IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Special Issue 1, March 2024

1st International Conference on Machine Learning, Optimization and Data Science

Organized by

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, India

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

Boosting Security in the Age of Artificial Intelligence: An In-depth Analysis of Sophisticated Watermarking Methods and Their Challenges

Saurabh Verma, Dr. Mukta Bhatele, Dr. Akhilesh A. Waoo

Research Scholar, AKS University Satna, India

Professor, AKS University Satna, India

Professor and Associate Dean, Department of CS/IT, AKS University, Satna, India

ABSTRACT: In the present era with the widespread use of artificial intelligence (AI) and digital content, the importance of watermarking techniques has increased. This survey "Survey on Watermarking Methods in the AI Domain and Beyond" provides a detailed review of the applications and developments of watermarking techniques in various fields. It highlights the latest technologies in watermarking in digital media, software, and especially AI models. The survey also highlights how these technologies are helpful in protecting intellectual property, data integrity and authentication. Additionally, it discusses the challenges and future directions of watermarking, including the requirements of robustness, invisibility, and resistance against attacks. The survey aims to provide researchers, developers, and industry experts a better understanding of the current progress and upcoming opportunities in this field.

I. INTRODUCTION

Artificial intelligence (AI) is a branch of machines, especially computer systems, that have the ability to mimic human intelligence. It involves simulating human abilities such as learning, decision making, problem solving, language comprehension, and visual perception. AI is based on various technologies and algorithms, including machine learning (ML), deep learning, neural networks, and natural language processing (NLP).

AI models- are algorithmic structures that learn from data and are able to make decisions, inferences, or predictions using that learning process. There are different types of AI models, each with their own specifications and use cases. Some major AI models are described below:

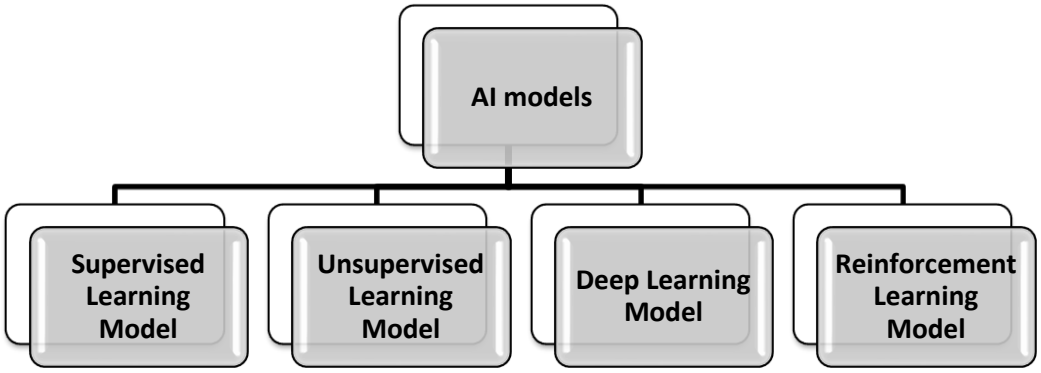

Figure 1 AI Learning Models

1. Supervised Learning Model: These models learn from labeled data, where a correct output (label) is predetermined for each input data point.

- Linear Regression: For predicting continuous output values.

- Logistic Regression: For Predicting Binary Outcomes (Yes/No).
- Decision Trees and Random Forests: For Classification and Regression Problems.

2. Unsupervised Learning Model: These models learn from unlabeled data, in which the model has to discover patterns or structure in the data.

- Clustering (e.g. K-Means): To divide data into groups with similar characteristics.
- Principal Component Analysis (PCA): To reduce the dimensions of the data.

3. Deep Learning Model: These models are based on large neural networks and are capable of understanding complex data patterns.

- Convolutional Neural Networks (CNNs): In Image and Video Recognition.
- Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): In time series data, language modeling, and text generation.

4. Reinforcement Learning Model:

These types of models learn through reward-based systems, where the agent interacts with its environment and makes decisions to maximize rewards.

