NIST AI 100-4

DRAFT FOR PUBLIC COMMENT

Reducing Risks Posed by Synthetic Content

An Overview of Technical Approaches to Digital Content Transparency

This publication is available free of charge from:
[DOI link TK]

April 2024

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-4

DRAFT FOR PUBLIC COMMENT

Reducing Risks Posed by Synthetic Content

An Overview of Technical Approaches to Digital Content Transparency

This publication is available free of charge from:
[DOI link TK]

Comments are especially requested on the completeness and clarity of the report, and:

- The current approaches outlined in the report and in the graphic highlighting key concepts.
- All "Additional Issues for Consideration" sections and whether they sufficiently address both technical and socio-technical concerns, and whether such issues and concerns are appropriate for future science-backed standards and technique development.
- Whether the report in its entirety, presents coverage of the digital content transparency technical landscape.
- Current state of the art for provenance data tracking techniques which may not be already addressed in the report, including watermarking techniques, as well as use cases for implementation.
- Testing, evaluation, and auditing techniques discussed in the report and technical literature references to expand on the techniques that are discussed.
- Technical mitigations for preventing and reducing harms from synthetic child sexual abuse material (CSAM) and Non-Consensual Intimate Images (NCII) beyond what is included in the report, as well as further evaluations and studies done on the efficacy of these various mitigations, including their application in open versus closed models.
- Potential development of standards and techniques on digital content transparency approaches.

Comments on NIST AI 100-4 may be sent electronically to NIST-AI-100-4@nist.gov with "NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency" in the subject line. Comments may also be submitted via www.regulations.gov: enter NIST-2024-0001 in the search field, click on the "Comment Now!" icon, complete the required fields, including "NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency" in the subject field, and enter or attach your comments. Comments containing information in response to this notice must be received on or before **June 2, 2024, at 11:59 PM Eastern Time.**

Acknowledgments: This report could not have been accomplished without the many helpful comments and contributions from the community and NIST staff and guest researchers: Bilva Chandra, George Awad, Yooyoung Lee, Peter Fontana, Razvan Amironesei, Mark Przybocki, Kamie Roberts, Elham Tabassi, Mat Heyman, and Jesse Dunietz.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. Any mention in the text of commercial, non-profit, academic partners, or their products, or references is for information only; it is not intended to imply endorsement or recommendation by any U.S. Government agency.

Table of Contents

1. Summary

Generative artificial intelligence (AI) technologies can generate realistic images, text, audio, video, as well as multimodal content. This enables novel applications with promising potential for good while also posing new risks to trust, safety, transparency, and credibility in digital information and communications.

This report examines the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for: authenticating content and tracking its provenance; labeling synthetic content, such as using watermarking; detecting synthetic content; preventing generative AI (GAI) from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual); testing software used for the above purposes; and auditing and maintaining synthetic content.

This report reflects public feedback and consultations with diverse stakeholders, including those who responded to a NIST Request for Information.

Digital content transparency, refers to the process of documenting and accessing information about the origin and history of digital content. Together, the approaches we discuss below can help manage and reduce risks related to synthetic content in four ways:

- Attesting that a particular system produced a piece of content,
- Asserting ownership of content,
- Providing tools to label and identify AI-generated content, and
- Mitigating the production and dissemination of AI generated child sexual abuse material and non-consensual intimate imagery of real individuals.

Digital content transparency provides a vehicle for individuals and organizations to access more information about the origin and history of content, which may contribute to trustworthiness, but does not guarantee it, and in some cases may actually undermine it. While transparency can help identify when content is being misrepresented, it can also create a false sense of trust, such as when a piece of content appears legitimate based on technical measures but is then manipulated through non-technical means (e.g., taking a legitimate piece of content out of context). Ultimately, the impact of transparency depends on the effectiveness of the technical methods used and on how people access and interact with digital content. With respect to the latter, digital information literacy as well as both formal and informal education can impact how individuals perceive content.

In this document, "synthetic content" refers to "information, such as images, videos, audio clips, and text, that has been significantly altered or generated by algorithms, including by AI."

This report provides an overview of technical approaches for provenance data tracking and synthetic content detection with issues for consideration, along with a review of the current testing and evaluation for digital content transparency techniques.

For selected techniques, the document identifies ongoing research and related research gaps. It also discusses technical mitigations for preventing and reducing the production and distribution of synthetic child sexual abuse material (CSAM) and non-consensual intimate images (NCII) and applies the concepts

discussed to the AI lifecycle as outlined in the NIST AI Risk Management Framework, or AI RMF (NIST AI 100-1).

The technical approaches described in this report provide building blocks that can be used to improve trust in digital content and the institutions and individuals who produce and disseminate it by indicating where AI techniques have been used to generate or modify digital content. None of these techniques offer comprehensive solutions on their own; and the value of any given technique is use-case and context specific and relies on effective implementation and oversight. Because this report focuses on technical approaches there may be normative, educational, regulatory, and market-based approaches not described in this report.

Science-backed standards forged through global actions, via international standards-setting bodies, several of which are mentioned in this report, can promote the adoption and interoperability necessary for these tools to have the desired impact.

There is no perfect solution to solve the issue of public trust and harms stemming from digital content, but additional, and improved approaches to synthetic content provenance, detection, labeling, and authentication techniques and processes are important capabilities to support trust between content producers, distributors, and the public.

2. Harms and Risks from Synthetic Content

Though synthetic content, may not be inherently harmful, it can accelerate and exacerbate pre-existing harms and negative impacts across the open information ecosystem, such as information integrity issues, synthetic child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII), fraud, and intellectual property and copyright issues. Taking a risk-based and human-centered approach to synthetic content, within the use case and context is important, given that there are benign use cases for synthetic content, and the approach adopted also depends on the audience.

The various issues that synthetic content presents affect how individuals consume information, can have negative effects on public safety and democracy. The negative consequences of synthetic content can uniquely and disproportionately impact individuals and communities who face intersectional discrimination and bias on the basis of gender, race and ethnicity, and other factors. The digital content transparency approaches discussed in this report on their own cannot comprehensively address the myriad of harms and risks that synthetic content poses but could be applied as tools to reduce harms and risks from synthetic content.

Synthetic content that supports misinformation and disinformation, synthetic CSAM and NCII, and fraud and financial schemes have concentrated or diffused effects depending on many factors. The spread of synthetic content that supports mis or disinformation narratives on social platforms is what makes it harmful and have diffused effects across a target population. Disinformation that is created by a malicious actor but is never disseminated across social platforms will not have its intended effects to shape perception. In comparison, synthetic CSAM and NCII is harmful at its creation with concentrated effects on specific individuals when such content depicts or appears to depict, real individuals, and could be used for sextortion schemes, blackmail, re-victimization, and more. Further, even when synthetic NCII and/or CSAM does not depict or appear to depict, real individuals, their generation and dissemination contribute to the normalization of gender-based violence and violence against children. Synthetic content could also produce concentrated harms by bolstering fraud and social engineering, and impose financial costs on victims of these schemes, while having diffused effects on wider markets, businesses, and the economy.

The harm and risks of synthetic content depend on factors including but not limited to *the severity of harm from the content itself, target audience for the synthetic content; context in which content is used or misused; and sophistication of the actor creating and/or disseminating the content; and any social, economic, and health-related (including mental health) costs incurred in association with the creation and/or dissemination of the content.*

Specific techniques may be suitable in reducing or limiting particular harms and risks. Provenance data tracking techniques that record the origin and history of digital content can be used to affirm both the authenticity of content, and in some cases, the authority of the entity who issued the content. Content authenticity does not directly translate to trustworthiness; authentic content that has provenance information available can still be harmful, depending on the content itself, nature of the source, and how it may be shared across platforms. However, these techniques may be useful for resourced good faith actors to secure their content and provide content transparency to their target audiences.

Synthetic content detection techniques may be more suitable for narrow use cases for analysts to determine whether specific adversarial content is AI-generated or not, or to detect covert watermarks in content. These techniques often have results that may be difficult for a layman or the wider public to interpret and may be more suitable as an approach for those conducting specific analyses, or for entities such as social media platforms, and specialized civil society organizations.

For high-risk or high-integrity applications, which could include election security, defense applications, CSAM/NCII investigations, and others, taking a defense-in-depth[1] approach by utilizing more than one method will likely be important for organizations, to mitigate potential overreliance on any one approach or technique. The application of digital content transparency approaches to mitigate harms and risks from synthetic content is still relatively new; these techniques will continue to evolve, and a variety of technical and sociotechnical evaluations are needed to guide their implementation.

Below is a table that outlines how different digital content transparency approaches and specific methods are currently applied and adopted, and how they could be used to mitigate harms and risks. Further information about these approaches and methods as well as their limitations are discussed in detail throughout the report. Lastly, more specific use case and context-based mitigations and controls for synthetic content are available in the Guidelines for Evaluating and Red-Teaming Generative AI Models and Systems and Dual Use Foundation Models[2] and the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.[3]

Digital Content Transparency approach	Example Methods	Current Applications	Current Adoption	Potential Use Cases to Mitigate Harms and Risks
Provenance data tracking	Metadata recording, digital watermarking	Determining content authenticity, the source or origin of content	Mainly for image and video, by high-resource software and media entities (with some hardware entities)	- IP protection via robust watermarks - Transparency about content origins and/or history
Synthetic Content Detection	Automated content-based detection, provenance data detection, human-assisted detection	Determining whether content is AI-generated, the presence and contents of provenance information	Diffused across industry, with some civil society adoption, mainly focused on deepfake detection.	- Analytical assessments of adversarial content through advanced multimodal detection - Public figure focused detection (deepfakes) - Detection of covert watermarks for developers and platforms

[1] Defense-in-depth refers to Information security strategy integrating people, technology, and operations capabilities to establish variable barriers across multiple layers and missions of the organization. See https://csrc.nist.gov/glossary/term/defense_in_depth

[2] TBA

[3] TBA

3. Current Approaches, Issues, and Opportunities

This section of the report describes current techniques and related issues and opportunities.

The most commonly used techniques to *directly disclose* to the audience how AI was used in the content creation process include:

- content labels (e.g., visual tags within content, warning labels, pre-roll or interstitial labels in video and/or audio, and typographical signals in text highlighting generated AI text with different fonts),
- visible watermarks (e.g., icons covering content indicating AI usage where the bigger the icon, the harder its removal), and
- disclosure fields (e.g., disclaimers and warning statements to indicate the role of AI in developing the content, and acknowledgments to provide more context to the AI contribution and credits to reviewers).

In contrast, *indirect disclosure* techniques require active effort to detect. These include:

- covert watermarks
- digital fingerprints, and
- embedded metadata.

They involve purposefully applying labels that are machine-readable and interpretable by technical systems. These are often identifiable by third-party entities and end users.

Both direct and indirect labels can be applied automatically during content creation or they can be applied post-generation.

Publishers and content platforms can use different types of labels to disclose content sources, such as clearly differentiated modalities of the generated content (image, text, audio, video).

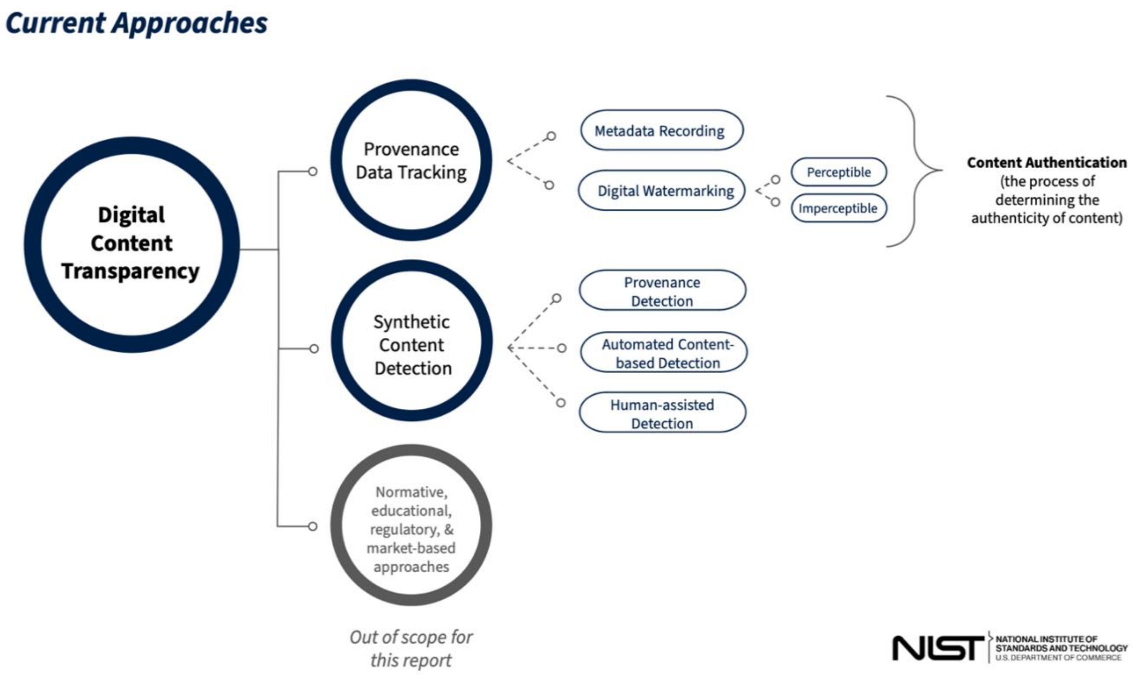

Current Approaches
Digital Content Transparency
Provenance Data Tracking
Metadata Recording
Digital Watermarking
Perceptible
Imperceptible
Content Authentication (the process of determining the authenticity of content)
Synthetic Content Detection
Provenance Detection
Automated Content-based Detection
Human-assisted Detection
Normative, educational, regulatory, & market-based approaches
Out of scope for this report
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Figure 1. Digital content transparency mechanisms can be broken down into provenance data tracking and synthetic content detection, each with multiple subcategories. Provenance Data Tracking can be applied to both synthetic and non-synthetic content where Synthetic Content Detection is employed to determines whether a given piece of content is synthetic or not.

This section highlights the current technical approaches for digital content transparency that this report covers in subsequent sections. The main techniques that the report discusses are provenance data tracking and synthetic content detection. Content authentication is not a technique, but rather a process that reveals the authenticity of all digital content (not just synthetic) by examining its origin and history. Therefore, it operationalizes provenance data tracking methods.

Content authentication	is a process that utilizes provenance data tracking methods (metadata recording and digital watermarking) to determine the authenticity of content by examining its origin and history. (i.e., the content has not been altered, or at least that the visual (or semantic) characteristics of the content are unchanged).
Provenance data tracking	records the origin and history for digital content, which assists in determinations about authenticity. It consists of techniques to record metadata as well as overt and covert digital watermarks on digital content. Provenance data tracking can help to establish the authenticity, integrity, and credibility of digital content.
Synthetic content detection	refers to techniques, methods, and tools used to classify whether a given piece of content is synthetic or not. Synthetic content detection may detect the existence of provenance information, such as digital watermarks, that was recorded, or it may look for other characteristics to help determine whether content has been generated or manipulated by AI.

3.1. Provenance Data Tracking

Provenance data tracking can help establish the authenticity, integrity, and credibility of digital content by recording the content's origin and history. It consists of techniques to embed or store metadata as well as overt and covert digital watermarks on digital content to indicate synthetic or authentic origins of content. Current methods for provenance data tracking include *digital watermarking* and *metadata recording*. These techniques vary in their implementation and their robustness across various types of content (images, text, audio, video).

3.1.1. Digital Watermarking

Digital watermarks have long been used to indicate content origins, with those shown on stock photography and other image previews being one popular usage. In terms of digital watermarking standards, The Advanced Television Systems Committee (ATSC), has produced well known set of standards for audio and visual content such as ATSC A/334 and ATSC A/335 (Appendix A).

Digital watermarking involves embedding information into content (image, text, audio, video) while making it difficult to remove. Such watermarking can assist in verifying the authenticity of the content or characteristics of its provenance, modifications, or conveyance. Watermarks can be either overt or covert depending on the content's audience. (See further digital watermark use cases and applications in the Appendix B and Appendix C.)

As an example, in an image watermarking system, a user would input an image, a watermark, and an embedding security key into an encoder to get a security key for extraction together with the watermarked image. The encoder algorithm controls how the embedding of the watermark will be applied to the image. On the other hand, a decoder uses the security key to extract the watermark from the watermarked image. Afterwards, the extracted watermark can be compared to the original watermark for verification.

Each type of watermarking has advantages and limitations.

Overt digital watermarks can be perceived directly by the human senses (e.g., a semi-transparent logo affixed to an image, text, or other audio, or video labels) by the audience of the content. An overt watermark may indicate the origin or source of content, including whether it was synthetically generated. If overt digital watermarks are limited to a small portion of the content, they can easily be cropped out or removed, diminishing their value and purpose. However, if these watermarks are applied across a large swath of the content, removing them can make that content too corrupted to be usable. In addition, overt watermarks may not be easily machine-readable, which can be a concern for identifying these watermarks at scale.

Covert digital watermarks are machine-readable watermarks involving subtle perturbations of the content that are hard for humans to detect. For example, a watermark can be embedded by altering the least significant bit (LSB) of some pixels in an image. These watermarks can be more secure than overt watermarks, as they are typically harder to remove. A covert watermark must be first detected to verify its presence by a digital watermark detector and then as applicable extract any information embedded within the watermark supporting the provenance of the content, which will have some nonzero probability of error. The effectiveness of a covert watermark is contingent on how accurately detectors can distinguish when the watermark is present and extract any additional information that may be included. Considerable research is focusing on how to embed *covert* watermarks into different types of digital content.

Digital watermarks are most effective when they possess the following attributes:

Low distortion	The watermark should not affect the quality of how a human would perceive the watermarked content compared to the original content
Robust	The watermark should be robust under various types of typical innocuous modifications—such as compression, filtering, or cropping—so that it can still be detected or extracted even after content has been altered.
Secure	The watermark should be secure against unauthorized attempts by malicious users to remove or tamper with the watermark information.
Sufficiently high-capacity	The watermark should have sufficient capacity to embed required amounts of information for its intended purpose, such as ownership information, copyright marks, or authentication data. A watermark may only need, in some use cases, to encode one bit (e.g., whether a given system generated the content). If more information is encoded in the watermark, it may be human-readable information, such as text or logos, or machine-readable information, such as binary codes or digital signatures. (In principle, a sufficiently high-capacity watermark could embed arbitrary metadata.)
Efficient	The watermarking process should be efficient and computationally feasible, allowing for fast and reliable embedding and detection of the watermark information.
Minimally disruptive	The watermarking process should be transparent to the user, meaning it should not require significant changes to the content creation or distribution process and maintain downstream uses.

This table highlights some of the design choices in utilizing watermarks for digital content:

Fragile or Robust	Watermarking techniques can be more or less robust to modifications and secure against attacks. Fragile watermarking methods are designed to become invalid in the face of any changes to the content, while robust methods are designed to withstand certain types of attacks or modifications.
Overt or Covert	Overt watermarks, such as logos or text overlayed on an image, are visible or audible to the content's normal audience, while covert methods are designed to be detectable only by those actively looking for them.
Blind or Non-Blind	Watermarking techniques can be blind or non-blind based on whether the original content is required for detecting the watermark. Blind watermarking methods do not require the original content for detection, while non-blind methods do. Non-blind watermarks add extra security to the content as it needs the original content to verify the watermark (such as for copyright use cases on licensed images), while

	blind watermarks can be more suitable for data hiding applications (such as covert communication) or even preventing the sharing of protected media online.
Private or Public	Watermarking techniques can be private or public based on the availability of the algorithms or cryptographic information needed to apply or validate the watermark.
Reversible or Irreversible	Reversible methods entail embedding the watermark into the digital content in such a way that the original content can be restored given the needed information to extract the watermark and retrieve the original content. In irreversible methods, the semantic distortion that is caused by modifications to the content cannot be reversed. (This distinction does not apply to watermarks that are applied during generation, where there is no original to revert to.)

3.1.1.1. Technical methods for covert watermarks

Methods for covert watermarking of GAI outputs must modify some characteristic of the content that can be subtly perturbed; this typically results in a change to the statistical properties of the content (such as perplexity measuring uncertainty in predicting the next word, and burstiness measuring the variation of sentences in language models). There also must be a systematic way of perturbing these characteristics so that the watermark can easily be generated, and a detector can recognize with high probability both when it is present and when it is not.

Below are some examples of *properties* that can be perturbed, along with the applicable types of content and examples that leverage these properties. These examples of properties that can be perturbed should be connected to the above design choices table to account for specific contextual use cases.

	Explanation	**Potentially applicable to**	**Examples**	**Stage of application**	**Risks and technical limitations**
Individual samples (pixels, audio samples)	Predictably chosen pixels or audio samples can be altered to embed content such as a watermark. To minimize perceptual distortion, modifications can be limited	Image, audio, video	LSB-based watermarking	Applied post-generation	Vulnerability to attacks (e.g., compression, cropping, filtering, scaling), overt distortions in the content, limited robustness, security concerns including a tradeoff between capacity and imperceptibility, and dependency

	to a small and relatively unimportant portion of each selected pixel or sample, such as the least significant bit (LSB).				on the host media (e.g., texture of images)
Frequency coefficients	Every piece of content that consists of samples laid out in time and/or space can be re-represented in terms of spatial or temporal frequencies instead of individual samples. The balance between some of these frequencies can be perturbed with minimal impact on human perception, much as JPEG compression discards some of these frequencies from images with little impact.	Image, audio, video	Discrete Cosine Transform (DCT) watermark, Discrete Fourier Transform (DFT) watermark	Applied post-generation	Vulnerable to geometric attacks (cropping, scaling, rotation), and requires high computing resources and processing time to run.

Initial noise output for diffusion models	Many recent GAI models are based on "diffusion models," which start from a full output that consists of random noise, then iteratively refine the noise into an output matching the prompt. The initial noise output can embed a predefined pattern, which can later be recovered by someone in possession of the model .	Image, text (in principle), audio, video	Tree ring watermark	Applied during generation	Robustness against GAI-based removal methods and attacks, Flexibility in message embedding (e.g., fixed vs dynamic messages), Security risks and privacy concerns, Computations and economic costs, Applicability to various modalities, and it is more suitable in private settings.
Token probabilities	Large language models typically generate text one "token" (or sub-word chunk) at a time. The probabilities of different tokens occurring can be used to embed information.	Text	LLM watermark	Applied during generation	Robustness against modifying text attacks, Perplexity and quality degradation, Coordination between LLM provider and detector, and scalability to long text contents

For many of these properties, a variety of techniques can be used to systematically perturb them into a watermark. Example methods include:

	Explanation	Potentially applicable to (properties)	Examples	Risks and Technical Limitations
Direct encoding	Where the element to be perturbed is a piece of data that is contained in the output, the watermark data can be embedded directly as replacement data.	Individual samples, frequency coefficients. For example, the LSBs of image pixels can be replaced with watermark information. (This would not work for methods that perturb the generation process, as that process is not directly encoded in the output.)	LSB-based watermarking, Discrete Cosine Transform (DCT) watermark, Discrete Fourier Transform (DFT) watermark	Can affect imperceptibility, Computational complexity, Detectable alterations in the signal, Security risks due to ease of watermark removal, Robustness risks due to transformations in the watermarked content, and low embedding capacity leading to inadequate embedding of information;
Cryptographic hashing or encryption	A cryptographic hash function can be used to generate a "hash value," a pseudo-random number, that determines how perturbations are performed.	Individual samples, frequency coefficients, token probabilities For audiovisual content, a hash of the original image, or data derived from it, can be embedded via direct encoding. Hashing can also be used for text watermarking: at each step, the hash value is used to designate "red" and "green" lists of tokens, and then the model preferentially selects a next token from the green list in a largely covert but statistically detectable way. To enable private operation, an	Robust hashing for visual watermarking, LLM Watermarking	Fragile to minor changes in the content as cryptographic hashes are highly sensitive, Limited amount of embedded information due to a fixed-size hash, and combining cryptographic hashes with watermarking adds more complexity in implementation compared to standalone watermarking techniques.

		encryption cipher with a key known only to the model operator can be used as the hash function.		
Machine learning	A machine learning system can be trained to perturb a piece of content in a way that is reliably detectable. The difference between this technique and direct encoding is that the perturbation happens during the data generation process, and not after the output is generated.	Any GAI systems can be fine-tuned to generate recognizable watermarking patterns in the course of generation. Machine learning perturbation methods usually require training an accompanying machine learning-based detector. These methods are easiest to use for private watermarks or where the watermarking algorithm is not known publicly.	Stable Signature, (commercial tool)	Computationally intensive, May introduce lack of interpretability/explainability of the embedding and detection process, Performance may degrade for data outside the training distribution, and vulnerable to deepfake generation networks to remove the watermarks.

3.1.1.2. Additional Issues for Consideration

Technical trade-offs: Watermarking techniques may require trade-offs between:

Robustness (the durability of a watermark) against adversarial uses and computational complexity (the resources required to implement watermarking). Less complex algorithms may not provide adequate durability against adversarial manipulations, with a negative impact on the security and performance of the watermark.

Robustness and low distortion: Another trade-off is between *robustness* and *low distortion*. A key challenge is ensuring that the watermark cannot be easily removed or altered while minimizing distortion. Typically, the mechanism by which a watermark is embedded entails the modification of components within digital content. As a result, the introduction of these changes in the digital content creates a verifiable signal that can be identified and extracted. Reaching a high degree of robustness can adversely impact content quality. Conversely, minimizing how much the watermark distorts the content could make it easier to remove the watermark (or, equivalently, harder to detect it which can lead to higher error rates)—which might decrease its robustness.

Capacity and quality trade-offs are also relevant. Embedding a watermark can reduce the quality of the content by making or inserting alterations that could corrupt the original digital content. Moreover, for some types of watermarks, increasing the capacity of the watermark further reduces the quality. Capacity refers to the amount of information that can be hidden in a watermark, without perceptibly distorting the digital content. Watermarks can be intrusive and negatively impact the visual and auditory components of audiovisual content or potentially the fluency or accuracy of text. For example, this is especially true of visible watermarks on color images, and even more so if the watermark itself is colorful: the watermark may interact in more complex ways with the color pattern in the host image than in a grayscale or binary image.

Adversarial tampering: Some watermarks, particularly those that are fragile or lack robustness, can be removed or tampered with, which may make them inadequate for purposes that require high integrity. As previously noted, overt watermarks applied to small portions of a piece of content can easily be edited out. Black-box attacks, or attacks conducted without watermark access, against digital watermarks using adversarial machine learning have demonstrated success in tampering with digital watermarks, even without knowledge or access of how the watermarking mechanism works. To date, researchers have shown the vulnerability of many covert watermarks to tampering and manipulation and how it is possible to evade or forge current watermarking methods. However, adversarial attacks against more robust forms of watermarking, such as watermarks that are cryptographically applied, are difficult to execute by comparison. There is some initial research that demonstrates how image watermarks can be designed to be robust against state of the art watermark attacks.

Most watermarking techniques involve software that must be run either after an output is generated or as an additional set of operations during the generation process. The watermarking behavior is not built into the model itself. Thus, if someone has access to the model's source code, they can easily modify that source code to disable watermarking; they do not typically need to retrain the model. In cases where the generative model itself has been trained to watermark, it may be possible to remove that behavior with limited additional training.

Trust: While digital watermarks can contribute to information integrity, they cannot guarantee it in a vacuum. Further research is needed on how digital watermarks may affect public perception or trust in digital media content. Also, false positives and false negatives can occur when using watermarks to authenticate the origin of content, reducing trust in the accuracy of watermarks and the watermarking process. Furthermore, watermarks can be exploited to create a false sense of security or trust in content if malicious actors are able to forge trusted watermarks or to add their own watermarks and apply them to misleading or untrustworthy content. Such an attack, if discovered, would likely also reduce trust in all similar content, impacting trust in the open information ecosystem. Synthetic content is a global phenomenon, which affects digital citizens around the world. There may be lower capacity to implement digital watermarking approaches in lower-resourced markets and regions, particularly for civil society organizations in these regions.

Scale: Many covert watermarking methods or protocols rely on unique, method-specific detectors. If different AI model developers created their own unique watermarking schemes, users may have to utilize multiple detection services created by these developers to know the source or origin of synthetic content created by different GAI tools, which can be inefficient and increase the burden on the audience, particularly for platforms. Open source standards and/or publicly available databased or resources that host multiple detectors would make the identification of watermarks streamlined, but come with security risks. There is also an educational barrier that must be addressed, for users to understand how to utilize detection tools for watermarks and interpret results. Furthermore, scaling is challenging for methods that involve cryptographic keys or machine learning tools. If one entity holds the keys or algorithms, they must be trusted and may become a bottleneck, either as a single point of failure or a process inefficiency. However, if the keys or algorithms are possessed by many entities, they could allow any actor to apply the watermark and permit bad actors to sidestep watermark generation by repeatedly generating content until they find an output that can fool the detector. Detection tools that are open source or otherwise not subject to rate limiting may be particularly susceptible to such an attack.

Opportunities for Further Development: Considerable research is focusing on how to embed covert watermarks such as statistical watermarks applied during generation, onto different types of digital content. More research is needed to understand how the application of digital watermark labels may affect public trust in digital content and the risks of inadvertent harms. For example, people with disabilities and those with limited language skills regularly using generative AI to create content may be discriminated against if the content they publish on platforms is labeled as AI-generated using watermarking, given potential existing issues of trust and credibility in AI-generated content online, and in relation to the context and use cases for generation. More research also is needed on emerging watermarking techniques such as statistical watermarks, ways to improve scale for detection techniques, how to improve watermark security through advanced cryptography that reveals minimal information to watermark detectors and future advanced cryptographic techniques such as zero-knowledge proofs, and developing best practices for implementation.

