Chapter 1
Enhancing Security and Ownership Protection of Neural Networks Using Watermarking Techniques: A Systematic Literature Review Using PRISMA

Roseline Oluwaseun Ogundokun, Christiana Oluwakemi Abikoye, Aditya Kumar Sahu, Akinyemi Omololu Akinrotimi, Akinbowale Nathaniel Babatunde, Peter O. Sadiku, and Omosola Jacob Olabode

Abstract The rise of artificial intelligence (AI) and machine learning (ML) has prompted concerns regarding the intellectual property (IP) protection of neural networks (NNs). A proposed solution is watermarking, which incorporates a unique identifier into a NN. However, the effectiveness of watermarking methods in enhancing privacy and ownership secrecy remains questionable. This study intended to evaluate the efficacy of watermarking techniques for enhancing the security and ownership protection (SOP) of NNs. An exhaustive search of scholarly databases for peer-reviewed journal articles and conference proceedings was conducted in accordance with PRISMA standards. Eligible papers evaluated the efficacy of watermarking techniques used to protect NNs. Twenty research articles using vari-

R. O. Ogundokun (✉)
Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania

Department of Computer Science, Landmark University, Omu Aran, Nigeria
e-mail: roseline.ogundokun@ktu.edu; ogundokun.roseline@lmu.edu.ng

C. O. Abikoye · P. O. Sadiku
Department of Computer Science, University of Ilorin, Ilorin, Nigeria

A. Kumar Sahu
Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amaravati, Andhra Pradesh, India

A. O. Akinrotimi
Department of Mathematics Sciences (Computer Science Program), Augustine University Ilara-Epe, Epe, Nigeria

A. N. Babatunde
Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria

O. J. Olabode
Department of Mathematical and Computing Science, Thomas Adewumi University, Oko-Irese, Nigeria

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

A. Kumar Sahu (ed.), *Multimedia Watermarking*,
https://doi.org/10.1007/978-981-99-9803-6_1

ous watermarking techniques, including digital watermarking (DW), reversible watermarking (REW), and robust watermarking (ROW), were analyzed. Various performance indicators, such as detection rate (DR), robustness, and distortion, were employed to evaluate the applicability of each method. The results demonstrated that watermarking techniques effectively protected the intellectual property of NNs with minimal impact on performance. However, the need for specialized apparatus and the difficulty of incorporating watermarks into deep neural networks (DNN) hampered their implementation. To improve the practicability and effectiveness of watermarking techniques, additional research is required. Researchers, professionals, and policymakers should consider watermarking to safeguard the intellectual property of NNs in a variety of domains, including finance, healthcare, and national security.

Keywords Watermarking · Neural network · Security and ownership protection · Systematic literature review · PRISMA

1.1 Introduction

Based on the swift advancement of computer efficiency, deep learning (DL) strategies have been intensively studied in numerous disciplines of study (Chaumont 2020; Adegun et al. 2021; Zou et al. 2019; Kalsi et al. 2018; Andi 2021; Awotunde et al. 2020; Zhang et al. 2020). The fundamental approaches of DL are the framework of a deep neural network (DNN) and the training of its predictive algorithm employing massive amounts of data. In academia and professional studies, the perpetuation of unified designs serves an essential purpose in advancing the field of DL. Alternatively, training models created with an integrated structure have distinct properties. Due to the significant quantity of processing equipment and information necessary for training a model of this type, safeguards to prevent unapproved replication of the trained model turn a considerable priority. Specifically, weight differences between neurons of a neural network continue to significantly impact the efficacy of a DNN-based structure (Sze et al. 2017).

As an intervention for illegal duplication of a generated DNN framework, the incorporation of watermarking into a DNN-based framework has been examined in (Wolfgang et al. 1999; Potdar et al. 2005) in prior studies (Wolfgang et al. 1999; Potdar et al. 2005). Following the methodology, an array of watermark evidence has been incorporated into various weights chosen from the particular layer of the convolutional neural network (CNN). Utilizing an undisclosed matrix, the trajectory of consequences gets modified during every training procedure so that extrapolation with the matrix approaches the array of the watermark (Ogundokun et al. 2021; Abikoye and Ogundokun 2021). The effectiveness of the incorporating watermark relies on the choice of the matrix, and the investigations quantify the negative impact caused by the incorporating action. The action is executed concurrently with

developing models employing supervised datasets. A binary cross entropy governs the efficiency of CNN-based systems concerning the loss function.

Furthermore, to the initial loss function, the incorporated loss function is taken into account during training to ensure that the weights are modified based on the watermark and supervised dataset. In (Lemma et al. 2003; Cox et al. 2002), the approach described in (Wolfgang et al. 1999; Potdar et al. 2005) for selecting the attribute vector is improved. Nevertheless, an attack has been identified in (Stefan and Fabien 2000). Following the incorporation of a watermark, the discrepancy of weight values increases, revealing details regarding a concealed communication in a particular CNN layer. Replacing additionally shows that the watermark is readily adjustable. The second intriguing strategy relies on the process of transparent watermarking described in (Brüggemann et al. 1995). Comparable to instances of opposition, the watermarked photo is categorized with a distinct tag due to the noticeable embedded watermark. The drawback of the approach is that it is vulnerable to similar watermark elimination assaults that have been studied for centuries.

The growing popularity of neural networks in daily existence reveals that proprietary information preservation is a significant problem and that watermarking is an efficient way of proving copyright. Watermarking NNs approaches are broadly categorized into weight-parameter-based strategies and trigger-set-based methods. The earliest category of NN watermarking techniques incorporates watermarks by minimally altering weight variables, and the incorporated watermark is extracted using a matrix production between the watermarked variables and key-derived matrices. As a result, the watermarked NN must appear as a white box during the watermark retrieval step. In reality, the individual who owns it may safeguard the NNs integrity by training it with a loss function that includes an additional watermark-extraction item. The insertion placements and shapes of distinct watermarking schemes vary owing to the vast number of weight variables used by the neural network.

Uchida et al. (2017) used a binary cross entropy regularization loss function to include the watermark bit into the parameters of convolutional layers. When obtained, the vector containing the watermark content is produced by multiplying the mean value of the respective variables by a pre-generated core matrix. Such watermarking necessitates getting specific details within the NN, which includes weight parameters and layer results, to probe the framework for provenance verification promptly. Watermark bits frequently incorporate identification information and could be found in the variability of parameters relating to weight (Wang and Kerschbaum 2019) or the output of activation layers (Darvish Rouhani et al. 2019).

To train the NN, the second type of watermarking technique needs to create a trigger set, which includes a couple of instances and their given labels. Once a trigger item is fed into the NN after training, the generated outcome must be the same as the associated ascribed label, and the trigger set is only known to the owner. If the owner wishes to confirm his ownership, he may provide the trigger set to an authorized third party who will validate the matching link between the trigger samples and the labels. Any underlying mechanism or weight parameter of the network is not exposed in this case, implying that the watermarked NN might be a black box

during the ownership authentication step. Given the benefits of trigger-set-based watermarking, the researchers have developed several strategies for constructing trigger sets. Adi et al. (2018) collected 100 abstract photos and labeled them with nothing to do with the images' substance. So, this is their trigger set, and they trained the classification NN using both the trigger set and the training set simultaneously. The trigger set was built using adversarial samples by Le Merrer et al. (2020). These opposing examples were created by introducing disruption to existing photos and could not be accurately identified by regular networks without trigger-set training. Hence, they are also ideal for trigger samples. Additionally, several trigger set construction strategies are given in (Zhang et al. 2018; Guo and Potkonjak 2018), comprising introducing unique patterns or signatures, combining Gaussian noise, and overlaying an exceptional picture linked to watermark information on the original image.

Although trigger-set-based approaches are more convenient than weight parameter-based approaches due to less inner information concerning the NN on the ownership authentication phase, an attacker with access to the NN may construct a phony trigger set to claim his ownership. The neural network thus contains two watermarks, and no one can tell the difference between the legal and unlawful watermarks. This means that the neural network's ownership will be muddled. The forging attack continues to be an issue with trigger-set-based watermarking solutions.