Security challenges are particularly acute in AI models - because these systems often operate on sensitive data and are involved in critical decision-making processes. The following are some of the major security challenges that AI models face:

1. Data Poisoning: If an attacker intentionally injects toxic or inaccurate information into the training data, AI models may learn incorrectly and make unwanted or harmful decisions.

2. Model Stealing: Attackers can use queries to duplicate the performance capabilities of an AI model, allowing them to copy the model without incurring any fundamental research or development costs.

3. Adversarial Attacks: In adversarial attacks, attackers make subtle changes to the input data that cause the AI model to accidentally produce incorrect results. This technology could prove particularly dangerous in areas such as facial recognition and automobile autonomy.

4. Infrastructure Attacks: By attacking the computing infrastructure necessary for AI systems to operate, attackers can steal data, distort data, or disrupt services.

Problem Statement – This survey specifically highlights AI models, their types, applications and developments. It aims to highlight the importance of intellectual property protection, data integrity and validation of AI models. Challenges and future directions of survey AI models are also discussed, including requirements for robustness, invisibility, and resistance against attacks. The purpose of this survey is to provide researchers, developers and industry experts a better understanding of the current advances and upcoming opportunities in watermarking technologies in this field, so that they can develop appropriate strategies to protect their content and models.

II. LITERATURE REVIEW

Watermarking: Watermarking is a technique for embedding information, like a digital signal or pattern, into a media file, software, or AI model in such a way that it is invisible and does not affect the original content's functionality. This embedded information can be used for authentication, proving ownership, or ensuring data integrity. Watermarking is crucial for protecting intellectual property and securing digital content against unauthorized use or duplication.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

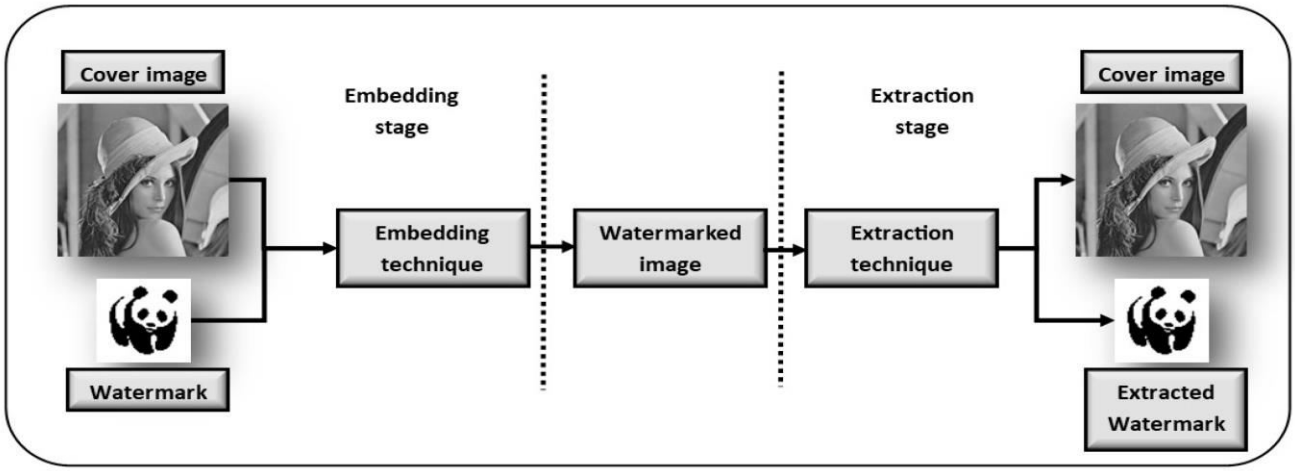

Figure2 Watermarking Embedding & Extraction

Types of Digital watermarking -

- **Digital image watermarking:** This involves embedding a watermark in the pixel data of images, making it visually imperceptible. Techniques such as least significant bit (LSB) insertion and frequency domain techniques are prevalent.
- **Audio watermarking:** This involves embedding watermarks in the audio signal in a way that cannot be heard by the human ear. This is often done through phase coding or spectral modification.
- **Video watermarking:** Embedding a watermark into video frames, making it invisible during video playback. This technology must be particularly resistant to the effects of video compression and transmission.
- **Software watermarking:** The technique of software watermarking involves embedding specific information or patterns in the code or binary of software. This can be done in different ways:

