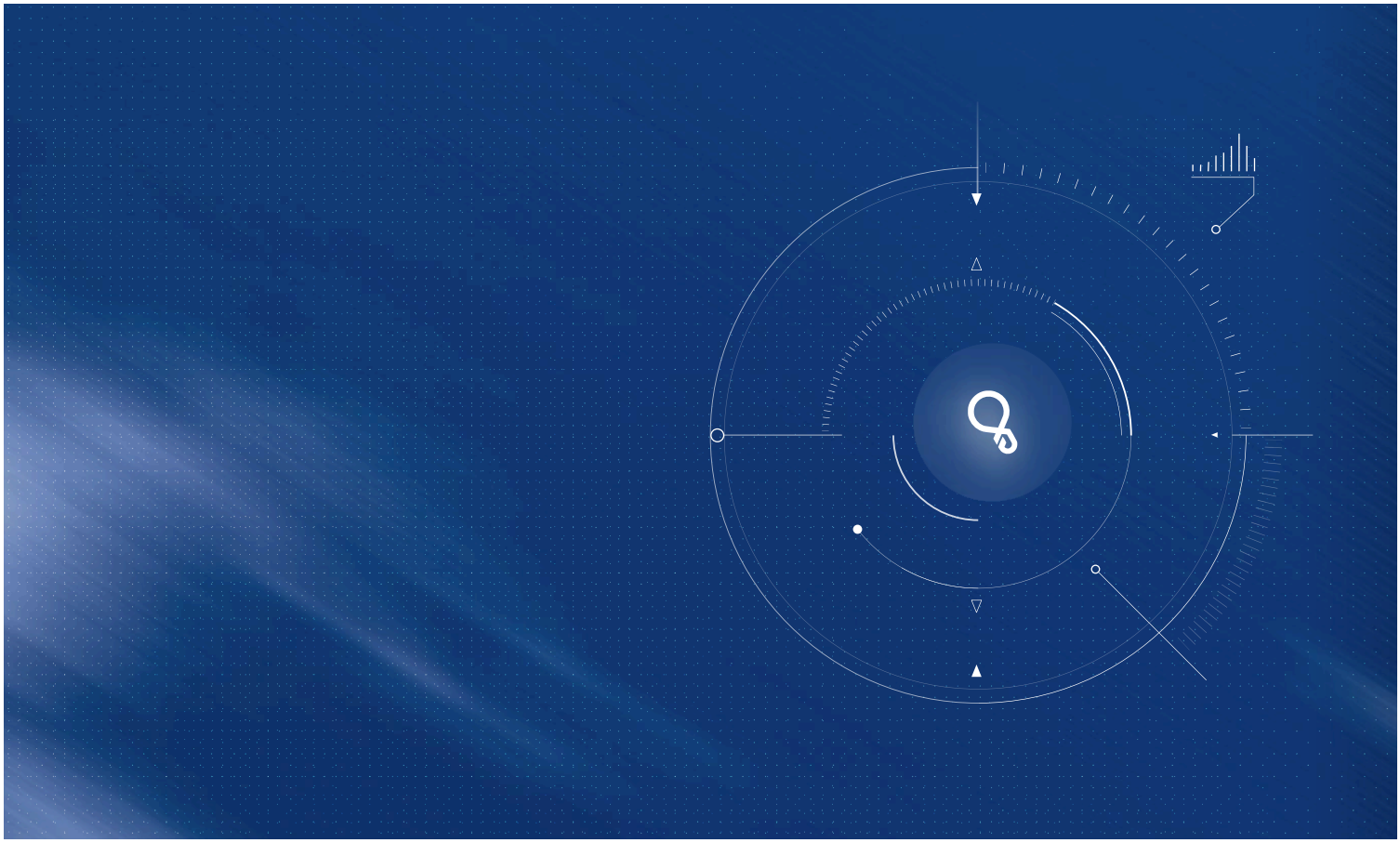
cypris

OCTOBER 2024



RESEARCH BRIEF

DATA FINGERPRINTING AND WATERMARKING

Prepared By

Cypris Team

info@cypris.ai

cypris

Cypris Insight Brief | For: Open Access

October 23, 2024

Table of Contents

Executive Summary

With the proliferation of AI systems, data ethics has become a critical concern. Modern AI models require vast amounts of data for training, often using information without explicit consent from individuals or corporations. Innovations in tracing leaked data have emerged, highlighting the misuse of proprietary information.

