer and Information Science. 125–130 (2011). https://doi.org/10.1007/978-3-642-18440-6_15.

3.Begum, M., Uddin, M.S.: Digital Image Watermarking Techniques: A Review. Information. 11, 110 (2020). https://doi.org/10.3390/info11020110.

4.Anand, A., Singh, A.K.: Watermarking techniques for medical data authentication: a survey. Multimedia Tools and Applications. (2020). https://doi.org/10.1007/s11042-020-08801-0.

5.Singh, A.K., Sharma, N., Dave, M., Mohan, A.: A novel technique for digital image watermarking in spatial domain. 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing. (2012). https://doi.org/10.1109/pdgc.2012.6449871.

6.Patvardhan, C., Kumar, P., Vasantha Lakshmi, C.: Effective Color image watermarking scheme using YCbCr color space and QR code. Multimedia Tools and Applications. 77, 12655–12677 (2017). https://doi.org/10.1007/s11042-017-4909-1.

7.Singh, R.K., Shaw, D.K., Jha, S.K., Kumar, M.: A DWT-SVD based multiple watermarking scheme for image based data security. Journal of Information and Optimization Sciences. 39, 67–81 (2017). https://doi.org/10.1080/02522667.2017.1372153.

8.Parah, S.A., Sheikh, J.A., Assad, U.I., Bhat, G.M.: Realisation and robustness evaluation of a blind spatial domain watermarking technique. International Journal of Electronics. 104, 659–672 (2016). https://doi.org/10.1080/00207217.2016.1242162.

9.Agarwal, N., Singh, A.K., Singh, P.K.: Survey of robust and imperceptible watermarking. Multimedia Tools and Applications. 78, 8603–8633 (2019). https://doi.org/10.1007/s11042-018-7128-5.

10.Ding, W., Ming, Y., Cao, Z., Lin, C.-T.: A Generalized Deep Neural Network Approach for Digital Watermarking Analysis. 6, 613–627 (2021). https://doi.org/10.1109/tetci.2021.3055520.

11.Tayyab, M., Marjani, M., Jhanjhi, N.Z., Hashem, I.A.T.: A Light-weight Watermarking-Based Framework on Dataset Using Deep Learning Algorithms. 2021 National Computing Colleges Conference (NCCC). 1–6 (2021). https://doi.org/10.1109/nccc49330.2021.9428845.

12.Chang, C.-Y., Su, S.-J.: A Neural-Network-Based Robust Watermarking Scheme. (2006). https://doi.org/10.1109/ICSMC.2005.1571521.

13.Le Merrer, E., Pérez, P., Trédan, G.: Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications. 32, 9233–9244 (2019). https://doi.org/10.1007/s00521-019-04434-z.

14.Wang, G., Chen, X., Xu, C.: Adversarial Watermarking to Attack Deep Neural Networks. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1962–1966 (2019). https://doi.org/10.1109/icassp.2019.8682351.

15.Wang, T., Kerschbaum, F.: Attacks on Digital Watermarks for Deep Neural Networks. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2622–2626 (2019). https://doi.org/10.1109/icassp.2019.8682202.

16.Colangelo, F., Neri, A., Battisti, F.: Countering Adversarial Examples by Means of Steganographic Attacks. 2019 8th European Workshop on Visual Information Processing (EUVIP). 193–198 (2019). https://doi.org/10.1109/euvip47703.2019.8946254.

17.Zhang, J., Chen, D., Liao, J., Zhang, W., Feng, H., Hua, G., Yu, N.: Deep Model Intellectual Property Protection via Deep Watermarking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 44, 4005–4020 (2022). https://doi.org/10.1109/TPAMI.2021.3064850.

18.Tartaglione, E., Grangetto, M., Cavagnino, D., Botta, M.: Delving in the loss landscape to embed robust watermarks into neural networks. 2020 25th International Conference on Pattern Recognition

(ICPR). 1243–1250 (2021). https://doi.org/10.1109/icpr48806.2021.9413062.

19.Nagai, Y., Uchida, Y., Sakazawa, S., Satoh, S.: Digital watermarking for deep neural networks. International Journal of Multimedia Information Retrieval. 7, 3–16 (2018). https://doi.org/10.1007/s13735-018-0147-1.

20.Luo, X., Zhan, R., Chang, H., Yang, F., Peyman Milanfar: Distortion Agnostic Deep Watermarking. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 13545–13554 (2020). https://doi.org/10.1109/cvpr42600.2020.01356.

21.Uchida, Y., Nagai, Y., Shigeyuki Sakazawa, Shin'ichi Satoh: Embedding Watermarks into Deep Neural Networks. arXiv (Cornell University). (2017). https://doi.org/10.1145/3078971.3078974.

22.Hitaj, D., Hitaj, B., Mancini, L.V.: Evasion Attacks Against Watermarking Techniques found in MLaaS Systems, https://ieeexplore.ieee.org/abstract/document/8768572/, last accessed 2020/08/20. https://doi.org/10.1109/SDS.2019.8768572.