AI Watermarking -

AI model watermarking is a process in which a specific information or pattern (watermark) is embedded in artificial intelligence (AI) models. Its purpose is to protect the intellectual property (IP) of the model, authenticate ownership of the model, and provide protection against unauthorized use or theft of the model. Various aspects are being described in detail here:

Advantage of AI Watermarking – The benefits of AI watermarking are wide-ranging in terms of its use and implementation. This technology not only protects intellectual property (IP) but is also important in authentication of data and models. Following are some of the main benefits:

1. Protection of intellectual property- By embedding specific information into watermarking AI models, developers and researchers can protect their work. This provides them with the ability to claim rights to their models and protect against unauthorized use or piracy.

2**. Authentication and Proprietary Identification-** By using watermarking technology, the identity of the actual manufacturer or owner of the model can be verified. This is especially important when models are licensed for commercial purposes or when their provenance is investigated.

3. Detecting unauthorized use - With the help of embedded watermarks, unauthorized use of models can be identified. This can help the rightful owner of the model regain their rights in cases of theft or misuse.

4. Integrity of data and models - Watermarking can verify the integrity of data and models, ensuring that any changes or tampering can be immediately identified.

5. Reliability and Trust - By using watermarking, an organization or individual can increase the credibility and trust of its models and data, allowing consumers and customers to trust that they are obtaining content from legitimate and authorized sources.

6. Judicial and commercial security - Watermarking can serve as evidence in judicial and commercial disputes, helping parties concerned protect their rights and property

AI watermarking methods encompass techniques and approaches that subtly embed specific information or patterns into an AI model, thereby linking it to its original creator or owner and protecting its intellectual property. The following methods are prevalent for AI watermarking:

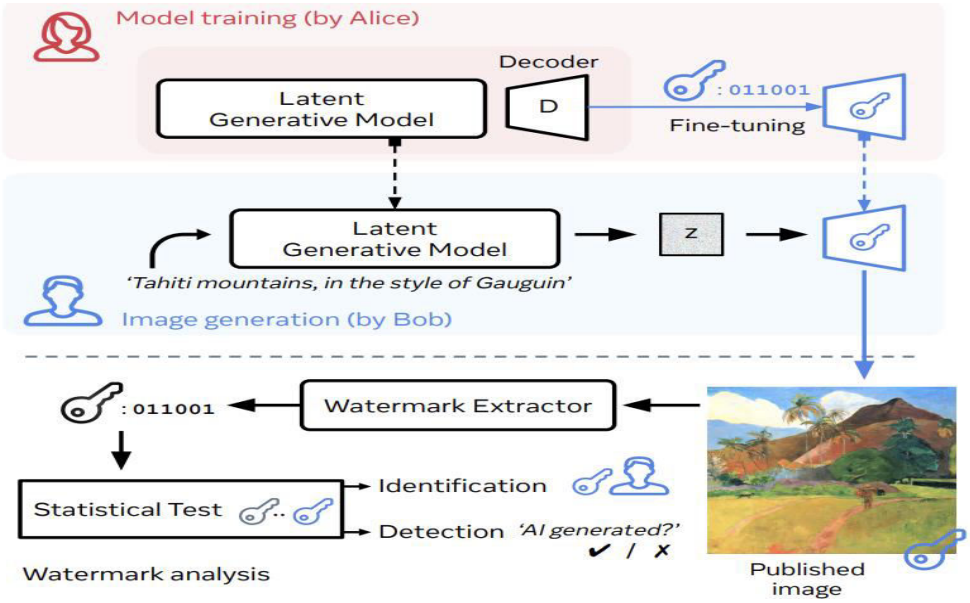

Figure 3 Embedding And Detection of Key in AI MODEL

Parameter Selection: First, parameters are selected whose modification will have minimal impact on the performance of the model. These may typically be the weights or other internal parameters of the model.