The intention of this Insight Brief is to provide insights into the latest state of the art in how data fingerprinting/watermarking is being done, for what applications, and by what entities. A comparison of the various data tracing techniques, as well as their use cases, and their strengths and weaknesses has also been provided.

Analyst Opinion

The use of data fingerprinting and watermarking in the AI and LLM space addresses critical concerns surrounding data ownership, security, and ethical usage, particularly in an era increasingly reliant on novel sources. From a strategic perspective, the advancements in fingerprinting and watermarking provide essential tools for AI developers, enterprises, and regulators to maintain control and traceability over data assets. These techniques, while often viewed as technical safeguards, are in fact evolving into key enablers for compliance with global data regulations such as GDPR and CCPA. The growing sophistication of adversarial attacks and data breaches makes the deployment of these technologies not only a defensive strategy but a necessary business imperative.

Fingerprinting and watermarking technologies also serve as a deterrent against data misuse, signaling to both external and internal actors that data provenance can and will be traced. In doing so, organizations safeguard their intellectual property, mitigate potential competitive losses, and manage the risks of AI systems being built or trained on illicit data. This is especially important for enterprises operating in industries where data is their most valuable asset—such as healthcare, finance, and legal sectors—where traceability of both data inputs and AI-generated outputs is crucial for maintaining accountability and trust.

Moreover, the role of these technologies extends beyond compliance and security. Fingerprinting and watermarking can foster new business models around data licensing and AI-as-a-service. With the ability to track how data and AI models are utilized, companies can explore more granular and secure monetization opportunities, offering customized models to different clients while maintaining control over model use through embedded tracking mechanisms. The innovations discussed in this Brief also highlight the growing intersection of AI ethics, data privacy, and intellectual property, where fingerprinting and watermarking act as foundational technologies supporting these principles.

While fingerprinting and watermarking are often regarded as technical tools, they play a critical role in shaping the future of AI governance, security, and business strategy. The capabilities they offer in data tracking, model security, and ethical AI deployment are set to be integral to the continued advancement of AI across industries. As the AI landscape continues to evolve, the importance of these technologies will only grow, demanding continuous innovation and adaptation to stay ahead of adversarial and regulatory challenges.

Research Methodology

In our research, we utilized the Cypris platform, third-party datasets, and broader internet searches to identify relevant data. Throughout this process, we refined our approach by adapting our keywords to synonyms and related terms to ensure comprehensive data collection within this sector. For our foundational query, we used Cypris' semantic searching functionality with the following search term: 'data fingerprinting watermarking detection'.

Introduction to Data Fingerprinting and Watermarking

In the rapidly evolving landscape of artificial intelligence (AI) and large language models (LLMs), data has emerged as a critical asset driving innovation and model performance. These models are heavily reliant on vast amounts of data to learn and generate human-like text, images, and other content. However, as the demand for data grows, so do the concerns about unauthorized usage and intellectual property infringement. Data fingerprinting and watermarking have emerged as vital techniques to trace leaked data back to its source, ensuring that proprietary or sensitive information remains secure.

Data fingerprinting involves embedding unique identifiers into datasets, allowing the originator to track who accesses or shares the data. Watermarking, on the other hand, subtly alters the data to include hidden information that signifies ownership without affecting its usability. In the context of AI and LLMs, these methods are particularly important because models are often trained on vast and sometimes proprietary datasets. Unauthorized distribution or use of this data can lead to intellectual property theft, competitive disadvantages, and legal complications. These same principles and techniques can also be applied towards detecting AI-generated content[1]. Relative to other approaches, watermarks are accurate and more robust to erasure and forgery. However, they are not foolproof, and a motivated actor can degrade watermarks in AI-generated content.

By implementing fingerprinting and watermarking, organizations can monitor the flow of their data, deter unauthorized usage, and protect their investments in data collection and curation. For AI developers, these techniques help ensure that their models are not unknowingly trained on illicit data, which could compromise the model's integrity and lead to ethical or legal issues. As AI continues to integrate into various sectors, the importance of safeguarding data through fingerprinting and watermarking becomes ever more critical to maintain trust, compliance, and the responsible advancement of technology.

Data Tracing Techniques and Applications

Data Fingerprinting

Data fingerprinting involves embedding unique, often imperceptible identifiers into each distributed copy of a dataset. By introducing slight variations, the data owner can trace the source of any leaked data based on these unique identifiers.