3.1.2. Metadata Recording

Metadata can provide information about a set of data and its content and contribute to digital content transparency. Metadata can be generated whenever digital content is created, uploaded, downloaded, or modified. It can be stored either internally in the same file or structure as the data (embedded metadata), or externally in a separate file. Almost all software applications use some type of metadata, including for document management, social media, emails, websites, databases, and geospatial objects.

Metadata can generally travel as part of the data it describes and provides information about the content's properties, structure, origin, purpose, time and date of creation, author, location, standards, file size, quality, versions, editing history, and other details. Thus, metadata can be applied to all media types (images, text, audio, video) but can be manipulated by anyone. These properties can improve the accuracy of metadata, since the metadata should be readily changed whenever data is changed. However, metadata is often stripped when files are shared, such as via social media platforms. Metadata generally can also easily be wiped, often for privacy reasons such as when content is uploaded on social media platforms or through adversarial tampering.

Metadata also can be used to help differentiate between authentic and synthetic or manipulated content, contributing to data integrity. Metadata recording approaches can also explicitly indicate synthetic origins of content.

Using Digital Fingerprints to Identify Metadata: Digital fingerprints, which are hashes that are predictably generated from the content itself, can also be used to generate unique identifiers to which metadata can be associated externally to the content itself. Digital fingerprints are commonly used across the technology industry to tag and identify known harmful, illegal and/or sensitive content, especially image content, through the sharing of content through hash databases between technology platforms, civil society, and other entities. Hashing (including both cryptographic and perceptual hashing) allows information about content to be shared without sharing the content itself, which serves to preserve privacy. Several databases and tools have been created to store hashes of harmful and/or illegal images and metadata about these images.

Two notable examples of the use of digital fingerprints are the Global Internet Forum to Counter Terrorism (GIFCT) and Tech Against Terrorism. The GIFCT uses its hash-sharing database to rapidly identify and share signals of terrorist and violent extremist activity with all of its member organizations, which include many large technology platforms. Tech Against Terrorism's Content Analytics Platform (TCAP) works similarly and automates the detection and removal of verified terrorist content on technology platforms, by hashing the content as well as via metadata about the content. (Later sections of this document on preventing and reducing the generation of synthetic CSAM and NCII discuss the use of hashes for content in greater detail.)

The most common types of metadata used for tagging or labeling digital content include:

Descriptive metadata	provides some descriptive information for discovery and identification such as file type, author, title, language, date created, and other specifications.
Structural metadata	provides logical and physical structural information about the containers of data and indicates how compound objects are put together—for example, how frames are ordered to form a video. It describes the types, versions, relationships, and other characteristics of digital materials.
Administrative metadata	provides information about the source of the content, its ownership, copyrights, licensing, and control permissions for easier management of the resource
Technical metadata	provides technical information like runtime, file type, size, resolution, color space, encoding format, compression algorithm, and other specifications.
Provenance metadata	provides information on the origins of a data resource, ownership, any transformation that the data may have undergone, usage of the data, and the archive of the data resource. This information helps track the lifecycle of a resource. Provenance metadata is generated whenever a new version of a data set is created and indicates the relationship between different versions of data objects. This allows users to query the relationship between versions and includes either or both fine- or coarse-grained provenance data on data resources.

Appendix C includes some of the most commonly known metadata standards across specific and multimodality data types.

3.1.2.1. Authenticating Metadata

Metadata can be cryptographically signed. A cryptographic or digital signature is "an electronic analogue of a written signature that provides assurance that the claimed signatory signed, and the information was not modified after signature generation." When metadata is signed with a digital signature, it can provide confidence about the contents of the metadata by determining the authenticity of electronically stored information.

A digital signature algorithm includes a signature generation process and a signature verification process, to provide assurance that the claimed signatory signed the given piece of information. A signatory uses the generation process to create a digital signature on data via a private key, which is kept secret. The verifier then uses the verification process via a public key that corresponds to the private key to verify the signature. In addition, the checksums and/or a digital signature can be embedded as metadata to verify the integrity of a digital content, allowing users to verify that the content has not been altered since its creation.

Utilizing digital signatures to sign metadata increases integrity, security, and tamper-evidence of the metadata. Unsigned metadata without verifiable credentials is not tamper-evident nor has it been stored with secure encryption. Metadata that is proactively embedded in content is more secure when it has been validated by digital or cryptographic signatures. (See below for additional considerations.)

3.1.2.2. Metadata and Content Authentication

Metadata can be used to verify the origins of content and how the history for a piece of content may change over time. Current entities creating specifications for metadata to verify content authenticity include the Coalition for Content Provenance and Authenticity (C2PA) and the International Press Telecommunications Council (IPTC). Further, secured metadata information that is disclosed to users can assist with information integrity and increase confidence in the content issuers' digital identity.

Provenance data tracking for metadata is only comprehensive if the software or hardware used to generate digital content and any other platform or tool used to modify or publish the content uses the same interoperable framework for retaining and securing metadata and establishing confidence that a particular entity issued the content.

For example, the IPTC has updated its Photo Metadata User Guide to include guidelines for using embedded metadata to signal "synthetic media" content created by GAI systems. They have developed the "digital source type" vocabulary, which now covers a range of AI-generated types such as:

Trained algorithmic media	is created using a model derived from sampled content.
Composite synthetic media	is a composite that includes synthetic elements.
Algorithmic media	is created entirely by an algorithm not based on any sampled training data (for example, an image created by software using a mathematical formula).

Some industry stakeholders (e.g., Google, Midjourney, Shutterstock) are starting to adopt the IPTC metadata in their outputs.

In an ideal interoperable digital environment, individuals would have access to a piece of content's chain of provenance information in order to maximize content transparency. For example, metadata attached to an AI-generated image would convey its origin and what time it was created, along with the artist's name (attached in an opt-in manner). When that image is posted on a social media platform that has opted into an interoperable framework for processing metadata, the metadata would be available to users interacting with the image. In practice, however, this is challenging to implement and scale for various reasons, such as platforms stripping metadata for privacy and data management reasons. Some representative research systems and prototypes were proposed in 2018, 2019, and 2021.

3.1.2.3. Additional Issues for Consideration

Privacy: Without privacy mechanisms and protections used in tandem with metadata recording, individuals and organizations could experience sensitive metadata leaks and violations of privacy. For example, if users are not aware that metadata is embedded at the capture or generation of synthetic or authentic content, they may inadvertently reveal private information about when and where an image was taken, and with what device. Furthermore, it is generally recommended that all metadata recording solutions include a process for users to opt-in and determine which metadata can be removed for privacy concerns. Systems that host metadata information should also ensure that privacy mechanisms are in place to prevent privacy leakage through the visibility of sensitive metadata across the network.

If metadata attached to content lacks these controls, then user privacy—especially for vulnerable populations—could be at risk. Malicious entities could co-opt metadata recording solutions to appear to promote transparency, while not providing any opt-in mechanism for tagging metadata and exploiting access to user information. Many platforms strip metadata from files on the Internet to prevent metadata leakage. Balancing the sharing of metadata for content transparency while also allowing users to take control over data that is shared is paramount.

Trustworthiness and Integrity: A recent study on provenance of digital content revealed users' lack of confidence in the trustworthiness of media when it did not have provenance information attached to it. Further, research shows that users do not clearly disambiguate provenance information from the credibility of the digital content itself, both of which demonstrate the limited and complicated role of provenance information in addressing the risks of mis- and disinformation in digital content. Lastly, the ability to tamper with metadata can undermine even the value it provides to people attempting to evaluate and understand content.

Security: Embedding metadata into content poses a wide array of concerns. Malicious attacks on metadata are possible even with secure infrastructure in place. Using a digital signature hardens the security posture of a metadata recording solution. The addition of cryptographic proofs for metadata can help prevent data tampering, as asymmetric encryption ensures that metadata has been secured by its signatory. However, encryption schemes and digital signatures are not foolproof. A variety of malicious attacks can be conducted on digital signatures to undermine their validity and trustworthiness. These attacks exploit parts of the digital signature creation system or the digital signature verification system. For example, the digital signature creation system does not necessarily protect the signer from signing a completely different document or piece of content. In this case, the attacker deceives the signer to sign a document that can benefit the attacker or be inconsistent with the signer's interests. Attackers could also modify information prior to the computation of the signature by adding or removing data before it is secured. These examples illustrate that even with digital signatures in place, vulnerabilities can be exploited to infringe on metadata security. Mitigation approaches for digital signature attacks include strong authentication measures, regularly updating digital signature software to ensure latest security patches, verifying the authenticity and validity of digital certificates before accepting digitally signed documents, encrypting sensitive data at storage and transmission times, and performing regular audits.

Metadata management and quality: The technical challenges for metadata management include the need for organizational processes, such as metadata management principles, and security solutions while optimizing for system performance and reducing latency. The value of embedded metadata is contingent on processes to create, input, and manage it. For example, systems to track metadata (or any provenance technique) will be more successful if they are interoperable across different platforms and metadata is not stripped. How external systems interact with a system that tracks metadata is an important consideration. Organizations may choose to establish principles to manage and secure metadata at an organizational level. Those principles could include guidance about how metadata descriptions can be constructed to be useful without being exhaustive; that also could help with scalability and the understanding of metadata labels. Exploring and deploying techniques such as digital signatures while reducing computational or communication overhead and/or latency costs can assist with implementation.

Furthermore, as hardware, software, and file formats become outdated, the need for continued accessibility necessitates migration of metadata to new platforms or systems. This may be addressed in part by storing metadata in formats that are resilient to technological changes and compatible with future systems. The compatibility of metadata is also bolstered by the use of a standard and open format for usable and reusable metadata.

Lastly, the completeness and accuracy of metadata is important in its management. Metadata completeness refers to the presence of all possible relevant attributes and information necessary to describe a digital resource adequately. This includes descriptive details, administrative information, structural relationships, and technical specifications. Incomplete metadata can result from manual entry errors, lack of standardized guidelines, or automated processes that fail to capture all relevant information. Metadata *accuracy* refers to the correctness and reliability of the information contained in the metadata. Accuracy can be compromised due to human error, outdated or incorrect information sources, or inconsistencies in metadata creation practices. Incomplete or inaccurate metadata can lead to unreliable descriptions of digital content and can weaken digital content transparency.

Opportunities for Further Development: Further research is needed to understand how metadata recording may impact user privacy and security, security of the metadata itself, how to mitigate adversarial uses and modifications of metadata recording, how it may impact trust and information integrity in digital content, the development of robust and open metadata standards, and how to develop best practices on completeness vs accuracy tradeoffs and scaling issues such as migrating metadata in new platforms or systems in metadata management.

3.1.3. Effectiveness of Provenance Data Tracking Techniques Across Different Types of Content

This section describes how provenance data tracking approaches vary in their current levels of robustness and effectiveness across different types of content and applications.

3.1.3.1. Images

Synthetic images are widely recognized as contributing greatly to harms from misinformation, disinformation, CSAM, and NCII. Provenance data tracking techniques are further developed for images than for any other medium, though adoption still remains low to reduce synthetic content harms. Frameworks such as C2PA, as well as ongoing digital watermarking research, largely focus on provenance data tracking approaches to images. Images can be manipulated in various ways, such as by altering pixels or by adding overlays, which also makes the medium better suited than others to provenance data tracking approaches.

Early research shows that even for robust covert watermarking protocols, it is possible to remove, alter, or generally manipulate watermarks. Some researchers also report that a family of regeneration attacks on invisible watermarks applied to images can render watermarks ineffective. Further, bad faith actors could apply watermarks to untrustworthy content, both authentic content and malicious synthetic content, to undermine information integrity. There are similar issues with the potential abuses of embedded metadata: actors could utilize existing provenance specifications to infringe on user privacy, and reduce information integrity broadly, as discussed in previous sections.

3.1.3.2. Text

Text is considered by far the most difficult modality when it comes to maintaining provenance given the nature of text—it is far easier to modify a pixel of an image with minimal distortion in comparison to a word. Provenance data tracking methods for text can be more challenging, given that structural modifications to text content could be easier to spot and subsequently removed. This can also be affected by the structure of written contracts, government official documents, blogs, news reports, and other text material. Much of the reported work on provenance data tracking for text focuses on differentiating synthetic text from human-written text. The main tracking methods proposed to deal with this issue include watermarking; perplexity estimation; negative log-likelihood curvature; stylometric variation methods (differentiating between human linguistic style and structure compared to AI text style); burstiness estimation (differentiating between the word choice and vocabulary size of humans compared to AI text outputs); and classifier-based approaches (building classifiers based on training data of human-written text and AI-generated text).

All provenance data tracking techniques discussed in this report when applied to text have limitations and can be vulnerable to tampering. For example, watermarking methods can be defeated or weakened through paraphrasing by humans or by machines. When it comes to perplexity and burstiness estimation, some research has shown that provenance techniques are not reliable metrics or indicators of human writing—especially in settings such as academic writing or with non-English languages. In many cases, the detection algorithm needs to keep track of specific features, which is computationally expensive and unrealistic to implement. Even with humans, each individual has their own writing style, and this can make it difficult to depend on a universal human writing style guide or feature set to support algorithmic detection. Finally, classifier-based methods generally target specific models by training on samples of their generated text or by utilizing the model itself, therefore their ability to classify new text from unknown models can be highly degraded.

3.1.3.3. Audio

The recent proliferation of AI-based synthetic audio has had great impact on applications such as voice assistants, text-to-speech, voice authentication, music, audiobooks, and podcasts. Meanwhile, synthetic voices created a new category of GAI models related to voice impersonation and synthetic audio recordings, raising concerns about the negative impact of audio deepfakes.

Several watermarking algorithms for audio have been explored. Most fit into two categories: *frequency domain* and *time domain* methods. The former takes watermarks and embeds them into transform coefficients, which are inverted to robustly conceal a watermark within an audio file. Time domain methods—where watermarks are embedded by modifying host signal samples—are simpler, but can lack robustness. The main issues with existing techniques for audio watermarks are robustness and computational costs, especially when considering long-duration audio. There are also newer techniques for audio watermarking, including using a trained neural network, which adds covert perturbations to the original audio in order to produce the watermark. Metadata can be added to audio files when AI-generated audio is created and can be cryptographically-secured—though as discussed in previous sections methods exist for manipulating embedded metadata.

3.1.3.4. Video

Risks regarding video authenticity have emerged as a public concern due to the rapid development of video generation tools. A digital video provides the appearance of movement across time. This makes digital video processing data intensive and requires significant bandwidth, processing power, and storage.

The process of extracting and finding evidence from a video to confirm its authenticity or integrity is known as video forensics. Many theories and methods used in video forgery detection are borrowed from image forensics. Although it is possible to analyze a video frame by frame using image forensics techniques, there are two reasons why this approach is ineffective: videos are more computationally demanding than images, and image-based methods may not be reliable for uses such as frame replications or deletions in videos.

Video tampering techniques generally can be divided into active and passive approaches. Watermarking and digital signatures are active techniques that verify content using features in the video. This data is then integrated into the video content at the moment of recording or capture and communicated to the receiver. However, tampering can occur before the digital signature or watermark is applied. On the other hand, with tamper-evident watermarks, in cases when the video is edited, this may suggest that the video has been manipulated. Another category of techniques actively enables devices (e.g., cameras, video recorders) to insert metadata information about the video source at the moment of capture. These are relatively new and not yet widely used, although they are expected to gain more attention in the near future. There are tradeoffs among watermark capacity, invisibility, and robustness. For example, increasing the capacity (i.e., embedding more information) requires altering more components in the host content which can compromise invisibility, while high robustness might require limiting the capacity to allow for more difficult to detect watermarks. On the other hand, increasing invisibility may require embedding watermarks in less obvious ways leading to potentially lower watermark robustness.

Opportunities for Further Development: Further research is needed to: understand how watermarks and metadata recording techniques can be abused by adversarial actors across all modalities of content, determine if provenance data tracking techniques for audio such as robust watermarking can be adopted, inform sociotechnical evaluations for how disclosures on audio content can be designed, and on how to improve the application of these techniques broadly for text content. There should also be further research on the sociotechnical effects and the effectiveness of labeling synthetic and authentic content, and any resulting impacts on the information environment.

3.1.4. Synthetic Content Detection

Synthetic content detection refers to techniques, methods, and tools used to classify whether a given piece of content or portion of content is synthetic or not. Synthetic content detection may rely on provenance information that was recorded, or it may look for other signals to help determine whether content has been generated or manipulated by AI. Reliable and robust methods for detecting synthetic content can mitigate and reduce harms and risks from the misuse of synthetic content when integrated within sound technical and social frameworks.

Detection methods relying on humans require extensive labor and high costs due to the large volume of data and are often subject to variations depending on individuals' lived experiences and expertise. The methods reflect a constant cat-and-mouse game between the detection and generation communities. As soon as a new detection method is created, models improve, and adversaries learn new ways to avoid detection. Furthermore, detectors are often tied to and may only perform well on specific generators.

Various tools are available to classify and detect synthetic content. Most are designed to detect content modifications or distinguish between AI-generated and human-produced outputs, and many utilize machine-learning and deep-learning detection techniques.

The DARPA Semantic Forensics (SemaFor) program takes a robust approach to detection by focusing and utilizing technologies that can detect, attribute, and characterize semantic inconsistencies in falsified multimodal media at scale. The DARPA SemaFor product also provides integrity scores to determine the probability that a piece of digital content is manipulated and also characterizes the "why," or what the intent of the multimodal media content could be.

Synthetic content detection techniques can broadly fit into three categories.

Automated content-based detection techniques are applied to identify synthetic content after it has been generated. These can include several different types of classification techniques that are designed to identify and separate synthetically-generated image, text, audio, and video, from authentic content across these modalities.

Provenance data detection techniques are used to identify digital watermarks (both overt and covert) embedded into synthetic content. (See earlier descriptions of provenance data detection.) Covert watermarks are machine-readable, while overt watermarks may be more difficult for detection algorithms to detect, given that they may not be machine-readable. Manipulations of digital content can also be traced utilizing metadata for synthetic content and deepfakes.

Human-assisted detection refers to the human-in-the-loop methods used in the detection process. It involves the cooperation of AI tools, crowd workers or data workers who handle and label data, and domain experts to improve the accuracy, explainability, and robustness of synthetic content detection techniques. Human-in-the-loop methods can be used for a wide range of contexts, including to validate and assess detection model outputs, though the evolving sophistication of GAI models may change the effectiveness of human labels in discerning whether content is synthetic or not. Human-in-the-loop methods may involve and augment content-based detection and provenance data detection methodologies above.

3.1.4.1. Issues for Consideration for Detection Techniques across Mediums

Issues for consideration for all detection techniques across modalities are summarized below.

Generalizability and Practicality: Incorporating diverse data, using ensemble models, and enabling continual learning are important strategies for improving the generalizability of detectors in real-world scenarios. As the amount of data is larger, computation power for such detection models still needs to improve for practical operational environments. Understanding the computational complexity of detectors is important for optimizing their performance and suitability for real-world applications.

Interpretability and explainability: Interpretability is crucial for synthetic content detection. Users must be able to form a coherent representation of the result that helps them understand how to act on it. (For example, they may need information about uncertainty.) In addition, it can be helpful for the results to be explainable so that end users are able to understand the mechanisms by which a model produced the decision.

Reproducibility: When using original data, code, and analysis, it is important for independent researchers to produce the same or similar results as the original experiment or method. The trend towards reproducible results can be promoted by providing the public with comprehensive datasets, human scores/reasons, experimental setups, and open-source tools/codes.

Comprehensive Data Inputs: There is a lack of benchmark datasets that can comprehensively evaluate existing detection technologies. These datasets should include real-world noises, diverse languages, compression, post-processing, and transmission methods. In addition, reusing synthetic content as input in subsequent model training can pose a challenge to detection technology. The accuracy and reliability of detectors can be improved by measuring the ambiguity of inputs and conducting further studies.

Robustness to security, privacy, intellectual property, and bias: The risks of synthetic content raise concerns in various domains related to security, privacy, intellectual property, and bias. GAI models rely on vast amounts of individual data, including sensitive information, which can lead to data breaches and unauthorized access to personal information. Adequate measures addressing those risk factors and developing robust detectors for synthetic content can help improve content integrity.

Incorporation of human-assisted techniques: Human collaborative decision-making is helpful in refining the task of synthetic content detection. For example, humans can help train and fine-tune detection AI models over time by providing feedback and correcting errors which can ultimately enhance the accuracy of their performance.

The detection methods for different modalities of content described below include techniques that fall into these various categories.

3.1.4.2. Synthetic Image Detection

Synthetic image detection refers to the process of identifying images that have either been generated by AI or manipulated using generative models such as Generative Adversarial Networks (GANs),diffusion models including their text-to-image products, neural radiance fields (NeRFs), variational autoencoder (VAE), among others. Given the rapid advances in image synthesis technology, there is a need to detect manipulated visual content in various application domains to preserve information integrity.

Synthetic Image Detection Techniques

Detection challenges arise from the highly realistic visual quality of synthetic images, and also the complexity of evolving AI and manipulation techniques. Systems for detection must be continuously improved to accurately detect synthetic images generated by rapidly advancing models. There are different ways to categorize synthetic image detection methods. The forensic community has often used a twofold of an active and passive detection method for identifying between authentic and synthetic images. Active detection methods focus on detecting whether an image is authentic or not or if it is forged by analyzing information hidden in an image at the time of its capture, using techniques such as watermarking, digital signatures, and cryptography. Passive detection methods, on the other hand, do not rely on any additional information in the image. Instead, they aim to find traces left (e.g., image pixel regularities or inconsistencies, tampering operations) during the image processing phases.

Recent studies in June 2020, February 2022, and May 2022 have focused on detecting deepfake images by using deep learning models, machine learning models, and statistical models. These methods describe some details in the Synthetic Video Detection Techniques below.

Some other researchers employ the following techniques for synthetic image detection.

Backbone models are pretrained networks that extract features from input images. These models comprise several layers of CNNs, including convolutional and pooling layers, and activation functions that are stacked to gradually minimize the spatial resolution of the input image while increasing their depth. These models can be used to differentiate between authentic and synthetic images.

Fake face detectors train models on face images and use differences in frequency statistics or global image features to distinguish between authentic and synthetic face images. General synthetic image detectors use special designs to classify general images, removing the limitation of face content. **Quality-based sampling detectors** involve training detectors on realistic synthetic images selected based on their quality scores according to a probabilistic quality estimation model. The method can lead to higher detection performance across various concept classes, such as training a detector on human faces and testing on synthetic animal images, thereby enhancing the overall effectiveness of synthetic image detectors. Furthermore, a practical guide discussed how adding synthetic images to object detection models can greatly improve their performance, especially when combined with authentic images. Utilizing this can enhance the performance of synthetic image detection models.

Manipulation Trace methods involve analyzing digital correction, overlapping, file format and structure analysis, metadata, and other enhancing effects to identify any inconsistencies or traces of manipulation. There are various traceable tools available such as Traces Extraction Network (AMTEN) and Manipulation Classification Network (MCNet) for detecting synthetic images.

Reverse Image Search/Trace methods involve searching for GAI model fingerprints using reverse image search engines, which predict network architecture and loss functions from the estimated fingerprints of the model used for synthetic images.

Synthetic Image Detection Performance

The accuracy of synthetic image detectors varies depending on the specific tool and the type of synthetic images being analyzed. The popular metrics used for image and video detection performance include accuracy metrics, as well as a graphical analysis such as ROC (receiver operating characteristic) curve and area under the ROC curve (AUC). AUC is a performance measure used to evaluate the classification capability of a model, especially when addressing imbalanced data, and is widely used to evaluate various AI models. Detection performance for synthetic images remains high without post-processing (accuracy between 52% and 76%, AUC between 75% and 93%). When synthetic and authentic images are post-processed (e.g., compressed and resized), as is common on social media platforms, detection accuracy decreases (accuracy is between 50% and 62%, AUC between 53% and 91%). The accuracy in this context indicates the proportion of true positives and true negatives among all evaluated detection cases. Experiments revealed that detecting a synthetic image by a specific generator is relatively straightforward. It can be achieved by training a binary classifier on a dataset comprising both authentic images and synthetic images created by that particular and only that generator, as the approach does not generalize well. Reported accuracy results on different training and test subsets using different methods range from 61% to 70%. Other performance measures demonstrate accuracy ranging from 50% to 62% and AUC from 52% to 91% with the post-processed images. This challenge is particularly prominent in real-world scenarios where the generator is often unknown during the training process, making it difficult to differentiate between authentic and synthetic images.

Additional Issues for Consideration

Robustness and Practicality: Synthetic data utilized for model training purposes, (which is distinct from synthetic content),used for detection may not fully capture the complexity and variability of authentic empirical data, which can limit the effectiveness of detection models trained on synthetic data. When detection models are trained on specific synthetic images they may not work well when applied to real-world scenarios, and may not be reliable if the images present artifacts that are significantly different from those seen during training. Post-processing, such as compression or resizing, exacerbates this challenge. The computational intensity of the detection models still needs to improve for practical operational environments. See details about synthetic image detection datasets in the Appendix D.

Societal Impact: In the application of synthetic image detection technologies, there may be wider societal implications and ethical considerations of the risks of synthetic content, and the design detection models to combat these risks. Some risks include impersonation, potential erosion of trust in institutions, synthetic CSAM and NCII, exacerbation of social divisions, threats to democracy and election integrity, and national security. When developing detection models, developers must consider these various risks and ensure that detection capabilities are built to detect harmful content that could have adverse societal effects and also work to improve detection accuracy for harmful images.

3.1.4.3. Synthetic Video Detection

Synthetic video refers to video manipulations, including deepfakes. A deepfake video is generated using machine learning or deep learning techniques to create realistic videos of real people in a malicious manner. Synthetic videos can also include manipulations to generate events that may not have ever occurred that could affect public safety, such as a false terror attacks or false natural disasters, or even fictional videos that are benign but do not reflect reality.

Adversaries can use available video manipulation tools for malicious impersonation, enabling fraud, creating misinformation and disinformation, and likely posing risks to democratic systems. Detecting deepfakes is becoming increasingly challenging due to their realistic nature and their rapid proliferation, leading to an “arms race” to develop new detection methods. Deepfakes generation may be categorized as involving: identity swap, attribute manipulation, expression swap, entire face synthesis, and source video.

Identity swap (or face swap)	is a method of replacing the face of a person in the target video with the face of another person in the source video.
Expression swap (or puppet master)	is a method of replacing the features of the mouth in the source image and producing a new face with the same identity but a different expression.
Attribute manipulation (face editing / retouching)	is a method of modifying some facial attributes (e.g., color of hair or skin, gender, age, adding glasses).
Entire face synthesis	is a method of generating a non-existing face or object.
Source video	is a method of analyzing the content of a source video to understand relevant attributes such as facial expressions and body language. The method then maps a voice recording to the video, making it appear as though the person in the video is speaking the words in the recording.

Synthetic Video Detection Techniques

Recent studies in June 2020, February 2022, and May 2022 have focused on detecting deepfake or manipulated videos by using deep learning models, machine learning models, and statistical models.

Deep Learning (DL) detectors	identify specific artifacts produced by their generation models. These models can extract or learn visual artifacts and features directly from the video frames. These features may include handcrafted features, spatio-temporal features, face landmarks, biological signal clues, among others, which help identify inconsistencies that may indicate the presence of a deepfake or manipulated video.
Machine Learning (ML) detectors	utilize feature selection algorithms to generate a feature vector, which is then used this vector as input to train a classifier to detect manipulations or deepfake videos.
Statistical-based	utilize different statistical measures, such as examining the shortest paths, photo response non-uniformity (PRNU) or mean normalized cross-correlation scores to distinguish between authentic and synthetic videos. Popular methods are Expectation-Maximization (EM) to extract a set of local features, Total Variational (TV) distance, Earth Mover's (EM) distance, Kullback-Leibler (KL) divergence, and Jensen-Shannon (JS) divergence, among others.
Forensics-based	detectors utilize the differences in frame-level features such as noise patterns or motion features. File structural analysis can be leveraged to determine the originality of a file employing unique device or GAI model characteristics

Spatial-based	detectors leverage the power of DNN models to capture the subtle differences or artifact clues between authentic and synthetic from the spatial or spatio-temporal domain.
Frequency-based	detectors investigate the differences or frequency artifacts between authentic and synthetic from the frequency domain.

Synthetic Video Detection Performance

Similar to image detection, the accuracy of synthetic video detectors varies depending on the specific method and the type of synthetic video. The performance is not robust to post-processing operations like compression, noisy effects, visible artifacts, among others. Although various studies show different performance results for synthetic video detectors, performance measures of deepfake detectors have shown for accuracy ranging from 62% to 99% and AUC from 82% to 98%. See details about synthetic video detection methods/results and datasets in Appendix D.

Additional Issues for Consideration

Generalizability: in general, synthetic video detection methods are trained for a certain data and compression level and demonstrate low generalization to unobserved datasets and scenarios, resulting in significant performance degradation.

Robustness: When dealing with low-quality videos, such as high levels of noise, low compression rates, or resizing, detection methods tend to perform lower when compared to high-quality videos. Adding a noise layer to the detection network that can account for different types of data degradation may improve system's robustness.

Computational cost: Processing time has become a critical factor due to the high volume of videos and media platforms for streaming. Future research should include how to develop efficient video detection techniques.

Benchmark and societal impact: there is a lack of standardized experimental methods that can facilitate meaningful comparisons among diverse datasets, scalability, and reliability of various detection methods. Additionally, there is a dearth of systematic or quantitative research on the perceptual and societal impact components that contribute to the deceptive nature of synthetic videos.

3.1.4.4. Synthetic Text Detection

The advancement of large language model (LLM) capabilities has made it difficult for humans to discern AI-generated text from human-written text, underlining a need for transparency about the use of LLMs in various contexts. LLMs are known for producing inaccurate or false outputs which have been called "hallucinations" or "confabulations," and they also can be used to generate false and/or misleading information at scale. For these reasons, being able to detect LLM-generated content is important to increase digital content transparency.