Neural networks (NNs) have become known as potent instruments in a variety of disciplines, such as recognition of images, speech processing (SP), and natural language processing (NLP), among others. In the past few decades, the continued existence of neural networks (NNs) has become more vital due to their extensive spectrum of applicability. Watermarking is one of the most intriguing approaches that can improve the privacy and copyright preservation of NNs.

Watermarking is incorporating an indication or signature within an electronic object, for example, an image, video, or audio file, to preserve the trademark and intellectual property rights (Guo and Potkonjak 2018; Bansal et al. 2022). In NNs, watermarking incorporates an identifier or signature within its structure to avoid illegal utilization or alteration. The watermark may contain an identification number, a digitally signed document, or any other concealed knowledge that may be utilized to ascertain the system's trademark and authenticity (Abikoye and Ogundokun 2021).

The requirement for neural network watermarking (NNW) has become more prevalent in the past few years as safety and confidentiality of data concerns have grown (Bansal et al. 2022; Rashidi 2010; Ono et al. 2014). Data intoxication, simulation tilting, and malicious hacking may jeopardize the privacy and authenticity of NNs. Watermarking approaches remedy these problems by facilitating the identification and avoidance of illegal structure alterations (Rashidi 2010; Guru and Damecha 2014).

The objective of this study is to explore and analyze the applications of watermarking techniques, the attributes associated with watermarking techniques, the attributes related to watermarking, the neural network algorithms utilized in the

field of watermarking, and the challenges as well as future directions about embedding watermarks in neural networks. By addressing these research questions, the study aims to provide a comprehensive understanding of the current state of watermarking techniques, their attributes, the neural network algorithms employed, and the potential challenges and prospects associated with watermarking in the content of neural networks. The research findings will contribute to the existing knowledge and their integration with neural networks:

To achieve the study's objectives, the following research questions are formulated.

1. What are the applications of watermarking techniques?
2. What are the attributes of watermarking?
3. What are the NN algorithms that have been used in the domain of watermarking?
4. What are the various challenges and future directions of embedding watermarks in NNs?

In this article, we investigated various watermarking approaches that can be utilized to improve the security and ownership protection of NNs. We will address the pros and cons of different watermarking techniques and evaluate their performance in various situations. In addition, we will discuss the difficulties and drawbacks associated with NNW and provide prospective study initiatives to surmount them. This article intends to offer a detailed description of the present situation of NNW and how it might be used in various disciplines. The main contributions of this study are presented as follows:

1. The review identifies and categorizes the diverse applications of watermarking techniques across various domains, providing a comprehensive overview of the practical uses of watermarking.
2. The review examines the attributes associated with watermarking, such as robustness, imperceptibility, capacity, and security. It provides insights into the key characteristics of watermarking techniques and their impact on watermark embedding and detection effectiveness.
3. The review investigates the neural network algorithms that have been employed in the domain of watermarking. It analyses the strength, limitations, and performance of different neural network models used for watermarking applications.
4. The review highlights the challenges encountered in embedding watermarks in NNs. It explores attack vulnerability, maintaining watermark integrity during training, and the trade-off between robustness and imperceptibility. By identifying these challenges, the review helps researchers understand the current limitations and potential areas for improvement.
5. Based on the analysis of existing literature, the review presents future directions and recommendations for embedding watermarks in NNs. It identifies potential research areas, novel techniques, and methodologies that can enhance the effectiveness and practicality of watermarking in the context of neural networks.
6. Through the review process, the study identifies gaps in the literature, areas where further research is needed, and potential research opportunities for advancing watermarking techniques in the context of neural networks. This information can guide future research efforts and stimulate further exploration.

The remaining chapter is structured as follows: Section 1.2 discusses existing studies relating to the present systematic review we are considering. Section 1.3 discusses the PRISMA methodology employed for the search of relevant publications analyzed in this study. Section 1.4 presents the findings discovered from the publications analyzed. The investigation was concluded in Sect. 1.5.

1.2 Related Works

Divers' approaches have been employed in protecting neural networks using Watermarking techniques. This section describes some of these techniques with their summary detailed in Table 1.1. Digital watermarking technology was used by Uchida et al. (2017) to safeguard intellectual property and identify infringement in trained models. The genius of their work was the creation of a framework that could contain a watermark. At the same time, a deep neural network is being trained from the beginning and during fine-tuning and distillation without impairing its performance. Mun, Nam, Jang, Kim, and Lee (Wang and Kerschbaum 2019) introduced a powerful blind watermarking method that requires less learning time than earlier strategies and is adaptable to various attacks. The transform was first designed in traditional watermarking research, followed by seizures and performance measurements. The suggested approach, however, optimized the watermarking domain through attack functions.

Modern deep neural network watermarking techniques' resilience and dependability were the main focus of Hitaj and Mancini's (2018) study by embedding a mechanism into the model so that the model owner could determine who owns the model externally. Namba and Sakuma (2019) developed a novel watermarking method that is tolerant of both model modification and query modification and prevents some malicious users from redistributing a neural network model. Li et al. (2019a) introduced a blind-watermark-based intellectual property protection (IPP) framework for watermarking deep neural networks that satisfy the security and practicality requirements. BlackMarks' performance on ImageNet datasets was further examined by Chen et al. (2019), who confirmed its effectiveness and robustness. A new ownership watermark system for DNN models was created by Li et al. (2019b) in 2019. It can achieve the crucial attribute of piracy resistance lacking in all previous watermark designs. Null embedding, a novel training technique that strongly correlates with a given watermark's classification accuracy when a model is still being trained, is the foundation of our watermark design. The classification space is constrained by null embeddings, which cannot be changed or added without impairing standard classification.

Deep convolutional neural networks were employed by Zhong et al. (2020) to construct an automated and reliable picture watermarking system. The generalization of image watermarking by deep neural networks is necessary for the proposed blind image watermarking approach. Three DNN-applicable watermark-generating

Table 1.1 Summary of related works

Authors/ year	Objective	Contribution	Limitation
Uchida et al. (2017)	Use of trained models for intellectual property protection and infringement detection	The creation of a framework that can integrate a watermark throughout the fine-tuning and distillation phases of a deep neural network's training without affecting the network's performance	The number of embedded bits was lower than the parameters since a single-layer perceptron was used
Mun et al. (2017)	Intellectual property protection and infringement detection	The technique was adaptive to the attacks and required only 1 day of learning time	Absence of an in-depth examination of the techniques' constraints or voids
Hitaj and Mancini (2018)	Verification of the reliability and robustness of cutting-edge deep neural network watermarking techniques	They showed that a malicious adversary could elude the legitimate owners' authentication even when the watermark is challenging to erase, preventing the discovery of model theft	We cannot tell in real-time whether a neural network querying instance is a valid or potential backdoor
Namba and Sakuma (2019)	Preventing neural network model redistribution and query modification	The model has a system built into it that allows the model's owner to confirm their ownership externally	It is not a real-time system, and updating the model can be difficult
Li et al. (2019a)	Defending against evasion attacks and directly dealing with false ownership claims made by rivals	They developed a system that can effectively verify ownership without having a material adverse impact on the primary duties and achieves an impressive level of undetectability against evasion attacks and unforgeability against false claims of privilege	Occurrence of over-fitting during the training of the NN
Chen et al. (2019)	Defending models from theft of intellectual property (IP)	The first end-to-end multi-bit watermarking framework that may be used in a black box situation (BlackMarks) was created	The system cannot function in a multi-user setting
Szyller et al. (2021)	Deterring model extraction IP theft	DAWN (dynamic adversarial watermarking of neural networks), the initial watermarking technique, was created	The user can try to avoid ownership demonstration for another user by ensuring that the watermark verification fails, which is easily accomplished

(continued)