Implementation for AI/LLMs:

[1] Detecting AI fingerprints: A guide to watermarking and beyond

- **Unique Data Variations [Categorical or Numerical]:** Slightly adjust numerical values within an acceptable error margin. Alternatively, swap rare categories or introduce benign anomalies.
- **Metadata Embedding:** Include hidden markers in the dataset's metadata, such as timestamps or version numbers.
- **Dummy Records:** Insert unique, non-essential records that act as identifiers if they appear elsewhere.

Strengths	Weaknesses
Traceability and Accountability: Fingerprinting allows data owners to identify precisely which copy of the data was leaked and who was responsible for the leak. This is achieved by embedding unique identifiers in each distributed copy of the dataset.	**Complexity in Managing Multiple Data Versions:** Distributing uniquely fingerprinted copies to a large number of recipients can be logistically challenging and resource-intensive.
Deterrence Against Unauthorized Distribution: Knowing that data can be traced back to them, recipients are less likely to distribute it without authorization, reducing the risk of data breaches.	**Vulnerability to Collusion Attacks:** If multiple recipients combine their distinct copies, they may be able to detect and remove the fingerprints, or create a version that cannot be traced.
Minimal Impact on Data Utility: Properly implemented fingerprinting introduces negligible changes to the data, ensuring that its utility for training AI models is not compromised.	**Potential for Fingerprint Removal or Alteration:** Sophisticated attackers may employ signal processing or machine learning techniques to identify and strip fingerprints from the data.
Customization for Different Data Types: Fingerprinting techniques can be adapted for various data types, including text, images, audio, and structured data, making it a flexible solution for diverse AI applications.	**Detection by Machine Learning Models:** Embedded fingerprints might be learned by AI models during training, leading to unintended biases or model behaviors.
Complementary to Other Security Measures: Fingerprinting can be used alongside encryption and access controls to provide a multi-layered security approach.	**Resource Intensiveness:** Generating and embedding fingerprints, especially for large datasets common in AI training, can require significant computational resources.

Applications of Data Fingerprinting in AI/LLMs

Data Security and Privacy

- **Fraud Detection:** In LLM-powered systems, digital fingerprints can help in detecting anomalous behaviors. This is particularly useful in applications like conversational AI, where interactions can be analyzed to ensure they align with legitimate patterns.
- **Unauthorized Model Usage:** Fingerprinting AI models allows organizations to track if their proprietary models are being used or copied without permission. Digital fingerprints tied to specific model instances help identify unauthorized usage.

- **Protecting User Data:** Digital fingerprinting can track how AI models interact with sensitive user data. It ensures that AI systems handle data securely, preventing leaks or misuse by identifying any unauthorized data access patterns.

Accountability and Provenance Tracking

- **Tracking Model Outputs:** Digital fingerprints allow for the tracking of model outputs, identifying which version of a model generated specific content. This is important for industries that require accountability for AI decisions, such as healthcare diagnostics or automated financial advice.
- **Content Verification:** By assigning digital fingerprints to the outputs generated by LLMs, systems can ensure that text, recommendations, or decisions are traceable back to their original source. This is particularly relevant in preventing misinformation or deepfake content.
- **AI Training Data Auditing:** Digital fingerprinting of datasets used for training can ensure that data sources are appropriately tracked. This is critical in AI ethics to confirm that models are trained using unbiased and legally compliant datasets.