Synthetic text detectors use parameters based on text features such as language, structure, perplexity, and burstiness. Perplexity measures how well the model is able to predict the next word in a sequence of words. Burstiness measures how predictable a piece of content is by the uniformity of sentence length and structure. Some detectors rely on language models similar to those used in AI writing tools to evaluate the predictability and language patterns of the text. These content detectors are being used by some educators to check students' writing, by businesses to ensure the originality of published content, and by individuals to verify the authenticity of text on the Internet. Other detectors rely on factual inconsistencies (e.g., fact-checking database and reasoning models) and metadata analysis (e.g., anomalies detection in content metadata such as timestamps, location, and author information). It should be noted that the efficacy of many detection tools like these is being debated.

Synthetic Text Detection Techniques

Techniques shown in the Appendix D can be classified into the categories described as follows:

Watermarking detectors	have two components: embedding and detection. Embedding inserts a watermarked text (e.g., a hidden signal or pattern) into the output of the LLMs, which assists with provenance data tracking, while detection identifies the watermark from the AI-generated text.
Zero-shot detectors	detect AI-generated text with no need for prior training on labeled data or fine-tuning samples. The technique uses distinctive features and statistics (e.g., grammatical analyses, word density, structural attributes, constituent length, inconsistencies) as key indicators in distinguishing AI-generated text from human-generated text.
Fine-tuning LM detectors	use a fine-tuned Language Model (LM) method in detecting LLM-generated text. This involves taking a pre-trained LM model and adapting it to a more specific dataset or task at hand. It optimizes specific sub-components of the model with a loss function to detect errors or inconsistencies in text. Most approaches require paired samples for supervised training processes.
LLMs as detectors	use Instruction Tuning of LLMs for document and sentence text detection, enabling LLMs to detect generated text by leveraging their pre-training knowledge. The method involves cross-examining one LLM with another to discriminate text generated by either themselves or other LLMs, leveraging fluency and errors in the text.
Adversarial learning detectors	differentiate between human- and LLM-generated text by exposing them to adversarial examples, thus improving their accuracy. This involves the configuration of an attack model alongside a detection model, with the iterative confrontation between the two culminating in enhanced detection.
Human-assisted detectors	leverage both human and machine discrimination capabilities to efficiently distinguish between human- and LLM-generated text, utilizing human prior knowledge and analytical skills as well as learning from the model's behavior.

Synthetic Text Detection Performance

A variety of metrics are used to measure the performance of synthetic text detectors depending upon their use in different scenarios. Detection performance varies depending on the methods and datasets used. Watermarking technology has significantly advanced in recent years and can now frequently label and identify text generated by language models. Zero-shot detectors can enhance detection accuracy, and some LLM-based detectors are capable of exhibiting superior detection performance, robustness, and resilience to various attacks. Fine-tuning language models often tends to overfit their training data or the source model's training distribution, leading to a decline in performance when dealing with cross-domain or unseen data. Moreover, language model-based detectors are limited in handling data generated by different models since the detectors are fine-tuned on specific datasets for a given task. Human-assist annotators can improve their performance over time but have limitations with handling large volumes of data.

Overall, detection performance significantly decreases with various attacks such as paraphrase attacks, adversarial attacks, prompt attacks, and due to data ambiguity. While some initial research has shown that retrieval-based detection methods could increase the robustness of AI-generated text detection against paraphrase attacks, further research is needed to defend against different kinds of attacks. In addition, it is essential to conduct benchmark studies in diverse testing scenarios. Rigorous testing and evaluation can improve understanding of detectors' capabilities and limitations and aid in developing more effective strategies for identifying LLM-generated text. The existing detection methods for text are a work in progress and need further evaluation and improvement to align claims of high performance with actual robustness, reliability, and generalizability.

Additional Issues for Consideration

Robustness and Detection Quality: Robustness and detection quality are current issues for synthetic text detection. Most detectors have been designed for English-language text, and there is a need to optimize their performance across various languages. The performance of detectors decreases in real-world scenarios, highlighting the need to improve their robustness for practical applications. The quality of LLM-generated text is also affected by the complexity or learning of the prompts used, which can make it difficult for detectors to accurately identify text generated via elaborate prompts. This also makes it challenging for evaluators to measure detector performance.

Some detectors identify AI-generated text by analyzing parameters such as word occurrence, positioning, frequency, and style. However, they may not be able to distinguish between different types of GAI models. Additionally, there is a limit to the range and diversity of benchmark datasets that can be used to comprehensively evaluate existing AI-generated text detection technologies; see details in the Appendix D.

High-risk applications: Socio-technical issues for consideration include the usage of immature text detection techniques in high-risk applications. These applications may be included in academic settings or in the detection of AI-generated misinformation and disinformation. When detectors have been used in academic settings to confirm the academic integrity of writing, given a lack of accuracy, students have been wrongfully accused of cheating with AI technology, putting their academic futures at risk. False positives with AI-based text detectors have been reported as a clear issue with dangerous consequences. Similarly, text detectors can be inaccurate and imperfect tools for determining whether content is synthetic or not as well as determining whether misinformation and/or disinformation narratives may be AI-generated. This is especially problematic in non-English languages, as most detectors have been designed for English-language applications. Language model sophistication is also rapidly increasing, making detection a bigger challenge. Moreover, some research shows that using some of these detectors may not be appropriate in various scenarios.

3.1.4.5. Synthetic Audio Detection

As the quality of synthetic voice generation advances, the challenges and complexities of detection are increasing. There are two types of synthetic audio fields: Text-to-speech (TTS)-based and imitation-based.

The **TTS-based** method transforms text into natural speech in real-time via two steps. First, clean and structured raw audio is collected, along with a text transcript of the audio. Second, the TTS model is trained using the collected data to build a synthetic audio-generation model.

The **imitation-based** method transforms source speech (secret audio) so that it sounds like another speech (target audio) without changing the linguistic content. Its primary purpose is of the secret audio. To replicate the attributes of a specific voice, the style, intonation, or prosody of the spoken signal may be adjusted. This can be useful for applications such as voice impersonation.

In addition to traditional audio generation methods, some generation techniques exhibit voice fingerprint artifacts and inconsistencies that can be captured through frequency domain analysis over a spectrogram. Mel Frequency Cepstral Coefficients (MFCCs) are commonly used in speech-processing techniques. Using MFCCs has been shown to produce better results for synthetic audio detection than directly feeding the raw audio signal into the model.

Synthetic Audio Detection Techniques

Detection techniques can be divided into the following ML and DL methods. DL is considered a subset of ML that uses multi-layered neural networks to enable machines to learn more complex representations of data in a human-like way. :

ML detectors involves identifying speech patterns or detecting anomalies in features that deviate from natural speech characteristics such as acoustic and spectral content, pronunciation errors, formant frequencies, pitch variations, and background noise and inconsistencies. The method is limited by scalability with large numbers of audio files due to excessive training and manual feature extraction which requires extensive labor to prepare the data.

DL detectors leverages features such as formant frequencies, pitch variations, and tone nuances to identify discrepancies that distinguish a synthetic voice. The method can use metadata and background noise patterns to differentiate between authentic and synthetic voices and it requires specific transformations (e.g., audio features such as spectrograms) on the audio files when DL algorithms were used.

See additional details about synthetic audio detection methods and datasets in the Appendix D.

Synthetic Audio Detection Performance

The performance of synthetic audio detectors varies depending on the specific detection methods, the type of datasets, and the audio preprocessing techniques used.

Performance measures for synthetic audio detection have been conducted based on three criteria, equal error rate (EER), Tandem Decision Cost Function (t-DCF), and accuracy. EER is the point at where the false positive rate and false negative rate are equal and t-DCF measures the reliability of decisions made by the detectors. EER for audio ranges from 0.43% to 42.5%, with t-DCF from 0.008 to 0.39, and accuracy from 50% to 99%. The methods employed to generate synthetic audio data can impact the performance of the detection methods. For instance, one of the methods that has a very low error rate compared to other methods when applied to TTS-based datasets, performs poorly when applied to imitation-based datasets.

In general, ML-based detection methods provide better explanations and interpretations of the detection results while DL-based detection methods such as Convolutional Neural Networks (CNN) are considered more stable and consistent than the ML-based detection methods with respect to the dataset and synthetic data type.

Additional Issues for Consideration

Non-English Language Coverage: Most current research is focused on developing detection methods for identifying synthetic voices speaking in English. A detection model developed for a specific language may not perform equally well for other languages or dialects, especially for languages or dialects that have limited available data. Most detection methods focus solely on identifying synthetic audio, without accounting for accents or dialects. A lack of language coverage for audio detection could increase disparities in other parts of the world such as the Global South, especially around election periods, and could result in the amplification of harmful audio deepfakes in non-English languages.

Detection in Real-World Scenarios: Due to the wide range of synthetic speech generation technologies, it is still difficult to recognize some families of synthetic voice tracks in an open-set situation. The open-set scenario refers to detecting a synthetic voice even if it was generated using a previously unseen model.

Opportunities for Further Development for All Detection Techniques: Existing detectors primarily emphasize discriminating between synthetic content and human-produced content. **Intent detection and characterization** is a connected issue where there needs to be more research as the detection and characterization of the intention behind manipulated or synthetic content can greatly influence individual opinions or behaviors and widely affect the misinformation and disinformation spaces. While the DARPA SemaFor program has made some progress in addressing this challenge, there is still room for improvement in the widespread development and adoption of semantic intent detection technologies. Further research should also include investigating how to effectively improve detection performance on synthetic content that was post-processed or corrupted by noise, transmission, compression, or reformatted by a different social media platform. Specifically for **audio detection**, future research should also investigate what occurs if voice recordings are corrupted by noise, coding, or transmission problems, as well as synthetic voice recordings posted on social media sites or utilized live during phone calls. Lastly, more research is likely needed to assess the effectiveness of **human-assisted techniques** to aid detection efforts, such as in determining the effectiveness of human labels.

4. Testing and Evaluating Provenance Data Tracking and Synthetic Content Detection Techniques

Measuring the effectiveness of provenance data tracking and synthetic content detection techniques through testing and evaluation can identify issues with digital content transparency techniques.

A *test* is "an activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component."

An evaluation is (1) "a systematic determination of the extent to which an entity meets its specified criteria; (2) action that assesses the value of something."

The testing and evaluation of digital transparency techniques described in this section focus on the testing and evaluation of provenance data tracking and synthetic content detection techniques. It is common practice to measure the loss in quality from system output accuracy of a system (using a loss function) during training. Many loss functions and metrics for training overlap with metrics and models for evaluation. While much of the testing and evaluation is automated, there is some testing that involves human or manual examination of results, especially in contexts where subject matter expertise is important. See Appendix E for more details.

4.1. Testing and Evaluating Provenance Data Tracking Techniques

4.1.1. Testing and Evaluating Digital Watermarking Techniques

Digital watermarks are typically tested by attacking and measuring their resilience. Different kinds of attacks include removal attacks (i.e., removing the watermark without breaking the encryption or security); distorting watermarks to fool a detector; cracking security measures to remove the watermark, and forging watermarks.

Different ways to construct experiments and measure the robustness of watermarks include experiments running a set of attacks and measuring the percentage of attacks that destroy watermarks or percentage of the watermarks not detected. Another approach is to use image quality metrics to compare the image similarities between and among an original unwatermarked image, a watermarked image, and an attacked/changed watermark image. This experimental approach checks to see if the images are similar after a benign change (such as a decompression) yet are dissimilar if the watermark is attacked. Another experiment design uses image distance metrics to compare the distance or difference between watermarked images to their benignly-changed images, similar to digital fingerprinting experiments; a threshold distance is then set to identify which images are considered to be different.

Image similarity metrics more specific to watermarking include hiding capacity (HC), the number of bits that can be hidden in an image; Peak Signal to Noise Ratio (PSNR); and Structural Similarity Index Measure (SSIM), and generic image similarity metrics and image distance metrics (including the L_2 norm) are sometimes used. Another metric related to watermarking is the bits per pixel in an image, which determines how much information can be embedded in the image as a tradeoff to security of the watermark.

4.1.2. Testing and Evaluating Metadata Recording Techniques

One way to record metadata securely is through attaching it to a digital fingerprint of the digital content; the digital fingerprint is commonly achieved via hashing, as described earlier. The concept is that different images should have different hashes (and not collide), yet two copies of the same image should have the same hashes.

In addition, work is being conducted on the human testing and verification of metadata. One method is for humans to check the provenance directly for accuracy, verifying a subset of the provenance data by hand. Statistical tests on the provenance of the entire dataset used in training can complement human verification. For example, a query can determine what fraction of each data sample has a particular attribute. Provenance-generating software (or software that automatically generates metadata) can be tested for functional correctness by verifying that certain automatically generated queries regarding the provenance metadata return expected results. Provenance can also be evaluated by running provenance generating models on a known use case and manually examining what is generated.

4.2. Testing and Evaluating Synthetic Content Detection Techniques

Testing and evaluating synthetic content detectors can help build trust in those systems. The most common way to measure and evaluate a synthetic content detection system (or discriminator) is to construct an evaluation dataset that has appropriately label human-generated (authentic) inputs (e.g., images, videos) and synthetic inputs. The detector is queried to detect which images are synthetic; in this experiment, the detector will give a real number for each input indicating how likely the input is synthetic, with a higher number indicating that the input is synthetic. Then, this output mirrors the experiment discussed in Appendix E of this report and is scored on an accuracy metric. For detection tasks, two particular metrics are the Area Under the Receiver Operator Curve (AUC) and the Detection Error Tradeoff (DET) curve. This experiment design can be used to test all of the different types of synthetic content detection techniques.

4.2.1. Testing and Evaluating Automated Content-Based Detection Techniques

Detection testing requires careful consideration of the training and evaluation data sets used in designing the test. It is important to balance the types and relevance of authentic and synthetic content to be tested, as well as the sources of inputs. Systems should be tested for their performance in detecting images when they are resized and compressed. For images, the size of the images as well as the size of the image regions tampered with are important considerations when evaluation data sets are constructed.

When designing these detection test experiments, there are particular considerations as to how people testing synthetic content detectors construct the training set and the evaluation set (the set of inputs the detector is asked to determine which are authentic and which are synthetic). These concerns include a balance of classes and a variety of different relevant situations that are context-dependent, so that the dataset used is a balanced representation of the situation. As a result, evaluation datasets are often custom-constructed for detection experiments, and multiple evaluation datasets may be used to evaluate a detector. One consideration is testing detectors when the training and test sets are from the same pool of inputs, and when the training and test sets are from different pools of inputs. Another consideration is constructing data that tests image detectors to check that the detectors are robust to images when they are post-processed such as resizing and compression. For images, the size of the images as well as the size of the image regions tampered are an important consideration used when making evaluation data sets.

For video content detectors have been tested in situations where the video had frames inserted, deleted, or duplicated. For copy-paste detection (objects are copied and pasted within specific frames, which are intra-frame forgeries), the accuracies of various detectors ranged from 54.9% to 99.3% depending on the system and the complexity of the forgery. For detecting inter-frame forgeries with insertions, duplications, and deletions of frames, the accuracies ranged from 57.9% to 99.3% depending on the system, the number of frames inserted/deleted/duplicated, and the complexity of the forgery. Another design involves constructing the evaluation dataset to have images produced from specific attacks to fool the detector.

4.2.2. Testing and Evaluating Provenance Detection Techniques

Watermarking detectors can be tested in a fashion similar to testing post-hoc detectors: A watermark detector is given watermarked images and non-watermarked images and is asked to detect the watermarks. Here is an example of such an experiment testing watermarking detector.

Specific to watermarking detection, rather than using an accuracy metric, a different experiment can be designed where the watermark detector is given a set of inputs and is asked to obtain the watermark. The watermarked image obtained is then measured by its pixel correlation to the original watermark. In another experiment where the watermarks are statistical, synthetic images are generated, and the bit error rate of the watermark detector is measured and compared to theoretical optimums.

One way to evaluate metadata detection techniques is to take authentic media with authentic metadata and then inject false metadata onto the media content. The metadata detector is then tested on whether it can spot the false metadata.

4.2.3. Testing and Evaluating Human-Assisted Detection Techniques

Human-assisted detectors can be tested in a variety of ways; how the detector is tested depends on the form of assistance. One kind of human-assisted detector is an automated detector that is assisted by human-annotated training data. In this case, it is possible to compare the correctness accuracy of the system trained on the human-annotated data to the system trained on the unannotated data. For human-assisted text detection, one source augmented the training of large language models with human-annotated descriptions of different text errors within LLM-generated text, though this was not used to differentiate LLM-generated and authentic text.

Another kind of human-assisted detector is where the human is assisted with an automated model (through a user interface), but the human makes the final decision. For these tasks, the evaluation measures the human output. The human is measured by the time taken to complete the tasks and a subjective difficulty rating. The different interfaces, models can be swapped with other models to compare the influence of different machine assistance. This strategy is used to evaluate other human-assisted software.

4.3. Additional Issues for Consideration

Scope: The variety of mainstream testing only tests for Validity & Reliability (accuracy), Safe, and Secure & Resilient. As discussed in the NIST AI Risk Management Framework, there are additional trustworthy characteristics including Fair - With Harmful Bias Managed, Privacy-Enhanced, Explainable & Interpretable, and Accountable & Transparent. Many harms can arise when software is used but not checked in these areas.

Context: These systems are tested in experiments that are sometimes isolated from context. For instance, in what use cases is an AUC of 0.95 effective and in what use cases is this number bad? When will an improvement in AUC score actually reduce the harms of bias, discrimination, fairness, and other issues? As the system's output is rarely the final decision, there is an entire scenario and context involving humans where inaccuracies of a system might have different impacts depending on how the humans use the system output when making decisions. Lastly, as software is used by humans in different contexts, the context and other non-technical concerns shape the impact of the accuracy of the system, for example if a system's outputs are used in high-risk use cases such as for employment, or utilized in battlefield environments, high accuracy across different metrics may be more important.

Quality: A third concern comes from the testing of attacks and defenses. As many evaluations measure the quality of techniques with an attack-and-defense style, there is a concern that the quality of defenses is indirectly measured by the quality of attacks. Consequently, there have been instances where defenses tested to be good by one series of attacks were broken by other attacks. As attacks get better and adapt to current defenses, new defenses may then be developed as they adapt to these newer attacks. This reflects the commonly-known cat and mouse game that can occur between attackers and defenders, particularly in the realm of detection techniques. Defense-in-depth strategies may be needed, where multiple approaches are applied depending on use case, with various security mechanisms in place to reduce the unauthorized access of watermarks or metadata, for example. Though some of these tests are being done, there is still a gap with the attacks that exist and the attacks that mainstream tests often cover.

Opportunities for Further Development: More socio-technical research and evaluations to understand how people interact with digital content transparency approaches across various types of systems and in varied environments across the Internet will be helpful to design and implement techniques effectively. There have been some initial studies done on how humans interpret provenance labels attached to content and how labels may affect the perception of content, such as research done on how disclosures that news media was AI-generated may affect perception of trustworthiness (and reduce trust in news), and another study on how provenance-enabled media is contingent on design choices, and how users may have difficulty in understanding provenance labels on content. More studies would be helpful to understand how various content authentication techniques can affect how people across various demographics interpret digital content, current societal disparities that may affect the adoption of provenance data tracking approaches, how provenance labels may affect (if at all) victims and survivors of synthetic CSAM and NCII content, and much more. The evaluations space for digital content transparency techniques in their application to synthetic content is relatively new, though applications of cryptography, authentication, provenance, and labeling concepts have been applied across different applications and use cases. A socio-technical perspective for evaluations, evaluating the human-centered design of approaches and how these techniques are affecting people is valuable to ensure that these techniques are being designed and implemented to improve digital content transparency for all.

5. Preventing and Reducing Harms from Synthetic Child Sexual Abuse Material and Non-Consensual Intimate Imagery

Child sexual abuse material (CSAM) and non-consensual intimate images (NCII) are not new forms of technology-facilitated abuse, but GAI tools allow for novel, and direct ways to create this content at scale causing new and growing harm to victims and survivors, both minors and adults—often with relative ease, and requiring few technical skills. It has been documented that some AI models have been trained on datasets containing confirmed, real CSAM. Various open-source tools developed by malicious actors—such as face and body swap apps and websites to build image generation models commonly trained on non-consensual intimate images—are expanding on the Internet and resulting in sextortion, monetization schemes, and/or the targeting and abuse of women, girls, and minors (in addition to the common use of NCII to stalk, harass, and humiliate victims, including by abusive partners). Editing tools, in which authentic images can be uploaded and subsequently manipulated with AI, are another way in which synthetic CSAM and NCII are proliferated online. The likenesses of political and public figures have been manipulated and generated using AI tools to create non-consensual intimate imagery, disproportionately targeting women and affecting the civic and political participation of women and the health of democracies. Lastly, the misuse of generative AI tools increases victim identification, re-victimization, and prevention issues for practitioners in this space. Victim identification is more difficult with photorealistic synthetic CSAM being distributed at scale, the distribution of this content exacerbates victim trauma, and prevention is difficult when known CSAM is in AI model training data. This is a major socio-technical challenge, with implications for democracy and individuals' safety.

5.1. Current Technical Mitigations to Prevent and Reduce Harms from Synthetic CSAM and NCII

5.1.1. Training Data Filtering

As noted above, the ability of GAI models to generate CSAM or NCII is made more likely by images included in its training data which can result in harmful outputs as a result of human prompts. Crowdsourcing data labeling often introduces biases and inaccuracies in human labels. Biases around children could also affect dataset labels. For example, a study shows that most adults view Black girls between the ages of 5-14 as more adult-like than their white peers. It is important to note that removing CSAM from training data can be uniquely difficult as effectively filtering and removing all harmful data from the training data is challenging when the data is scraped systematically from the Internet, and also given that it is generally illegal for entities to possess CSAM, with a few exceptions respective of the right legal protocols in place for reporting. Neither human review nor automated filters or a combination of the two are effective enough to classify and capture all harmful and illegal content, including known CSAM. One example is the LAION-5B dataset that was confirmed to contain CSAM and was assembled from Common Crawl data, an open repository.

Filtering too little data allows the model to be trained on harmful content, but filtering too much could affect the quality of the model's outputs and reduce its sophistication or quality of outputs. Filtering training data to prevent unsafe outputs and designing various safety classifiers to clean up datasets before conducting model training can be useful.

Designing filters for training data could involve training ML-based classifiers using images of known and vetted CSAM and NCII content (safety classifiers) and any other generally sexually explicit content, testing this classifier on large datasets to determine precision and recall rates, and then using the classifier to identify harmful content in training data, which could then be removed prior to model training. Deep-learning based classifiers could inform image and video classification tasks for detection of sexually explicit images. For example, lightweight nudity detection techniques consisting of neural nets for image classification are publicly available. Developers could also remove harmful or illegal content from training data by filtering content from websites that are known to host CSAM and NCII. Another method for reducing CSAM in training data is by training models only on vetted data, such as licensed stock images and data in the public domain, given that all image data would be vetted for licensing and/or exists in the U.S. public domain, though this may be costly for training and may not be sufficient for training larger diffusion models.

5.1.1.1. Challenges and Limitations of Training Data Filtering

Key challenges and limitations of filtering training data include the subjective nature of safety labels, resulting in ineffective filters and potential opportunity costs with the quality of model outputs.

Creating a safety filter can involve a human labeling process to classify different types of content as violative and the type and severity of violation. For example, for the LAION-5B dataset, developers attempted to remove sexually explicit and harmful content from the original training dataset, but the safety filters used did not classify and capture all of the harmful or illegal content, including known CSAM. It may be difficult to remove NCII from training data because consent – the defining feature of NCII—may not be evident or decipherable in the content itself.

Context matters when implementing a safety filter to remove harmful content from training data. When crafting internal content policies to train safety classifiers, determining the severity of types of sexual content can be challenging, especially if it is in a legal gray area. For example, an image of a toddler wearing a bathing suit on a beach is generally quite harmless in training data, but it also means that the model was trained on an image of a minor’s body, which can then enable the model to generate harmful, illegal outputs such as synthetic CSAM using that data, given that generative AI models generate outputs based on training data inputs. Labels can be used to improve and clean training data but cannot fully translate context within a product. Inaccurate labels and even accurate labels taken out of context could result in harmful model outputs.

There are also potential opportunity costs to consider when filtering or limiting training data. Filtering out data more conservatively could improve the safety of a system but could also reduce its functionality for benign use cases. For example, removing all images of individuals wearing revealing clothing at beaches in order to exclude any images of women and children in bathing suits could result in a model that is not capable of generating high quality images of people in beach settings Research shows large-scale data filtering could have unexpected side effects on model performance and reduce the quality of image generation across different tasks.

5.1.2. Input Data Filtering

Input data filtering can be applied at the prompt level for text-to-image models and is used in the machine learning safety pipeline to prevent harmful generations. This form of moderation is conducted after a training run for a GAI model and occurs at the product level, when users type in prompts to generate images. Input data filtering can block malicious content that a user is intentionally attempting to generate through violative or harmful prompts.

Input data filtering includes machine learning safety or moderation classifiers trained on text data. These classifiers can be trained on a variety of different text prompts in order to classify different categories of violative content. With respect to this section, text classifiers can be designed to detect sexual prompts of different severity. For example, this can include benign intimate activity such as two people kissing, all the way to synthetic CSAM and NCII, which is of the highest severity. Several companies provide moderation classifiers at the prompt level. Similar to the classifiers discussed for training data filtering, these classifiers are also contingent on human data labels and internal content policies to determine what types of content are violative and at what severity levels.

A second, more simplistic, type of input filter used by platforms and developers is a keywords filter, also known as a keywords block list—an internally-managed database of violative keywords that prevents the generation of images when a violative keyword is entered as an input prompt. This approach identifies low-hanging fruit or known egregious content, such as commonly known CSAM terminology or sexual terms. A keywords block list can be less sophisticated than safety classifiers and used more as a blunt instrument. Nuance is difficult to achieve when the safety architecture operates in a binary; either a prompt is blocked because it contains a violative keyword, or it is not blocked because it does not contain violative keywords.

5.1.2.1. Challenges and Limitations of Input Data Filtering

There are a variety of challenges and limitations with input data filtering techniques, mainly related to accuracy and robustness. Keyword filters on open-source image generation models can be bypassed easily, and open-source models generally also have less technical restraints on the creation of harmful content. Because text safety classifiers are contingent on both robust and nuanced human labels across many types of content, and on gray-area content that even humans can disagree on, they may not always be accurate in their classifications. Further, it is a socio-technical challenge to determine the statistical confidence threshold at which certain types of content should be blocked. For example, should a CSAM classifier block the generation of content at a lower confidence level, such as 60%, to ensure that the false negative rate is lower? What kinds of benign content would be blocked if it were to be set at that threshold? Similar to issues with training data filtering, context is important. On the other hand, the effectiveness of keyword blocking can be limited on terms that have both harmful and benign meanings, and could result in false positives. Malicious actors could also easily evade keyword blocks and violate content moderation policies by adding different characters in between words, using trial and error to find phrases that are not blocked, or utilizing visual synonyms to generate explicit imagery.

5.1.3. Image Output Filtering

Image output filtering is a method used to directly block the generation of an image based on any violations or harms coming from the image output itself. Different output filtering techniques are used by AI developers to help prevent harmful generations, although there is no publicly-available information on how they are trained or the content that they block. Image output filters can also be known as image classifiers. Image classifiers utilize labeled images that feed into a neural network, which then conducts image classification and predicts a specific label or class depending on the original labeled images. For example, an AI developer can create a training dataset with a variety of harmful sexual images and detailed labels. This dataset can then be used to train a machine learning classifier to help identify similar content at scale. Once the classifier has been tested and evaluated, it can be included as a moderation mitigation within a product: When the classifier is triggered at a specific confidence level indicating harmful content (e.g., 0.9 or above), then the generation can be blocked at the user level.

5.1.3.1. Challenges and Limitations of Image Output Filtering

Image output filtering challenges and limitations are similar to those of other methods that are applied after training GAI models; they cannot entirely prevent harmful synthetic CSAM and NCII outputs if training data contains this harmful content. The effectiveness of image output filters depends on their training data and how well it covers a wide range of sexual content. Similar to other classifiers, these classifiers could become conservative or far too lenient in blocking content, and it is a challenge to set filter thresholds and confidence levels for content that exists in a gray area, can be subjective, or simply may not be covered as violative within an organization's internal policies. These classifiers can also be informed by the real-world abuse of tools by malicious actors and in order to be effective, should be constantly updated to reflect empirical abuse cases. This is another significant implementation challenge, given that new forms of abuse may be difficult to detect if classifiers were not designed to detect novel abuse content. Image output filtering may be more effective for content that contains clear and evident nudity, given that nudity classifiers exist, but could be less helpful for unseen harmful sexual depictions or content that exists in a gray area. Lastly, possessing a training data set with explicit content could be a risk in and of itself for industry and academic researchers.

5.1.4. Hashing Confirmed Synthetic CSAM and NCII

Hashing confirmed synthetic CSAM and NCII after it has been created, and then appropriately sharing these hashes with platforms, civil society, and law enforcement, as appropriate, can help track its dissemination across the Internet and curtail further spread.

Cryptographic hashes make use of the "avalanche effect," which states that even a slight alteration to the input data would produce a vastly different cryptographic hash. When a single letter in a written document or a single pixel in an image is altered, the new cryptographic hash will not resemble the original one. For example, if a CSAM image is cryptographically-hashed, and that exact image is posted on a social media platform that participates in the hash-sharing database containing the original image, the platform should be able to identify the match.

Cryptographic hashing is currently used for service providers and platforms to prevent the redistribution of synthetic CSAM and NCII content, in order to identify exact matches of egregious content. However, it is not impervious to hacking and adversarial attacks.