Table 1.1 (continued)

Authors/ year	Objective	Contribution	Limitation
Li et al. (2019b)	Building piracy-resistant watermarks into DNNs	Proposed watermarks successfully prevent piracy and achieve other watermark properties across various tasks and models	Since the model owner must select the watermark before training a model and any changes to the watermark necessitate retraining the model from scratch, the developed watermark technique is not time efficient. Additionally, watermarks cannot be directly added to already trained models; instead, the owner must completely retrain the model. Additionally, the most significant models cannot be used to test the watermark
Zhong et al. (2020)	Algorithms for image watermarking that are more generic	It does not require prior knowledge of potential distortions on the marked image to achieve its robustness attribute	Performances for individual common attacks were not evaluated
Zhang et al. (2018)	Using a watermark for verification of model ownership	Put the watermark into the network's produced images using a sub-network	The proposed idea will not be able to identify a leaked model if it is used as an internal service rather than being provided as an online service. As a result, the copycat cannot make money off the stolen models directly
Darvish et al. (2019)	Defending the models from theft of intellectual property (IP)	Before sharing, developers can methodically add digital watermarks to the intended DL model	Has a significantly low overall runtime and resource utilization
Wang et al. (2020)	The study involved safeguarding the ownership of the DL models that will be published	Introduced a distinct neural network that enables the usage of particular weights for watermarking	An attacker may use a new independent neural network to embed their watermark, which will arouse ambiguity regarding the watermark authentication
Guo and Potkonjak (2018)	Enabling DNN watermarking on embedded devices for the protection of intellectual property	This investigation includes the author's signature during the DNN training process	Limited to only authors' signatures for the authentication process

(continued)

Table 1.1 (continued)

Authors/ year	Objective	Contribution	Limitation
Yang et al. (2021)	Introduces a novel DNN framework that maximizes the resiliency of the embedded watermarks	They used the fault attack concept to incorporate a watermark for IP protection into a DNN model	The strategy's particular drawbacks or void are not disclosed

techniques were examined by Zhang et al. (2018). They also presented a watermark implanting method to embed a watermark within the models to establish their ownership. They also created a remote verification system to establish model ownership. WM embedding in neural network activations is made possible by an effective end-to-end paradigm proposed by Darvish et al. (2019). Wang et al. (2020) created a watermarking method for DNNs by including a separate neural network. Error back-propagation is used to implant the watermark into the host DNN. As a result of the experimental findings, it is clear that the suggested work can offer greater levels of fidelity, robustness, and capacity than earlier watermarking methods. Yang et al. (2021) devised a robust watermarking approach to maximize the embedded watermarks' ability to safeguard intellectual property.

1.3 Method

We performed a systematic literature review (SLR). According to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) methodology (Moher et al. 2009a), the SLR was carried out (Liberati et al. 2009; Moher et al. 2009b). PRISMA offers a four-phase structure diagram and a 27-item protocol for preparing and submitting SLRs and meta-analyses. PRISMA is acknowledged and utilized in a vast array of educational settings.

1.3.1 Data Sources and Search Approaches

We utilized the structure of a systematic literature review (SLR) proposed by Kitchenham et al., along with Liberati et al.'s reporting guidelines (Liberati et al. 2009) for systematic reviews and meta-analyses. Our paper adheres to the PRISMA guidelines for reporting. The study's review protocol has been previously published by Kitchenham (2004) and Kitchenham and Brereton (2013). To search, we applied a predefined set of search terms such as "Security," "Ownership," "Protection," "Neural Network," and "Watermarking" to the primary scientific repositories employed for this study.

1.3.2 Inclusion and Exclusion Indicators

Scopus and Web of Science (WoS) were the two primary scientific repositories for this SLR. The two libraries mentioned above are internationally recognized and contending city libraries (Zhu and Liu 2020). The search used the following terms: Watermarking, Security, Ownership Protection, and Neural Network. Following a preliminary examination, both sources' outcomes were confined to English-language research papers from 2013 to 2022 (10 years). In the assessment of this research, solely published scientific journals and conference papers were considered; publications in production were omitted. Non-English articles, other writings, and administration study reports were not examined. On April 1, 2023, an exploration was carried out. Table 1.2 shows the number of articles retrieved from the initial search of the databases with their search strings, while Table 1.3 shows the databases with their query links.

1.3.3 Process Exploration and Research Selection

To identify relevant studies, we conducted a thorough search process. We utilized specific search terms and applied them to relevant digital databases. The search was completed within a particular timeframe, taking into account the inclusion and exclusion criteria outlined for our study. The initial search yielded 513 records, supplemented with keywords such as ownership protection and watermarking technique. By reviewing the references of the studies, we eliminated four duplicate entries from the total of 513 records. Based on examining their titles, abstracts, keywords, and conclusions and our inclusion and exclusion criteria, we excluded 457 records. Subsequently, we evaluated the remaining 52 articles by examining their full text. Of these, eight studies were not aligned with the research question and were consequently excluded. Ultimately, we thoroughly reviewed and qualitatively analyzed the full texts of the remaining 39 studies included in the survey after following the four PRISMA processes, as shown in Fig. 1.1.

Table 1.2 Number of articles retrieved per database

Database	No of articles	Search String
Scopus	498	TITLE-ABS-KEY (Watermarking OR Security OR Ownership Protection AND Neural Network) AND (LIMIT-TO (PUBYEAR,2022) OR LIMIT-TO (PUBYEAR,2021) OR LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015) OR LIMIT-TO (PUBYEAR, 2013))
Web of Science (WoS)	15	ALL = (Watermarking AND Security AND Ownership Protection AND Neural Network)

Table 1.3 Query link per database

Database	Query link
Scopus	https://www.scopus.com/results/results.uri?sort=plf-f&src=s&nlo=&nlr=&nls=&sid=15de79917172740c88227f0e221123e4&sot=a&sdt=cl&cluster=scopubyr%2c%222022%22%2ct%2c%222021%22%2ct%2c%222020%22%2ct%2c%222019%22%2ct%2c%222018%22%2ct%2c%222017%22%2ct%2c%222016%22%2ct%2c%222015%22%2ct%2c%222014%22%2ct%2c%222013%22%2ct&sl=69&s=Watermarking+AND+Security+AND+Ownership+Protection+AND+Neural+Network&origin=resultslist&zone=leftSideBar&editSaveSearch=&txGid=bbdf6afe4f17389fdaf7f1db661d3e5e
WoS	https://www.webofscience.com/wos/woscc/summary/6ae5073d-24a4-40a2-bcfb-1ddc71fdfb7e-88ffae55/relevance/1

