Watermarking Models and Artificial Intelligence

B. M. El-den[1], Marwa M. Eid[2]

[1]Department of Electronics and Communication Engineering, Faculty of Engineering, Delta University for Science& Technology, International Coastal Road, Gamasah City, Mansoura, Dakhliya, Egypt, Deltauniv.edu.eg,

[2]Faculty of Artifcial Intelligence, Delta University for Science and Technology, Mansoura, Egypt

Emails: Basant_moheyelden@yahoo.com ; mmm@ieee.org

Abstract

Machine learning and deep learning are good bets for solving various intelligence-related problems. While it has practical applications in watermarking, it performs less well on more standard tasks like prediction, classification, and regression. This article offers the results of a thorough investigation into watermarking using modern tools like AI, ML, and DL. Watermarking's origins, some historical context, and the most fascinating and practical applications are also covered briefly.

Keywords: Steganography; digital watermarking; data hiding applications; fingerprint; Deep Neural Networks

1. Introduction

Data hiding means the using of data concealing techniques [1] and an unremarkable cover medium, users can communicate, authenticate, identify, copy-protect, etc., in secret. As more and more multimedia content moves to the cloud, techniques like digital watermarking and steganography become more vital to protecting the integrity of this content. Since the main requirements of these methods—invisibility, robustness, security, and capacity—are complementary, many data hiding systems strive to deliver optimal performance.

Although adding a watermark to a model won't stop it from being stolen, it can help rightful owners track down their property if it does get stolen. So, if the model is stolen, the rightful owner can use the watermark to prove ownership even after the fact. For instance, even if a watermarking system does not provide any link between the watermark and the identity of the original model owner, the legitimate model owner may be able to identify stolen model instances. But if the owner wishes to assert copyright before a third party, like a company, it's not very helpful. Therefore, it is essential to assess the situation and select an appropriate watermarking approach properly. In this part, we proposedly ide a taxonomy to help classify watermarking techniques along five d. Classifications like these might help find and contrast different watermarking methods that can meet the needs of a specific circumstance [1]

In general, and especially for watermarking methods, it is essential to evaluate the security of a system in a well-defined threat space that characterizes the attacker's knowledge, capabilities, and intentions in addition to the underlying security goals. This helps explain in detail what characteristics the watermarking scheme needs to have to accomplish its goal of protecting the model in a specific scenario. For instance, a watermarking scheme designed to protect an ML model directly distributed to the users in a white-box fashion will likely need to possess different characteristics than a scheme applied to a model that is only accessible in a black-box fashion [2]. A thorough understanding of when and how an attacker would try to circumvent the watermark requires a detailed characterization of the attack surface. Regardless of the attacker type, the critical factor is whether the model was stolen in a black or white-box fashion.

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

2. Related Work

Digital watermarking: There are a plethora of contexts in which data hiding techniques could be helpful. In addition to broadcast monitoring, content authentication, tamper detection, device control, owner identification or transaction tracking, copy management, legacy enhancement, and content authentication, digital watermarking is the most widely used copyright protection technique. In contrast, steganography is concerned with hidden messages for the military, political dissidents, or criminal organizations. Since steganography has found usage in the military and the criminal underworld, there has been a rise in scholarly interest in steganalysis, or the procedures used to decipher stenographic communications [2]. Recently, new use cases for data hiding have emerged [3] in addition to the aforementioned classic ones. Steganography has found new applications in areas such as malware injection, network steganography, voice-over-IP steganography, privacy-preserving transaction tracking, digital watermarking for forensics, and network flow watermarking. Digital watermarking provides post-decryption security for multimedia files, allowing access only to authorized users. It gradually alters the source material (host signal) by masking the embedded identification data (watermark). In the future, this data can be used to verify the carrier signal's origin and determine who is responsible for it. The first two processes of a digital watermarking system are embedding and extracting the watermark. To create a watermarked signal, an embedding algorithm will append a watermark to the host signal; the extracted signal will then be free of the watermark. If the signal wasn't tampered with during transmission, the watermark will still be there and can be wiped out Embedding and extracting the watermark both employ a secret key to prevent theft or tampering. On the other hand, watermark detection can only verify ownership, whereas watermark extraction can prove ownership. It's essential to consider the following [4] qualities while deciding on a watermarking method: Imperceptibility refers to how similar the watermarked and unwatermarked versions of digital content appear to the human eye. There can be no loss of quality due to the inserted watermark distortion.