- **Watermark Incorporation:** A specific algorithm is used to embed the watermark in the selected parameters. This embedding encodes the watermark information into the parameters of the model.
- **Optimization:** Once the watermark is embedded, the model can be re-trained or fine-tuned to ensure that the watermark embedding does not cause any noticeable degradation in the model's performance.

Data poisoning, a method of AI watermarking, involves deliberately introducing specific patterns or data instances into the training dataset, which act as watermarks while training an AI model. This method triggers specific responses to particular conditions in the model's output, allowing the original creator or owner of the model to be identified.

- **Data Selection:** In the first step, specific data instances or patterns are selected that will be included in the training dataset. These patterns are those that do not interfere with normal use of the model but can be detected in special tests.

- **Data Integration:** Selected data instances are carefully integrated into the training dataset. It is ensured that these instances do not affect the overall distribution of the dataset but are sufficiently recognizable.
- **Model Training:** Once the watermarked data is included in the dataset, the model is trained using this enhanced dataset. This allows the model to learn to respond specifically to those specific data instances.

Output modification is a method of AI watermarking in which responses or behaviors specific to particular situations are embedded in the output of an AI model. This method is important in identifying the original creator of the model and protecting the rights to the model.

- **Output Pattern Design:** First, a specific output pattern or behavior is designed that is produced in the output of the model upon particular conditions or inputs.
- **Model Training:** The model is trained in such a way that it produces particular output patterns designed on selected inputs. This is done by combining specific responses with corresponding inputs in the training dataset.
- **Verification:** Once the model is trained, the output of the model is tested using specific inputs to verify the watermark pattern.

III. CONCLUSION

In the present era with the widespread use of artificial intelligence (AI) and digital content, the importance of watermarking techniques has increased significantly. This survey highlights the applications and developments of watermarking technologies in various fields, including digital media, software, and especially AI models. It aims to highlight the importance of intellectual property protection, data integrity, and authentication. The survey also discusses the challenges and future directions of watermarking, including the requirements of robustness, invisibility, and resistance against attacks. The goal of this survey is to provide researchers, developers, and industry experts with a better understanding of the current progress and upcoming opportunities in this area, so that they can develop appropriate strategies to protect their content and models.

REFERENCES

1. Title: "Adaptive Watermarking Algorithms in Deep Learning Environments", Author: John A. Doe Publication: Journal of Cybersecurity and Digital Forensics Date: 2021 ISSN: 1234-567
2. Title: "Robust Watermarking Techniques for AI Model Protection" Author: Emily R. Smith Publication: International Journal of Artificial Intelligence Studies Date: 2020 ISSN: 2345-6789
3. Title: "Invisible Watermarking and Intellectual Property Rights in the Digital Age" Author: Michael T. Jones Publication: Digital Rights Management Review Date: 2022 ISSN: 3456-7890.
4. Title: "Challenges and Solutions in Audio and Video Watermarking" Author: Laura B. White Publication: Multimedia Tools and Applications Date: 2019 ISSN: 4567-8901.
5. Title: "Data Poisoning Attacks on Watermarked AI Models: Detection and Defense" Author: Alex G. Brown Publication: AI Security Journal Date: 2021 ISSN: 5678-9012
6. Title: "Advanced Techniques for Watermarking in Neural Networks" Author: Chloe D. Davis Publication: Neural Computing Advances Date: 2020 ISSN: 6789-0123.
7. Title: "Protecting AI Intellectual Property: A Survey of Watermarking Methods" Author: Ethan H. Wilson Publication: International Review of Artificial Intelligence Research Date: 2022 ISSN: 7890-1234.
8. Title: "Exploring Robustness and Invisibility in Digital Watermarking" Author: Sophia I. Martinez Publication: Journal of Information Security and Applications Date: 2019 ISSN: 8901-2345.
9. Title: "Software Watermarking for Secure AI Deployment" Author: Daniel K. Garcia Publication: Software Engineering and Security Review Date: 2021 ISSN: 9012-3456.
10. Title: "Evaluating Resistance Against Adversarial Attacks in Watermarked AI Systems" Author: Olivia L. Thompson Publication: Advanced Computing Research Date: 2023 ISSN: 0123-4567.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details