Perceptual hash algorithms output similar hashes for comparable input files as seen by humans; the hash value is contingent on the content and stays approximately the same if the content is not significantly changed, such as if modifications are made to compression, brightness, orientation, or color. The objective is to use the distance between and similarity of the perceptual hashes to approximate the degree of similarity between input files. There is still a chance that perceptual hashes will produce false positives and false negatives, meaning that different input files may have hashes that are same or comparable, while similar input files may have hashes that are different.

Research shows that perceptual hashing can have greater benefits in multimedia formats given that it produces hash matches based on similarity and tolerates differences in format and quality. For example, if a confirmed synthetic CSAM image is edited with a color filter and posted on a platform, it would retain the hash match to the original content.

If safety experts at AI companies are able to examine these forms of synthetic content and identify images as CSAM and NCII, (though the classification of images itself may be a legal and policy challenge) it could be hashed (both with cryptographic and perceptual hashes) and stored in shared databases for known CSAM and NCII for other entities to identify. This would enable the detection and mitigation of the content across different platforms and websites. This solution may not *prevent* the harms of synthetic CSAM and NCII but could *reduce* the impact and severity of these harms by stymying the dissemination of this content and reducing further exposure of those depicted without their consent.

5.1.4.1. Challenges and Limitations of Hashing Confirmed Synthetic CSAM and NCII

There are coordination, policy, and technical challenges for hashing confirmed CSAM and NCII.

Coordination across organizations to share this content safely and effectively can be difficult. There are established norms and laws about reporting CSAM, as well as established organizations that conduct this work, such as the National Center on Missing and Exploited Children (NCMEC), which has also started to hash reported synthetic CSAM. However, efforts for synthetic CSAM and NCII are still in early stages, and can be better coordinated and standardized between AI developers, social media platforms, messaging platforms, and other Internet providers in order to track the dissemination of this content and report it to law enforcement effectively and proactively.

Policy challenges of understanding context also apply to hashing synthetic CSAM and NCII. The explicit depiction of minors in images is a felony offense, and synthetic CSAM can represent a visual depiction of sexually explicit conduct involving a minor, which has been facilitated using GAI technologies. However, there currently is no unified classification system for synthetic content to shed light on how an authentic image may be modified by AI, if the image is completely synthetic, whether the image shows a minor, and what kind of explicit conduct is shown in the image. Another key issue is determining whether an image that is uploaded and modified by AI was NCII or not. It is difficult to adjudicate consent for widely-used GAI tools, unless the tool itself is malicious and trained on authentic images of people, and consent is also difficult to discern on social media platforms. Consent may also be limited to particular contexts, for example, there could be consent for the use of a person's image to create a new GAI image, but not for distribution of that image. Further, even if these policy gray areas are standardized across industry and civil society and clarified through regulations and legal action, vetting and hashing this content would still have to be done at scale. At this point in time, human vetting is still a requirement for accurate labeling, which also takes a toll on the mental health of reviewers vetting this content. These policy considerations are vital to understand since labeling and assigning severity levels for hashed content is not straightforward.

Lastly, it is important to note the technical limitations of hashing—both perceptual and cryptographic and their vulnerabilities. Hashing can have robustness issues such as hash collisions, and though they can be a helpful security measure, they can also be attacked and manipulated. Malicious modifications of hashes, particularly perceptual hashes, are a concern since malicious modifications are possible without distinguishing them from legitimate distortion. Bad actors could modify an image in a manner that is not distinguishable from legitimate or benign distortions (such as compression), thus affecting the integrity of tracking the original image. Perceptual hashing can allow significant data leakage. The same properties that make the technique robust can allow inference of information about underlying content from that content's hash, introducing serious privacy risks.

Lastly, databases hosting hashes should be secured properly. Insufficiently secured hash-sharing algorithms can allow for further exploitation, which could harm victims if hash-sharing databases contain confirmed CSAM or other sensitive content.

5.1.5. Provenance Data Tracking Techniques for Synthetic CSAM and NCII

Provenance data tracking techniques for synthetic content, such as digital watermarks and metadata recording, could be used to reduce synthetic CSAM and NCII harms. They could dissuade malicious actors from using tools that disclose all synthetic content, including synthetic CSAM and NCII as AI-generated.

Malicious actors who create synthetic CSAM and NCII might find tools less appealing for exploitation if those tools include provenance information about the origin of an image, or a watermark that shows an image is AI-generated, which can quickly assist in debunking any claims that synthetic CSAM and NCII images are authentic. Most malicious actors generating this content on the Internet utilize open-source tools or can even build their own smaller models based on existing open-source code, given that they can easily remove safeguards. Implementing provenance data tracking approaches that utilize robust watermarks and/or embed cryptographically-signed and secure metadata could add barriers for malicious actors looking to quickly spin up and even monetize synthetic CSAM and NCII. This method may reduce how much synthetic CSAM and NCII is created using tools that include provenance data tracking techniques, though there needs to be more research to support this assertion.

Directly designating synthetic CSAM and NCII as AI-generated through provenance labels can allow for the streamlined identification of this content by practitioners tracking these harms. The benefit of streamlined identification would likely apply when content is generated and disseminated by less sophisticated actors, who do not strategically use tools without provenance data tracking techniques, or actors who are not aware of methods to remove watermarks or metadata. Harmful content created by GAI tools that use provenance data tracking techniques—such as digital watermarking and metadata recording—could be identified more easily in an interoperable ecosystem, when various content providers and platforms are able to detect watermarks and/or preserve metadata.

5.1.5.1. Challenges and Limitations of Provenance Data Tracking Techniques for Synthetic CSAM and NCII

The challenges and limitations of provenance data tracking techniques for synthetic CSAM and NCII include uncertainties about efficacy, robustness issues, and potential for adversarial abuse.

There is a lack of research and evidence about whether and how provenance labels are effective in reducing harms from synthetic CSAM and NCII. Survivors and victims whose images are altered without their consent through AI experience, harm, humiliation, and degradation regardless of whether the content has overt labels and metadata attached to it.

Robustness issues with provenance data tracking techniques are also a concern. As mentioned in previous sections, even the most robust frameworks for metadata recording and digital watermarks can be vulnerable to manipulation and modification. Covert and overt watermarks can be removed from digital content, and embedded metadata could be stripped. All of the provenance issues discussed in this report apply to its use for synthetic CSAM and NCII. Given the level of sensitivity and harm with this type of content, robustness can affect the identification of this content at scale by practitioners, as well as victims of these harms.

Lastly, issues of robustness can create avenues for the adversarial abuse of provenance data tracking. Initial research shows how malicious actors can remove watermarks and metadata from synthetic CSAM and NCII. If so, they could undermine the benefits of labels on this content. Furthermore, the section on embedded metadata shows how adversarial attacks can be conducted on metadata that is both unsigned and signed; cryptography does not guarantee complete defense against adversarial attacks.

5.1.6. Red-Teaming and Testing for CSAM and NCII

Red-teaming and testing for synthetic CSAM and NCII prior to the deployment of GAI models could provide further safeguards. As defined in EO 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, red-teaming refers to a "structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI." Red-teaming is a narrow type of evaluation method. Currently, standardized red-teaming for GAI models does not exist, as the space is emergent. However, a baseline level of red-teaming—like inputting various types of adversarial prompts to generate synthetic CSAM and NCII—could be used. By scoping the Internet and internal systems for known prompts used to generate or attempt to generate synthetic CSAM and NCII, developers of AI models can develop initial assessments of a model's propensity toward generating this content. An established and uniform red-teaming protocol or guidelines for synthetic CSAM and NCII could assist with the future measurement of this content.

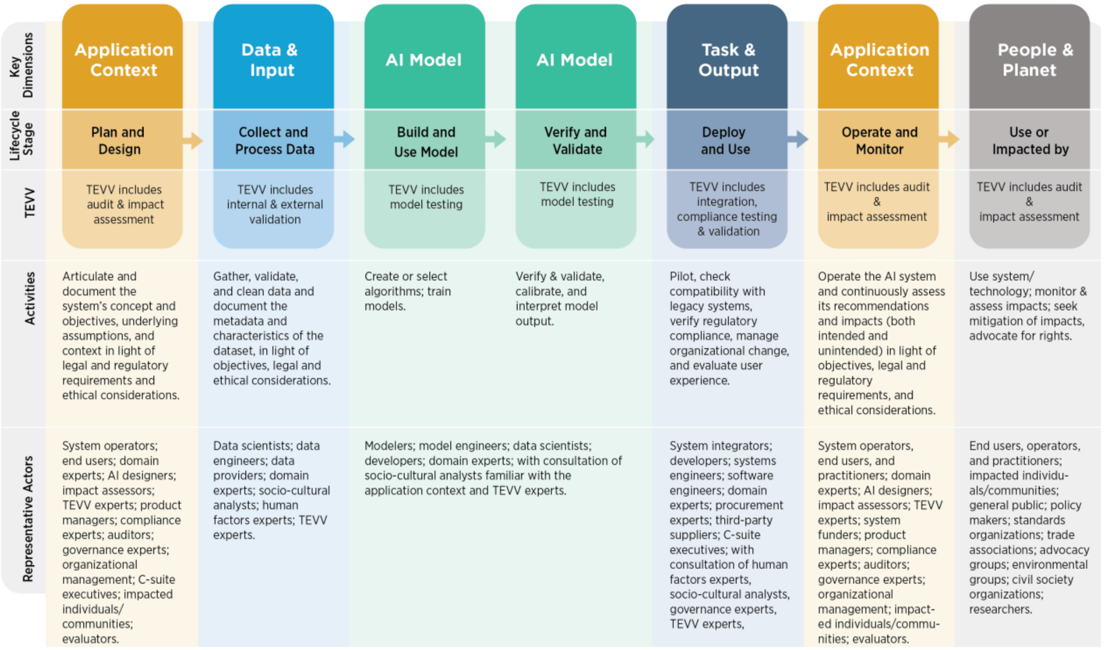
5.1.6.1. Challenges and Limitations of Red-Teaming and Testing for CSAM and NCII

Red-teaming and testing cannot effectively make up for issues in training data. These methods are also contingent on how testing is conducted, and therefore are biased toward testing for known vulnerabilities in an AI system. As mentioned throughout this section, training datasets without CSAM and NCII data can help reduce the generation of this content. Additional safeguards applied after the initial training run may not prove to be sufficient if the data itself is polluted. Lastly, by probing the model with prompts that are already established as harmful and/or capable of creating synthetic CSAM or NCII, a developer may not have coverage of new adversarial prompts that could bypass model safeguards.

Opportunities for Further Development: More research and development is needed for designing effective red-teaming strategies to catch synthetic CSAM and NCII outputs, determining the effectiveness of provenance data tracking techniques on this content in reducing harm, designing classifiers and filters to remove CSAM and NCII from training data as well as at the input and output model levels, and developing coordination between civil society, industry, law enforcement, and other relevant entities to hash synthetic CSAM and NCII. Further research is also needed to examine the viability of privacy-preserving perceptual hashing.

6. Application of Concepts to the NIST AI Risk Management Framework Lifecycle

The NIST AI Risk Management Framework (NIST AI RMF) states, “measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage; some risks may be latent at a given point in time and may increase as AI systems adapt and evolve.” Different AI actors (both actors who are building and/or utilizing AI models) will often have different risk perspectives and may find certain provenance data tracking or synthetic content detection techniques more useful contingent on use case, product, and organizational goals.



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impacted individuals/communities; evaluators.	End users, operators, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

Figure 2. AI actors across AI lifecycle stages. From NIST AI 100-1 AI RMF 1.0

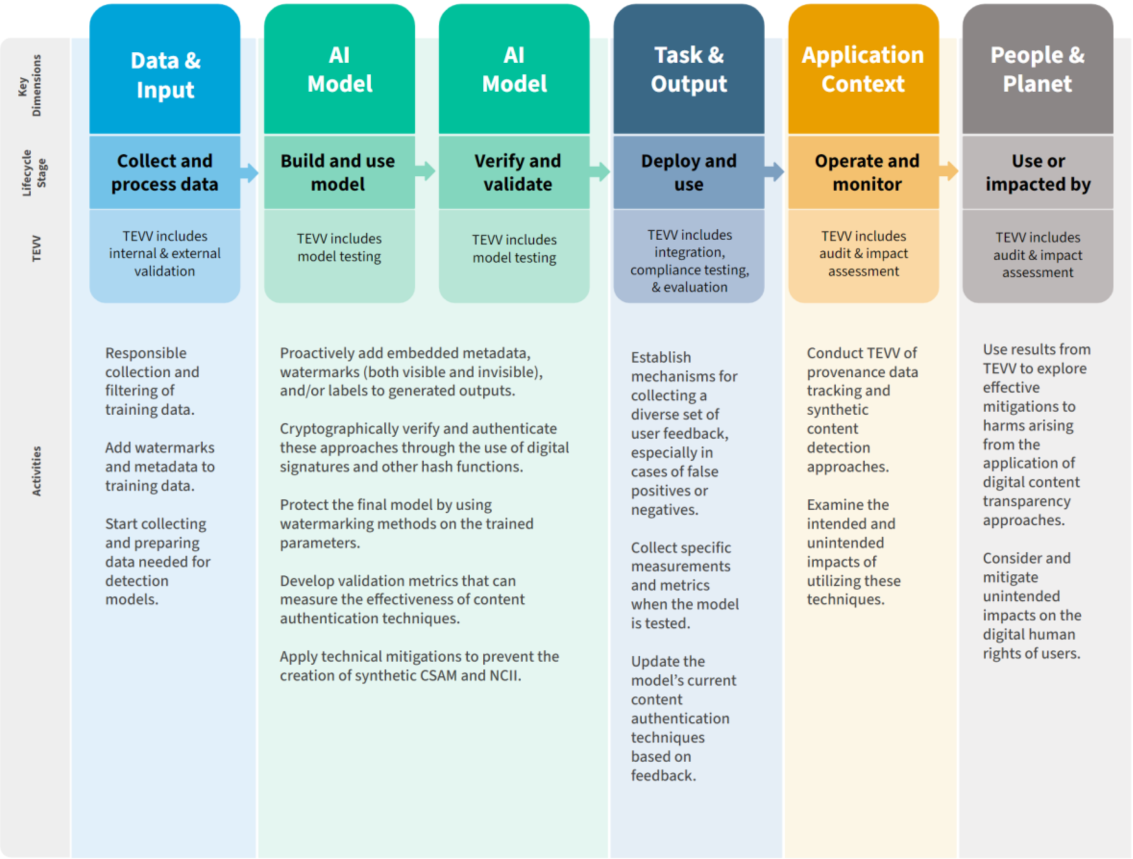
Data & Input: Collect and Process Data: The responsible collection and filtering of training data could help reduce and/or prevent the harms of synthetic CSAM and NCII outputs in this phase. Provenance data tracking techniques such as watermarking and metadata could be added to training data to preserve the provenance of datasets used in training. In this phase, data and input needed to design detection models to classify synthetic content can also be collected.

AI Model: Build and Use Model, Verify and Validate: During the build and use, and verify and validate phases, provenance data tracking techniques such as metadata or watermarks can be proactively added to model outputs at the time of generation. To apply these provenance approaches securely, they can be cryptographically verified and authenticated in their application, through the use of a digital signature, or other types of hash functions. Also, the final model can be protected by watermarking the model weights or parameters. The effectiveness of provenance data tracking techniques, such as accuracy in detecting watermarks, or correctly identifying manipulated and synthetic content, prior to deployment needs to be verified. Mitigation mechanisms that prevent the creation of synthetic CSAM and NCII (as discussed in previous sections) may be proactively applied during the model building phase.

Task and Output: Deploy and Use: Establishing mechanisms for collecting a diverse set of user feedback—especially in cases of false positives (e.g., disclosing a content as AI-generated while it is actually human generated) or negatives (e.g., missing a disclosure of AI-generated content)—supports content authentication and synthetic content detection efforts. The measurements and metrics used are highly dependent on the use case, context, systems being tested, and the application.

Application Context: Operate and Monitor: The broader impact of digital content transparency approaches may be examined during the operate and monitor phase of the AI lifecycle in light of objectives, legal and regulatory requirements, and ethical considerations.

People and Planet: Use or Impacted By: By sharing the results from TEVV conducted across the AI lifecycle with various representative actors such as AI developers, civil society entities, and end users, effective mitigations to potential harms can be explored. A focus on digital rights for all and safety-by-design remains important in this last phase, and closing gaps where groups may be denied access to their digital human rights to digital content transparency, through factors such as a lack of Internet access, or information literacy resources to understand labels on content, or as a result of the malign or unintended use of provenance data tracking techniques to negatively impact user privacy should be considered for AI actors across the content lifecycle.



Digital content transparency approaches, parsed by the NIST AI RMF lifecycle

Key Dimensions	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Collect and process data	Build and use model	Verify and validate	Deploy and use	Operate and monitor	Use or impacted by
TEVV	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing, & evaluation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Responsible collection and filtering of training data. Add watermarks and metadata to training data. Start collecting and preparing data needed for detection models.	Proactively add embedded metadata, watermarks (both visible and invisible), and/or labels to generated outputs. Cryptographically verify and authenticate these approaches through the use of digital signatures and other hash functions. Protect the final model by using watermarking methods on the trained parameters. Develop validation metrics that can measure the effectiveness of content authentication techniques. Apply technical mitigations to prevent the creation of synthetic CSAM and NCII.		Establish mechanisms for collecting a diverse set of user feedback, especially in cases of false positives or negatives. Collect specific measurements and metrics when the model is tested. Update the model's current content authentication techniques based on feedback.	Conduct TEVV of provenance data tracking and synthetic content detection approaches. Examine the intended and unintended impacts of utilizing these techniques.	Use results from TEVV to explore effective mitigations to harms arising from the application of digital content transparency approaches. Consider and mitigate unintended impacts on the digital human rights of users.

Figure 3. Digital content transparency approaches, across the AI lifecycle described in NIST AI RMF

7. Conclusion

This report is intended to enhance understanding of technical approaches to synthetic content and digital content transparency as a key step in reducing related risks. It focuses on and provides an overview of technical approaches to digital content transparency, which is key to achieving the goal of reducing AI risks involving synthetic image, text, audio, and video content. This report provides specific information about synthetic content related to child sexual abuse material (CSAM) and non-consensual intimate images (NCII) and describes techniques being used or considered to prevent and reduce related harms.

This report describes technical approaches that are being used and offered commercially or are available today as well as those that are being explored. After explaining the advantages and issues with each technique, this document highlights selected opportunities for further development.

Each of the approaches described in this report holds the promise of helping to improve trust by clearly and readily indicating where AI techniques have been used to generate or modify digital content. *Yet each has important limitations that are both technical and social in nature.* It is vital to note that none of these techniques can be considered as comprehensive solutions; the value of any given technique is use-case and context specific. In order for digital content transparency to succeed, the application of provenance data tracking and synthetic content detection approaches must occur in tandem with various social efforts and initiatives to affirm content authenticity.

Collaboration and coordination across the content value chain—and consideration of social factors—are needed to ensure adoption of effective digital content transparency approaches. That includes the need for science-backed standards forged through global actions; this report cites several of those initiatives for particular techniques aimed at fostering digital content transparency.

While there is no silver bullet to solve the issue of public trust in and safety concerns posed by digital content, the consideration of the various approaches for provenance data tracking and synthetic content detection across different modalities of content is important, and research on these approaches can be developed further. This report is a resource to promote understanding and help to lay the groundwork for the development of additional, improved technical approaches to advancing synthetic content provenance, detection, labeling, and authentication.

8. Bibliography

Arranged by section.

Current Approaches: Provenance Data Tracking

“FIPS PUB 200: Minimum Security Requirements for Federal Information and Information Systems,” NIST, DOC, March 2006. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.200.pdf.

Geisler, E. et al. “Information integrity: an emerging field and the state of knowledge.” PICMET '03: Portland International Conference on Management of Engineering and Technology, Technology Management for Reshaping the World, 2003. https://doi.org/10.1109/PICMET.2003.1222797.

Akbari, Y. “Digital forensic analysis for source video identification: A survey.” Forensic Sci Int: Digital Investig 41, 2022. https://researchportal.northumbria.ac.uk/en/publications/digital-forensic-analysis-for-source-video-identification-a-surve.

Aldweesh, A. “The impact of blockchain on digital content distribution: a systematic review.” Wireless Networks, 2023. https://link.springer.com/article/10.1007/s11276-023-03524-0.

Almutairi, Z. et al. “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions.” Algorithms 2022, 15 (5), 155. https://doi.org/10.3390/a15050155.

Amrit, P. et al. “Survey on watermarking methods in the artificial intelligence domain and beyond.” Computer Communications Volume 188, 15, April 2022: 52-65. https://www.sciencedirect.com/science/article/pii/S0140366422000664.

“A multi-dimensional approach to disinformation – Report of the independent High level Group on fake news and online disinformation,” European Commission, Directorate-General for Communications Networks, Content and Technology, 2018. https://data.europa.eu/doi/10.2759/739290.

Andrews, M. “Emerging best practices for disclosing AI-generated content.” Kontent.ai, August 2023. https://kontent.ai/blog/emerging-best-practices-for-disclosing-ai-generated-content/.

“Artificial Intelligence Risk Management Framework (AI RMF 1.0),” National Institute of Standards and Technology, Department of Commerce, January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Ashenfelder, M. “Social Media Networks Stripping Data from Your Digital Photos.” Library of Congress Blogs, April 2013. https://blogs.loc.gov/thesignal/2013/04/social-media-networks-stripping-data-from-your-digital-photos/.

Athalye, A. “Inverting PhotoDNA.” December 2021. https://anishathalye.com/inverting-photodna/.

“Authenticating AI-Generated Content: Exploring Risks, Techniques & Policy Recommendations.” ITIC, January 2024. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf.

Begum, M. et al. “Digital Image Watermarking Techniques: A Review.” Information, 11 (2) (2020): 110. https://doi.org/10.3390/info11020110.

Bender, W. et al “Techniques for data hiding.” IBM Systems Journal, 35 (3.4) (1996): 313-336. https://ieeexplore.ieee.org/document/5387237.

Bentzen, N. “Computational Propaganda Techniques.” European Parliamentary Research Service, October 2018. https://www.europarl.europa.eu/RegData/etudes/ATAG/2018/628284/EPRS_ATA(2018)628284_EN.pdf

Boenisch, F. “A Systematic Review On Model Watermarking For Neural Networks.” arXiv, 2021. https://arxiv.org/pdf/2009.12153.pdf.

Boujerfaoui, Said et al. "Image Watermarking between Conventional and Learning-Based Techniques: A Literature Review." Electronics 12 (1) (2023) 74: https://doi.org/10.3390/electronics12010074.

Bourouis, S. “Recent advances in digital multimedia tampering detection for forensics analysis.” Symmetry 12 (11) (2020):1811. https://www.mdpi.com/2073-8994/12/11/1811.

Bradshaw, S. et al “Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation.” Oxford Internet Institute, 2018. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/07/ct2018.pdf.

Brassil, J. et al. “Electronic marking and identification techniques to discourage document copying.” IEEE, June 1994. https://ieeexplore.ieee.org/document/337544.

“Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure,” Partnership on AI, December 2023. https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/.

Busch, K.E. “Blockchain: Novel Provenance Applications.” Congressional Research Service, April 2022. https://crsreports.congress.gov/product/pdf/R/R47064.

“C2PA Security Considerations,” C2PA, n.d. https://c2pa.org/specifications/specifications/1.0/security/Security_Considerations.html.

Cecco, L. “Death of the narrator? Apple unveils suite of AI-voiced audiobooks.” The Guardian, January 2023. https://www.theguardian.com/technology/2023/jan/04/apple-artificial-intelligence-ai-audiobooks.

Chakraborty, M. et al. “Counter Turing Test CT^ 2: AI-Generated Text Detection is Not as Easy as You May Think--Introducing AI Detectability Index.” arXiv, 2023. https://arxiv.org/abs/2310.05030.

Chakraborty, S. et al. “On the Possibilities of AI-Generated Text Detection.” arXiv, 2023. https://arxiv.org/abs/2304.04736.

Chia, A. “Metadata 101: Definition, Types & Examples.” Splunk, June 2023. https://www.splunk.com/en_us/blog/learn/metadata-types.html.

Clegg, N. “Labeling AI-Generated Images on Facebook, Instagram and Threads.” Meta, February 2024. https://about-fb-com.cdn.ampproject.org/c/s/about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/amp/.

“Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence.” U.S. Copyright Office, March 2023. https://www.copyright.gov/ai/ai_policy_guidance.pdf.

Cox, I.J. et al. "Watermarking applications and their properties." Proceedings International Conference on Information Technology: Coding and Computing (Cat. No.PR00540) (2000): https://ieeexplore.ieee.org/document/844175.

“Cryptography: Overview.” NIST, DOC, n.d. https://www.nist.gov/cryptography.

Croman, K. et al. “On scaling decentralized blockchains.” International Conference on Financial Cryptography and Data Security (2016): 106–125. https://increscent.org/writing/assets/bitcoin_scaling.pdf.

“Digital Signature Standard (DSS).” NIST, DOC, February 2023. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.186-5.pdf.

Dittmann, Jana et al. "Using Cryptographic and Watermarking Algorithms." IEEE MultiMedia, 2001. http://ivizlab.sfu.ca/arya/Papers/IEEE/Multimedia/2001/Oct/Cryptography%20and%20Watermarking.pdf.

Earnshaw, N. et al. “Fighting Misinformation with Authenticated C2PA Provenance Metadata.” Proceedings of the 2023 NAB Broadcast Engineering and Information Technology (BEIT) Conference, 2023. https://drive.google.com/file/d/1-KdLn5n-k_9bAILsnGAi81BW8c0zibee/view.

Eastlake, D. et al. “US secure hash algorithm 1 (SHA1).” IETF, 2001. https://datatracker.ietf.org/doc/rfc3174/.

El-Shafai, W. “A comprehensive taxonomy on multimedia video forgery detection techniques: challenges and novel trends.” Multimed Tools Appl. (2023): 1-67. https://pubmed.ncbi.nlm.nih.gov/37362636/.

England, P. et al. “Amp: Authentication of media via provenance.” Proceedings of the 12th ACM Multimedia Systems Conference (July 2021): 108-121. https://arxiv.org/pdf/2001.07886.pdf

Epstein, Z. et al. “What label should be applied to content produced by generative AI?” Preprint, July 2023. https://doi.org/10.31234/osf.io/v4mfz.

Fazli, S. et al. “Trade-Off between Imperceptibility and Robustness of LSB Watermarking Using SSIM Quality Metrics.” 2009 Second International Conference on Machine Vision (2009): 101-104. https://ieeexplore.ieee.org/document/5381093.

“Federal Information Processing Standards Publication Secure Hash Standard (SHS).” NIST, DOC, August 2015. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf.

Fei, C. et al. “Analysis and design of secure watermark-based authentication systems.” IEEE transactions on information forensics and security, 1 (1) (2006): 43-55. https://ieeexplore.ieee.org/document/1597134.

Feng, K.J. et al. “Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions.” Proc. ACM Hum.-Comput. Interact. (October 2023). https://dl.acm.org/doi/pdf/10.1145/3610061.

Fisher, T. “What Is a Cryptographic Hash Function?” Lifewire, July 2022. https://www.lifewire.com/cryptographic-hash-function-2625832.

Fridrich, J. Steganography in Digital Media Principles, Algorithms, and Applications. Cambridge University Press, 2009. https://doi.org/10.1017/CBO9781139192903.

“From Deepfakes to TikTok Filters: How Do You Label AI Content?” Partnership on AI, May 2021. https://medium.com/partnership-on-ai/from-deepfakes-to-tiktok-filters-how-do-you-label-ai-content-ba61747bc457.

Ghosal, Soumya Surva et al. "Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey." arXiv, October 2023. https://arxiv.org/pdf/2310.15264.pdf.

"GIFCT's Hash-Sharing Database." Global Internet Forum to Counter Terrorism, n.d. https://gifct.org/hsdb/.

Gonzalez-Perez, C. "Metainformation." Information Modelling for Archaeology and Anthropology: Software Engineering Principles for Cultural Heritage (2018): 181–189. https://link.springer.com/book/10.1007/978-3-319-72652-6.

"Guidelines for Embedded Metadata within DPX File Headers for Digitized Motion Picture Film." Federal Agencies Digital Guidelines Initiative (FADGI) Audio-Visual Working Group, 2017. https://www.digitizationguidelines.gov/audio-visual/documents/DPX_Embed_Guideline_20170814.pdf.

Harran, M. et al. "A method for verifying integrity & authenticating digital media." Applied Computing and Informatics 14, no. 2 (July 2018): 145-158. https://doi.org/10.1016/j.aci.2017.05.006.

Hartung, F. et al. "Fast public-key watermarking of compressed video." Proceedings of International Conference on Image Processing, 1 (October 1997): 528-531. https://ieeexplore.ieee.org/document/647966.

Hasan, H.R. et al. "Combating Deepfake Videos Using Blockchain and Smart Contracts." IEEE Access, Vol. 7 (2019): 41596–41606. https://ieeexplore.ieee.org/document/8668407.

Heidari, A. et al. "Deepfake detection using deep learning methods: A systematic and comprehensive review." Wiley Interdisciplinary Reviews, November 2023. https://doi.org/10.1002/widm.1520.

Hernandez-Ardieta, J.L. et al. "A taxonomy and survey of attacks on digital signatures." Computers & Security Volume 34 (May 2013): 67-112. https://doi.org/10.1016/j.cose.2012.11.009.

"ID3.org." Archived November 2011, accessed April 2020. https://web.archive.org/web/20111111055609/http://www.id3.org/.

Ioini, N. et al. "A Review of Distributed Ledger Technologies." OTM 2018 Conferences, Cloud and Trusted Computing (2018). https://www.researchgate.net/publication/328475892_A_Review_of_Distributed_Ledger_Technologies
.