- Resilience, or the ability to decipher the watermark using just commonplace signal-processing techniques. The level of security provided by digital watermarking can be classified as either robust, fragile, or semi-fragile. A watermark's durability must withstand signal-processing operations (at least below some distortion threshold).
- Capacity: the number of bits a watermark can encode in a specified time (or space in the case of still images). Protection from outside intrusion or attack; security. Watermarking algorithms must be secure because the encoded information cannot be read by an attacker or extracted. Authorized individuals should only access the information included in a watermark.
- Digital watermarking has been effectively implemented in various contexts, such as copyright protection, transaction tracking, content authentication, and heritage enhancement. There have been hardly any presented watermarking systems.
- Watermarking audio and video data have gained attention in recent years for the aforementioned wide range of applications.

Multimedia fingerprinting: Also known as transaction tracking is possible to track down the identity of the pirates after discovering an unlawful copy, in contrast to digital watermarking, which cannot identify the source of piracy.

This traceability is made possible by adding a special user-specific piece of data, called a fingerprint, to several copies of the same content. A multimedia fingerprinting algorithm is a three-step technique that connects the client and the content owner and enables the tracking of pirates from illegally obtained or coerced copies. The following restrictions are anticipated to be addressed by a multimodal fingerprinting system [5]:

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

- Robustness: The chosen watermark embedding method affects how resilient a fingerprint is to signal processing activities. A reliable watermarking algorithm must be used to identify an illegal re-distributor after the digital content has been altered by conventional signal processing techniques.
- Collusion resistance: Even though digital fingerprinting might be effective at locating a single adversary, many dishonest consumers could unite to launch effective collusion attacks against the fingerprinting system. By comparing their numerous versions, the conspirators can identify the regions carrying the fingerprint signal, delete the data from those areas, and then create a copy that cannot be linked to any of the originals. A fingerprinting mechanism must be made to survive these collusion attacks.
- Quality tolerance: Fingerprinted content ought to be aesthetically pleasing and perceptually accurate to the original.
- Embedding capacity: This parameter controls how much fingerprint space is allotted to each user. The fingerprint is a binary string, which may be quite long. A digital fingerprint system must have sufficient embedding capacity to hold an entire fingerprint.

Since the content owner learns personally identifiable information about the client during the embedding process, a conventional fingerprinting protocol between the consumer and the content owner is unfavorable from the client's standpoint.

A malicious content owner might then use the customer's identifying information to prosecute them for unauthorized distribution of the embedded content. To solve this problem, cryptographic approaches inspired the development of anonymous fingerprinting systems.

An anonymous fingerprinting system that is both thorough and secure is expected to provide features such as buyer frame proofness, traceability, anonymity, non-repudiation, and the ability to unlink [6]. There has been an uptick in the proposal of collusion-resistant and anonymous fingerprinting techniques for multimedia material.

***Steganography and watermarking*:** Data can be hidden in a piece of media via steganography or watermarking. The information is not an identifier of the object's creator but rather a coded message. This secret communication must be sent without being uncovered, intercepted, or decoded. Unlike encryption, the primary goal of steganography is to maintain the cover media's original structure after data has been hidden inside it [7]. It is essential that the actual content of the cover media is still accessible to the public while the embedded messages remain undetectable. Steganography is used anywhere secret communication is required for safety concerns.