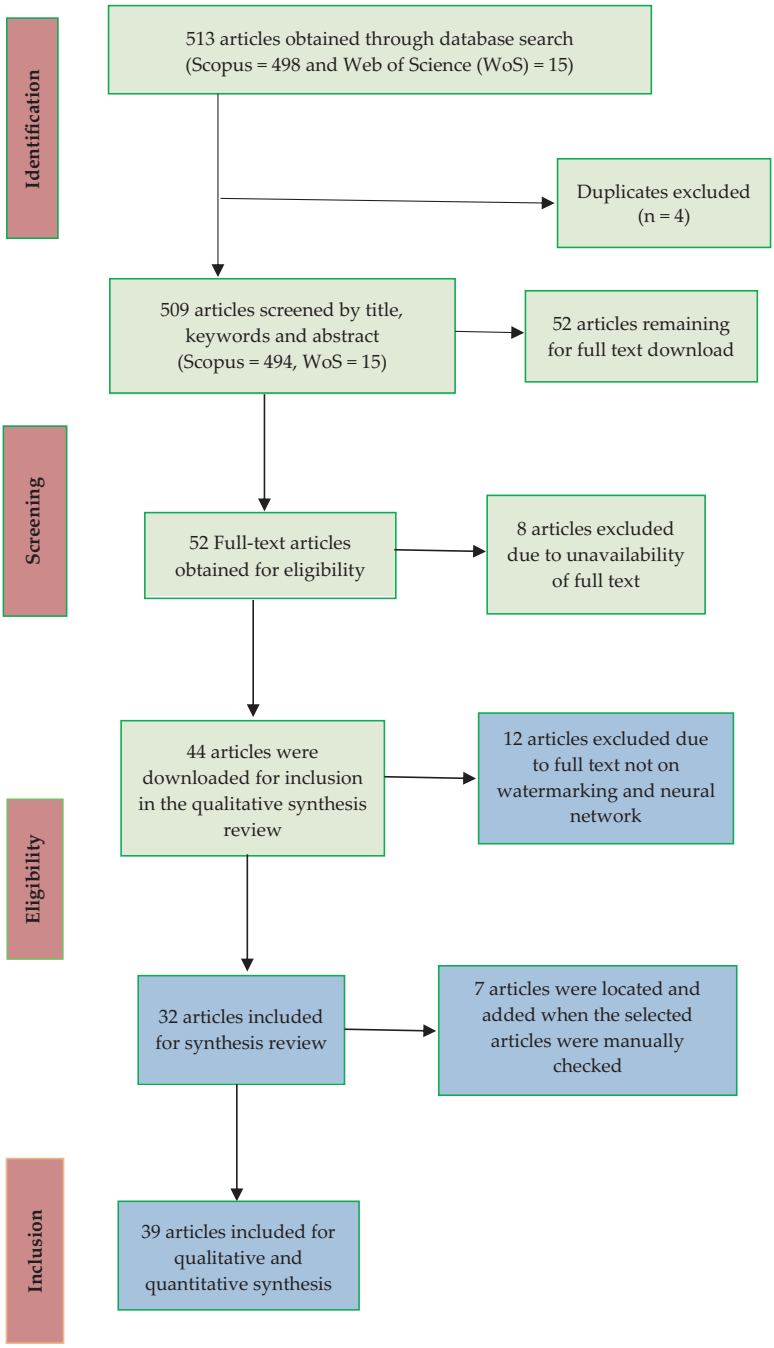


Fig. 1.1 PRISMA flow diagram

1.3.4 Data Gathering Procedure

We gathered specific information from each article, including the author's name, publication year, source (journal, conference, or preprint papers), and reference. Additionally, we recorded details such as the main focus, keywords, abstract, request questions, and study results. We also categorized whether the study was of a technical or non-technical nature and noted any challenges, security implementation, neural network techniques, and watermark techniques discussed. The authors of this study independently read all 39 articles, and in cases where there were different options, ROO made the final decision. Figure 1.2 shows the number of articles published per publication source, and it can be seen that 51% (20) of articles were conference, journals were 31% (12 articles), and the remaining 18% (7) were preprint papers. Figure 1.3 shows the number of articles published per year, and it was

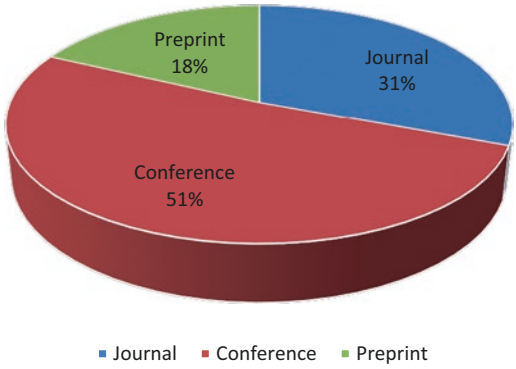


Fig. 1.2 Articles published per publication source

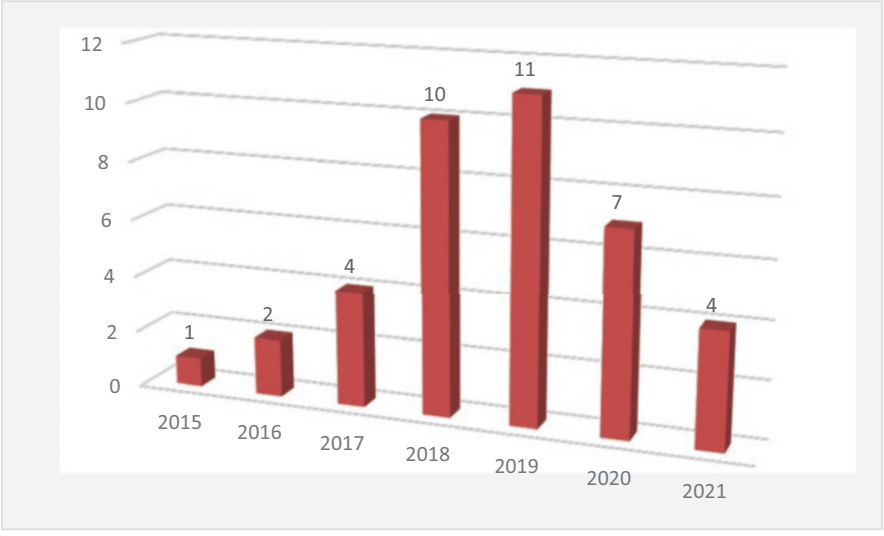


Fig. 1.3 Articles published per year

illustrated that most of the articles were published in 2019, with 11 papers, followed by the year 2018, with nine reports. The next was in 2020 with seven papers; in the years 2021 and 2017, 4 papers were published, 2 in 2016, and finally, 1 article was published in 2015.

1.4 Result

In this section, each formulated research question is attended to from the analysis of the included qualitative articles. They are into four subsections.

1.4.1 What Are the Applications of Watermarking Techniques?

1.4.1.1 Content Verification

Technological advancements have made modifying electronic multimedia material simpler while rendering it nearly impossible to ascertain the genuineness associated with such content. Content authentication is the procedure that verifies the authenticity of watermarked material to ensure that it has not been altered.

Digital rights administration frameworks, monitoring and telemetry applications, electronic claims tracking verification, and reputable recording devices employ watermarking for verification. Watermarks are utilized in security camera footage to ensure that all single media entries come from permissible resources. In such applications, a watermark describing the content is occasionally employed. The document's synopsis must be distinctive and challenging for an adversary to acquire (Zhou 2021; Rizzo et al. 2016).

1.4.1.2 Concealment Fortification

Computerized medical services permit individual medical data collection and digital storage. Most such documents contain personally identifiable information, which must be rigorously safeguarded. Personal data associated with healthcare photographs is typically maintained in a distinct record that can be retrieved alongside the pictures. Nevertheless, this type of data is not consistently safeguarded so attackers may discover it at any moment. The concerns of protecting confidentiality and accessibility management were sparked by a recent occurrence involving obtaining information about patients for employment vetting in Toronto's University healthcare system. Electronic watermarking is an effective way to encrypt documents with highly confidential data. The benefit of watermarking is that the content remains with the initial record, thus minimizing the likelihood of unintentional destruction or alteration. In addition, because

watermarking provides just slight discrepancies, it will not create a problem for clinicians who lack permission to view the patient's records (Patel and Bhatt 2015; Alqassab and Alanezi 2021).

1.4.1.3 Proprietorship Proclamation and Identification

Information on the internet could be protected from unlawful replication and dissemination using watermarking. In a controversy, scholars may employ watermarks to demonstrate the proprietorship of the content they have developed. The possession can be determined by retrieving the incorporated content from the watermarked file. Robustness is crucial for such a program considering the integrated watermark should not be readily eradicated by minor communication flaws.

1.4.1.4 Transmission Monitoring

Typically carried out through computerized tracking locations, it constitutes one of the watermarked applications that has been successfully commercialized. To find or recognize preserved or transmitted information, programs that involve information surveillance and tracing demand additional robustness. Instances include the computerized tracking of radio broadcasts for financial gain and the use of internet crawlers to identify photographs on the internet. To identify such programs, watermarks need to be readily extracted and relatively impervious to standard techniques for processing data, including transforming formats and compressing. Observing broadcasts enables the details incorporated into multimedia broadcasts to be employed for various purposes. The data can be utilized to confirm how much appropriate airtime distribution has taken place, design a computer-based royalty-gathering mechanism for intellectual property transmitted by streaming administrators, or compile data about the number of individuals tuned to a particular transmission (Ruiz et al. 2019; Hongbin et al. 2017).