"IPTC photo metadata." Google for Developers. n.d., https://developers.google.com/search/docs/appearance/structured-data/image-license-metadata.

"IPTC publishes metadata guidance for AI-generated 'synthetic media." IPTC, May 2023. https://iptc.org/news/iptc-publishes-metadata-guidance-for-ai-generated-synthetic-media/.

Johnston, P. "A review of digital video tampering: from simple editing to full synthesis." Digit Investig 29 (2019): 67–81. https://www.sciencedirect.com/science/article/abs/pii/S1742287618304146.

Jin, Sheng et al. "Unsupervised semantic deep hashing." Neurocomputing 351 (July 2019): 19–25. https://doi.org/10.1016/j.neucom.2019.01.020.

Kapoor, S. "How to Prepare for the Deluge of Generative AI on Social Media." Knight First Amendment Institute, June 2023. https://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media.

Katsis, C. et al. "Real-time Digital Signatures for Named Data Networking." ICN '20: Proceedings of the 7th ACM Conference on Information-Centric Networking (September 2020): 149–151. https://doi.org/10.1145/3405656.3420227.

Kirchenbauer, J. et al. “A watermark for large language models.” arXiv, 2023. https://arxiv.org/abs/2301.10226.

Kirchenbauer, John et al. "On the Reliability of Watermarks for Large Language Models." arXiv, June 2023. https://arxiv.org/abs/2306.04634.

Knibbs, K. “Researchers Tested AI Watermarks—and Broke All of Them.” WIRED, 2023. https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

Leibowicz, C. “Why watermarking AI-generated content won’t guarantee trust online.” MIT Technology Review, August 2023. https://www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/.

Lin, CY. “Watermarking and digital signature techniques for multimedia authentication and copyright protection.” Columbia University, PhD thesis, 2001. https://www.ee.columbia.edu/ln/dvmm/publications/PhD_theses/cylin-thesis.pdf.

Liu, A. et al. “A Survey of Text Watermarking in the Era of Large Language.” arXiv, 2023. https://arxiv.org/abs/2312.07913.

Liu, H. et al. “Deep supervised hashing for fast image retrieval.” Conference on Computer Vision and Pattern Recognition (2016): 2064–2072. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Liu_Deep_Supervised_Hashing_CVPR_2016_paper.pdf.

Lüthi, Philipp et al. “Distributed Ledger for Provenance Tracking of Artificial Intelligence.” arXiv, February 2020. https://arxiv.org/pdf/2002.11000.pdf.

Maesen, P. et al. “On the Effectiveness of Dataset Watermarking in Adversarial Settings; Image Watermarking for Machine Learning Datasets.” ACM Digital Library, December 2023. https://doi.org/10.1145/3600046.3600048.

“Metadata and tags you should include in your website.” Search.gov, n.d. https://search.gov/indexing/metadata.html.

“Midjourney and Shutterstock AI sign up to use of IPTC Digital Source Type to signal generated AI content.” IPTC, May 2023. https://www.iptc.org/news/midjourney-and-shutterstock-ai-sign-up-to-use-of-iptc-digital-source-type-for-generated-ai-content/.

Michel-Villarreal, R. et al. “Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT.” Educ. Sci. 2023, 13(9), 856. https://doi.org/10.3390/educsci13090856.

Mitchell, E. et al. “Detectgpt: Zero-shot machine-generated text detection using probability curvature.” arXiv, 2023. https://arxiv.org/abs/2301.11305.

Newman, Lily Hay. “A New Tool Protects Videos from Deepfakes and Tampering.” WIRED, 2019. https://www.wired.com/story/amber-authenticate-video-validationblockchain-tampering-deepfakes/.

“NIST AI RMF Playbook.” NIST, DOC. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

Nguyen, T.T. et al. “Deep learning for deepfakes creation and detection: A survey.” Computer Vision and Image Understanding 223, October 2022. https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114.

Noah, B. “Public Perceptions Towards Synthetic Voice Technology.” Proceedings of the 2021 HFES 65th International Annual Meeting (2021): 1448. https://journals.sagepub.com/doi/pdf/10.1177/1071181321651128.

Numbers Protocol. n.d. https://docs.numbersprotocol.io/introduction/numbers-protocol.

“Overview of Perceptual Hashing Technology.” Ofcom, November 2022. https://www.ofcom.org.uk/__data/assets/pdf_file/0036/247977/Perceptual-hashing-technology.pdf.

“PAI’s Responsible Practices for Synthetic Media A Framework for Collective Action.” Partnership on AI, February 2023. https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf.

Patel, J. “Passive video forgery detection techniques to detect copy move tampering through feature comparison and RANSAC.” Cyber security and digital forensics (January 2022): 161–177. https://www.researchgate.net/publication/355030746_Passive_Video_Forgery_Detection_Techniques_to_Detect_Copy_Move_Tampering_Through_Feature_Comparison_and_RANSAC.

Patil RD et al. "Fragile video watermarking for tampering detection and localization." ICACCI (2015): 1661-1666. https://ieeexplore.ieee.org/document/7275852.

Pickholtz, R. et al “Theory of spread-spectrum communications-a tutorial.” IEEE Transactions on Communications, 30 (5) (1982): 855-884. https://ptabdata.blob.core.windows.net/files/2017/IPR2017-01024/v36_Ex.%201036%20-%20Pickholtz.pdf.

Pizzolante, R. et al. “Protection of Microscopy Images through Digital Watermarking Techniques.” 2014 International Conference on Intelligent Networking and Collaborative Systems, September 2014. https://ieeexplore.ieee.org/document/7057071.

Prokos, J. et al. “Squint Hard Enough: Attacking Perceptual Hashing with Adversarial Machine Learning.” USENIX Security Symposium, 2023. https://www.usenix.org/system/files/sec23summer_146-prokos-prepub.pdf.

Qasim, A. et al. “Digital watermarking: Applicability for developing trust in medical imaging workflows state of the art review.” Computer Science Review 27 (February 2018): 45-60. https://www.sciencedirect.com/science/article/abs/pii/S157401371730148X.

Qiong, D. et al. “Exposing frame-based video tampering by Fourier analysis of MCEA difference.” Proceedings of the 2012 Second International Conference on Electric Information and Control Engineering (ICEICE) (2012): 686–689. https://www.researchgate.net/publication/262239357_Exposing_Frame-Based_Video_Tampering_by_Fourier_Analysis_of_MCEA_Difference.

Quiring, E. et al. “Adversarial Machine Learning Against Digital Watermarking.” 2018 26th European Signal Processing Conference (EUSIPCO) (2018): 519-523. 10.23919/EUSIPCO.2018.8553343. https://ieeexplore.ieee.org/document/8553343.

Regazzoni, F. et al. “Protecting artificial intelligence IPs: a survey of watermarking and fingerprinting for machine learning.” CAAI Transactions on Intelligence Technology, 6 (2) (April 2021): 180-191. https://doi.org/10.1049/cit2.12029.

Rivest, R. “The MD5 message-digest algorithm.” IETF, 1992. https://www.ietf.org/rfc/rfc1321.txt.

Sartori, L. et al. “A sociotechnical perspective for the future of AI: narratives, inequalities, and human control.” Ethics Inf Technol 24, 4 (2022). https://doi.org/10.1007/s10676-022-09624-3.

Scholl, M. et al. “Safeguarding Data Using Encryption.” NIST, DOC, 2014. https://csrc.nist.gov/CSRC/media/Presentations/HIPAA-2014-Safeguarding-Data-Using-Encryption/images-media/scholl_hipaa_2014_day1.pdf.

Shaik, A.S., et al. “A review of hashing based image authentication techniques.” Multimed Tools Appl 81, 2489–2516 (2022). https://doi.org/10.1007/s11042-021-11649-7.

Shaliyar, M. et al “Watermarking approach for source authentication of web content in online social media: a systematic literature review.” Multimedia Tools and Applications (2023): 1-53. https://www.researchgate.net/publication/376081297_Watermarking_approach_for_source_authentication_of_web_content_in_online_social_media_a_systematic_literature_review.

Sharma, H. “An ontology of digital video forensics: classification, research gaps & datasets.” International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (2019): 485–491. https://www.researchgate.net/publication/339410290_An_Ontology_of_Digital_Video_Forensics_Classification_Research_Gaps_Datasets.

Sharma, R. et al. “A Unique Approach towards Image Publication and Provenance using Blockchain.” Third International Conference on Smart Systems and Inventive Technology (ICSSIT)(2020). https://ieeexplore.ieee.org/document/9214203.

Sheybani, N. et al. "ZKROWNN: Zero Knowledge Right of Ownership for Neural Networks." arXiv, September 2023. https://arxiv.org/pdf/2309.06779.pdf.

Shi, C. et al. “Review of Image Forensic Techniques Based on Deep Learning.” Mathematics 11 (14) (2023): 3134. https://doi.org/10.3390/math11143134.

Shumailov, I. et al. “The Curse Of Recursion: Training On Generated Data Makes Models Forget.” arXiv, 2023. https://arxiv.org/abs/2305.17493.

Simonite, T. “Synthetic Voices Want to Take Over Audiobooks.” WIRED, January 2022. https://www.wired.com/story/audiobooks-synthetic-voices/.

Singal, A. “Detecting the Deceptive: Unmasking Deep Fake Voices.” Hugging Face, October 2023. https://huggingface.co/blog/Andyrasika/deepfake-detect.

Singh, P. et al. “Robust Homomorphic Image Hashing.” Workshop on Media Forensics at CVPR (2019): 11–18. https://www.semanticscholar.org/paper/Robust-Homomorphic-Image-Hashing-Singh-Farid/aee1d46829efa7f46dd3d6af90f3095173b316c2.

Singh, RD et al. “Optical flow and pattern noise-based copy–paste detection in digital videos.” Multimedia Systems 27 (3) (2021): 449–469. https://dl.acm.org/doi/abs/10.1007/s00530-020-00749-3.

Sobti, R. et al. “Cryptographic Hash Functions: A Review.” International Journal of Computer Science Issues, 9 (2) (March 2012): 461 - 479. https://www.researchgate.net/publication/267422045_Cryptographic_Hash_Functions_A_Review.

Srinivasan, S. “Detecting AI fingerprints: A guide to watermarking and beyond.” Brookings, January 2024. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

Stern, J. “I Cloned Myself With AI. She Fooled My Bank and My Family.” Wall Street Journal, April 2023. https://www.wsj.com/articles/i-cloned-myself-with-ai-she-fooled-my-bank-and-my-family-356bd1a3.

“Technical Specification FG DLT D1.1 Distributed ledger technology terms and definitions.” International Telecommunication Union, 2019. https://www.itu.int/en/ITU-T/focusgroups/dlt/Documents/d11.pdf.

“Terrorist Content Analytics Platform.” TCAP, n.d. https://terrorismanalytics.org/.

“Truepic Unveils Watershed Gen-AI Transparency Directly on Devices Powered by Snapdragon Mobile Platform.” Global Newswire, October 2023. https://www.globenewswire.com/news-release/2023/10/24/2765978/0/en/Truepic-Unveils-Watershed-Gen-AI-Transparency-Directly-on-Devices-Powered-by-Snapdragon-Mobile-Platform.html.

Tulchinskii, E. et al. “Intrinsic dimension estimation for robust detection of ai-generated texts.” Advances in Neural Information Processing Systems, 36 (2024). https://arxiv.org/abs/2306.04723.

Tyagi, S. et al. “A detailed analysis of image and video forgery detection techniques.” Vis Comput 11 (39) (2022): 813–833. https://doi.org/10.1007/s00371-021-02347-4.

Uchida, Y., et al. “Embedding watermarks into deep neural networks.” Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (2017): 269–277. https://doi.org/10.1145/3078971.3078974.

Vasse’i, R. et al. “In Transparency We Trust? Evaluating the Effectiveness of Watermarking and Labeling AI-Generated Content.” Mozilla, February 2024. https://foundation.mozilla.org/en/research/library/in-transparency-we-trust/research-report/.

Viglas, S.D. “Data Provenance and Trust.” Data Science Journal, 12 (2013). https://doi.org/10.2481/dsj.GRDI-010.

Wadheraa, S. et al. “A Comprehensive Review on Digital Image Watermarking.” arXiv, 2022. https://arxiv.org/abs/2207.06909.

Wan, W. et al. “A comprehensive survey on robust image watermarking.” Neurocomputing, 488 (1) (June 2022): 226-247. https://www.sciencedirect.com/science/article/abs/pii/S0925231222002533#b0310.

Wang, A. “The Shazam music recognition service.” Communications of the ACM, 49 (8) (August 2006): 44–48. https://doi.org/10.1145/1145287.1145312.

Weng, L. et al. “A Secure Perceptual Hash Algorithm for Image Content Authentication.” Communications and Multimedia Security (2011): 108–121. https://link.springer.com/content/pdf/10.1007/978-3-642-24712-5_9.pdf.

“What is Digital Metadata and Why should I care about it?” ioMoVo, June 2023. https://iomovo.medium.com/what-is-digital-metadata-and-why-should-i-care-about-it-388accf6e0b.

Xiang, Z. et al. “Forensic Analysis of Video Files Using Metadata.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2021): 1042-1051. https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/Xiang_Forensic_Analysis_of_Video_Files_Using_Metadata_CVPRW_2021_paper.html.

Yaga, D. et al. “Blockchain Technology Overview.” NIST, DOC, October 2018. https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf.

Yahya, A. Steganography Techniques for Digital Images. Springer: 2019. https://link.springer.com/book/10.1007/978-3-319-78597-4.

Yamni, M. “An efficient watermarking algorithm for digital audio data in security applications.” Nature, 2023. https://www.nature.com/articles/s41598-023-45619-w.pdf.

Yi, J. “Audio Deepfake Detection: A Survey.” Journal Of Latex Class Files, Vol. 14, No. 8, (August 2023): 1. https://arxiv.org/pdf/2308.14970.pdf.

Zaiane, O. et al. “Digital Watermarking: Status, Limitations and Prospects.” University of Alberta, 2002. https://doi.org/10.7939/R3FF3M64K.

Zauner, C. “Implementation and Benchmarking of Perceptual Image Hash Functions.” July 2010, https://www.phash.org/docs/pubs/thesis_zauner.pdf.

Zhang, H. et al. “Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models.” arXiv, November 2023. https://arxiv.org/pdf/2311.04378.pdf.

Zhang, J. et al. “Do You Know Where Your Data’s Been? – Tamper-Evident Database Provenance.” MITRE, 2009. https://www.mitre.org/sites/default/files/pdf/09_1348.pdf.

Zhang, Lijun. "Robust Image Watermarking using Stable Diffusion." arXiv, January 2024. https://arxiv.org/html/2401.04247v1.

Zhao, X. et al. “Invisible Image Watermarks Are Provably Removable Using Generative AI.” arXiv, August 2023. https://arxiv.org/pdf/2306.01953.pdf.

Current Approaches: Synthetic Content Detection

Akhtar, Z. “Deepfakes Generation and Detection: A Short Survey.” Journal of Imaging 9, no. 1 (January 13, 2023): 18. https://doi.org/10.3390/jimaging9010018.

Albertoni, R. et al. “Reproducibility of Machine Learning: Terminology, Recommendations and Open Issues.” arXiv, February 2023. https://arxiv.org/pdf/2302.12691.pdf.

Almutairi, Z. et al. “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” n.d. https://www.mdpi.com/1999-4893/15/5/155.

“Artem9k/Ai-Text-Detection-Pile · Datasets at Hugging Face.” Accessed February 8, 2024. https://huggingface.co/datasets/artem9k/ai-text-detection-pile.

“Authenticating AI-Generated Content January 2024 Exploring Risks, Techniques & Policy Recommendations.” Information Technology Industry Council, January 2024. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf.

Bird, J. et al. “CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images.” IEEE Access 12 (2024): 15642–50. https://doi.org/10.1109/ACCESS.2024.3356122.

Brock, A. et al. “Large Scale GAN Training for High Fidelity Natural Image Synthesis.” arXiv, February 25, 2019. http://arxiv.org/abs/1809.11096.

Broniatowski, D. “Psychological Foundations of Explainability and Interpretability in Artificial Intelligence.” NIST, DOC, April 2021. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf.

“Contextualizing Deepfake Threats to Organizations,” n.d. https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF.

Corvey, W. “Semantic Forensics (SemaFor).” DARPA, n.d. https://www.darpa.mil/program/semantic-forensics.

Corvi, R. et al. “On The Detection of Synthetic Images Generated by Diffusion Models.” In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. Rhodes Island, Greece: IEEE, 2023. https://doi.org/10.1109/ICASSP49357.2023.10095167.

Cuccovillo, L. et al. “Open Challenges in Synthetic Speech Detection.” In 2022 IEEE International Workshop on Information Forensics and Security (WIFS), 1–6, 2022. https://doi.org/10.1109/WIFS55849.2022.9975433.

Gerstner, C. et al. “Deepfakes: Is a Picture Worth a Thousand Lies?,” The Next Wave, p. 41, 2021. https://media.defense.gov/2021/Jul/06/2002756456/-1/-1/0/TNW_23-1.PDF.

Dhariwal, P. et al. “Diffusion Models Beat GANs on Image Synthesis,” n.d. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

“Eurasia Group | The Top Risks of 2023.” Accessed February 10, 2024. https://www.eurasiagroup.net/issues/top-risks-2023.

Fowler, G. “What to do when you’re accused of AI cheating.” Washington Post, August 2023. https://www.washingtonpost.com/technology/2023/08/14/prove-false-positive-ai-detection-turnitin-gptzero/.

Gu, S. et al. “Vector Quantized Diffusion Model for Text-to-Image Synthesis.” In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10686–96. New Orleans, LA, USA: IEEE, 2022. https://doi.org/10.1109/CVPR52688.2022.01043.

Guo, Z. et al. “Fake Face Detection via Adaptive Manipulation Traces Extraction Network.” arXiv, December 16, 2020. http://arxiv.org/abs/2005.04945.

Hadi, M.U. et al. “A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage.” Preprint, July 10, 2023. https://doi.org/10.36227/techrxiv.23589741.v1.

Heikkilä, M. “AI image generator Midjourney blocks porn by banning words about the human reproductive system.” MIT Tech Review, February 2023. https://www.technologyreview.com/2023/02/24/1069093/ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system/.

Juefei-Xu, F. et al. “Countering Malicious DeepFakes: Survey, Battleground, and Horizon.” International Journal of Computer Vision 130, no. 7 (July 2022): 1678–1734. https://doi.org/10.1007/s11263-022-01606-8.

Klee, M. “She Was Falsely Accused of Cheating With AI—And She Won’t Be the Last.” Rolling Stone, June 2023. https://www.rollingstone.com/culture/culture-features/student-accused-ai-cheating-turnitin-1234747351/.

Koopman, M. et al. “Detection of Deepfake Video Manipulation.” Proceedings of the 20th Irish Machine Vision and Image Processing conference, 2018. https://www.researchgate.net/profile/Zeno-Geradts/publication/329814168_Detection_of_Deepfake_Video_Manipulation/links/5c1bdf7da6fdccfc705da03e/Detection-of-Deepfake-Video-Manipulation.pdf.

Krishna, K. et al. “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense.” arXiv, 2023. https://arxiv.org/abs/2303.13408.

Leibowicz, C. et al. “The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media.” In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 736–44. AIES ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3461702.3462584.

“LLM - Detect AI Generated Text.” Accessed February 8, 2024. https://kaggle.com/competitions/llm-detect-ai-generated-text.

Masood, M. et al.“Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward.” Applied Intelligence 53, no. 4 (February 2023): 3974–4026. https://doi.org/10.1007/s10489-022-03766-z.

“Labeling AI-Generated Images on Facebook, Instagram and Threads,” Meta, February 6, 2024. https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/.

Midjourney. “Midjourney.” Accessed February 10, 2024. https://www.midjourney.com/home.

Moranchel, T. ““Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection,” June 2020. https://arxiv.org/abs/2001.00179.

Nichol, A. et al. “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models.” arXiv, March 8, 2022. http://arxiv.org/abs/2112.10741.

“Overview - C2PA.” Accessed February 10, 2024. https://c2pa.org/.

Rana, M. et al. “Deepfake Detection: A Systematic Literature Review.” IEEE Access 10 (2022): 25494–513. https://doi.org/10.1109/ACCESS.2022.3154404.

Rombach, R. et al. “High-Resolution Image Synthesis with Latent Diffusion Models.” In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10674–85. New Orleans, LA, USA: IEEE, 2022. https://doi.org/10.1109/CVPR52688.2022.01042.

Sadasivan, V. et al. “Can AI-Generated Text Be Reliably Detected?” arXiv, March 17, 2023. http://arxiv.org/abs/2303.11156.

Shamsolmoali, P. et al. “Image Synthesis with Adversarial Networks: A Comprehensive Survey and Case Studies.” arXiv, December 26, 2020. http://arxiv.org/abs/2012.13736.

Ting, K.M. “Confusion Matrix.” Encyclopedia of Machine Learning. Springer (2011): 209. https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_157.

Uhl, N. “Synthetic Media Transparency Methods: Indirect Disclosure.” Partnership on AI, December 19, 2023. https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/.

WhatIs. “What Is Deepfake AI? A Definition from TechTarget.” Accessed February 10, 2024. https://www.techtarget.com/whatis/definition/deepfake.

Wu, J. et al. “A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions.” arXiv, October 24, 2023. http://arxiv.org/abs/2310.14724.

“Wukong.” Accessed February 10, 2024. https://xihe.mindspore.cn/modelzoo/wukong.

Yang, X. et al. “A Survey on Detection of LLMs-Generated Content.” arXiv, October 24, 2023. https://doi.org/10.48550/arXiv.2310.15654.

Yu, I. et al. “Manipulation Classification for JPEG Images Using Multi-Domain Features.” IEEE Access 8 (2020): 210837–54. https://doi.org/10.1109/ACCESS.2020.3037735.

Testing and Evaluating Provenance Data Tracking and Synthetic Content Detection Techniques

Abdusalomov, A. et al. “Evaluating Synthetic Medical Images Using Artificial Intelligence with the GAN Algorithm.” Sensors 23, no. 7 (January 2023): 3440. https://doi.org/10.3390/s23073440.

Agnese, J. et al. “A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis.” WIREs Data Mining and Knowledge Discovery 10, no. 4 (2020): e1345. https://doi.org/10.1002/widm.1345.

AlBadawy, E. et al. “Detecting AI-Synthesized Speech Using Bispectral Analysis.” In CVPR 2019, 2019. https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19b.pdf.

Amini, M. et al. “Multichannel color image watermark detection utilizing vector-based hidden Markov model.” 2017 IEEE International Symposium on Circuits and Systems (ISCAS) (2017): 1-4. https://ieeexplore.ieee.org/abstract/document/8050596.

Beekhof, F. et al. “Content Authentication and Identification under Informed Attacks.” In 2012 IEEE International Workshop on Information Forensics and Security (WIFS), 133–38, 2012. https://doi.org/10.1109/WIFS.2012.6412638.

Birhane, A. et al. “AI auditing: The Broken Bus on the Road to AI Accountability.” arXiv, January 2024. https://arxiv.org/abs/2401.14462.

Boato, G. et al. “TrueFace: A Dataset for the Detection of Synthetic Face Images from Social Networks.” In 2022 IEEE International Joint Conference on Biometrics (IJCB), 1–7, 2022. https://doi.org/10.1109/IJCB54206.2022.10007988.

Borji, A. “Pros and Cons of GAN Evaluation Measures.” Computer Vision and Image Understanding 179 (February 1, 2019): 41–65. https://doi.org/10.1016/j.cviu.2018.10.009.

Bradley, A. “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms.” Pattern Recognition 30, no. 7 (July 1, 1997): 1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

Carlini, N. et al. “On Evaluating Adversarial Robustness,” February 18, 2019. https://doi.org/10.48550/arXiv.1902.06705.

Carlini, N. et al. “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods.” In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 3–14. AISec ’17. New York, NY, USA: Association for Computing Machinery, 2017. https://doi.org/10.1145/3128572.3140444.

Carlini, N. et al. “Towards Evaluating the Robustness of Neural Networks.” arXiv, March 22, 2017. https://doi.org/10.48550/arXiv.1608.04644.

Caruana, R. et al. “An Empirical Comparison of Supervised Learning Algorithms.” In Proceedings of the 23rd International Conference on Machine Learning, 161–68. ICML ’06. New York, NY, USA: ACM, 2006. https://doi.org/10.1145/1143844.1143865.

"Cataloguing LLM Evaluations." Infocomm Media Development Authority, AI Verify Foundation, n.d. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf.

Chen, M. et al. "Evaluating Large Language Models Trained on Code." arXiv, July 14, 2021. https://doi.org/10.48550/arXiv.2107.03374.

Ciftci, U. et al. "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 1–1. https://doi.org/10.1109/TPAMI.2020.3009287.

Clark, E. et al. "All That's `Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 7282–96. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.acl-long.565.

Dou, Y. "Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1 (2022). https://aclanthology.org/2022.acl-long.501/.

Dowson, D. et al. "The Fréchet Distance between Multivariate Normal Distributions." Journal of Multivariate Analysis 12, no. 3 (September 1, 1982): 450–55. https://doi.org/10.1016/0047-259X(82)90077-X.

Dugan, L. et al. "RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, edited by Qun Liu and David Schlangen, 189–96. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.25.

Epstein, Ziv, et al. "What Label Should Be Applied to Content Produced by Generative AI?" PsyArXiv, July 2023. https://osf.io/preprints/psyarxiv/v4mfz.

Friedman, S. et al. "Provenance as a Substrate for Human Sensemaking and Explanation of Machine Collaborators." In 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1014–19, 2021. https://doi.org/10.1109/SMC52423.2021.9659278.

Gafurov, D. et al. "Spoof Attacks on Gait Authentication System." IEEE Transactions on Information Forensics and Security 2, no. 3 (September 2007): 491–502. https://doi.org/10.1109/TIFS.2007.902030.

Gebru, T. et al. "Datasheets for Datasets." CACM (December 2021). https://arxiv.org/abs/1803.09010.

"Glossary." Trustworthy & Responsible AI Resource Center, NIST, DOC, n.d. https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary.

Goodfellow, Ian. "NIPS 2016 Tutorial: Generative Adversarial Networks." arXiv:1701.00160 [Cs], April 3, 2017. http://arxiv.org/abs/1701.00160.

Gragnaniello, D. "Detection of AI-Generated Synthetic Faces." In Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks, edited by Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch, 191–212. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-87664-7_9.

Hämäläinen, P. et al. “Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study.” In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–19. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3544548.3580688.

Hartmann, J. et al. “The Power of Generative Marketing: Can Generative AI Reach Human-Level Visual Marketing Content?” SSRN Scholarly Paper. Rochester, NY, July 12, 2023. https://doi.org/10.2139/ssrn.4597899.

Hernandez, J.R. et al. “Statistical analysis of watermarking schemes for copyright protection of images.” Proceedings of the IEEE, 87 (7) (July 1999). https://ieeexplore.ieee.org/document/771069.

Heusel, M. et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.” In Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Hong, Y. et al. “How Generative Adversarial Networks and Their Variants Work: An Overview.” ACM Computing Surveys 52, no. 1 (February 13, 2019): 10:1-10:43. https://doi.org/10.1145/3301282.

Isola, P. et al. “Image-to-Image Translation with Conditional Adversarial Networks.” In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967–76, 2017. https://doi.org/10.1109/CVPR.2017.632.

Jalui, K. et al. “Synthetic Content Detection in Deepfake Video Using Deep Learning.” In 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), 01–05, 2022. https://doi.org/10.1109/GCAT55367.2022.9972081.

Japkowicz, N. et al. Evaluating Learning Algorithms A Classification Perspective. Cambridge University Press, 2014. http://www.cambridge.org/us/academic/subjects/computer-science/pattern-recognition-and-machine-learning/evaluating-learning-algorithms-classification-perspective.

Japkowicz, N. et al. “Performance Evaluation for Learning Algorithms: Techniques, Application and Issues.” International Conference on Machine Learning (ICML) 2012, June 26, 2012. http://www.mohakshah.com/tutorials/icml2012/Tutorial-ICML2012/Tutorial_at_ICML_2012.html.

Karsh, R. et al. “Image Authentication Based on Robust Image Hashing with Geometric Correction.” Multimedia Tools and Applications 77, no. 19 (October 1, 2018): 25409–29. https://doi.org/10.1007/s11042-018-5799-6.

Khrulkov, V. et al. “Geometry Score: A Method For Comparing Generative Adversarial Networks.” In Proceedings of the 35th International Conference on Machine Learning, 2621–29. PMLR, 2018. https://proceedings.mlr.press/v80/khrulkov18a.html.

Kosuru, S. et al. “Efficiency Evaluation Parameters of Digital Image Watermarking Techniques.” In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 1:1077–82, 2023. https://doi.org/10.1109/ICACCS57279.2023.10113036.

Kreps, S. et al. “All the News That’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation.” Journal of Experimental Political Science 9, no. 1 (March 2022): 104–17. https://doi.org/10.1017/XPS.2020.37.

Lage, I. et al. “Human Evaluation of Models Built for Interpretability.” Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7 (2019). https://ojs.aaai.org/index.php/HCOMP/article/view/5280.

Lee, S. et al. “Towards Better Understanding of Training Certifiably Robust Models against Adversarial Examples.” In Advances in Neural Information Processing Systems, 34:953–64. Curran Associates, Inc., 2021. https://proceedings.neurips.cc/paper/2021/hash/07c5807d0d927dcd0980f86024e5208b-Abstract.html.

Leibowicz, C. et al. “The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media.” In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 736–44. AIES ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3461702.3462584.

Liang, P. et al. “Holistic Evaluation of Language Models,” November 16, 2022. https://doi.org/10.48550/arXiv.2211.09110.

Long, J. et al. “Fully Convolutional Networks for Semantic Segmentation,” 3431–40, 2015. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

Longpre, S. et al. “The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI.” arXiv, November 4, 2023. https://doi.org/10.48550/arXiv.2310.16787.