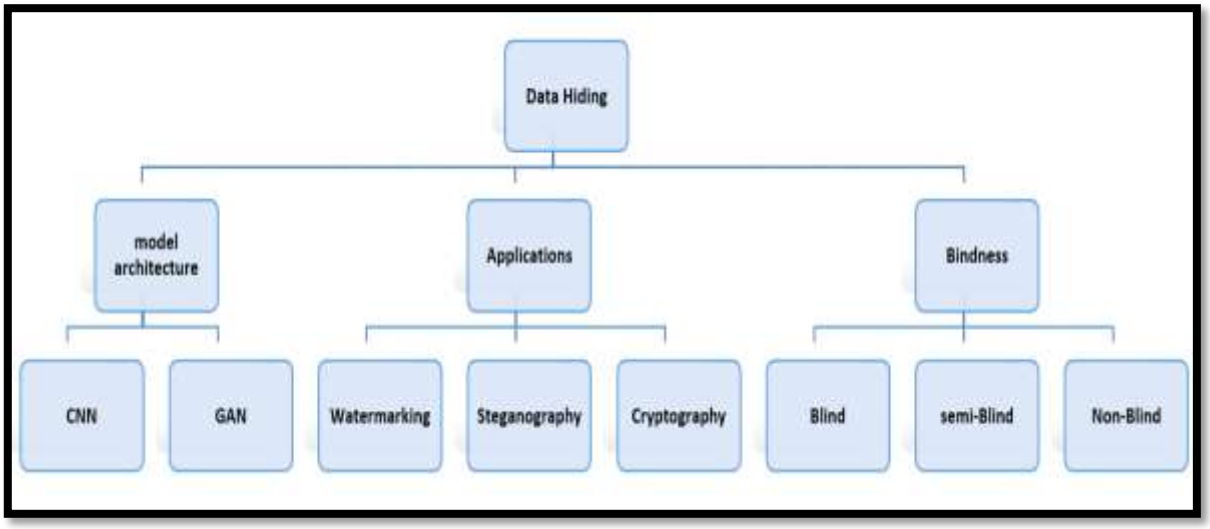

Figure 1: is a block diagram illustrating various approaches to categorizing deep learning-based data concealing techniques.

**Model architectures based on *GAN* and *CNN* can be separated from one another. Attack training is the practice of planning and producing attacks during model training. Blindness refers to the operation of the data masking technique.

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

Because deep neural networks are so good at representing data, new developments in the field of deep learning affect a wide range of businesses. Deep learning models offer adaptive, generic frameworks for a range of applications in watermarking and steganography in the realm of data concealment. Compared to conventional watermarking or steganography algorithms, these machine-learning models can create sophisticated embedding patterns that are more effectively resistant to various attacks [9]. The majority of recent studies in this field focus on image-based data concealing; hence these studies will be contrasted in this survey.

To securely embed data, provide communication and content authentication services, and interact with a range of cover media types, universal frameworks for data concealing must be developed [10, 11]. Deep data concealment techniques may help increase the security of the embedded messages. The advantage of the deep learning technique is that networks may be retrained to resist new attack types or to prioritize particular goals like payload capacity or imperceptibility without building specialized algorithms for each new application [12]. Due to the considerable non-linearity of deep neural networks, it is difficult for an adversary to obtain the underlying data.

Deep learning-based approaches are more adaptable to many applications and secure than previous approaches and provide improved resistance to adversarial assaults and distortions. Additionally, they can accomplish more subtle forms of data embedding.

A. Deep Learning-Based Data Hiding Techniques

Deep learning-based data concealing techniques teach models to reliably and stealthily conceal data by utilizing encoder-decoder-decoder network designs. Compared to more conventional data concealment methods, their flexibility and adaptability make them far more advantageous. Due to their "black box" nature, deep learning models can increase security without requiring specialized training to implement data concealment methods.

Two methods of hiding information that rely on deep learning are steganography and watermarking. In Figure 2, we see a taxonomy of the data masking methods that rely on deep learning discussed in this article. It's worth noting that adversarial-training CNNs are different from GAN-based approaches in this respect. Adversarial training, in this sense, refers to the rigorous analysis of encoded and cover pictures by a discriminator to increase embedding imperceptibility, in contrast to GAN-based approaches, which use trained CNNs for noise injection during the assault simulation step. [13]