Personalization and User Identification

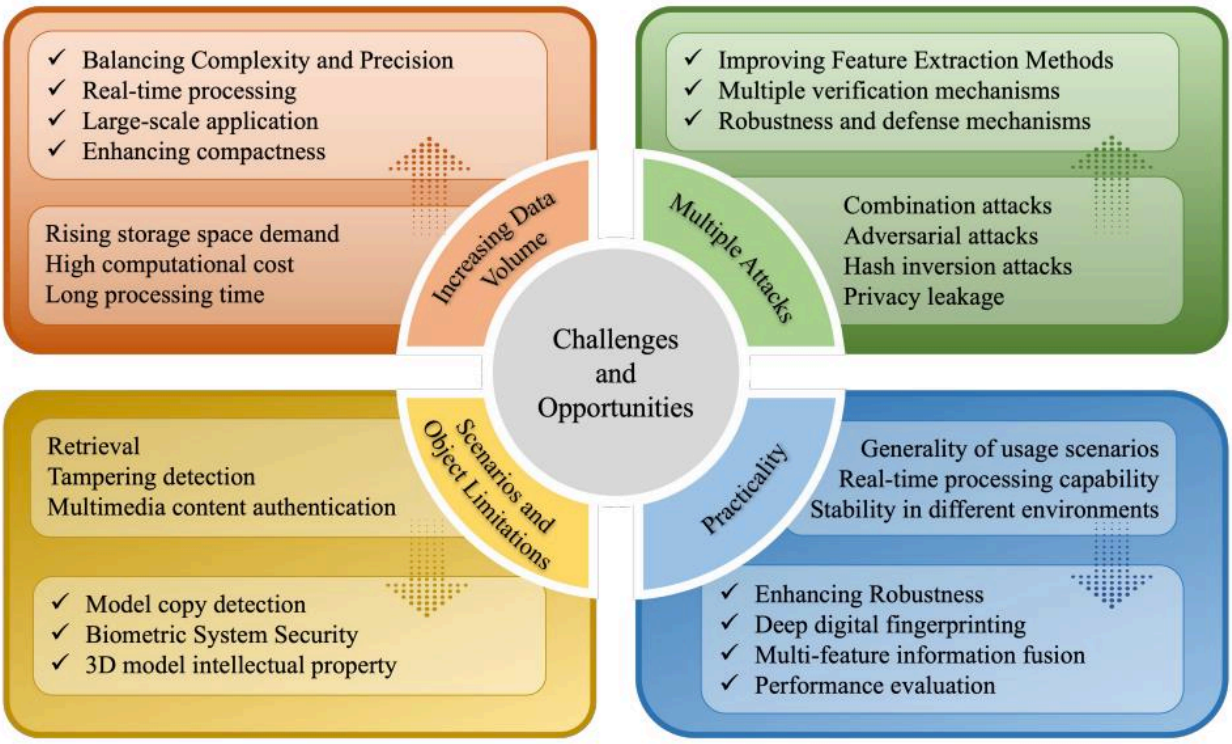
- **Customized AI Interactions:** Fingerprinting can track users across sessions, ensuring personalized responses from LLMs. This is particularly useful in customer service bots or AI tutors, where maintaining context from prior interactions is crucial.
- **Enhanced User Experience:** With digital fingerprints, AI systems can recognize individual users and adapt content to their preferences, learning from their behavior without the need for invasive data collection.
- **Authentication in AI Systems:** By identifying users or devices through digital fingerprints, AI-powered systems can offer additional layers of authentication. This can be used in secure applications, such as virtual assistants managing sensitive tasks (e.g., financial transactions or medical information).

Legal and Ethical Accountability

- **Compliance with Regulations:** Fingerprints can be used to audit AI systems to ensure they comply with regulations such as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA). These regulations often require traceability of how user data is processed.
- **Model Auditing for Bias:** Digital fingerprints can help track which datasets or models led to specific decisions. This allows for better auditing and identification of biased outputs or decisions made by AI models, ensuring ethical AI deployment.
- **Evidentiary Use in Litigation:** Fingerprints can serve as evidence to support claims in legal proceedings, particularly in cases where AI outputs (e.g., automated decisions) need to be verified and attributed to a specific AI model or instance.

Combatting Adversarial Attacks

- **Detecting Model Manipulation:** Adversarial actors often attempt to manipulate or reverse-engineer AI models. Digital fingerprints embedded within models can detect when they've been tampered with, ensuring the integrity of the system.
- **Preventing Model Stealing:** LLMs are vulnerable to model extraction attacks, where a malicious actor tries to replicate a model's functionality. Fingerprints can detect and prevent these attempts, safeguarding proprietary technology.



Challenges and Opportunities of Digital Fingerprinting, adapted from Chen et al., 2024

Data Watermarking

Watermarking embeds hidden information into the data or models, signifying ownership without affecting their usability or performance. Unlike fingerprinting, watermarks are usually consistent across all copies.

Implementation for AI/LLMs:

- **Dataset Watermarking:**
  - *Image Data:* Embed subtle patterns or noise in images that are imperceptible to the human eye but detectable algorithmically.
  - *Text Data:* Introduce specific linguistic patterns, synonyms, or formatting that serve as markers.
  - *Audio Data:* Embed inaudible signals or modulate frequencies slightly.
- **Model Watermarking:**
  - *Backdoor Triggers:* Train the model to produce a specific output when presented with a unique input (trigger).

- *Weight Perturbation:* Slightly adjust model weights to encode information without affecting performance.