1.4.1.5 Transaction Tracking

This is composed of concealed and undetectable details concerning the individual using it that a watermark device can discover; as a result, it is simpler to trace unlawful reproductions of electronically stored content once they are dispersed. In addition, registered versions that are associated with a specific person are identified, settling any potential proprietorship disputes regarding a particular piece of proprietary information. The recognition of cinema theaters where an illicit recording of a film was made with a portable camera happens to fall within the scope of this kind of application (Frattolillo 2020; Huang et al. 2017).

1.4.1.6 Fingerprinting

Fingerprinting's primary function is to safeguard customers. Fingerprinting may hinder the illicit distribution of an item that was obtained lawfully. This could be accomplished by incorporating a singular resilient watermark for every recipient throughout the interaction.

Medical applications: Watermarking is crucial in the healthcare sector to safeguard healthcare data from unidentified individuals, exceptionally patient reports. Confidentiality and information authentication are becoming more critical in the healthcare sector because electronic data is readily disseminated online.

1.4.2 What Are the Attributes of Watermarking?

The fundamental criteria for electronic watermarking may be considered traits or features. A variety of applications need distinct watermarking features. The various watermarking properties have varied roles in designing applications. The following are the fundamental characteristics of watermarking:

1.4.2.1 Fidelity

Current watermarking approaches try to preserve how well the network performed on the initial assignment to ensure fidelity. Based on the system, this may be accomplished in various ways, such as by slightly modifying the initial choice of boundaries (Le Merrer et al. 2020) or adding the watermark to earlier convergent weights for the models (Wang et al. 2020).

1.4.2.2 Robustness

The system executes two jobs if the trigger dataset has a considerably different allocation than the initial data (Adi et al. 2018; Fan et al. 2019). As a result, they may be extracted individually or removed without affecting the forecasting ability of the framework. For optimal resilience, watermarking techniques must execute steps to establish a link between both functions and bind the watermark to the system so that it cannot be readily withdrawn.

1.4.2.3 Dependability

Specific situations may affect the reliability of a watermark. Firstly, comparable to resilience, if the trigger dataset is derived from an alternate distribution than the initial dataset, an intruder could mask responses of the hijacked algorithm to watermark triggers. Second, any approaches that depend entirely on white-box confirmation may be

less reliable since entry to all possibly stolen simulation cases could occasionally not be a reasonable assumption, possibly hindering adequate assurance.

1.4.2.4 Integrity

Measuring the authenticity of watermarking techniques is a complex problem since it entails determining how (possibly all other) non-watermarked algorithms behave on the supplied trigger dataset. A suitable trigger dataset is distinguished by the distinctiveness of the watermarked algorithm's forecasts on it, such that no legitimate individuals with comparable algorithms are accused of stealing.

1.4.2.5 Capacity

The number of details that may be contained in the watermark is expressed as capability. Applying multi-bit watermarks to enable specialized activities, including proprietor verification, is widespread (Zhang et al. 2018; Guo and Potkonjak 2018).

1.4.2.6 Secrecy

For instance, watermarking approaches that modify the model's settings in a discernible manner (Uchida et al. 2017) breach the confidentiality criteria. Watermark recognition may be avoided by including them in the dynamically generated contents (Rouhani et al. 2018b) or by implementing efforts to compel the parameters associated with the algorithms to remain nearly the same (Li et al. 2019a).

1.4.2.7 Efficiency

This may be measured in terms of imprinting and confirmation duration, which includes the cost involved in training and the calculation time or the variety of requests required to validate the watermark. Most previous research fails to analyze the computing expense of their methodologies clearly. Jia et al. (2021) performed one of the rare effectiveness studies. They concluded that the trigger dataset should have over 50% of the number of data elements as the initial data source for their technique. As a result, the algorithm must be trained using 150–200% of the initial data. This could end up in significant additional costs, particularly for big datasets.

1.4.2.8 Generality

Some current approaches are immediately generalizable across all datasets; for instance, Chen et al. (2019) require a separate watermark embedding approach for each dataset. An absence of generalization characterizes such conduct.

1.4.3 What Are the NN Algorithms That Have Been Used in the Domain of Watermarking?

Uchida et al. (2017) initially suggest that deep neural networks (DNNs) be marked with a variable in the regularizer that allows the watermark to be inserted within the model-building process. Wang et al. (2020) extend Uchida et al.'s approach by including a self-sustaining NN to project weights of networks to the watermark region. It cannot, nevertheless, evade the unclear nature of the threat. Fan et al. (2019) present innovative passport-founded DNN possession authentication systems that resist network alterations and vagueness attacks. Rouhani et al. (2018a) present DeepSigns, an entire intellectual property safeguarding platform that allows designers to methodically embed electronic watermarks in the relevant DL architecture before releasing it. Zhang et al. (2018) put forward a watermark integrating technique to incorporate watermarks into DL algorithms, as well as an automated confirmation procedure for verifying possession, allowing DL algorithms to obtain customized watermarks during training and stimulate with predetermined forecasts whenever watermark variations are observed during inductive reasoning.

Adi et al. (2018) provide a backdoor-centered watermarking approach that incorporates the contributor's fingerprint during the training of DNNs. When supplied with any signature materials, the resultant watermarked DNN acts in a distinct, predetermined fashion, thereby demonstrating copyright. Shafieinejad et al. (2021) additionally provide a backdoor strategy for watermarking DNNs, which is adequate for general classification requirements and is capable of being paired with existing algorithmic methods for learning, indicating that incorporated watermarks are somewhat vulnerable to elimination attempts, including fine-tuning and pruning. Further comparable works may be found at (Namba and Sakuma 2019; Li et al. 2019a; Chen et al. 2019). There is no question that the aforementioned watermarking methods have accelerated the IP preservation of DL approaches. In real-life situations, nevertheless, a corporation or individual could steal a DL approach and afterward give the API functionality to consumers for revenue without authorization. In this situation, the proprietor cannot determine authorship unless the watermark had earlier been incorporated into the DL framework by deliberately altering the network parameters, given that they lack visibility to the network parameters and architecture. Conventional watermarking technologies or antagonistic networks (Goodfellow et al. 2020) could possibly be used to analyze the output for modeling preservation (Zhang et al. 2020).

Nevertheless, since the procedure is not dependent on proprietary information, it cannot safeguard the architecture if the internet connection is compromised. Although trigger-set categorization indicators may help, the watermarking capability cannot deliver suitable content. Furthermore, the techniques are frequently confined to classification tasks and cannot be immediately employed in typical dynamic applications like visual alteration. We offer a unique watermarking methodology to safeguard DNNs in this paper, which is suitable for systems that output pictures as outcomes, which includes image colorization and modification. All images

produced by a DNN-generated authors' approach must have an identifiable watermark. To retrieve the watermark, they create a fingerprint-extraction structure with a loss dependent on its accuracy. The secured DNN (named the host network) and the watermark-extraction structure receive training in tandem. The server network's settings will be modified throughout the training process based on its own loss and the watermark loss. At the same time, just the watermark loss is being utilized to alter the signature-extraction system's configuration variables. Only the network that serves as the host will be revealed during training, while the watermark-extraction system will remain hidden. Investigations have proven that this approach is prosperous and stable on various picture assignments, such as colorization, super-resolution, photomanipulation, and semantic segmentation.