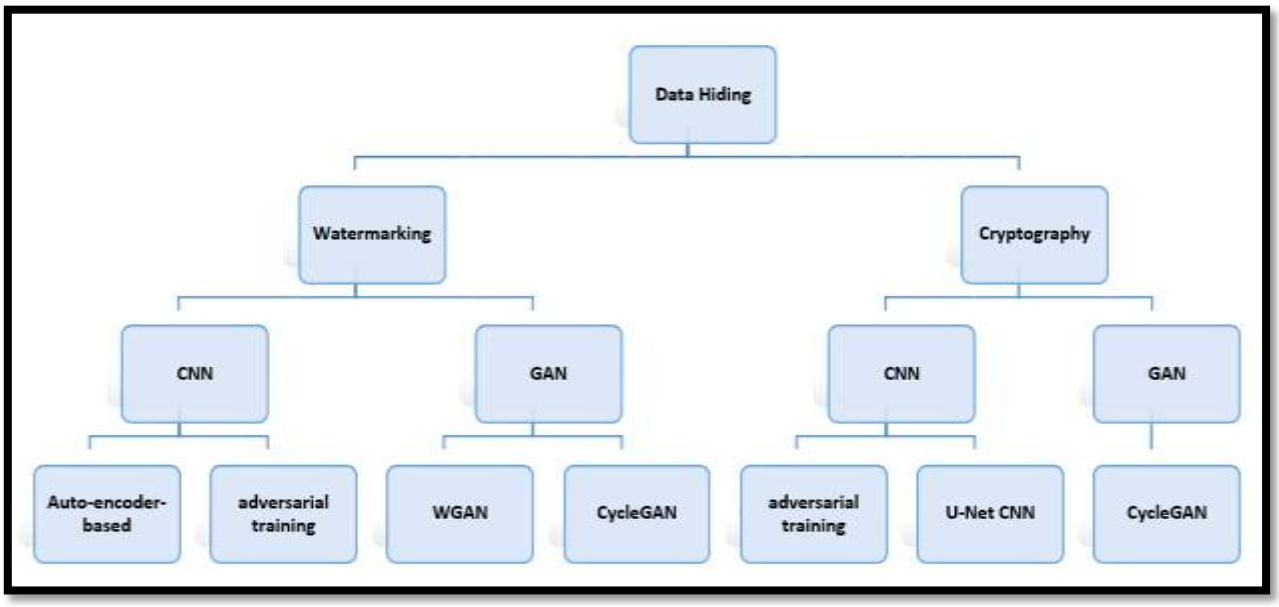

Figure 2: A block diagram illustrating the classification of models for data concealing based on deep learning shown during training, "adversarial training" refers to attack simulation, which includes attacks produced by trained *CNNs* that use noise as their primary attack type.

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

Most modern watermarking models currently use encoder-decoder architectures based on convolutional neural networks (*CNNs*). Figure.3 provides a straightforward schematic illustrating the deep learning-based data concealing procedure.

In these models, the encoder embeds data in a piece of cover media, which is then put through an assault simulation before the decoder network extracts the data. The embedding technique grows increasingly resistant to the attacks used during simulation through the iterative learning process, while the extraction procedure enhances the integrity of the recovered data. This method has an advantage over earlier traditional algorithms in that it does not require specialized programming knowledge, and it can be retrained for various applications and attack types rather than being created from scratch. The embedding system's intricate workings are unknown and are not detectable because the system is a black box with strong non-linearity.

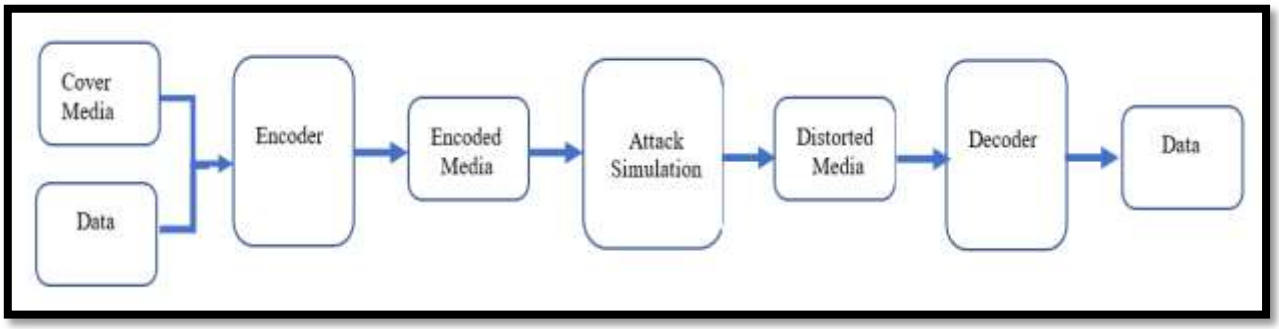

Figure 3: A generic encoder-decoder architecture for digital watermarking therefore, solutions based on deep learning can be used securely in a wide range of applications.