Strengths	Weaknesses
Proof of Ownership: Watermarks serve as tangible evidence of ownership in legal disputes over intellectual property infringement. By embedding a unique identifier into the data or model, the rightful owner can demonstrate that the content originated from them.	**Vulnerability to Watermark Removal Attacks:** Sophisticated attackers can employ various techniques to detect and remove watermarks without significantly degrading the data, such as using filtering, compression, or geometric transformations
Robustness to Common Data Manipulations: Well-designed watermarks can withstand common data transformations such as compression, resizing, cropping	**Potential Degradation of Data Quality:** Embedding a watermark may introduce slight distortions or artifacts, which can be problematic in applications requiring high fidelity, such as medical imaging or high-quality audio.
Imperceptibility: Watermarks are typically designed to be invisible or inaudible, ensuring that the quality and usability of the data are not noticeably affected	**Difficulty in Applying to Certain Data Types:** Watermarking text data without affecting readability or meaning is complex, and methods for structured data like databases are less developed compared to multimedia data.
Versatility Across Data Types: Watermarking techniques can be applied to various forms of data including images, audio, video, text, and even neural network models	**Limited Capacity for Embedded Information:** There is a trade-off between the amount of information embedded and the imperceptibility and robustness of the watermark. High-capacity watermarks are more susceptible to detection and removal.

Applications of Data Watermarking in AI/LLMs

Ownership and Copyright Protection

- **Watermarking AI-Generated Text**: LLMs like GPT-4 or other generative models produce vast amounts of content, including text, code, and creative outputs. Watermarking the content generated by these models ensures that the original creator can claim ownership, making it useful in fields like automated journalism, creative writing, and content creation.[2]
- **Preventing Unauthorized Redistribution**: Watermarking allows companies to protect their proprietary AI models and content by embedding watermarks that can trace the source of any unauthorized redistribution. For instance, if an organization creates marketing material or product descriptions using an LLM, watermarks ensure that stolen or plagiarized content can be traced back to its original source.
- **Creative Industries**: In sectors like film, publishing, and music, AI is being used to generate creative content. Watermarking the outputs ensures that the creators retain ownership

[2] Takale et al., 2024

rights, especially when these models produce large-scale creative works such as books, scripts, or compositions.

Tracking and Authenticating Model Outputs

- **Provenance Tracking in AI Systems**: Watermarking enables organizations to track the origin of generated outputs and verify which model instance or version was used to produce specific content. This application is crucial in sectors like legal, medical, and financial services, where AI-generated outputs need to be verified for accuracy and accountability.
- **Content Authenticity**: In fields like news media, where AI is used to draft articles or reports, watermarking can help confirm the authenticity of the text, ensuring that it comes from a trusted source and has not been tampered with or manipulated post-generation.
- **Preventing Deepfakes**: Watermarking techniques can be applied to visual or audio outputs from AI models to prevent the creation of deepfakes or other misleading content. Watermarks embedded in images, videos, or audio files can help verify whether the content was genuinely AI-generated and ensure its ethical use.

Model Ownership and Licensing Control

- **Watermarking of AI Models**: Developers and companies often license AI models to third parties or distribute them via cloud platforms. Watermarking AI models helps track and control their usage, ensuring that only licensed users are operating the models. This is particularly relevant in industries like cloud AI, where various organizations deploy pretrained LLMs for commercial applications.
- **Preventing Model Theft**: Watermarks can be embedded in the parameters or outputs of AI models to detect whether the model has been copied or cloned illegally. This is important in cases where intellectual property theft of AI models could result in significant financial or competitive harm to organizations.

Regulatory Compliance and Auditing

- **GDPR and CCPA Compliance**: Watermarking can ensure that data used in training and the outputs generated by LLMs comply with regulatory requirements like the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA). Watermarks can track how specific datasets or models interact with personal data, providing a clear audit trail.
- **AI Governance**: In high-stakes environments such as healthcare or autonomous systems, it is critical to ensure that AI models adhere to safety and ethical standards. Watermarking provides a mechanism to trace and audit AI decisions or outputs, ensuring that they align with prescribed rules and governance standards.
- **Ethical AI Auditing**: AI models, particularly those used in sensitive applications, must be transparent about their decision-making processes. Watermarking can embed traceable information within model outputs to allow external auditors to verify whether the AI system adhered to ethical principles during operation.