Yuan et al. (2019) offers training for the algorithm to be safeguarded using a collection of specially created inputs to allow the system to attribute an identifiable category to those inputs. In essence, their technique is comparable to trojaning (Adi et al. 2018): activating a particular designation in response to constructed information serves as evidence of possession for the researchers. Similarly, Adi et al. (2018) openly employ NN trojaning (i.e., backdooring) for an identical goal. Conversely, Rouhani et al. (2018a) uses unique loss algorithms to integrate the watermark into the layer parameters explicitly. Nevertheless, unlike (Uchida et al. 2017), whereby the parameters explicitly comprise the watermark (which consequently becomes inaccessible in a black box scenario), parameters in (Rouhani et al. 2018a) are adjusted at execution to achieve intended stimulation provided specific inputs are made accessible.

Interestingly, the researchers demonstrate that watermarking algorithms developed in a black box setting may be applied in a white box one. Guo and Potkonjak (2018) suggest adopting a black box competent watermark retrieval to integrated gadgets in an increasingly limited arrangement. Tramer et al. (2016) show how to get an undetectable clone of a centrally operated strategy from several internet machine learning APIs. Scheduling attacks on side channels can additionally be used to infringe on the computational capacity of neural networks (Duddu et al. 2018). Papernot et al. (2017) demonstrated that online simulation assaults are viable, resulting in incorrect outputs. The investigators of Sethi and Kantardzic (2018) use the reverse development of online classification algorithms to focus on assaults that involve avoiding a CAPTCHA examination.

Further attacks target model hyperparameters stolen from APIs (Wang and Gong 2018; Oh et al. 2019); Oh et al. (2019) attempt to deduce core hyperparameter values (e.g., number of layers, type of non-linear activation) of a remote neural network architecture by analyzing its characteristic reactions to specific inputs. Lastly, techniques in Uchida et al. (2017) offer to identify simulation manipulation by simple API requests; evaluated assaults include trojaning, contraction, fine-tuning, and the watermarking approach suggested in this research. In this paper, we offer an API-compliant watermarking technique because it only uses the basic classification request to communicate with the online application. Table 1.4 shows the summary of NN algorithms that have been used in the domain of watermarking.

Table 1.4 Summary of neural networks algorithms that have been used in the domain of watermarking

Authors	Models	Contributions	Limitations
Uchida et al. (2017)	DNN	The investigation suggested marking DNNs with a variable in the regularizer for inserting watermarks during the model-building process	The investigation does not address the unclear nature of threats
Wang et al. (2020)	Self-sustaining NN	The study extended Uchida et al.'s approach by including a self-sustaining NN to project weights of networks to the watermark region	It does not entirely evade the unclear nature of threats
Fan et al. (2019)	Passport-founded DNN possession authentication systems	Introduced innovative systems for DNN possession authentication that are reasonably resistant to network alterations and vagueness attacks	The study only addressed passport-founded DNN possession authentication systems
Rouhani et al. (2018a)	DeepSigns	The author presented DeepSigns, an intellectual property safeguarding platform that allows the systematic embedding of electronic watermarks in DL architecture	The study only considered the embedding of electronic watermarks in the DL network
Zhang et al. (2018)	DL algorithms	The authors proposed a watermark integrating technique for DL algorithms and an automated confirmation procedure for verifying possession	It does not address safeguarding the architecture if the internet connection is compromised
Adi et al. (2018)	DNN	The research developed a backdoor-centered watermarking approach incorporating the contributor's fingerprint during DNN training	The study does not discuss limitations

Table 1.4 (continued)

Authors	Models	Contributions	Limitations
Shafieinejad et al. (2021)	DNN	The study provided a backdoor strategy for watermarking DNNs adequate for general classification requirements, susceptible to elimination attempts	Incorporated watermarks are somewhat vulnerable to elimination attempts, including fine-tuning and pruning
Other works (Namba and Sakuma 2019; Li et al. 2019a; Chen et al. 2019)	DL approaches	Mentioned additional comparable watermarking methods for DL approaches	Not specified
Goodfellow et al. (2020)	Conventional watermarking technologies or antagonistic networks	Possible use in analyzing the output for modeling preservation	Not able to safeguard the architecture if the internet connection is compromised
Zhang et al. (2020)	Output analysis for modeling preservation	Possible application of conventional watermarking technologies or antagonistic networks in analyzing the output for modeling preservation	Not dependent on proprietary information. Watermark capability is insufficient for delivering suitable content
Wu et al. (2020)	Image colorization, super-resolution, image editing, semantic segmentation	Developed a unique watermarking methodology for safeguarding DNNs, particularly in systems that output pictures like image colorization and modification	Not able to safeguard the architecture if the internet connection is compromised. Watermark capability may be insufficient for delivering suitable content. Techniques may be limited to classification tasks and not immediately applicable to dynamic applications such as visual alteration
Yuan et al. (2019)	Not specified	The study developed a technique similar to trojaning for watermarking, where activating a specific designation in response to constructed inputs serves as evidence of possession	The study does not include particular watermarking approaches or limitations

(continued)

Table 1.4 (continued)

Authors	Models	Contributions	Limitations
Adi et al. (2018)	NN trojaning (backdooring)	The study used NN trojaning (backdooring) for watermarking and possession attribution	The study employed limited studies of other watermarking technologies and their drawbacks
Rouhani et al. (2018a)	DeepSigns	The study used unique loss algorithms to explicitly integrate the watermark into the layer parameters, with adjustments made at execution to achieve the intended stimulation	Parameters in (Mun et al. 2017) are adjusted at execution, which may limit the watermark accessibility in a black box scenario
Guo and Potkonjak (2018)	Deep neural network	The study suggested adopting a black box competent watermark retrieval to integrated gadgets in an increasingly limited arrangement	Considered only employing the black box
Tramer et al. (2016)	Machine learning (ML)	The study demonstrated how to get an undetectable clone of a centrally operated strategy from several internet machine learning APIs	The study only employed machine learning in this study
Duddu et al. (2018)	Backdoor routes for NN theft	The study proposed a scheduling attack on side channels that infringe on the computational capacity of neural networks	The study emphasizes on temporal information, and the assessment of watermarking methods is restricted
Papernot et al. (2017)	Black box attacks against ML	The study demonstrated online simulation attacks resulting in incorrect simulation outputs	The study makes no explicit reference of watermarking methods or its constraints
Sethi and Kantardzic (2018)	The study is experimentally informed by data assaults against black box models	The study used reverse development of online classification algorithms to focus on attacks involving avoiding a CAPTCHA examination	The study does not access watermarking methods or their drawbacks in detail

(continued)

Table 1.4 (continued)

Authors	Models	Contributions	Limitations
Wang and Gong (2018); Oh et al. (2019)	ML hyperparameter theft	The authors proposed attacked model hyperparameters stolen from APIs, attempting to deduce core hyperparameter values of a distant NN architecture	The study is mostly concerned with stealing hyperparameters, with just a cursory examination of watermarking approaches
Le Merrer and Gilles (2019)	Reverse-engineering black box neural networks	The study offered techniques to identify simulation manipulation by simple API requests, including trojaning, contraction, fine-tuning, and watermarking	The study does not evaluate approaches to watermarking or its constraints explicitly
Le Merrer et al. (2020)	CNN, IRNN, MLP	The study offered an API-compliant watermarking technique that only uses basic classification requests to communicate with the online application	Algorithms that can resist alternative types of attacks were not considered

1.4.4 What Are the Various Challenges and Future Directions of Embedding Watermarks in NNs?

This section describes the current constraints and prospective future research paths based on assessing existing schemes and their security needs. The fundamental disadvantage of watermarking is its passive nature since it can only identify theft after it has happened rather than prevent it. Some research has looked at sending security alerts when an ML model is on the cusp of releasing enough information for possible attackers to extract its functionality (Kesarwani et al. 2018). Additional research has focused on constructing models that retain high accuracy only when authorized users perform queries (Chen and Wu 2018). Another area of research is developing more complicated models to steal, such as returning just hard labels without probability per output class, perturbing prediction outputs (Orekondy et al. 2019), or creating weight-sensitive networks that make it difficult for attackers to steal and modify them (Zentannai et al. 2019).