Convolutional auto-encoders and CNNs with adversarial training components improve upon the illustratively straightforward CNN encoder-decoder method. The U-Net CNN architecture finds widespread use in steganographic applications with superior image segmentation capabilities. The GAN structure is used by several models [14].

The GAN framework combines a generative model and a discriminative model. In deep data hiding, the discriminator network must classify encoded and unaltered picture versions. As the generative model improves its data embedding abilities, it can produce samples that are harder to identify, while the discriminative model improves its ability to recognize encoded images. The training is complete when the discriminator only succeeds half the time in identifying legally encoded images, at which point it is merely making educated guesses. Steganography and watermarking rely heavily on discriminative networks due to their ability to increase data's invisibility significantly. CycleGANs and Wasserstein GANs are just two examples of a subset of the GAN framework (WGANs). The CycleGAN architecture efficiently transforms one image into another by combining two generative and discriminative models. One of the primary advantages of CycleGANs is that the model may be trained without requiring paired examples. Instead, the first generator creates images in domain A, while the second creates images in domain B; both use an image from the other domain to execute the translation. Discriminator determines whether the photographs are genuine or fraudulent based on the photos from domain A as input and the images produced by the generator as output. The resulting structure is functional for rapidly navigating between images.

3. Artificial Intelligence Model Watermarking

As computers and servers have gotten better at mimicking human performance, artificial intelligence (AI) has flourished in recent years. Artificial intelligence has allowed the resolution of many problems in speech recognition, image recognition, and natural language processing. It has also helped to increase the security of documents by including a digital watermark in the original versions sent over the internet [15-20]. In recent years, machine learning techniques have been applied to ensure the safety of data transmitted over the internet. Predicting the optimal embedding strength, striking a balance between robustness and security, speeding up the training set, and designing and optimizing the maximum likelihood relation set are all tasks that can be accomplished with the help of machine learning techniques in watermarking algorithms. Several deep learning algorithms with different uses were discovered over time [21-24].

In recent years, deep learning with a powerful learning capability has emerged as the most economical watermarking technique due to its efficient tradeoff between quality and robustness. Several deep learning methods with varying degrees of utility have been discovered. Watermarking techniques that efficiently compromise quality and resilience are more common, and deep learning with a strong learning power has emerged as the most successful method .

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

4. Conclusion

In this study, we map out the landscape of digital picture watermarking in settings that include machine learning. In addition, we talk about the reasons behind, targets of, methods for, and models of each strategy in detail.

In this study, deep learning-based models for data concealment using steganography or watermarking techniques are thoroughly categorized and compared.

Presents the future path for deep learning-based data concealing research and a thorough discussion and comparison of the various objective strategies, evaluation measures, and training datasets used in state-of-the-art deep data hiding systems today.

References

[1] W. Bender, D. Gruhl, N. Morimoto, A. Lu Techniques for data hiding IBM Syst. J., 35 (1996), pp. 313-336 CrossRefView Record in Scopus Google Scholar

[2] Cox, I.; Miller, M.; Jeffrey, A.; Fridrich, J.; Kalker, T. Digital Watermarking and Steganography, 2nd ed.; Morgan Kaufmann: Burlington, MA, USA, 2008. [CrossRef]

[3] Megías, D. Data hiding: New opportunities for security and privacy? In Proceedings of the European Interdisciplinary Cybersecurity Conference (EICC 2020), Rennes, France, 18 November 2020; Article No.: 15. pp. 1–6. [CrossRef].

[4] Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; Kalker, T. Digital Watermarking and Steganography, 2nd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2007.

[5] Katzenbeisser, S.; Petitcolas, F.A. Information Hiding Techniques for Steganography and Digital Watermarking, 1st ed.; Artech House, Inc.: Norwood, MA, USA, 2000.

[6] Ibrahim, Abdelhameed, and El-Sayed M. El-kenawy. "Applications and datasets for superpixel techniques: A survey." Journal of Computer Science and Information Systems 15, no. 3 (2020): 1-6.

[7] Eid, Marwa M., El-Sayed M. El-kenawy, and Abdelhameed Ibrahim. "A binary sine cosine-modified whale optimization algorithm for feature selection." In 2021 National Computing Colleges Conference (NCCC), pp. 1-6. IEEE, 2021.