Accountability for AI-Generated Content

- **Verifiable AI Decisions**: Watermarking ensures that the decisions or outputs of an AI system are traceable back to the model version or configuration that generated them. This is critical in scenarios like financial advising, legal AI services, or medical diagnostics, where accountability for decisions is paramount.
- **Preventing AI-Driven Misinformation**: As AI-generated content proliferates, there is a risk of creating and spreading misinformation, particularly through fake news or manipulated media. By watermarking the outputs of AI-generated content, companies can demonstrate that their AI systems are producing verified and trustworthy information, helping combat the spread of false or harmful content.

Watermarking in Collaborative AI Workflows

- **Version Control and Attribution**: Watermarking allows different versions of AI-generated content to be tracked and attributed to specific collaborators, ensuring proper recognition and accountability in team-based environments. This is particularly useful in creative industries such as film production or collaborative writing, where multiple contributors may use AI tools to generate or enhance content.
- **Secure Collaboration in AI Projects**: Watermarks can also ensure that AI models or datasets shared between organizations in collaborative projects remain secure and are not used beyond agreed-upon terms. This is crucial in industries like pharmaceuticals, where multiple companies may work together on research projects involving sensitive AI models.

Combined and Advanced Techniques

The latest innovations in data fingerprinting and watermarking focus on more sophisticated, resilient, and imperceptible techniques for tracking and protecting digital assets, particularly in AI and machine learning systems. These advances enable secure embedding of identifiable markers into data and models without compromising performance or quality, enhancing detection of unauthorized use and ensuring provenance. New methods also improve robustness against tampering, ensure efficient retrieval of embedded information, and provide stronger defenses against adversarial attacks, making them crucial for applications in content protection, model security, and regulatory compliance in AI-driven industries.

Spread Spectrum Watermarking

Spread spectrum watermarking is widely used for embedding imperceptible watermarks into multimedia data such as images, audio, and video. It spreads the watermark signal across a wide frequency spectrum, making it resilient to common signal distortions, compression, and even some deliberate attacks. This technique is robust due to its noise-like properties, making the watermark harder to detect and remove.[3]

Use Cases

- Copyright protection for digital media (e.g., audio, video, images)
- Tracking and verifying ownership of media content
- Secure communication systems to protect transmitted data

[3] Cox et al., 1997

Strengths	Weaknesses
Robustness: Resistant to various signal processing attacks, including compression, noise addition, and filtering	**Lower Embedding Capacity:** The amount of data that can be embedded as a watermark is limited compared to other methods.
Imperceptibility: Watermarks are embedded subtly, without noticeable degradation to the original content.	**Potential Degradation of Data Quality:** Embedding a watermark may introduce slight distortions or artifacts, which can be problematic in applications requiring high fidelity, such as medical imaging or high-quality audio.
Wide Application: Effective across different media formats, including audio, video, and images	

Quantization Index Modulation (QIM)

QIM is a widely used method in watermarking and data hiding. It works by quantizing the host signal into different sets (called cosets) and embedding watermark information through these coset indices. The goal is to achieve efficient trade-offs among rate-distortion-robustness in embedding information.[4]

Use Cases

- **Digital Watermarking:** QIM is extensively used in multimedia watermarking, particularly in image, audio, and video processing, where it helps to embed hidden watermarks for copyright protection and authentication purposes.
- **Secure Communications:** QIM is applied in secure data transmission scenarios where sensitive information is hidden within non-sensitive data for secure transmission without being easily detected.
- **Steganography:** This technique is often used for embedding covert messages in digital files.

Strengths	Weaknesses
High Robustness: QIM is resistant to a wide range of signal distortions, including noise, compression, and certain forms of tampering. This makes it effective in maintaining the integrity of the embedded watermark under various processing conditions.	**Vulnerability to Quantization Attacks:** QIM can be vulnerable to certain types of attacks, such as scaling and requantization, which can degrade the embedded watermark. These attacks can sometimes remove or alter the watermark without being detected.
Efficient Trade-offs: QIM provides an efficient balance between embedding rate, distortion to the host signal, and robustness, making it an attractive option for multimedia watermarking.	**Perceptual Quality Degradation:** While QIM strives to minimize distortion, embedding watermarks may still cause slight perceptual degradation in the host media, especially in cases where high embedding rates are used.

[4] Liu, 2023

Low Complexity: Compared to other watermarking techniques, QIM is computationally efficient, allowing for real-time embedding and detection in many cases.	**Fixed Embedding Distortion:** The distortion introduced by QIM is fixed and may not be optimal for certain content types, making it less adaptive to various host signal problems.