Lucic, M. et al. “Are GANs Created Equal? A Large-Scale Study.” In Proceedings of the 32nd International Conference on Neural Information Processing Systems, 698–707. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018. https://arxiv.org/abs/1711.10337.

Lüthi, P. et al. “Distributed Ledger for Provenance Tracking of Artificial Intelligence Assets.” In Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers, 411–26. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-42504-3_26.

Martin, A. et al. “The DET Curve in Assessment of Detection Task Performance.” In Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997. Rhodes, Greece, 1997. http://www.isca-speech.org/archive/eurospeech_1997/e97_1895.html.

Mitchell, M. et al. “Model Cards for Model Reporting.” FAT* '19 (2019). https://dl.acm.org/doi/10.1145/3287560.3287596.

“Model Card Guidebook.” Hugging Face, n.d. https://huggingface.co/docs/hub/main/model-card-guidebook.

Morales-García, J. et al. “Evaluation of Synthetic Data Generation for Intelligent Climate Control in Greenhouses.” Applied Intelligence 53, no. 21 (November 1, 2023): 24765–81. https://doi.org/10.1007/s10489-023-04783-2.

Nakagawa, T. et al. “How Provenance Helps Quality Assurance Activities in AI/ML Systems.” In Proceedings of the Second International Conference on AI-ML Systems, 1–9. AIML Systems ’22. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3564121.3564801.

Nangia, N. et al. “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models.” arXiv:2010.00133 [Cs], September 30, 2020. http://arxiv.org/abs/2010.00133.

Neekhara, P. et al. “FaceSigns: Semi-Fragile Neural Watermarks for Media Authentication and Countering Deepfakes.” arXiv, April 4, 2022. https://doi.org/10.48550/arXiv.2204.01960.

Olsson, C. et al. “Skill Rating for Generative Models.” arXiv:1808.04888 [Cs, Stat], August 14, 2018. http://arxiv.org/abs/1808.04888.

Vassilev, A. et al. “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (Draft).” National Institute of Standards and Technology, March 8, 2023. https://csrc.nist.gov/pubs/ai/100/2/e2023/final.

Papernot, N. et al. “Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples.” arXiv, May 23, 2016. https://doi.org/10.48550/arXiv.1605.07277.

Papernot, N. et al. “Practical Black-Box Attacks against Machine Learning.” In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 506–19. ASIA CCS ’17. New York, NY, USA: Association for Computing Machinery, 2017. https://doi.org/10.1145/3052973.3053009.

Phillips, P. J. et al. “FERET Evaluation Methodology for Face-Recognition Algorithms.” NIST IR. National Institute of Standards and Technology, October 1, 1998. https://www.nist.gov/publications/feret-evaluation-methodology-face-recognition-algorithms.

Phillips, P. J. et al. “The FERET Evaluation Methodology for Face-Recognition Algorithms.” IEEE Transactions on Pattern Analysis and Machine Intelligence 22, no. 10 (October 2000): 1090–1104. https://doi.org/10.1109/34.879790.

Pillutla, K. et al. “MAUVE: Measuring the Gap Between Neural Text and Human Text Using Divergence Frontiers.” In Advances in Neural Information Processing Systems, 34:4816–28. Curran Associates, Inc., 2021. https://proceedings.neurips.cc/paper/2021/hash/260c2432a0eecc28ce03c10dadc078a4-Abstract.html.

Preu, E. et al. “Perception vs. Reality: Understanding and Evaluating the Impact of Synthetic Image Deepfakes over College Students.” In 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 0547–53, 2022. https://doi.org/10.1109/UEMCON54665.2022.9965697.

Pu, J. et al. “Deepfake Text Detection: Limitations and Opportunities.” In 2023 IEEE Symposium on Security and Privacy (SP), 1613–30, 2023. https://doi.org/10.1109/SP46215.2023.10179387.

Qin, C. et al. “A Novel Image Hashing Scheme with Perceptual Robustness Using Block Truncation Coding.” Information Sciences 361–362 (September 20, 2016): 84–99. https://doi.org/10.1016/j.ins.2016.04.036.

Raji, I. et al. “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.” FAT ’20 (January 27–30, 2020). https://dl.acm.org/doi/pdf/10.1145/3351095.3372873.

Richards, J, et al. “A Human-Centered Methodology for Creating AI FactSheets.” Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2021. http://sites.computer.org/debull/A21dec/p47.pdf.

Roman, A. et al. “Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments.” arXiv, December 2023. https://arxiv.org/abs/2312.06153.

Salimans, T. et al. “Improved Techniques for Training GANs.” In Advances in Neural Information Processing Systems, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, 29:2234–42. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Salvi, D. et al. “Synthetic Speech Detection through Audio Folding.” In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, 3–9. MAD ’23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3592572.3592844.

Salvi, D. et al. “TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection.” IEEE Access 11 (2023): 50851–66. https://doi.org/10.1109/ACCESS.2023.3276480.

Shaik, A. et al. “A Review of Hashing Based Image Authentication Techniques.” Multimedia Tools and Applications 81, no. 2 (January 1, 2022): 2489–2516. https://doi.org/10.1007/s11042-021-11649-7.

Singh, R. et al. “Video Content Authentication Techniques: A Comprehensive Survey.” Multimedia Systems 24, no. 2 (March 1, 2018): 211–40. https://doi.org/10.1007/s00530-017-0538-9.

Singla, S. et al. “Second-Order Provable Defenses against Adversarial Attacks.” arXiv, June 1, 2020. https://doi.org/10.48550/arXiv.2006.00731.

Sokolova, M. et al. “A Systematic Analysis of Performance Measures for Classification Tasks.” Information Processing & Management 45, no. 4 (July 2009): 427–37. https://doi.org/10.1016/j.ipm.2009.03.002.

Srivastava, A. et al. “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models.” arXiv, June 12, 2023. https://doi.org/10.48550/arXiv.2206.04615.

Steinebach, M., et al. “StirMark Benchmark: Audio Watermarking Attacks.” In Proceedings International Conference on Information Technology: Coding and Computing, 49–54, 2001. https://doi.org/10.1109/ITCC.2001.918764.

Szegedy, C. et al. “Intriguing Properties of Neural Networks.” arXiv, February 19, 2014. https://doi.org/10.48550/arXiv.1312.6199.

Theis, L. et al. “A Note on the Evaluation of Generative Models.” arXiv, April 24, 2016. https://doi.org/10.48550/arXiv.1511.01844.

Toff, B. et al. ““Or they could just not use it?”: The Paradox of AI Disclosure for Audience Trust in News.” arXiv, December 2023. https://osf.io/preprints/socarxiv/mdvak.

Uchendu, A. et al. “TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation.” In Findings of the Association for Computational Linguistics: EMNLP 2021, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 2001–16. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.172.

Uchendu, A. et al. “Tutorial on Artificial Text Detection.” Tutorial, The 15th International Language Generation Conference (INLG 2022), July 18, 2022. https://artificial-text-detection.github.io/.

Vives, C. “NoiseLearner: An Unsupervised, Content-Agnostic Approach to Detect Deepfake Images.” Virginia Tech, 2022. http://hdl.handle.net/10919/109381.

Voloshynovskiy, S. et al. “Attack Modelling: Towards a Second Generation Watermarking Benchmark.” Signal Processing, Special section on Information theoretic aspects of digital watermarking, 81, no. 6 (June 1, 2001): 1177–1214. https://doi.org/10.1016/S0165-1684(01)00039-1.

Voloshynovskiy, S. et al. “Attacks on Digital Watermarks: Classification, Estimation Based Attacks, and Benchmarks.” IEEE Communications Magazine 39, no. 8 (August 2001): 118–26. https://doi.org/10.1109/35.940053.

Wang, Z., et al. “Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures.” IEEE Signal Processing Magazine 26, no. 1 (January 2009): 98–117. https://doi.org/10.1109/MSP.2008.930649.

Weng, L. et al. “Towards an Understanding and Explanation for Mixed-Initiative Artificial Scientific Text Detection.” arXiv, April 2023. https://arxiv.org/pdf/2304.05011.pdf.

Willis, C. et al. “Analysis and Synthesis of Metadata Goals for Scientific Data.” Journal of the American Society for Information Science and Technology, 63 (8) (August 2012): 1505-1520. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22683.

Wu, J. et al. "A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions." arXiv, October 2023. https://arxiv.org/pdf/2310.14724.pdf.

Xiang, Z. et al. “Forensic Analysis of Video Files Using Metadata.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2021): 1042-1051. https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/Xiang_Forensic_Analysis_of_Video_Files_Using_Metadata_CVPRW_2021_paper.html.

Xu, Q. et al. “An Empirical Study on Evaluation Metrics of Generative Adversarial Networks.” arXiv, August 16, 2018. https://doi.org/10.48550/arXiv.1806.07755.

Yang, X. et al. “A Survey on Detection of LLMs-Generated Content.” arXiv, October 24, 2023. https://doi.org/10.48550/arXiv.2310.15654.

Yu, D. et al. “Watermark Detection and Extraction Using Independent Component Analysis Method.” EURASIP Journal on Advances in Signal Processing (January 2022). https://link.springer.com/article/10.1155/S111086570200046X.

Zellers, R. et al. “Defending Against Neural Fake News.” In Advances in Neural Information Processing Systems, Vol. 32. Curran Associates, Inc., 2019. https://papers.neurips.cc/paper_files/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Zhang, X. et al. “High-Resolution Remote Sensing Image Integrity Authentication Method Considering Both Global and Local Features.” ISPRS International Journal of Geo-Information 9, no. 4 (April 2020): 254. https://doi.org/10.3390/ijgi9040254.

Zhao, Y. et al. “Perceptual Image Hashing Based on Color Structure and Intensity Gradient.” IEEE Access 8 (2020): 26041–53. https://doi.org/10.1109/ACCESS.2020.2970757.

Zimmermann, R. et al. “Increasing Confidence in Adversarial Robustness Evaluations.” arXiv, June 28, 2022. https://doi.org/10.48550/arXiv.2206.13991.

Zingarini, G. et al“M3Dsynth: A Dataset of Medical 3D Images with AI-Generated Local Manipulations.” arXiv, September 14, 2023. https://doi.org/10.48550/arXiv.2309.07973.

Zobaed, S. et al. “DeepFakes: Detecting Forged and Synthetic Media Content Using Machine Learning.” In Artificial Intelligence in Cyber Security: Impact and Implications: Security Challenges, Technical and Ethical Issues, Forensic Investigative Challenges, edited by Reza Montasari and Hamid Jahankhani, 177–201. Advanced Sciences and Technologies for Security Applications. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-88040-8_7.

Preventing and Reducing Harms from Synthetic CSAM and NCII

Balaji, K. et al. “Medical Image Analysis With Deep Neural Networks.” Deep Learning and Parallel Computing Environment for Bioengineering Systems (2019). https://www.sciencedirect.com/topics/engineering/image-classification.

Belanger, A. “Toxic Telegram group produced X’s X-rated fake AI Taylor Swift images, report says.” Ars Technica, January 2024. https://arstechnica.com/tech-policy/2024/01/fake-ai-taylor-swift-images-flood-x-amid-calls-to-criminalize-deepfake-porn/.

“DALL·E 3 System Card.” OpenAI, 2023. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf.

Davis, J. et al. “The Relationship Between Precision-Recall and ROC Curves.” Proceedings of the 23rd International Conference on Machine Learning (2006). https://www.biostat.wisc.edu/~page/rocpr.pdf.

“Detailed Class Descriptions – Text Moderation.” Hive, n.d. https://docs.thehive.ai/docs/detailed-class-descriptions-text-moderation.

Edelman, B. et al. “Watermarking in the sand.” Kempner Institute, Harvard University, November 2023. https://www.harvard.edu/kempner-institute/2023/11/09/watermarking-in-the-sand/.

Epstein, R. et al. “Girlhood Interrupted: The Erasure of Black Girls’ Childhood.” Georgetown Center on Poverty and Inequality, July 2017. https://genderjusticeandopportunity.georgetown.edu/wp-content/uploads/2020/06/girlhood-interrupted.pdf.

“Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” The White House, October 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

"FBI and Partners Issue National Public Safety Alert on Sextortion Schemes." U.S. Attorney's Office, Southern District of Indiana, January 2023. https://www.justice.gov/usao-sdin/pr/fbi-and-partners-issue-national-public-safety-alert-sextortion-schemes.

"Gendered Disinformation: Tactics, Themes, and Trends by Foreign Malign Actors." Global Engagement Center, U.S. Department of State, March 2023. https://www.state.gov/gendered-disinformation-tactics-themes-and-trends-by-foreign-malign-actors/.

Gorro, K. et al. “An Experimental Approach for Hybrid Content-based Web Page Detection” Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City (December 2022): 54–59. https://doi.org/10.1145/3582197.3582206.

“How Hashing and Matching Can Help Prevent Revictimization.” Thorn, August 2023. https://www.thorn.org/blog/hashing-detect-child-sex-abuse-imagery/.

Khrystyna. “Guide to Licensing your Photos online properly.” Fair Licensing, March 2023. https://www.fairlicensing.com/en/blog/guide_to_licensing_your_photos_online_properly.

Lakatos, S. "A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business." Graphika, December 2023. https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf.

Lanxon, N. "Why Are Deepfakes Everywhere? Can They Be Stopped?" Bloomberg, February 2024. https://www.bloomberg.com/news/articles/2024-02-09/fighting-deepfakes-whats-being-done-biden-robocalls-to-taylor-swift-ai-images.

Li, G. et al. "Warfare: Breaking the Watermark Protection of AI-Generated Content." arXiv, December 2023. https://arxiv.org/pdf/2310.07726.pdf.

"Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook." The Center for Equity, Gender and Leadership at the Haas School of Business, University of California, Berkeley, July 2020. https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf.

"Moderation." Platform.openai.com, n.d. https://platform.openai.com/docs/models/moderation.

Nichol, Al. et al. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models." Proceedings of the 39th International Conference on Machine Learning (2022). https://proceedings.mlr.press/v162/nichol22a/nichol22a.pdf.

"Open Licenses." Resources.data.gov, n.d. https://resources.data.gov/open-licenses/.

Qu, Y. et al. "Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models." ACM Conference on Computer and Communications Security (November 2023). https://arxiv.org/pdf/2305.13873.pdf.

Rando, J. "Red-Teaming the Stable Diffusion Safety Filter." ML Safety Workshop, 36th Conference on Neural Information Processing Systems (2022). https://arxiv.org/pdf/2210.04610v5.pdf.

Friedler, S. et al. "AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability." Data & Society, October 2023. https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf/.

Tabone, A. et al. "Pornographic content classification using deep-learning." DocEng (2021). https://dl.acm.org/doi/pdf/10.1145/3469096.3469867.

Thiel, D. "Investigation Finds AI Image Generation Models Trained on Child Abuse." Stanford Cyber Policy Center, December 2023. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

Thiel, D. et al. "Cross-Platform Dynamics of Self-Generated CSAM." Stanford Cyber Policy Center, July 2023. https://stacks.stanford.edu/file/druid:jd797tp7663/20230606-sio-sg-csam-report.pdf.

Upadhyay, D. et al. "Investigating the Avalanche Effect of Various Cryptographically Secure Hash Functions and Hash-Based Applications." IEEE Access, vol. 10 (2022): 112472-112486. https://ieeexplore.ieee.org/document/9923931.

Vincent, James. "Anyone can use this AI art generator – that's the risk." The Verge, September 2022. https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data.

"Visual Moderation." Hive, n.d. https://docs.thehive.ai/docs/visual-content-moderation/.

“What is Azure AI Content Safety?” Microsoft, December 2023. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview.

Application of Concepts to the NIST AI Risk Management Framework Lifecycle

“AI Risk Management Framework (AI RMF).” NIST, DOC. January 2023. https://www.nist.gov/itl/ai-risk-management-framework.

Synthetic Image Detection (Appendix D)

Sohan, M et al. (2023). “A survey on deepfake video detection datasets,” Indonesian Journal of Electrical Engineering and Computer Science. 32. https://www.researchgate.net/publication/374142887_A_survey_on_deepfake_video_detection_datasets.

Wu, J. et al. (2023). Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions,” http://arxiv.org/abs/2310.14724

Zhu, M. et al, (2023). “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image,” https://arxiv.org/pdf/2306.08571

Appendix A. Current Standards

A.1. Synthetic Content Standards and Guidelines

There are various hardware, software, and risk management standards for AI systems that are pertinent to authentication of synthetic content. The table below provides a non-exhaustive list.

Standard	Domain	Purpose	Organization(s)
ISO/IEC 38505-1:2017	Governance of data	Evaluating, directing, and monitoring the handling and use of data in organizations.	International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC)
ISO/IEC 23894:2023	AI risk management	Integrating risk management into AI-related activities and functions.	ISO and IEC
ISO/IEC JTC 1/SC 29	Audio, picture, multimedia, and hypermedia	Coding of digital information such as multimedia, environment, and user-related metadata, media security, privacy management, source authentication, and integrity verification.	ISO and IEC
IPTC Photo Metadata standard	Photos	Defining metadata structure, properties, and fields so that images are optimally described and easily accessed later.	International Press Telecommunications Council (IPTC)
ISO/IEC 2022:2021	Information security management systems	Measuring a software product based on internal security, reliability, performance efficiency, and maintainability.	ISO and IEC
ISO/IEC/IEEE 29119	Software testing	Testing across the AI lifecycle and for black box systems, which are directly useful in the context of GAI systems.	ISO, IEC, and the Institute of Electrical and Electronics Engineers (IEEE)

ISO/IEC 22989:2022	AI concepts and terminology	Establishing terminology for AI and describing concepts in the field of AI.	ISO and IEC
ISO/IEC 42001:2023	AI–Management system	Specifying requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations.	ISO and IEC
ISO/IEC TR 24027:2021	Bias in AI systems and AI-aided decision making	Describing measurement techniques and methods for assessing bias, with the aim to address and treat bias-related vulnerabilities.	ISO and IEC
ISO/IEC TS 4213:2022	Assessment of machine learning classification performance	Specifying methodologies for measuring classification performance of machine learning models, systems, and algorithms.	ISO and IEC
SMPTE 2112-10	Open Binding of Content Identifiers standard	Describes a method of binding content identifiers to media, utilizing audio watermarking, allowing the content to be identified both electronically and acoustically	Society of Motion Picture and Television Engineers (SMPTE) Technology Committee on Television and Broadband (24TB)
ATSC A/334	Audio Watermarking	Specifies the VP1 audio watermark for use with systems conforming to the ATSC 3.0 family of specifications and the format in which the audio watermark resides in a PCM audio signal	Advanced Television Systems Committee (ATSC)
ATSC A/335	Video Watermarking	Describes a video watermarking technology to robustly embed ancillary data in the transmitted pixels of a video signal	Advanced Television Systems Committee (ATSC)

A.2. Selected NIST Practices and Guidelines

NIST's past work in the realms of AI, privacy, and cybersecurity is useful with regard to authentication of synthetic content. Selected guidelines are noted below.

Framework	Description
NIST AI Risk Management Framework	Foundation for what organizations should do to manage risk for AI systems.
NIST AI RMF Playbook	Foundation for how to implement the NIST AI RMF.
AI RMF Core	Outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibly develop trustworthy AI systems.
Security and Privacy Controls for Information Systems and Organizations	A catalog of security and privacy controls for information systems and organizations to protect organizational operations and assets, individuals, other organizations, and the Nation from a diverse set of threats and risks.
Digital Identity Guidelines and 2022 Initial Public Draft (IPD) for Digital Identity Guidelines	Technical requirements for federal agencies implementing digital identity services. The 2022 Initial Public Draft (IPD) for Digital Identity Guidelines enhances fraud prevention measures from previous versions.
Privacy Framework	A tool for improving privacy through enterprise risk management.

A.3. Metadata Standards

EXIF (Exchangeable Image File Format) Metadata: A standard that specifies the formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners, and other systems. EXIF data includes details about the camera model used, shutter speed, the creation date, and location information.

IPTC (International Press Telecommunications Council) Metadata: A standard for exchanging metadata in images, particularly those used in journalism. It includes fields for information such as captions, keywords, location, and copyright.

XMP (Extensible Metadata Platform) Metadata: An ISO standard, originally created by Adobe Systems Inc., for the creation, processing, and interchange of standardized and custom metadata for digital documents (e.g., images, videos, PDFs).

ANSI/NISO Z39.87-2006 (R2017) Data Dictionary - Technical Metadata for Digital Still Images: Defines a set of metadata elements for raster digital images to enable users to develop, exchange, and interpret digital image files. The dictionary has been designed to facilitate interoperability between systems, services, and software, as well as to support the long-term management of and continuing access to digital image collections.

textMD: An XML Schema that details technical metadata for text-based digital objects. It most commonly serves as an extension schema used within the Metadata Encoding and Transmission Schema (METS) administrative metadata section. However, it could also exist as a standalone document. While textMD is attached to text files, individual document pages may additionally be defined as distinct objects with their own metadata.

ISO/IEC 11179 Metadata Registry (MDR): A standard for the management of metadata registries, designed to ensure interoperability across different systems.

Dublin Core Metadata Initiative (DCMI): One of the most-used digital metadata standards. A straightforward and adaptable set of metadata elements is offered by the DCMI, which can be utilized to characterize different kinds of digital resources. Titles, creators, subjects, descriptions, dates, formats, and identifiers are among the essential elements it provides. These components offer a basis for interoperability amongst various metadata systems and can be used to construct informative metadata.

Metadata Object Description Schema (MODS): An XML-based metadata system created by the Library of Congress. It offers specific elements for various content types, including music, photos, videos, documents, and maps, enabling a more detailed description of resources. Furthermore, it facilitates the encoding of intricate relationships among resources, making it possible to depict collections, series, or hierarchical organizations.

The Metadata Encoding and Transmission Standard (METS): A standard expressed in XML for encoding descriptive, administrative, and structural metadata regarding objects within a digital library that provides the means to convey the metadata necessary for both the management of digital objects within a repository and the exchange of such objects between repositories (or between repositories and their users).

Appendix B. Technical Tools

Selected technical tools related to digital content transparency.

Tool Name	Domain	Modality	Description
Google Deepmind SynthID	Watermarking and identification	Image Audio	Tool for watermarking and identifying AI-generated content
C2PA Tool	Content authentication	Image Video Audio Documents	Open-source tools for content authenticity and provenance
HIVE Classification APIs	Detection	Images Text	Identify AI-generated or modified images and text
AISEO	Detection	Text	Identify human text and AI-generated text
Photoguard	Deepfakes	Image	Prevents unauthorized image manipulation
Sensity	Deepfakes	Image Video	Detect Deepfake images and videos
GPTzero	Detection	Text	Detect AI-generated text
Turnitin	Detection	Text	Detect AI-generated text; specialized for student writing
RADAR	Detection	Text	A framework for AI-generated text
Resemble AI	Detection	Audio	Detect AI-generated audio and deepfakes

Tool Name	Domain	Modality	Description
Truepic Lens	Content Authentication	Image Video	Mobile camera SDK powered with C2PA standard
Serelay	Content Authentication	Image Video	Verify authenticity of captured images/videos
Attestiv	Blockchain-based authentication	Image Video Documents	Media validation and fraud detection
Copyleaks	Detection	Text Source Code	Detect AI-generated content including source code plagiarism
Azure AI Content Safety	Content Moderation	Text Image	Detects harmful user-generated and AI-generated content in applications and services
Reality Defender	Deepfakes	Text Image Video Audio	Detect deepfakes and generative content
Verify	Authentication	Image Video Audio	Inspect and verifies the content credentials of a digital content
FakeNet AI	Deepfakes	Video	Detects synthetic media
PhotoDNA	CSAM	Image Video	Detects CSAM content
CSAI Match	CSAM	Video	Detects CSAM videos

Tool Name	Domain	Modality	Description
NeuralHash	CSAM	Image	Detects CSAM on client devices
PDQ TMK+PDQF	CSAM	Image Video	Detects CSAM content
eGLYPH	Harmful Content	Audio Image Video	Alerting system to social media platforms
GIFCT	Harmful Content	Image Video	Shared hashing database to identify terrorism materials
TinEye	Retrieval	Image	Search and retrieves perceptual similar images including image source
Google reverse image search	Retrieval	Image	Search and retrieves perceptual similar images including image source
Steg.AI	Watermarking	Image Video Documents	Secures and authenticate digital assets using forensic watermarks
SAFE	Watermarking	Digital assets	digital watermark embedding and detection tool for digital assets
ZIRCON	Watermarking	Internet of Things (IOT)	a novel zero-watermarking approach to establish end-to-end data trustworthiness in an IoT network
Imatag	Watermarking	Image Video	Digital watermarking to embed secure and robust invisible watermarks during

Tool Name	Domain	Modality	Description
			the image generation process
WinstonAI	Detection	Text	AI content detection tool for text generated by LLMs
ZeroGPT	Detection	Text	AI content detection tool for text generated by LLMs

Appendix C. Provenance Data Tracking

C.1. Example Digital Watermark Use Cases

Steganography: Watermarking can be used to conceal or hide a message (text, file, image, or video) into another piece of digital content by altering textual information, altering the pixel values in an image, or inserting discrete sounds in an audio file that are covert to human (visual or auditory) detection.

Invisible forensic watermarking: Dataset watermarking, model watermarking, and steganography can be combined to secure digital assets such as tracing back content to specific models.

Copyright protection: Watermarking has been used to protect digital media content, such as images, audio, and video, from unauthorized use or distribution by embedding ownership or copyright information. This may discourage piracy and unauthorized distribution, either because the content can be detected as belonging to someone else or because an overt watermark renders the content unusable. Watermarks have also been applied by global news organizations to track and monitor the distribution of digital media content across channels or platforms, with the goal of fighting copyright infringement.

Content authentication: Watermarking can affirm the authenticity of the origin and integrity of digital content, while minimizing the chances that the content has not been tampered with or altered.

C.2. Watermarking Applications Prior to Content Creation

The application of watermarks can span from datasets and trained models till digital content generation. For example, here are some types and categories being applied:

Dataset watermarking	is a technique that embeds a unique identifier that traces the provenance of a dataset and acts as proof of ownership of digital content.
Model watermarking	is a technique that embeds a unique identifier in a model and acts as proof of ownership of digital content while preventing unauthorized uses and distribution.
Differential watermarking	is a technique that embeds a unique identifier between two data points (pixel values or features of a data table/tabular dataset).

C.3. Current Provenance-Related Initiatives

Framework	Description	Techniques Discussed	Type
Coalition for Content Provenance and Authenticity (C2PA)	The C2PA framework is an interoperable specification that "enables the authors of provenance data to	Metadata embedding Digital signatures Watermark (with Content Credential feature)	Framework

	securely bind statements of provenance data to instances of content using their unique credentials”		
The Starling Framework for Data Integrity	A set of tools and principles utilizing Web3 technology in order to store, capture, and verify content. The framework has also utilized the C2Pa specification.	- Blockchain/Web3 - Digital fingerprinting - Embedded metadata	Framework
The Numbers Protocol	“Numbers Protocol is the Decentralized Provenance Standard. It secures digital media provenance through a decentralized ecosystem and blockchain technology.” It utilizes existing standards such as the IPTC and C2PA framework as well.	- Blockchain/Web3 - Digital fingerprinting - Embedded metadata	Framework
Interoperable Digital Media Indexing	A method to record, discover and retrieve digital media on Ethereum Virtual Machine-compatible blockchains.	- Blockchain - Digital fingerprinting	Method
Partnership on AI’s Responsible Practices for Synthetic Media	Responsible practices and recommendations regarding synthetic media for three stakeholders: builders, creators, and distributors / publishers. Core concepts are consent, disclosure, and transparency.	- Watermarking - Embedded metadata	Best Practices

Swear Framework	The patented framework fingerprints and maps digital media within a Web3.0 blockchain network. Every pixel and soundbite are protected and authenticated.	- Blockchain - Digital fingerprinting - Metadata - Watermarking	Framework

Appendix D. Synthetic Content Detection

D.1. Synthetic Image Detection Datasets

The datasets below are popular detection datasets for synthetic images with the real-fake size which categorizes image content as general, face, and art.