[8] Qureshi, A.; Megías, D.; Rifà-Pous, H. Framework for Preserving Security and Privacy in Peer-to-Peer Content Distribution Systems. Expert Syst. Appl. 2015, 42, 1391–1408. [CrossRef]

[9] Mehdi Hussain, Ainuddin Wahid Abdul Wahab, Yamani Idna Bin Idris, Anthony T.S. Ho, and Ki-Hyun Jung. 2018. Image steganography in spatial domain: A survey. Signal Processing: Image Communication 65 (2018), 46–66. https://doi.org/10.1016/j.image.2018.03.012

[10] Inas Jawad Kadhim, Prashan Premaratne, Peter James Vial, and Brendan Halloran. 2019. Comprehensive survey of image steganography: Techniques, Evaluations, and trends in future research. Neurocomputing 335 (2019), 299–326.

[11] El-Sayed Towfek, M., and M. Saber El-kenawy. "Reham Arnous. An Integrated Framework to Ensure Information Security Over the Internet." International Journal of Computer Applications 178, no. 29 (2019): 13-15.

[12] Haribabu Kandi, Deepak Mishra, and Subrahmanyam R.K. Sai Gorthi. 2017. Exploring the learning capabilities of convolutional neural networks for robust image watermarking. Computers & Security 65 (2017), 247–268. HTTPS: //doi.org/10.1016/j.cose.2016.11.016

[13] Xiyang Luo, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. 2021. DVMark: A Deep Multiscale Framework for Video Watermarking. (04 2021).

[14] Innfarn Yoo, Huiwen Chang, Xiyang Luo, O. Stava, Ce Liu, P. Milanfar, and Feng Yang. 2021. Deep 3D-to-2D Watermarking: Embedding Messages in 3D Meshes and Extracting Them from 2D Renderings. ArXiv abs/2104.13450 (2021).

[15] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. Hidden: Hiding Data with Deep Networks. In Proceedings of the European Conference on Computer vision (ECCV 2018). 657–672.

[16] El-Kenawy, El-Sayed M., Marwa Eid, and Alshimaa H. Ismail. "A New Model for Measuring Customer Utility Trust in Online Auctions." International Journal of Computer Applications 975: 8887.

[17] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. 2020. Distortion Agnostic Deep Watermarking. Computer Vision Foundation (2020). arXiv:2001.04580 [cs.MM].

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. Advances in Neural Information Processing Systems 3, 11 (2014). arXiv:1406.2661 [stat.ML]

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022

[19] Das, Kajaree, and Rabi Narayan Behera. "A survey on machine learning: concept, algorithms and applications." International Journal of Innovative Research in Computer and Communication Engineering 5, no. 2 (2017): 1301-1309.
[20] El-kenawy, El-Sayed M., Marwa M. Eid, and Abdelhameed Ibrahim. "Anemia estimation for covid-19 patients using a machine learning model." Journal of Computer Science and Information Systems 17, no. 11 (2021): 2535-1451.
[21] Kavitha, R.S.; Eranna, U.; Giriprasad, M.N. DCT-DWT Based DigitalWatermarking and Extraction using Neural Networks. In 556Proceedings of the 2020 International Conference on Artificial Intelligence and Signal Processing (AISP); IEEE: Amaravati, India, 5572020; pp. 1–5.
[22] Ibrahim, Abdelhameed, Seyedali Mirjalili, Mohammed El-Said, Sherif SM Ghoneim, Mosleh M. Al-Harthi, Tarek F. Ibrahim, and El-Sayed M. El-Kenawy. "Wind speed ensemble forecasting based on deep learning using adaptive dynamic optimization algorithm." IEEE Access 9 (2021): 125787-125804.
[23] D. Sukheja, J. A. Shah, G. Madhu, K. Sandeep Kautish, F. A. Alghamdi et al., "New decision-making technique based on hurwicz criteria for fuzzy ranking," Computers, Materials & Continua, vol. 73, no.3, pp. 4595–4609, 2022.
[24] Frattolillo, F. A Watermarking Protocol Based on Blockchain. Applied Sciences 2020, 10, 7746.

Doi: https://doi.org/10.54216/JAIM.010203
Received: February 27, 2022 Accepted: July 02, 2022