Code-Based Fingerprinting

Code-based fingerprinting is a method used to embed unique identifiers into content or code, which allows the tracking and identification of individual users or copies of a product. This technique is widely used in digital content distribution, particularly for tracking the distribution of media or software, ensuring security in cloud services, and combating piracy.[5]

Use Cases

- **Digital Content Protection**: Code-based fingerprinting is heavily used to protect intellectual property (IP) in digital content distribution. For example, companies distributing movies or music embed unique fingerprints into each copy distributed, allowing them to trace any pirated versions back to the original source.
- **Software Licensing and Anti-Piracy**: Software companies embed fingerprints in distributed software copies to ensure that licensed versions can be tracked, enabling the detection of unauthorized or pirated copies.
- **Cloud Services and API Usage**: Cloud platforms use code-based fingerprinting to track API usage and manage resource access. This allows companies to monitor misuse or breaches more effectively.

Strengths	Weaknesses
Traceability: Code-based fingerprinting allows precise tracking of distributed content, helping businesses identify leaks or unauthorized sharing.	**Vulnerability to Collusion Attacks**: In some cases, multiple unauthorized users could compare their versions of the content to try and detect or remove the fingerprint, diminishing its effectiveness.
Granularity: The ability to embed unique fingerprints in individual copies ensures detailed insights into how each version of the content or software is being used.	**Overhead**: Embedding fingerprints can introduce overhead in terms of processing time or resource use, particularly in large-scale media files or complex software.
Piracy Deterrence: By embedding these fingerprints, companies can discourage piracy, as they can trace unauthorized versions back to the original user or distribution point.	**Difficulty in Large-Scale Identification**: While fingerprints can be used to track individual copies, identifying specific alterations in very large datasets can be computationally intensive.

[5] Vector space data digital fingerprint method based on GD-PBIBD coding

Permutation-Based Fingerprinting

Permutation-based fingerprinting is a method that embeds unique identifiers into a host data set by permuting its elements, allowing for the identification of different copies of data distributed to users. This technique is widely used in digital rights management (DRM), where different permutations are applied to distributed content to trace unauthorized copies.[6]

Use Cases

- **DRM**: Trace the origin of leaked digital media or documents.
- **Software Licensing**: Detect pirated versions of software by identifying tampered copies.

Strengths	Weaknesses
Robustness: Effective against common tampering attempts, as small changes in the data set are unlikely to affect the permutation-based fingerprint.	**Collusion Attacks**: Multiple attackers can compare versions of the same data set to identify and neutralize the fingerprint.
Low Overhead: Minimal impact on the original data set, as the permutations do not noticeably alter its properties.	**Complexity**: Requires careful design and computation to ensure uniqueness without overly disrupting the data.

Perceptual Hashing

Perceptual hashing is a technique used to generate hash values for multimedia files (images, videos, audio) in a way that is tolerant to minor variations in the content. Unlike cryptographic hashing algorithms like SHA-256, which are highly sensitive to even the slightest change in the input data, perceptual hashes generate similar hash values for inputs that "look" or "sound" similar, making them ideal for applications like detecting duplicates, copyright violations, and image comparison.[7]

Use Cases

- **Duplicate Detection in Multimedia**: Companies like Google and Facebook use perceptual hashing to detect duplicate or similar images and videos across their platforms. For example, Google's "Reverse Image Search" and Facebook's efforts to detect and block non-consensual sharing of sensitive media rely on perceptual hashing to identify altered versions of the same content, such as a cropped image or a video with slightly adjusted colors.
- **Content Moderation and Copyright Protection**: The **PhotoDNA** technology developed by Microsoft, used by organizations like the National Center for Missing & Exploited Children (NCMEC), uses perceptual hashing to detect and block child sexual abuse material online. Even if the images have been resized, cropped, or slightly modified, PhotoDNA can still identify them as matches.

[6] Thanh et al., 2012
[7] Farid, 2021

- **Cybersecurity:** In cybersecurity, perceptual hashing is applied to detect phishing websites that look visually similar to legitimate websites. By analyzing the appearance of web pages, security tools can flag spoofed sites used for phishing, even if the HTML code is different.