Dataset	Image Content	(Generator Category)		Public Availability	Real Images	Fake Images
		GAN	Diffusion			
UADFV	Face	✓	x	x	241	252
FakeSpotter	Face	✓	x	x	6,000	5,000
DFFD	Face	✓	x	✓	58,703	240,336
APFDD	Face	✓	x	x	5,000	5,000
ForgeryNet	Face	✓	x	✓	1,438,201	1,457,861
DeepArt	Art	x	✓	✓	64,479	73,411
CNNSpot	General	✓	x	✓	362,000	362,000
IEEE VIP Cup	General	✓	✓	x	7,000	7,000
DE-FAKE	General	x	✓	x	20,000	60,000
CiFAKE	General	x	✓	✓	60,000	60,000
GenImage	General	✓	✓	✓	1,331,167	1,350,000

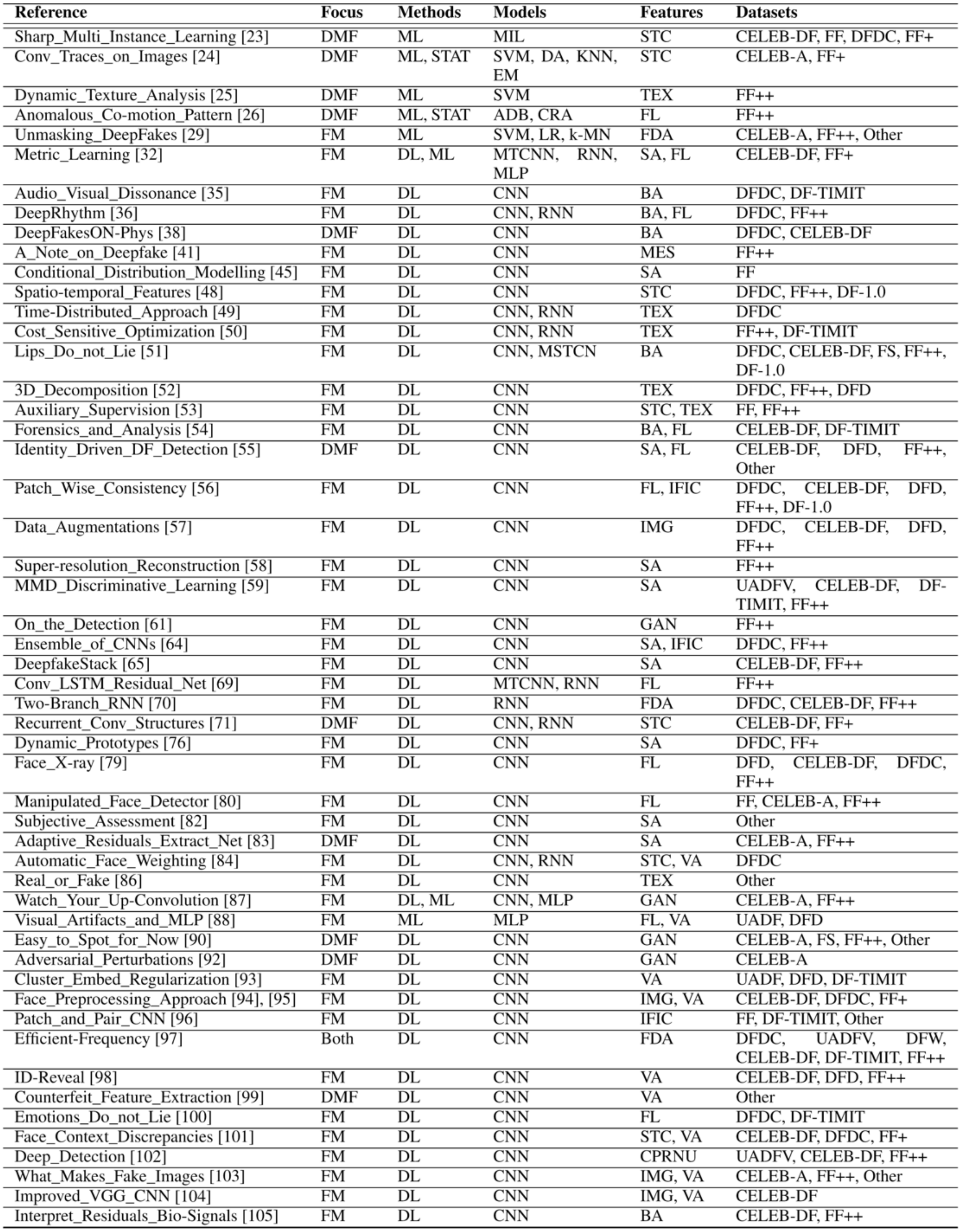
Zhu et al., “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image” Table 1, https://arxiv.org/pdf/2306.08571

D.2. Synthetic Video Detection Datasets

There are various deepfake detection datasets used in numerous studies for training and testing purposes. Deepfake detection datasets have enabled rapid advances in the field. However, there is a limit to those datasets: Authentic videos in these datasets are filmed with volunteer actors in limited scenes, while synthetic videos are created by researchers using a few deepfake tools available.

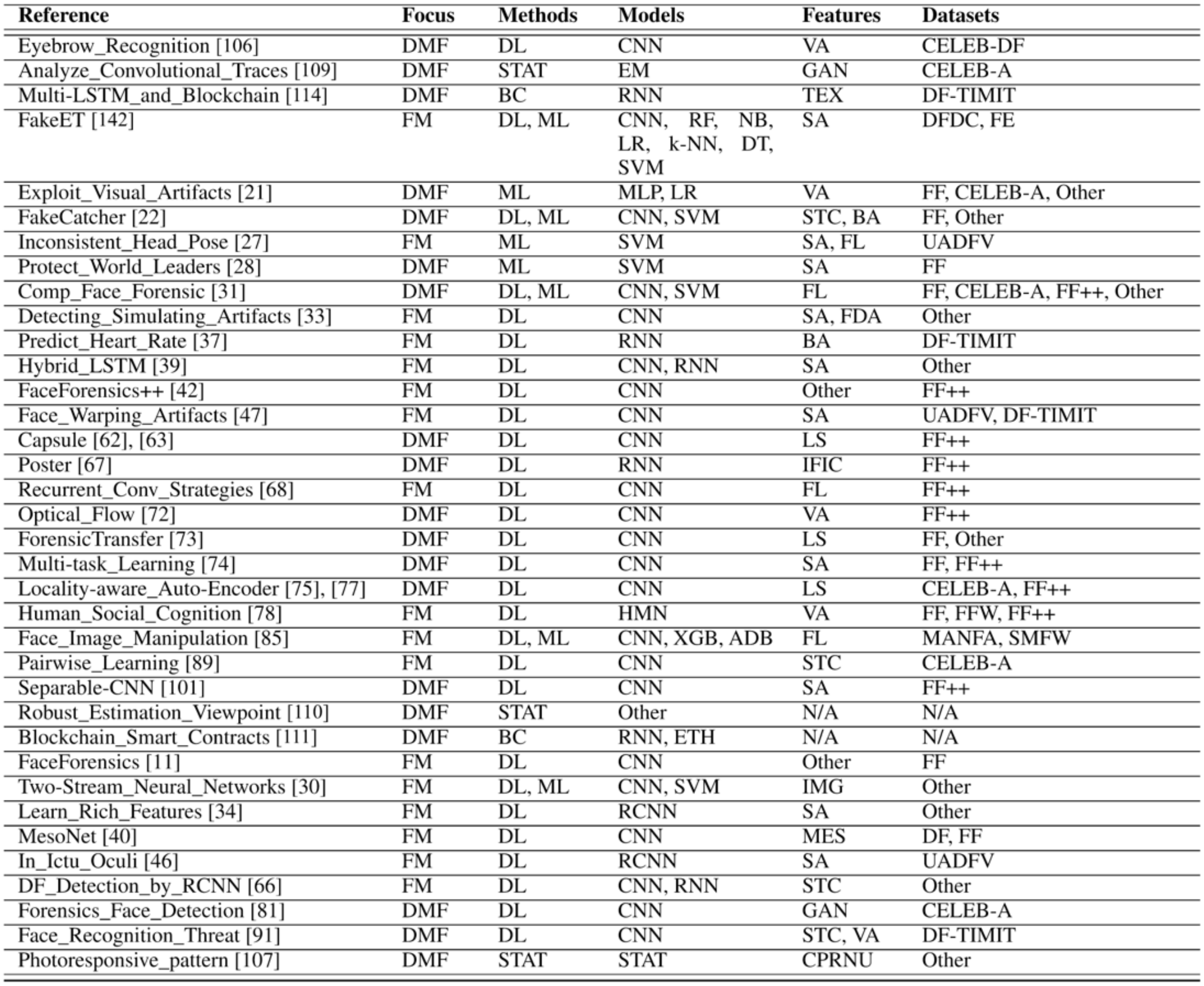
D.3. Synthetic Video (Deepfakes) Detection Methods and Results

The tables below summarize recent deepfake detection methods and their DL- and ML-based results.



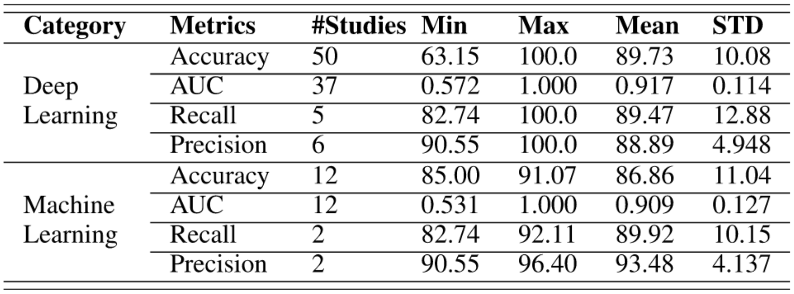
Reference	Focus	Methods	Models	Features	Datasets
Sharp_Multi_Instance_Learning [23]	DMF	ML	MIL	STC	CELEB-DF, FF, DFDC, FF+
Conv_Traces_on_Images [24]	DMF	ML, STAT	SVM, DA, KNN, EM	STC	CELEB-A, FF+
Dynamic_Texture_Analysis [25]	DMF	ML	SVM	TEX	FF++
Anomalous_Co-motion_Pattern [26]	DMF	ML, STAT	ADB, CRA	FL	FF++
Unmasking_DeepFakes [29]	FM	ML	SVM, LR, k-MN	FDA	CELEB-A, FF++, Other
Metric_Learning [32]	FM	DL, ML	MTCNN, RNN, MLP	SA, FL	CELEB-DF, FF+
Audio_Visual_Dissonance [35]	FM	DL	CNN	BA	DFDC, DF-TIMIT
DeepRhythm [36]	FM	DL	CNN, RNN	BA, FL	DFDC, FF++
DeepFakesON-Phys [38]	DMF	DL	CNN	BA	DFDC, CELEB-DF
A_Note_on_Deepfake [41]	FM	DL	CNN	MES	FF++
Conditional_Distribution_Modelling [45]	FM	DL	CNN	SA	FF
Spatio-temporal_Features [48]	FM	DL	CNN	STC	DFDC, FF++, DF-1.0
Time-Distributed_Approach [49]	FM	DL	CNN, RNN	TEX	DFDC
Cost_Sensitive_Optimization [50]	FM	DL	CNN, RNN	TEX	FF++, DF-TIMIT
Lips_Do_not_Lie [51]	FM	DL	CNN, MSTCN	BA	DFDC, CELEB-DF, FS, FF++, DF-1.0
3D_Decomposition [52]	FM	DL	CNN	TEX	DFDC, FF++, DFD
Auxiliary_Supervision [53]	FM	DL	CNN	STC, TEX	FF, FF++
Forensics_and_Analysis [54]	FM	DL	CNN	BA, FL	CELEB-DF, DF-TIMIT
Identity_Driven_DF_Detection [55]	DMF	DL	CNN	SA, FL	CELEB-DF, DFD, FF++, Other
Patch_Wise_Consistency [56]	FM	DL	CNN	FL, IFIC	DFDC, CELEB-DF, DFD, FF++, DF-1.0
Data_Augmentations [57]	FM	DL	CNN	IMG	DFDC, CELEB-DF, DFD, FF++
Super-resolution_Reconstruction [58]	FM	DL	CNN	SA	FF++
MMD_Discriminative_Learning [59]	FM	DL	CNN	SA	UADFV, CELEB-DF, DF-TIMIT, FF++
On_the_Detection [61]	FM	DL	CNN	GAN	FF++
Ensemble_of_CNNs [64]	FM	DL	CNN	SA, IFIC	DFDC, FF++
DeepfakeStack [65]	FM	DL	CNN	SA	CELEB-DF, FF++
Conv_LSTM_Residual_Net [69]	FM	DL	MTCNN, RNN	FL	FF++
Two-Branch_RNN [70]	FM	DL	RNN	FDA	DFDC, CELEB-DF, FF++
Recurrent_Conv_Structures [71]	DMF	DL	CNN, RNN	STC	CELEB-DF, FF+
Dynamic_Prototypes [76]	FM	DL	CNN	SA	DFDC, FF+
Face_X-ray [79]	FM	DL	CNN	FL	DFD, CELEB-DF, DFDC, FF++
Manipulated_Face_Detector [80]	FM	DL	CNN	FL	FF, CELEB-A, FF++
Subjective_Assessment [82]	FM	DL	CNN	SA	Other
Adaptive_Residuals_Extract_Net [83]	DMF	DL	CNN	SA	CELEB-A, FF++
Automatic_Face_Weighting [84]	FM	DL	CNN, RNN	STC, VA	DFDC
Real_or_Fake [86]	FM	DL	CNN	TEX	Other
Watch_Your_Up-Convolution [87]	FM	DL, ML	CNN, MLP	GAN	CELEB-A, FF++
Visual_Artifacts_and_MLP [88]	FM	ML	MLP	FL, VA	UADF, DFD
Easy_to_Spot_for_Now [90]	DMF	DL	CNN	GAN	CELEB-A, FS, FF++, Other
Adversarial_Perturbations [92]	DMF	DL	CNN	GAN	CELEB-A
Cluster_Embed_Regularization [93]	FM	DL	CNN	VA	UADF, DFD, DF-TIMIT
Face_Preprocessing_Approach [94], [95]	FM	DL	CNN	IMG, VA	CELEB-DF, DFDC, FF+
Patch_and_Pair_CNN [96]	FM	DL	CNN	IFIC	FF, DF-TIMIT, Other
Efficient-Frequency [97]	Both	DL	CNN	FDA	DFDC, UADFV, DFW, CELEB-DF, DF-TIMIT, FF++
ID-Reveal [98]	FM	DL	CNN	VA	CELEB-DF, DFD, FF++
Counterfeit_Feature_Extraction [99]	DMF	DL	CNN	VA	Other
Emotions_Do_not_Lie [100]	FM	DL	CNN	FL	DFDC, DF-TIMIT
Face_Context_Discrepancies [101]	FM	DL	CNN	STC, VA	CELEB-DF, DFDC, FF+
Deep_Detection [102]	FM	DL	CNN	CPRNU	UADFV, CELEB-DF, FF++
What_Makes_Fake_Images [103]	FM	DL	CNN	IMG, VA	CELEB-A, FF++, Other
Improved_VGG_CNN [104]	FM	DL	CNN	IMG, VA	CELEB-DF
Interpret_Residuals_Bio-Signals [105]	FM	DL	CNN	BA	CELEB-DF, FF++

Rana et al. "Deepfake Detection: A Systematic Literature Review" Table 6, https://doi.org/10.1109/ACCESS.2022.3154404



Reference	Focus	Methods	Models	Features	Datasets
Eyebrow_Recognition [106]	DMF	DL	CNN	VA	CELEB-DF
Analyze_Convolutional_Traces [109]	DMF	STAT	EM	GAN	CELEB-A
Multi-LSTM_and_Blockchain [114]	DMF	BC	RNN	TEX	DF-TIMIT
FakeET [142]	FM	DL, ML	CNN, RF, NB, LR, k-NN, DT, SVM	SA	DFDC, FE
Exploit_Visual_Artifacts [21]	DMF	ML	MLP, LR	VA	FF, CELEB-A, Other
FakeCatcher [22]	DMF	DL, ML	CNN, SVM	STC, BA	FF, Other
Inconsistent_Head_Pose [27]	FM	ML	SVM	SA, FL	UADFV
Protect_World_Leaders [28]	DMF	ML	SVM	SA	FF
Comp_Face_Forensic [31]	DMF	DL, ML	CNN, SVM	FL	FF, CELEB-A, FF++, Other
Detecting_Simulating_Artifacts [33]	FM	DL	CNN	SA, FDA	Other
Predict_Heart_Rate [37]	FM	DL	RNN	BA	DF-TIMIT
Hybrid_LSTM [39]	FM	DL	CNN, RNN	SA	Other
FaceForensics++ [42]	FM	DL	CNN	Other	FF++
Face_Warping_Artifacts [47]	FM	DL	CNN	SA	UADFV, DF-TIMIT
Capsule [62], [63]	DMF	DL	CNN	LS	FF++
Poster [67]	DMF	DL	RNN	IFIC	FF++
Recurrent_Conv_Strategies [68]	FM	DL	CNN	FL	FF++
Optical_Flow [72]	DMF	DL	CNN	VA	FF++
ForensicTransfer [73]	DMF	DL	CNN	LS	FF, Other
Multi-task_Learning [74]	DMF	DL	CNN	SA	FF, FF++
Locality-aware_Auto-Encoder [75], [77]	DMF	DL	CNN	LS	CELEB-A, FF++
Human_Social_Cognition [78]	FM	DL	HMN	VA	FF, FFW, FF++
Face_Image_Manipulation [85]	FM	DL, ML	CNN, XGB, ADB	FL	MANFA, SMFW
Pairwise_Learning [89]	FM	DL	CNN	STC	CELEB-A
Separable-CNN [101]	DMF	DL	CNN	SA	FF++
Robust_Estimation_Viewpoint [110]	DMF	STAT	Other	N/A	N/A
Blockchain_Smart_Contracts [111]	DMF	BC	RNN, ETH	N/A	N/A
FaceForensics [11]	FM	DL	CNN	Other	FF
Two-Stream_Neural_Networks [30]	FM	DL, ML	CNN, SVM	IMG	Other
Learn_Rich_Features [34]	FM	DL	RCNN	SA	Other
MesoNet [40]	FM	DL	CNN	MES	DF, FF
In_Ictu_Oculi [46]	FM	DL	RCNN	SA	UADFV
DF_Detection_by_RCNN [66]	FM	DL	CNN, RNN	STC	Other
Forensics_Face_Detection [81]	DMF	DL	CNN	GAN	CELEB-A
Face_Recognition_Threat [91]	DMF	DL	CNN	STC, VA	DF-TIMIT
Photoresponsive_pattern [107]	DMF	STAT	STAT	CPRNU	Other

Rana et al. “Deepfake Detection: A systematic Literature Review” Table 6 continued, https://doi.org/10.1109/ACCESS.2022.3154404



Category	Metrics	#Studies	Min	Max	Mean	STD
Deep Learning	Accuracy	50	63.15	100.0	89.73	10.08
	AUC	37	0.572	1.000	0.917	0.114
	Recall	5	82.74	100.0	89.47	12.88
	Precision	6	90.55	100.0	88.89	4.948
Machine Learning	Accuracy	12	85.00	91.07	86.86	11.04
	AUC	12	0.531	1.000	0.909	0.127
	Recall	2	82.74	92.11	89.92	10.15
	Precision	2	90.55	96.40	93.48	4.137

Rana et al. “Deepfake Detection: A Systematic Literature Review” Table 9, https://doi.org/10.1109/ACCESS.2022.3154404

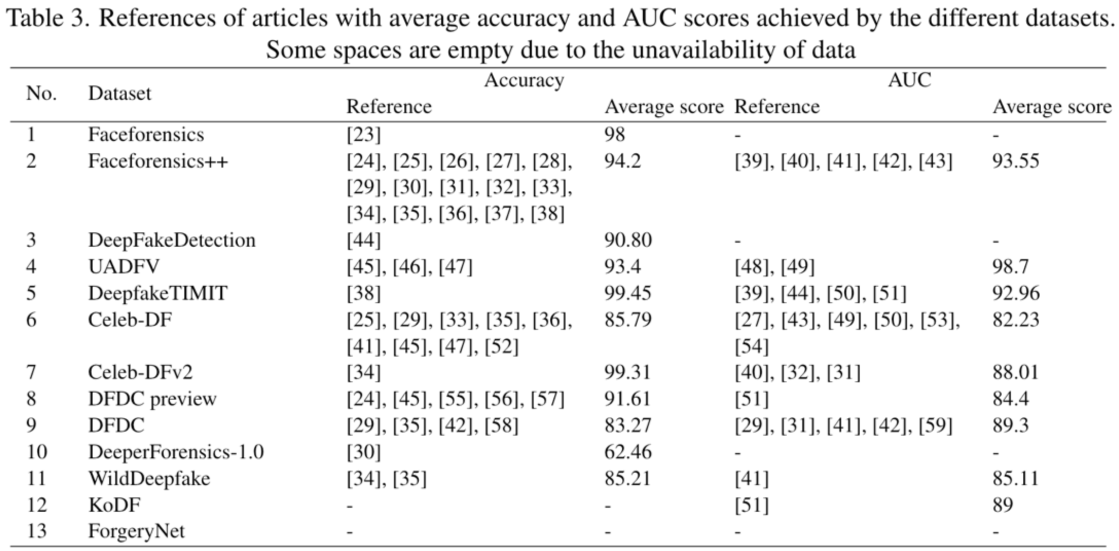
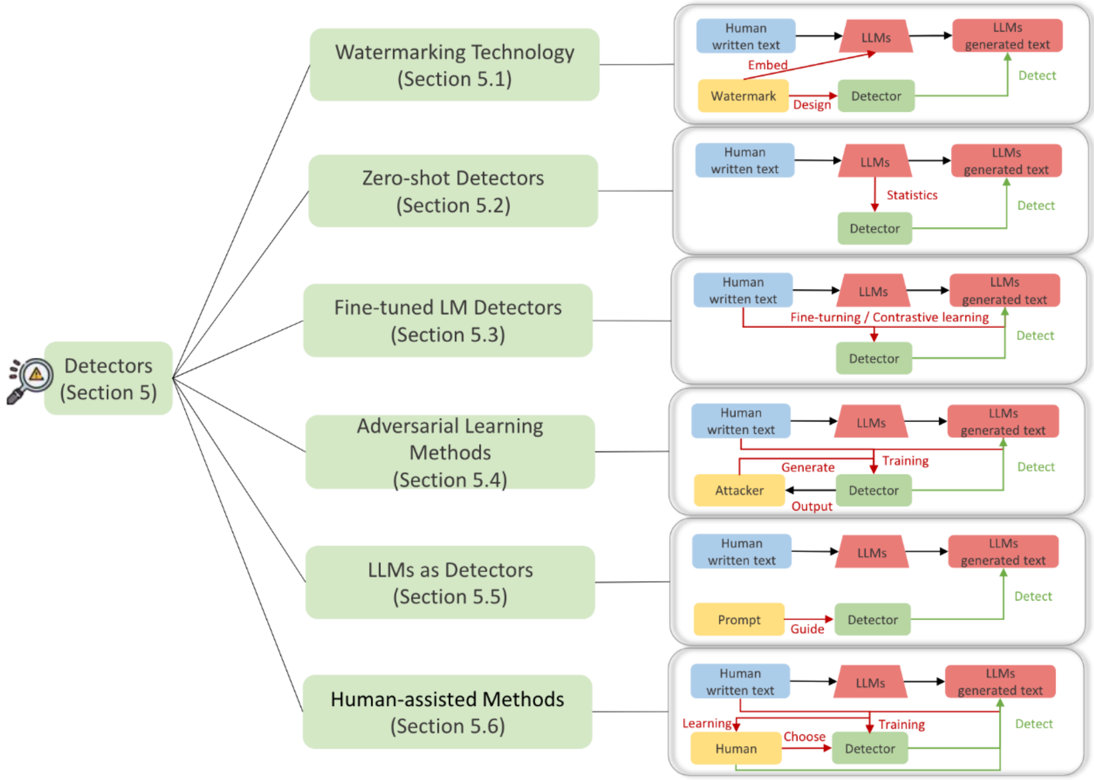


Table 3. References of articles with average accuracy and AUC scores achieved by the different datasets. Some spaces are empty due to the unavailability of data

No.	Dataset	Accuracy		AUC	
		Reference	Average score	Reference	Average score
1	Faceforensics	[23]	98	-	-
2	Faceforensics++	[24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38]	94.2	[39], [40], [41], [42], [43]	93.55
3	DeepFakeDetection	[44]	90.80	-	-
4	UADFV	[45], [46], [47]	93.4	[48], [49]	98.7
5	DeepfakeTIMIT	[38]	99.45	[39], [44], [50], [51]	92.96
6	Celeb-DF	[25], [29], [33], [35], [36], [41], [45], [47], [52]	85.79	[27], [43], [49], [50], [53], [54]	82.23
7	Celeb-DFv2	[34]	99.31	[40], [32], [31]	88.01
8	DFDC preview	[24], [45], [55], [56], [57]	91.61	[51]	84.4
9	DFDC	[29], [35], [42], [58]	83.27	[29], [31], [41], [42], [59]	89.3
10	DeeperForensics-1.0	[30]	62.46	-	-
11	WildDeepfake	[34], [35]	85.21	[41]	85.11
12	KoDF	-	-	[51]	89
13	ForgeryNet	-	-	-	-

Sohan, M. et al. “A survey on deepfake video detection datasets” Table 3, https://www.researchgate.net/publication/374142887_A_survey_on_deepfake_video_detection_datasets

D.4. Synthetic Text Detection Methods Summary


Detectors (Section 5)
Watermarking Technology (Section 5.1)
Zero-shot Detectors (Section 5.2)
Fine-tuned LM Detectors (Section 5.3)
Adversarial Learning Methods (Section 5.4)
LLMs as Detectors (Section 5.5)
Human-assisted Methods (Section 5.6)
Human written text
LLMs
LLMs generated text
Embed
Watermark
Design
Detector
Detect
Statistics
Fine-turning / Contrastive learning
Generate
Training
Attacker
Output
Prompt
Guide
Learning
Choose
Human

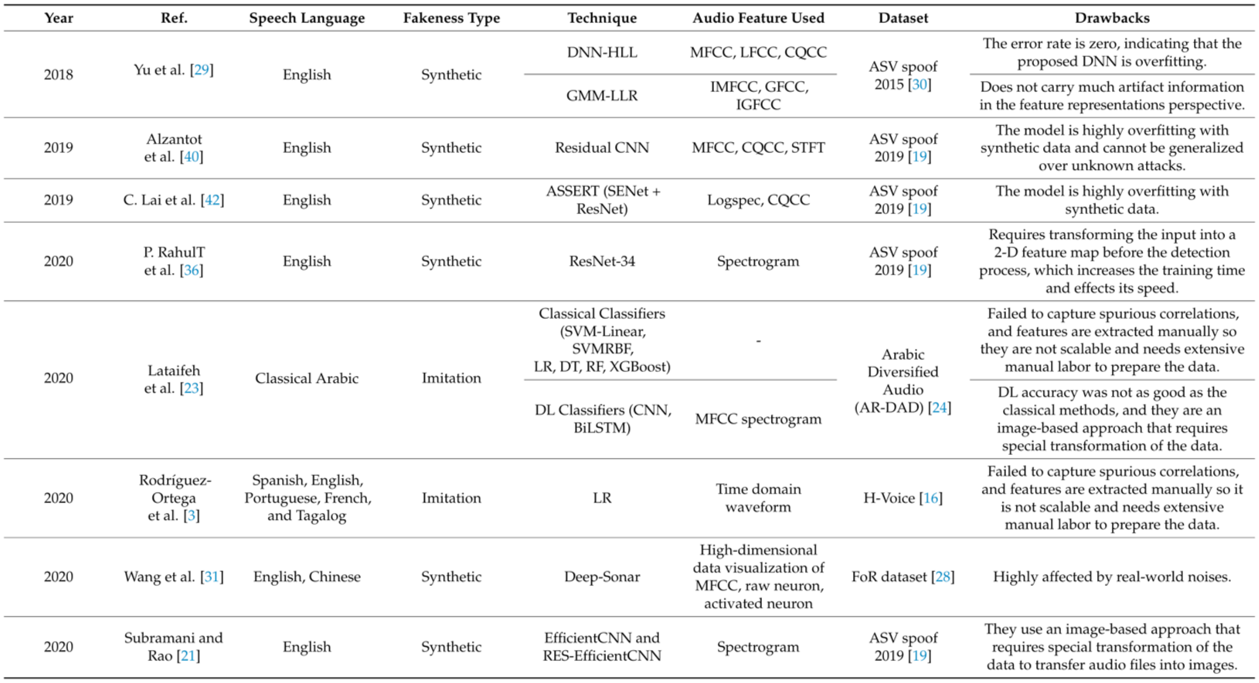
Wu, J. et al. “A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions” Figure 4, http://arxiv.org/abs/2310.14724

D.5. Benchmark Datasets for LLMs-generated Text Detection

Synthetic text datasets support the detection of synthetic text content due to their ground truth labels. For example, one dataset focuses on detecting AI-generated text using LLMs trained on a vast amount of text and code, while the other dataset is designed for long-form text and essays, containing samples of both human and AI-generated text from various language models. Studies (Wu J. et al, Tables 5 and 6) and Yang et al, Table 1) summarizes popular benchmark datasets for LLM-generated text detection. Various benchmark text corpora include synthetic and human text datasets from different domains, such as finance, medicine, news articles, web, and academic-related writings to support detection.

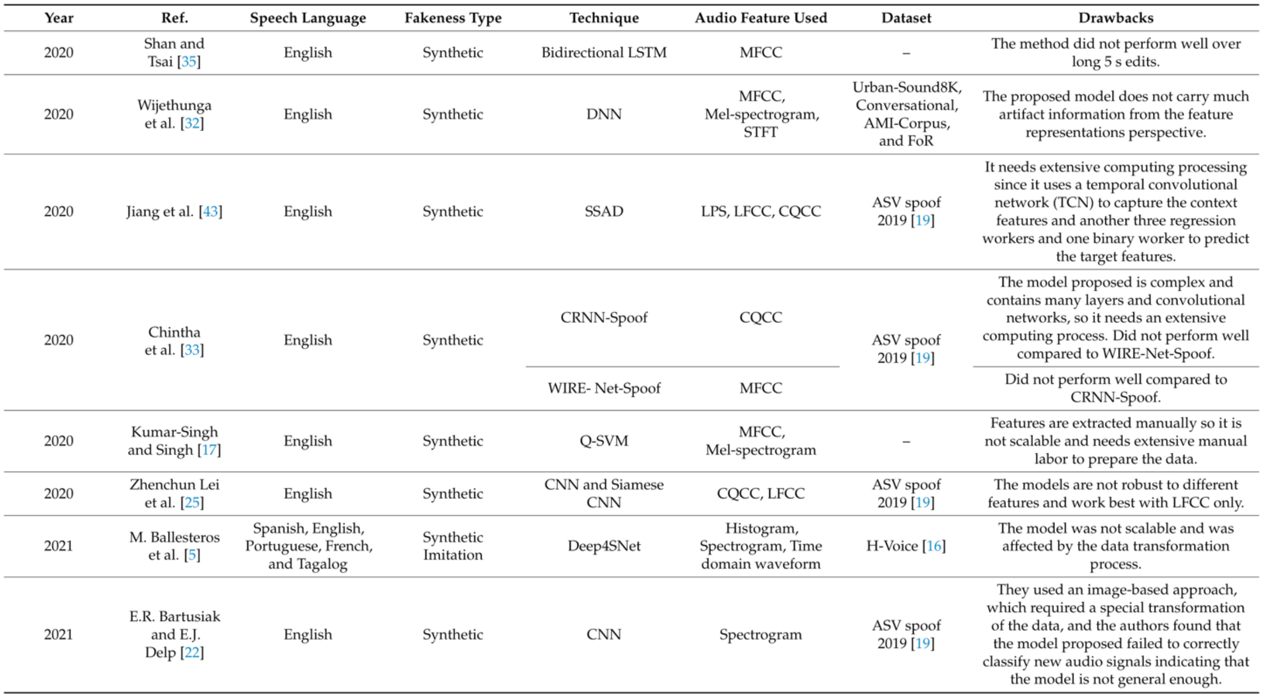
D.6. Synthetic Audio Detection Methods Summary

Table 1. Summary of AD detection methods studies surveyed.