23.Zhang, J., Chen, D., Liao, J., Fang, H., Zhang, W., Zhou, W., Cui, H., Yu, N.: Model Watermarking for Image Processing Networks. 12805–12812 (2020). https://doi.org/10.48550/arXiv.2002.11088.

24.Aiken, W., Kim, H., Woo, S., Ryoo, J.: Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. Computers & Security. 106, 102277 (2021). https://doi.org/10.1016/j.cose.2021.102277.

25.Masoumeh Shafieinejad, Lukas, N., Wang, J., Li, X., Kerschbaum, F.: On the Robustness of Backdoor-based Watermarking in Deep Neural Networks. arXiv (Cornell University). (2021). https://doi.org/10.1145/3437880.3460401.

26.Zhang, J., Gu, Z., Jang, J., Wu, H., Marc Ph. Stoecklin, Huang, H., Molloy, I.M.: Protecting Intellectual Property of Deep Neural Networks with Watermarking. Computer and Communications Security. (2018). https://doi.org/10.1145/3196494.3196550.

27.Deeba, F., Tefera, G., Kun, S., Memon, H.: Protecting the Intellectual Properties of Digital Watermark Using Deep Neural Network. 91–95 (2019). https://doi.org/10.1109/icise.2019.00025.

28.Li, M., Zhong, Q., Zhang, L.Y., Du, Y., Zhang, J., Xiang, Y.: Protecting the Intellectual Property of Deep Neural Networks with Watermarking: The Frequency Domain Approach. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 402–409 (2020). https://doi.org/10.1109/trustcom50675.2020.00062.

29.Guan, X., Feng, H., Zhang, W., Zhou, H., Zhang, J., Yu, N.: Reversible Watermarking in Deep Convolutional Neural Networks for Integrity Authentication. Proceedings of the 30th ACM International Conference on Multimedia. 2273–2280 (2020). https://doi.org/10.1145/3394171.3413729.

30.Namba, R., Sakuma, J.: Robust Watermarking of Neural Network with Exponential Weighting. (2019). https://doi.org/10.1145/3321705.3329808.

31.Itay Mosafi, David, E.O., Netanyahu, N.S.: Stealing Knowledge from Protected Deep Neural Networks Using Composite Unlabeled Data. 2022 International Joint Conference on Neural Networks (IJCNN). 1–8 (2019). https://doi.org/10.1109/ijcnn.2019.8851798.

32.Li, D., Yang, Y.: The Multi-Watermarks Attack of DNN Watermarking. 2020 4th International Conference on Advances in Image Processing. 178–184 (2020). https://doi.org/10.1145/3441250.3441279.

33.Quan, Y., Teng, H., Chen, Y., Gao, H.: Watermarking Deep Neural Networks in Image Processing. IEEE transactions on neural networks and learning systems. 32, 1852–1865 (2021). https://doi.org/10.1109/tnnls.2020.2991378.

34.Wang, J., Wu, H., Zhang, X., Yao, Y.: Watermarking in deep neural networks via error back-propagation. Electronic Imaging. 022-1022-8 (2020). https://doi.org/10.2352/issn.2470-1173.2020.4.mwsf-022.

35.Wu, H., Liu, G., Yao, Y., Zhang, X.: Watermarking Neural Networks With Watermarked Images. IEEE transactions on circuits and systems for video technology. 31, 2591–2601 (2021). https://doi.org/10.1109/tcsvt.2020.3030671.

36.Minoru Kuribayashi, Tanaka, T., Suzuki, S., Yasui, T., Nobuo Funabiki: White-Box Watermarking Scheme for Fully-Connected Layers in Fine-Tuning Model. (2021). https://doi.org/10.1145/3437880.3460402.

37.Gupta, M., Parmar, G., Gupta, R., Saraswat, M.: Discrete wavelet transform-based color image watermarking using uncorrelated color space and artificial bee colony. International Journal of Computational Intelligence Systems. 8, 364 (2015). https://doi.org/10.1080/18756891.2015.1001958.

38.Pandey, M.K., Parmar, G., Gupta, R., Sikander, A.: Non-blind Arnold scrambled hybrid image watermarking in YCbCr color space. Microsystem Technologies. 25, 3071–3081 (2018). https://doi.org/10.1007/s00542-018-4162-1.

39.Pandey, M.K., Parmar, G., Gupta, R., Sikander, A.: Lossless robust color image watermarking using lifting scheme and GWO. International Journal of System Assurance Engineering and Management. 11, 320–331 (2019). https://doi.org/10.1007/s13198-019-00859-w.

40.Patsariya, S., Dixit, M.: A New Block Based Non-Blind Hybrid Color Image Watermarking Approach Using Lifting Scheme and Chaotic Encryption Based on Arnold Cat Map. Traitement du Signal. 39, 1159–1168 (2022). https://doi.org/10.18280/ts.390408.

41.Patsariya, S., Dixit, M.: Entropy Based Secure and Robust Image Watermarking Using Lifting Wavelet Transform and Multi-Level-Multiple Image Scrambling Technique. Traitement du Signal. 39, 1751–1759 (2022). https://doi.org/10.18280/ts.390533.