Strengths	Weaknesses
Robust to Minor Modifications: Perceptual hashing can detect variations of the same content even when minor edits have been made, such as changes in resolution, color adjustments, or watermarking. This makes it ideal for detecting duplicates or near-duplicates of images and videos.	**False Positives and False Negatives**: While perceptual hashing is resilient to minor changes, it is not perfect. Two very different pieces of content could, on occasion, generate similar perceptual hashes, leading to false positives. Similarly, certain modifications that significantly alter the perceptual characteristics could result in false negatives, where a hash fails to detect content as being a modified version of the original.
Efficient Search and Comparison: Perceptual hashes are typically small, compact representations of the content, making them computationally efficient for searching and comparing large datasets of media files.	**Limited Resistance to Significant Alterations**: If a piece of media undergoes substantial transformation, such as heavy filtering, significant cropping, or the addition of unrelated content, the perceptual hash may no longer match. This limitation makes perceptual hashing less effective in scenarios where more sophisticated or malicious alterations are used to evade detection.
Effective for Copyright and Licensing: Content creators and distributors can track the usage of their media across the web and platforms through perceptual hashing, ensuring that even altered versions of their work are detectable.	**Vulnerability to Adversarial Attacks**: Perceptual hashing algorithms can be manipulated by adversaries who understand how the hash is generated. By carefully altering the content to avoid generating a matching hash while maintaining the perceptual similarity, attackers can circumvent detection.

Backdoor Watermarking

Backdoor watermarking is a technique used in machine learning, particularly in neural networks, to embed a secret trigger or pattern (the "backdoor") into the model during training. When this trigger is presented during inference, the model behaves in a predetermined way, typically outputting a specific, incorrect result. Backdoor watermarking can also refer to embedding a hidden watermark that serves as a signature to identify or authenticate a model.[8]

Use Cases

[8] Li et al., 2020

- **Model Intellectual Property (IP) Protection**: Backdoor watermarks are embedded into machine learning models to serve as a proprietary signature, helping companies protect their IP. If a company suspects that their model has been stolen or illegally reused, they can use the hidden watermark to prove ownership. For instance, watermarks have been proposed for protecting deep neural networks (DNNs) trained by companies like Google or OpenAI, who invest heavily in AI model development.
- **Digital Rights Management for AI Models**: Watermarking has also been used to prevent unauthorized redistribution or modifications of machine learning models. When a model is redistributed, the watermark serves as a way to trace the origin and ensure it hasn't been tampered with. This is especially relevant for models sold on platforms like Hugging Face or GitHub, where model piracy is a concern.
- **Trigger-based Adversarial Attacks**: Backdoor watermarks have been weaponized in adversarial attacks to modify the behavior of a model under specific conditions. Attackers inject a backdoor during the training process such that, during inference, a particular input (e.g., a specific image or pattern) activates the backdoor, leading to targeted misclassification.
- **Authentication of Federated Learning Models**: In federated learning, models are trained across multiple distributed devices. Backdoor watermarking allows the central authority (server) to verify that the models submitted by users are legitimate and that malicious entities haven't inserted backdoor attacks. Each legitimate model has a distinct, identifiable watermark.

Strengths	Weaknesses
IP Protection and Ownership Proof: Backdoor watermarking provides a robust way for companies to prove ownership of AI models. The embedded watermark is often invisible and does not affect the model's normal functionality but can be revealed upon the presentation of a specific input.	**Vulnerability to Reverse Engineering**: If adversaries gain access to the model and understand the watermarking technique, they may be able to detect and remove the watermark, thereby circumventing its protective functions. This is especially true for less sophisticated watermarking methods that are not designed with robust cryptographic techniques.
Minimal Performance Degradation: Well-implemented backdoor watermarks typically do not significantly degrade the performance of a model on normal inputs. The model behaves normally for standard tasks but exhibits different behavior when the backdoor is triggered.	**Backdoor Exploitation in Adversarial Attacks**: A significant risk arises when backdoor watermarking is used in adversarial attacks. If an attacker gains access to the model and can modify the training data or inject a backdoor, they can force the model to behave maliciously under specific conditions. These types of attacks can be hard to detect and may only be revealed in rare circumstances.

Scalability in Distributed Learning: In federated learning or distributed AI environments, backdoor watermarks offer a scalable way to verify that all models comply with security standards. Watermarks can be used to detect if any model contributor has inserted malicious modifications.	**Difficulty in Watermark Detection Without Special Inputs**: Detecting backdoor watermarks often requires specialized knowledge or input triggers, which means that if the watermark is not designed well, it can be hard to detect even by the rightful owner. In cases where watermarks are meant for IP protection, this can complicate the legal process of proving model ownership.