Future studies might focus on incorporating watermarking into active protection measures against model theft. Furthermore, most known watermarking systems have been predominantly applied to picture data, with just a few outliers, such as the one shown by Jai et al. (2021), indicating the application to various data types. It is vital to study the generality and universal application of current systems and, if required, modify or expand them. Furthermore, existing watermarking algorithms primarily address classification problems, with little research on watermarking in

other ML areas, such as reinforcement learning (Behzadan and Hsu 2019), data creation using GANs (Ong et al. 2021), or picture captioning (Lim et al. 2020). As a result, establishing watermarking systems for diverse ML applications represents a fascinating future challenge.

Furthermore, current watermarking algorithms have primarily been deployed and assessed on small research datasets such as MNIST (LeCun et al. 2010). The scalability of these techniques remains unresolved since solutions requiring training with substantially more giant datasets may not be appropriate in every circumstance. As a result, future work should analyze the practical applicability of current watermarking systems to more enormous real-world datasets and check if their qualities exhibited on research datasets (training efficiency, suitable trigger dataset size, integrity, and so on) stay true.

Finally, once watermarking techniques have met all the technical criteria, adapting them to real-world operations provides a new issue. Legal and organizational procedures must be modified to permit the assertion of ownership rights based on the existence of watermarks.

1.5 Conclusion

Machine learning (ML) is currently utilized in various applications. With an increasing number of implemented designs, the use of watermarks to safeguard the intellectual property embedded in those frameworks has evolved into a significant concern in both industry and academia. This systematic literature review has shed light on the diverse applications, attributes, neural network algorithms, challenges, and future directions of watermarking techniques in embedding watermarks in NNs. By examining a wide range of research articles and synthesizing their findings, this review has made several significant contributions to the field.

First, it has provided a comprehensive understanding of the applications of watermarking techniques, demonstrating their relevance and effectiveness in various domains. Watermarking has proven to be a versatile and valuable tool, from digital media copyright protection to biometrics data authentication. Second, the review selected and assessed watermarking properties such as robustness, imperceptibility, capacity, and security. Understanding these characteristics is critical for developing watermarking approaches that fulfill particular needs and aims. Third, examining NN algorithms employed in watermarking highlighted the strengths and weaknesses of various models. This information may help researchers choose good algorithms for watermarking jobs and optimize their performance.

Furthermore, the paper underlined the difficulties of incorporating watermarks in NNs. These difficulties include being vulnerable to assaults, preserving the integrity of watermarks, and achieving the correct mix of robustness and imperceptibility. Researchers may focus on discovering novel answers to these challenges by recognizing them. Furthermore, the research gave valuable insights into future approaches and suggestions for NN watermarking. It has revealed study topics that deserve

more investigation, proposed new methods and procedures, and provided the potential to improve the efficacy and practicality of watermarking.

Overall, this thorough assessment of the literature on watermarking methods and their integration with neural networks has consolidated current information. It is a comprehensive resource for scholars, practitioners, and industry experts, laying the groundwork for future advances in the discipline. As the subject of watermarking evolves, and new issues arise, this overview lays the groundwork for future research efforts. By building upon the insights gained from this study, researchers can strive to develop more robust, secure, and efficient watermarking techniques that meet the growing demands of digital media protection and data authentication in neural networks.

This study's findings may provide a solid basis for examining current watermarking systems, inventing new ones, or selecting appropriate alternatives for particular cases. As a result, it may be utilized as a resource for academics and ML professionals across all fields.

References

Abikoye OC, Ogundokun RO (2021) Efficiency of LSB steganography on medical information. Int J Electr Comput Eng 11(5):4157–4164

Adegun AA, Viriri S, Ogundokun RO (2021) Deep learning approach for medical image analysis. Comput Intell Neurosci 2021:1–9

Adi Y, Baum C, Cisse M, Pinkas B, Keshet J (2018) Turning your weakness into a strength: watermarking deep neural networks by backdooring. In: 27th {USENIX} security symposium ({USENIX} security 18), pp 1615–1631

Alqassab A, Alanezi M (2021) Relational database watermarking techniques: a survey. J Phys 1818(1):012185

Andi HK (2021) Estimating the role of Blockchain, deep learning, and cryptography algorithms in cloud security. J Trends Comput Sci Smart Technol 3(4):305–313

Awotunde JB, Ogundokun RO, Ayo FE, Matiluko OE (2020) Speech segregation in background noise based on deep learning. IEEE Access 8:169568–169575

Bansal A, Chiang PY, Curry MJ, Jain R, Wigington C, Manjunatha V, Goldstein T et al (2022) Certified neural network watermarks with randomized smoothing. In: International conference on machine learning. PMLR, London, pp 1450–1465

Behzadan V, Hsu W (2019) Sequential triggers for watermarking of deep reinforcement learning policies. arXiv preprint arXiv:1906.01126v1

Brüggemann HH, Gerhardt-Häckl W, Caronni G (1995) Assuring ownership rights for digital images. In: Verläßliche IT-Systeme: proceedings der GI-Fachtagung VIS'95, pp 251–263

Chaumont M (2020) Deep learning in steganography and steganalysis. In: Digital media steganography. Academic Press, New York, pp 321–349

Chen M, Wu M (2018) Protect your deep neural networks from piracy. In: IEEE international workshop on information forensics and security (WIFS). IEEE, Piscataway, NJ, pp 1–7. https://doi.org/10.1109/WIFS.2018.8630791

Chen H, Rouhani BD, Koushanfar F (2019) Blackmarks: Blackbox multi-bit watermarking for deep neural networks. arXiv preprint arXiv:1904.00344

Cox I, Miller M, Bloom J, Honsinger C (2002) Digital watermarking. J Electr Imaging 11(3):414–414

Darvish Rouhani B, Chen H, Koushanfar F (2019) Deepsigns: an end-to-end watermarking framework for ownership protection of deep neural networks. In: Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems, pp 485–497

Duddu V, Samanta D, Rao DV, Balas VE (2018) Stealing neural networks via timing side channels. arXiv preprint arXiv:1812.11720

Fan L, Ng KW, Chan CS (2019) Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks. Adv Neural Inf Process Syst 32

Frattolillo F (2020) A watermarking protocol based on Blockchain. Appl Sci 10(21):7746

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144

Guo J, Potkonjak M (2018) Watermarking deep neural networks for embedded systems. In: In 2018 IEEE/ACM international conference on computer-aided design (ICCAD). IEEE, Piscataway, NJ, pp 1–8

Guru J, Damecha H (2014) A review of watermarking algorithms for digital images. Int J Innov Res Comput Commun Eng 2(9):5701–5708

Hitaj D, Mancini LV (2018) Have you stolen my model? Evasion attacks against deep neural network watermarking techniques. arXiv preprint arXiv:1809.00615

Hongbin Z, Bohan R, Keqing W, Xiaoxi X (2017) The application of multiple texts watermarking algorithm in the transmission of power information security under the internet. In: 2017 IEEE conference on energy internet and energy system integration (EI2). IEEE, Piscataway, NJ, pp 1–5

Huang JC, Jeng FG, Chen TH (2017) A new buyer-seller watermarking protocol without multiple watermark insertion. Multimed Tools Appl 76:9667–9679

Jia H, Choquette-Choo CA, Chandrasekaran V, Papernot N (2021) Entangle watermarks as a defense against model extraction. In: 30th {USENIX} security symposium {USENIX} security 21, p 18

Kalsi S, Kaur H, Chang V (2018) DNA cryptography and deep learning using genetic algorithm with NW algorithm for key generation. J Med Syst 42:1–12

Kesarwani M, Mukhoty B, Arya V, Mehta S (2018) Model extraction warning in Mlaas paradigm. In: Proceedings of the 34th annual computer security applications conference on—ACSAC '18. ACM, New York, pp 371–380. https://doi.org/10.1145/3274694.3274740

Kitchenham B (2004) Procedure for undertaking systematic reviews. Computer science department, Keele University (TRISE-0401) and National ICT Australia ltd (0400011T. 1), joint technical report, 33

Kitchenham B, Brereton P (2013) A systematic review of systematic review process research in software engineering. Inf Softw Technol 55(12):2049–2075

Le Merrer E, Gilles T (2019) Tampernn: efficient tampering detection of deployed neural nets. In: 2019 IEEE 30th international symposium on software reliability engineering (ISSRE). IEEE, Piscataway, NJ, pp 424–434

Le Merrer E, Perez P, Trédan G (2020) Adversarial frontier stitching for remote neural network watermarking. Neural Comput Applic 32:9233–9244

LeCun Y, Cortes C, Burges C (2010) MNIST Handwritten Digit Database. AT&T Labs, Florham Park, NJ

Lemma AN, Aprea J, Oomen W, van de Kerkhof L (2003) A temporal domain audio watermarking technique. IEEE Trans Signal Process 51(4):1088–1097

Li Z, Hu C, Zhang Y, Guo S (2019a) How to prove your model belongs to you: a blind-watermark-based framework to protect the intellectual property of DNN. In: Proceedings of the 35th annual computer security applications conference, pp 126–137

Li H, Wenger E, Shan S, Zhao BY, Zheng H (2019b) Piracy-resistant watermarks for deep neural networks. arXiv preprint arXiv:1910.01226

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Moher D et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 151(4):W-65

Lim JH, Chan CS, Ng KW, Fan L, Yang Q (2020) Protect, show, attend, and tell: image captioning model with ownership protection. arXiv preprint arXiv:2008.11009

Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group (2009a) Reprint—preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Phys Ther 89(9):873–880

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group (2009b) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151(4):264–269

Mun SM, Nam SH, Jang HU, Kim D, Lee HK (2017) A robust blind watermarking using a convolutional neural network. arXiv preprint arXiv:1704.03248

Namba R, Sakuma J (2019) Robust watermarking of neural network with exponential weighting. In: Proceedings of the 2019 ACM Asia conference on computer and communications security, pp 228–240

Ogundokun RO, Awotunde JB, Adeniyi EA, Ayo FE (2021) Crypto-Stegno based model for securing medical information on IOMT platform. Multimed Tools Appl 80:31705–31727

Oh SJ, Schiele B, Fritz M (2019) Towards reverse-engineering black-box neural networks. In: Explainable AI: interpreting, explaining and visualizing deep learning, pp 121–144

Ong DS, Chan CS, Ng KW, Fan L, Yang Q (2021) Protecting intellectual property of generative adversarial networks from ambiguity attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3630–3639

Ono S, Maehara T, Nakai K, Ikeda R, Taniguchi K (2014) Semi-fragile watermark design for detecting illegal two-dimensional barcodes by evolutionary multi-objective optimization. In: Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation, pp 175–176

Orekondy T, Schiele B, Fritz M (2019) Prediction poisoning: towards defenses against Dnn model stealing attacks. arXiv preprint arXiv:1906.10908v2

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp 506–519

Patel R, Bhatt P (2015) A review paper on digital watermarking and its techniques. Int J Comput Appl 110(1):10–13

Potdar VM, Han S, Chang E (2005) A survey of digital image watermarking techniques. In: INDIN'05. 2005 3rd IEEE international conference on industrial informatics, 2005. IEEE, Piscataway, NJ, pp 709–716

Rashidi H (2010) A novel watermarking scheme for detecting and recovering distortions in database tables. arXiv preprint arXiv:1009.0827

Rizzo SG, Bertini F, Montesi D (2016) Text authorship verification through watermarking. In: 2016 European intelligence and security informatics conference (EISIC). IEEE, Piscataway, NJ, pp 168–171

Rouhani BD, Chen H, Koushanfar F (2018a) Deepsigns: a generic watermarking framework for IP protection of deep learning models. arXiv preprint arXiv:1804.00750

Rouhani BD, Riazi MS, Koushanfar F (2018b) Deepsecure: scalable provably-secure deep learning. In: 2018 55th ACM/ESDA/IEEE design automation conference (DAC). IEEE, Piscataway, NJ, p 6. https://doi.org/10.1109/dac.2018.8465894

Ruiz P, Dill O, Raju G, Cossairt O, Walton M, Katsaggelos AK (2019) Visible transmission imaging of watermarks by suppression of occluding text or drawings. Digit Appl Archaeol Cult Herit 15:e00121

Sethi TS, Kantardzic M (2018) Data driven exploratory attacks on black box classifiers in adversarial domains. Neurocomputing 289:129–143

Shafieinejad M, Lukas N, Wang J, Li X, Kerschbaum F (2021) On the robustness of backdoor-based watermarking in deep neural networks. In: Proceedings of the 2021 ACM workshop on information hiding and multimedia security, pp 177–188

Stefan K, Fabien AP (2000) Information hiding techniques for steganography and digital watermarking. Artech House Books, Norwood, MA

Sze V, Chen YH, Yang TJ, Emer JS (2017) Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE 105(12):2295–2329

Szyller S, Atli BG, Marchal S, Asokan N (2021) Dawn: dynamic adversarial watermarking of neural networks. In: Proceedings of the 29th ACM international conference on multimedia, pp 4417–4425

Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T (2016) Stealing machine rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks. In: 25th USENIX security symposium (USENIX security 16), pp 601–618

Uchida Y, Nagai Y, Sakazawa S, Satoh SI (2017) Embedding watermarks into deep neural networks. In: Proceedings of the 2017 ACM on International Conference on multimedia retrieval, pp 269–277

Wang B, Gong NZ (2018) Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy (SP). IEEE, Piscataway, NJ, pp 36–52

Wang T, Kerschbaum F (2019) Attacks on digital watermarks for deep neural networks. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Piscataway, NJ, pp 2622–2626

Wang J, Wu H, Zhang X, Yao Y (2020) Watermarking in deep neural networks via error back-propagation. Electron Imaging 2020(4):22–21

Wolfgang RB, Podilchuk CI, Delp EJ (1999) Perceptual watermarks for digital images and video. Proc IEEE 87(7):1108–1126

Wu H, Liu G, Yao Y, Zhang X (2020) Watermarking neural networks with watermarked images. IEEE Trans Circuits Syst Video Technol 31(7):2591–2601

Yang P, Lao Y, Li P (2021) Robust watermarking for deep neural networks via bi-level optimization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14841–14850

Yuan X, He P, Zhu Q, Li X (2019) Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 30:2805–2824

Zentannai K, Al-Afandi J, Horváth A (2019) Mimosanet: an unrobust neural network preventing model stealing. arXiv preprint arXiv:1907.01650v1. In: Presented at CVPR workshop: Adversarial Machine Learning in Real-World Computer Vision Systems

Zhang J, Gu Z, Jang J, Wu H, Stoecklin MP, Huang H, Molloy I (2018) Protecting the intellectual property of deep neural networks with watermarking. In: Proceedings of the 2018 on Asia conference on computer and communications security, pp 159–172

Zhang, J., Chen, D., Liao, J., Fang, H., Zhang, W., Zhou, W., Yu, N et al. (2020) Model watermarking for image processing networks. Proceedings of the AAAI conference on artificial intelligence (34,7, 12805–12812)

Zhong X, Huang PC, Mastorakis S, Shih FY (2020) An automated and robust image watermarking scheme based on deep neural networks. IEEE Trans Multimed 23:1951–1961

Zhou Z (2021) Research on image verification code based on watermarking technology. In: 2021 2nd international seminar on artificial intelligence, networking and information technology (AINIT). IEEE, Piscataway, NJ, pp 80–83

Zhu J, Liu W (2020) A tale of two databases: the use of web of science and Scopus in academic papers. Scientometrics 123(1):321–335

Zou Y, Zhang G, Liu L (2019) Research on image steganography analysis based on deep learning. J Vis Commun Image Represent 60:266–275