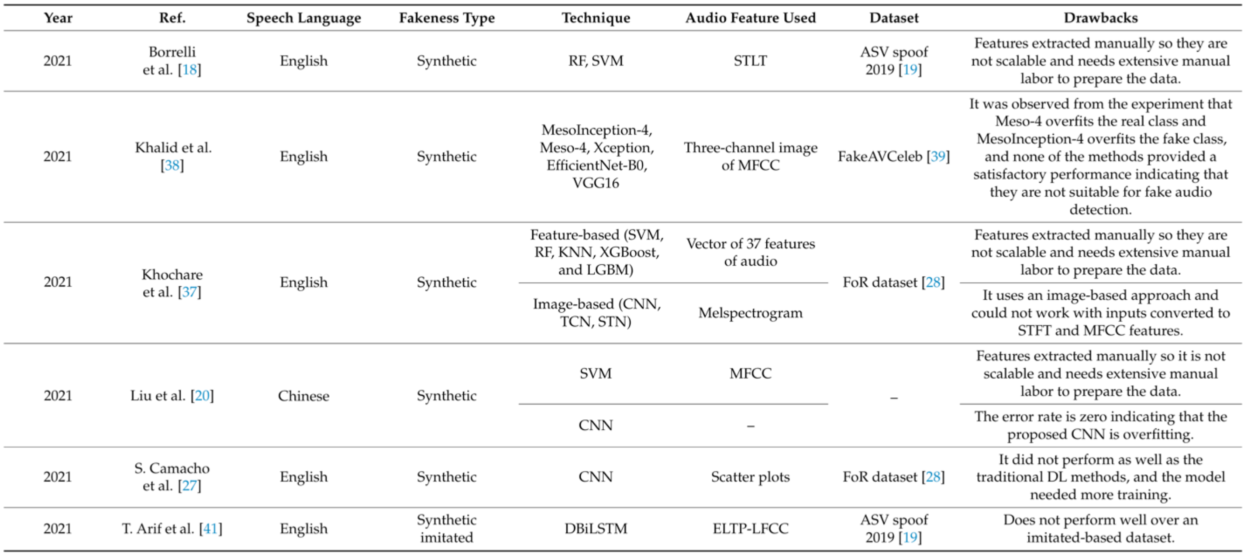
Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
2018	Yu et al. [29]	English	Synthetic	DNN-HLL	MFCC, LFCC, CQCC	ASV spoof 2015 [30]	The error rate is zero, indicating that the proposed DNN is overfitting.
				GMM-LLR	IMFCC, GFCC, IGFCC		Does not carry much artifact information in the feature representations perspective.
2019	Alzantot et al. [40]	English	Synthetic	Residual CNN	MFCC, CQCC, STFT	ASV spoof 2019 [19]	The model is highly overfitting with synthetic data and cannot be generalized over unknown attacks.
2019	C. Lai et al. [42]	English	Synthetic	ASSERT (SENet + ResNet)	Logspec, CQCC	ASV spoof 2019 [19]	The model is highly overfitting with synthetic data.
2020	P. RahulT et al. [36]	English	Synthetic	ResNet-34	Spectrogram	ASV spoof 2019 [19]	Requires transforming the input into a 2-D feature map before the detection process, which increases the training time and effects its speed.
2020	Lataifeh et al. [23]	Classical Arabic	Imitation	Classical Classifiers (SVM-Linear, SVMRBF, LR, DT, RF, XGBoost)	-	Arabic Diversified Audio (AR-DAD) [24]	Failed to capture spurious correlations, and features are extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
				DL Classifiers (CNN, BiLSTM)	MFCC spectrogram		DL accuracy was not as good as the classical methods, and they are an image-based approach that requires special transformation of the data.
2020	Rodríguez-Ortega et al. [3]	Spanish, English, Portuguese, French, and Tagalog	Imitation	LR	Time domain waveform	H-Voice [16]	Failed to capture spurious correlations, and features are extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
2020	Wang et al. [31]	English, Chinese	Synthetic	Deep-Sonar	High-dimensional data visualization of MFCC, raw neuron, activated neuron	FoR dataset [28]	Highly affected by real-world noises.
2020	Subramani and Rao [21]	English	Synthetic	EfficientCNN and RES-EfficientCNN	Spectrogram	ASV spoof 2019 [19]	They use an image-based approach that requires special transformation of the data to transfer audio files into images.

Almutairi Z. et al. "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Table 1.



Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
2020	Shan and Tsai [35]	English	Synthetic	Bidirectional LSTM	MFCC	–	The method did not perform well over long 5 s edits.
2020	Wijethunga et al. [32]	English	Synthetic	DNN	MFCC, Mel-spectrogram, STFT	Urban-Sound8K, Conversational, AMI-Corpus, and FoR	The proposed model does not carry much artifact information from the feature representations perspective.
2020	Jiang et al. [43]	English	Synthetic	SSAD	LPS, LFCC, CQCC	ASV spoof 2019 [19]	It needs extensive computing processing since it uses a temporal convolutional network (TCN) to capture the context features and another three regression workers and one binary worker to predict the target features.
2020	Chintha et al. [33]	English	Synthetic	CRNN-Spoof	CQCC	ASV spoof 2019 [19]	The model proposed is complex and contains many layers and convolutional networks, so it needs an extensive computing process. Did not perform well compared to WIRE-Net-Spoof.
				WIRE- Net-Spoof	MFCC		Did not perform well compared to CRNN-Spoof.
2020	Kumar-Singh and Singh [17]	English	Synthetic	Q-SVM	MFCC, Mel-spectrogram	–	Features are extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
2020	Zhenchun Lei et al. [25]	English	Synthetic	CNN and Siamese CNN	CQCC, LFCC	ASV spoof 2019 [19]	The models are not robust to different features and work best with LFCC only.
2021	M. Ballesteros et al. [5]	Spanish, English, Portuguese, French, and Tagalog	Synthetic Imitation	Deep4SNet	Histogram, Spectrogram, Time domain waveform	H-Voice [16]	The model was not scalable and was affected by the data transformation process.
2021	E.R. Bartusiak and E.J. Delp [22]	English	Synthetic	CNN	Spectrogram	ASV spoof 2019 [19]	They used an image-based approach, which required a special transformation of the data, and the authors found that the model proposed failed to correctly classify new audio signals indicating that the model is not general enough.

Almutairi Z. et al. “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Table 1 continued.



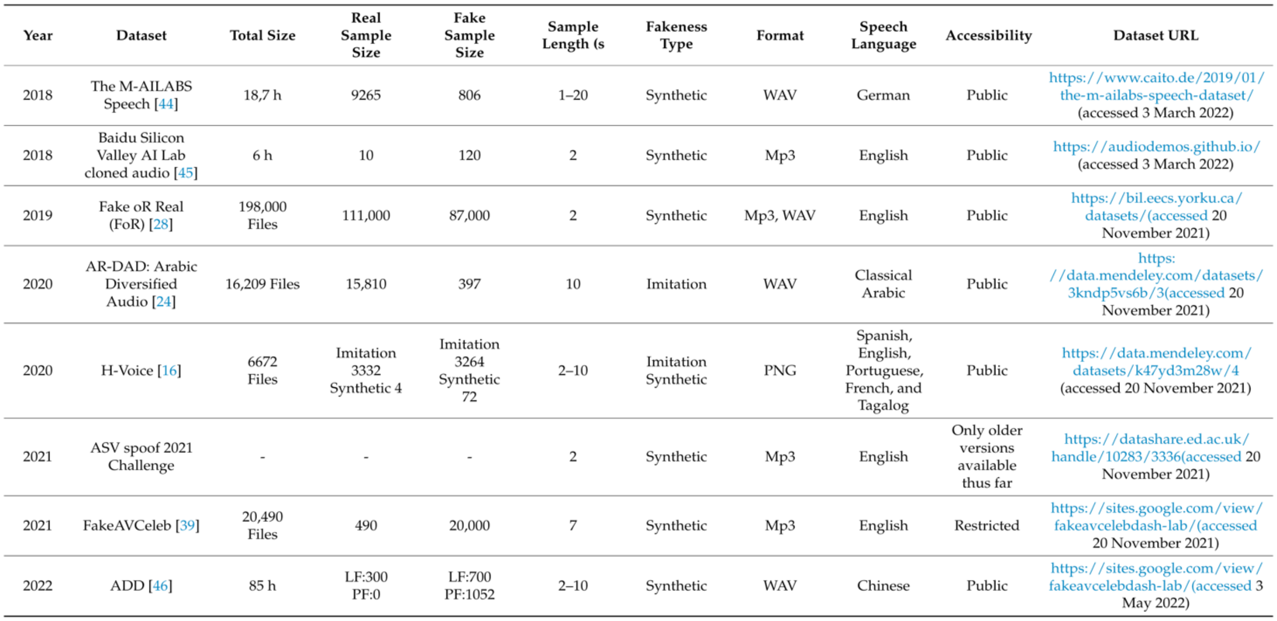
Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
2021	Borrelli et al. [18]	English	Synthetic	RF, SVM	STLT	ASV spoof 2019 [19]	Features extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
2021	Khalid et al. [38]	English	Synthetic	MesoInception-4, Meso-4, Xception, EfficientNet-B0, VGG16	Three-channel image of MFCC	FakeAVCeleb [39]	It was observed from the experiment that Meso-4 overfits the real class and MesoInception-4 overfits the fake class, and none of the methods provided a satisfactory performance indicating that they are not suitable for fake audio detection.
2021	Khochare et al. [37]	English	Synthetic	Feature-based (SVM, RF, KNN, XGBoost, and LGBM)	Vector of 37 features of audio	FoR dataset [28]	Features extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
				Image-based (CNN, TCN, STN)	Melspectrogram		It uses an image-based approach and could not work with inputs converted to STFT and MFCC features.
2021	Liu et al. [20]	Chinese	Synthetic	SVM	MFCC	–	Features extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
				CNN	–		The error rate is zero indicating that the proposed CNN is overfitting.
2021	S. Camacho et al. [27]	English	Synthetic	CNN	Scatter plots	FoR dataset [28]	It did not perform as well as the traditional DL methods, and the model needed more training.
2021	T. Arif et al. [41]	English	Synthetic imitated	DBiLSTM	ELTP-LFCC	ASV spoof 2019 [19]	Does not perform well over an imitated-based dataset.

Almutairi Z. et al. “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Table 1 continued.

D.7. Synthetic Audio Detection Datasets and Results Summary

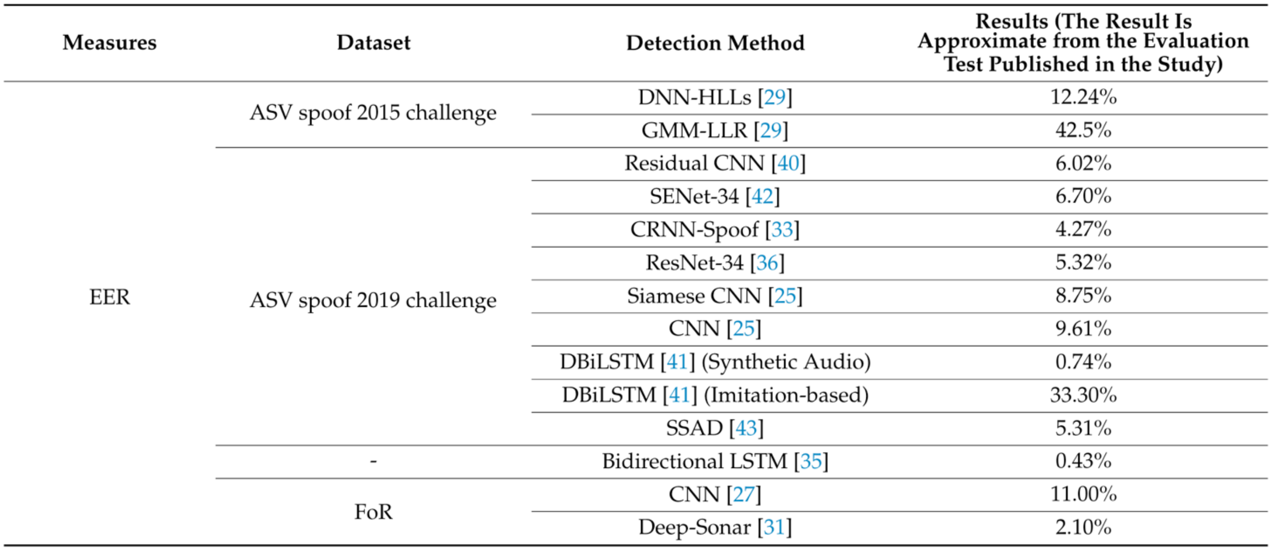
The latest datasets have been created for the purpose of synthetic audio detection methods.

The table below summarizes recent datasets. These datasets include various language speakers and use a neural voice cloning tool. The dataset that has been created can be used in the detection model to identify both imitation-based and synthetic-based audios with minimal preprocessing and training time. However, it is still necessary to create a new dataset to further enhance the detection of synthetic audio.



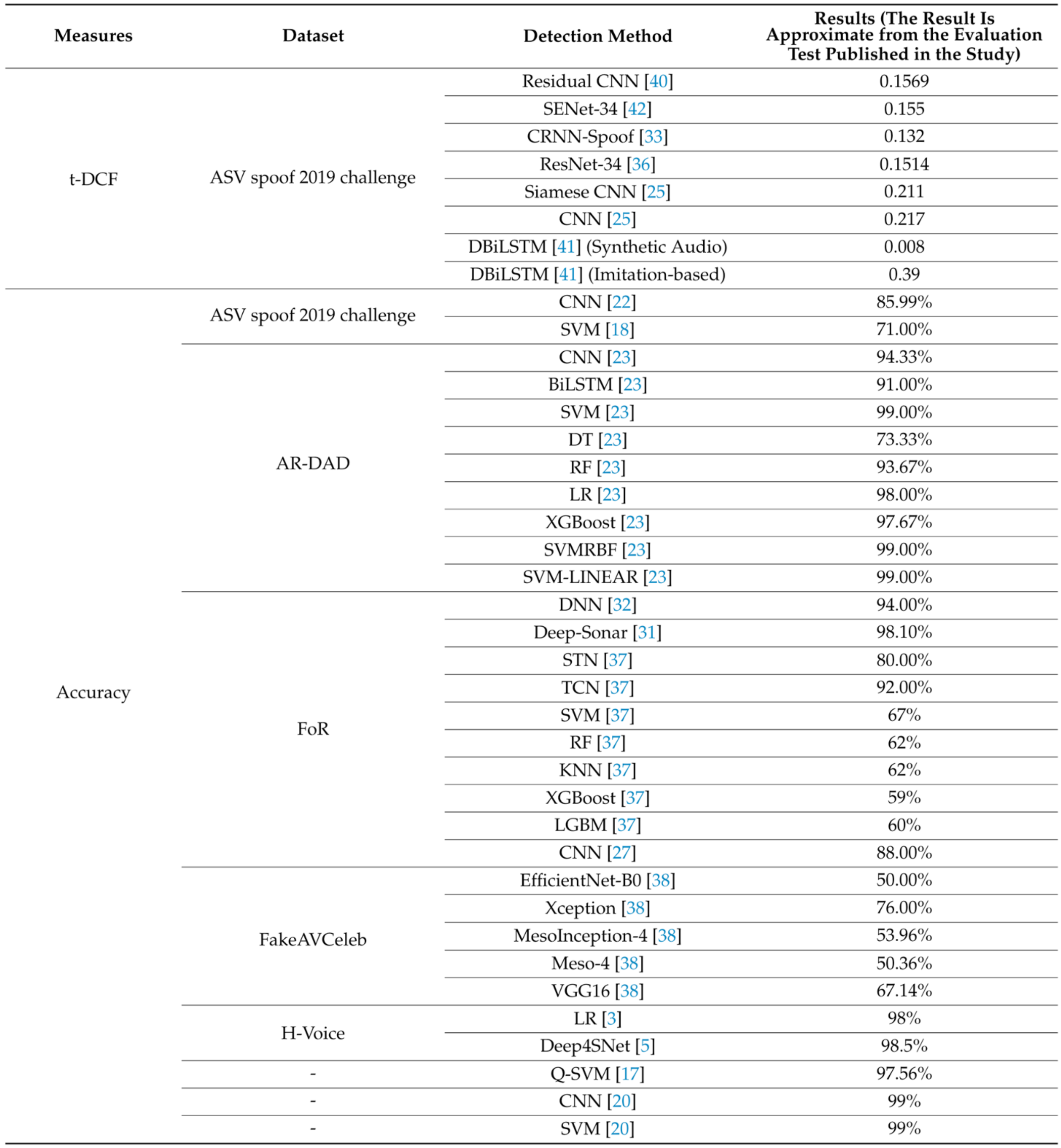
Year	Dataset	Total Size	Real Sample Size	Fake Sample Size	Sample Length (s	Fakeness Type	Format	Speech Language	Accessibility	Dataset URL
2018	The M-AILABS Speech [44]	18,7 h	9265	806	1–20	Synthetic	WAV	German	Public	https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/ (accessed 3 March 2022)
2018	Baidu Silicon Valley AI Lab cloned audio [45]	6 h	10	120	2	Synthetic	Mp3	English	Public	https://audiodemos.github.io/ (accessed 3 March 2022)
2019	Fake oR Real (FoR) [28]	198,000 Files	111,000	87,000	2	Synthetic	Mp3, WAV	English	Public	https://bil.eecs.yorku.ca/datasets/(accessed 20 November 2021)
2020	AR-DAD: Arabic Diversified Audio [24]	16,209 Files	15,810	397	10	Imitation	WAV	Classical Arabic	Public	https: //data.mendeley.com/datasets/3kndp5vs6b/3(accessed 20 November 2021)
2020	H-Voice [16]	6672 Files	Imitation 3332 Synthetic 4	Imitation 3264 Synthetic 72	2–10	Imitation Synthetic	PNG	Spanish, English, Portuguese, French, and Tagalog	Public	https://data.mendeley.com/datasets/k47yd3m28w/4 (accessed 20 November 2021)
2021	ASV spoof 2021 Challenge	-	-	-	2	Synthetic	Mp3	English	Only older versions available thus far	https://datashare.ed.ac.uk/handle/10283/3336(accessed 20 November 2021)
2021	FakeAVCeleb [39]	20,490 Files	490	20,000	7	Synthetic	Mp3	English	Restricted	https://sites.google.com/view/fakeavcelebdash-lab/(accessed 20 November 2021)
2022	ADD [46]	85 h	LF:300 PF:0	LF:700 PF:1052	2–10	Synthetic	WAV	Chinese	Public	https://sites.google.com/view/fakeavcelebdash-lab/(accessed 3 May 2022)

Almutairi Z. et al. "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Table 2.



Measures	Dataset	Detection Method	Results (The Result Is Approximate from the Evaluation Test Published in the Study)
EER	ASV spoof 2015 challenge	DNN-HLLs [29]	12.24%
		GMM-LLR [29]	42.5%
	ASV spoof 2019 challenge	Residual CNN [40]	6.02%
		SENet-34 [42]	6.70%
		CRNN-Spoof [33]	4.27%
		ResNet-34 [36]	5.32%
		Siamese CNN [25]	8.75%
		CNN [25]	9.61%
		DBiLSTM [41] (Synthetic Audio)	0.74%
		DBiLSTM [41] (Imitation-based)	33.30%
		SSAD [43]	5.31%
	-	Bidirectional LSTM [35]	0.43%
	FoR	CNN [27]	11.00%
		Deep-Sonar [31]	2.10%

Almutairi Z. et al. "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Table 3.



Measures	Dataset	Detection Method	Results (The Result Is Approximate from the Evaluation Test Published in the Study)
t-DCF	ASV spoof 2019 challenge	Residual CNN [40]	0.1569
		SENet-34 [42]	0.155
		CRNN-Spoof [33]	0.132
		ResNet-34 [36]	0.1514
		Siamese CNN [25]	0.211
		CNN [25]	0.217
		DBiLSTM [41] (Synthetic Audio)	0.008
		DBiLSTM [41] (Imitation-based)	0.39
Accuracy	ASV spoof 2019 challenge	CNN [22]	85.99%
		SVM [18]	71.00%
	AR-DAD	CNN [23]	94.33%
		BiLSTM [23]	91.00%
		SVM [23]	99.00%
		DT [23]	73.33%
		RF [23]	93.67%
		LR [23]	98.00%
		XGBoost [23]	97.67%
		SVMRBF [23]	99.00%
		SVM-LINEAR [23]	99.00%
	FoR	DNN [32]	94.00%
		Deep-Sonar [31]	98.10%
		STN [37]	80.00%
		TCN [37]	92.00%
		SVM [37]	67%
		RF [37]	62%
		KNN [37]	62%
		XGBoost [37]	59%
		LGBM [37]	60%
		CNN [27]	88.00%
	FakeAVCeleb	EfficientNet-B0 [38]	50.00%
		Xception [38]	76.00%
		MesoInception-4 [38]	53.96%
		Meso-4 [38]	50.36%
		VGG16 [38]	67.14%
	H-Voice	LR [3]	98%
		Deep4SNet [5]	98.5%
	-	Q-SVM [17]	97.56%
	-	CNN [20]	99%
	-	SVM [20]	99%

Almutairi Z. et al. “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Table 3 continued.

Appendix E. Testing and Evaluation

E.1. Background: Testing, and Evaluating Synthetic Content Generators

Many of the experimental methods and practices that test the quality of synthetic content are also used to test, and evaluate digital transparency techniques. We summarize these techniques here. The tests covered here are accuracy tests that are numerically scored using a numerical accuracy metric scored by a computer.

E.2. Background: A Common Testing Experiment Framework

Many of the experiments testing synthetic content authentication techniques have a common design. often used to test a supervised machine learning classifier, which we now describe in more detail. First, these experiments use a test or evaluation dataset of media (e.g., images, text segments, audio segments, video segments, code). Then, the experiment provides input to the AI system. These inputs are either a single content piece or a pair of content pieces. When the input is a single content piece, the system will be asked to say how likely it is that that content is of the "positive" class. The answer is often expressed as a real number between 0 and 1. A value of 0 means the content is considered to be "negative," while a value of 1 means the content piece is considered to be certainly a positive. The higher the real number, the more likely the system believes that the content piece is of the "positive" class. The meaning of the positive class varies: For detection, a 1 indicates that the input is fake (i.e., AI-generated). For pairs of content, the system will be given two images and asked to show how often the content pair is of a positive class, again with a real number from 0 to 1.

In both of these contexts, each trial has a ground truth of 0 or 1 and a system real number between 0 and 1. This output can be scored as machine learning classification tasks are often scored, using a performance accuracy metric. A variety of classification accuracy metrics, including fraction correct, precision, and recall, are defined. Two particular visualizations used are the Receiver Operator Curve (ROC) and the Detection Error Tradeoff (DET) curve, and the metrics used to score from these visualizations are the Area under the ROC (AUC) and the minimum of a Decision Cost Function (DCF) (a DCF is a weighted sum of misses and false alarms). These visualizations and metrics not only measure the system's accuracy as it makes decisions but also how its accuracy (in terms of misses and false alarms) change when the system changes its threshold to become more lenient or stricter. This framework is used quite often for the testing and measuring of content authentication techniques.

E.3. Background: Frameworks for Model and Data Transparency

One form of testing software is for humans to manually spot-check or check properties of the synthetic content systems. Having transparency into the system and its models, the training data, and the data used to test the system can provide helpful information to users as they spot-check and test the different synthetic content systems. There are a variety of frameworks that provide ways to disclose key details about the model, as well as any data used in training or testing. Various frameworks include model cards, data sheets, a model card guidebook, and AI fact sheets.

E.4. Adversarial Attacks and Defenses on Synthetic Content

A common framework used to measure the quality of synthetic content is to construct attacks and defenses on the system. There are a variety of adversarial attacks that exist, but the ones typically used to evaluate systems involve adding carefully-crafted data inputs either to the training set or the test set so that the system will mishandle or misclassify images in the test set. As attacks and defenses are used relevant to the context of the system and how it is being used, one way the context is represented is by defining or restricting attacks relative to a threat model. One example of a threat model is: all attacked images are images altered from the test set where the alteration does not change the true class and the maximum distance (such as a norm) from that original image is at most a small, fix value. This style of experiment is a common way to evaluate attacks and defenses, as seen in studies from December 2013, February 2016, and May 2016.

When constructing attacks and defenses, specific strategies are used. The first such strategy (and a baseline strategy) is to construct an attack on the data, and then train a defense specific to treating that attack, and using that as the (attack, defense) pair. As this strategy does not show how the defense generalizes relative to other attacks, a second strategy has been developed. This strategy is a transferability analysis. A transferability analysis measures if attacks and defenses trained on one model and on one dataset can be successful on other datasets and situations. There is both intra-algorithm transferability (where the attack and defense are trained on one dataset but the system now must handle the same attack on a different dataset); as well as inter-algorithm transferability (using adversarial attacks from one trained model to fool a completely different algorithm, sometimes trained on the training dataset). The third strategy takes testing defenses on new attacks further and is a binarization test. This strategy uses a custom-designed machine learning classification to generate additional attacks to test the robustness of given defenses.

E.5. Theoretical Proofs To Support Synthetic Content Techniques

Mathematical proofs can guarantee success or establish properties supporting the correctness, efficiency, or effectiveness of synthetic content techniques. By proving specific components of a synthetic content generator correct, it can give evidence of the generator's effectiveness in certain situations. For defenses against adversaries fooling classifiers with tampered images, one such proof is a robustness certificate. In more detail, a robustness certificate gives a guarantee that that no attempt to alter image by at most a pre-specified small amount (according to a distance metric) can fool the system into misclassifying the altered image.

E.6. Similarity and Distance Metrics

Having a way to compare the quality of generated images to regular images is important. As human labels are expensive, having an automated way to compute image similarity (or image distance) can be efficient. In particular, there is a belief that AI-generated images are considered better or higher quality if they are more similar to human-generated images in the data. Similarity metrics are often from 0 to 1, where identical images have value 1 and completely different inputs have value 0. Distance metrics are such that an input has distance 0 to itself; the more different the two inputs are, the higher the distance. There are a variety of these metrics, automated similarity, and distance metrics used for specific use cases are in the Table below. Although this table aims to provide examples used in sources, this table should be viewed neither as representative nor as all-inclusive. Similarity metrics specifically for text also exist; these metrics are sometimes evaluated by comparing these scores to human judgements in separate experiments.

Use Case	Example Metrics Used
Measuring Quality of Automatically-Generated Images	the Inception Score (IS), Fréchet Inception Distance (FID, and based of the Fréchet distance), the Structural Similarity Index (SSIM)
Distance Metrics used to Show robustness of defenses	Lp distance norms (including Euclidean distance)
Measuring the Quality of AI-Generated Text	Self-BLEU Score, Mauve Score
Measuring the Quality of Watermark Extraction	Pixel correlation to original watermark
Measuring the Quality of Digital Fingerprinting (Hash Distance Metrics)	Hamming Distance, Euclidean Distance (L2 norm), Correlation Coefficient

Appendix F. Glossary

AI Content Detection: Determining whether content is AI-generated or not.

AI System: An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (NIST AI RMF)

Audit: "Independent review and examination of records and activities to assess the adequacy of system controls, to ensure compliance with established policies and operational procedures." (NIST SP 1800-15B under Audit from NIST SP 800-12 Rev. 1)

Authentication: Verifying the identity of a user, process, or device, often as a prerequisite to allowing access to resources in an information system. (FIPS 200 under AUTHENTICATION)

Authenticity: With respect to digital content transparency, it refers to the quality of being genuine, with trustworthiness about its source or origin.

Best practices: "A procedure that has been shown by research and experience to produce optimal results and that is established or proposed as a standard suitable for widespread adoption." (NIST SP 1800-15B from Merriam-Webster NIST SP 1800-15C from Merriam-Webster)

Content authentication: utilizes provenance data tracking methods to determine the authenticity of content (i.e., to indicate non-synthetic origins).

CSAM: Child sexual abuse material.

Digital content transparency: refers to the ability to obtain access and exposure to information regarding the origin and history of digital content. Transparency does not directly imply trust, but rather provides a vehicle for individuals, organizations, and other entities to have greater information access.

Digital signature: The result of a cryptographic transformation of data that, when properly implemented, provides a mechanism for verifying origin authentication, data integrity, and signatory non-repudiation. (FIPS 186-5)

Digital watermarking: involves embedding information into content (image, text, audio, video) in order to make it difficult to remove. The goal of such watermarking is to assist in verifying the authenticity of the content or characteristics of its provenance, modifications, or conveyance. (White House EO, 2023)

Evaluation: systematic determination of the extent to which an entity meets its specified criteria; (2) action that assesses the value of something (https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary citing ISO/IEC 24765)

Hash function: A hash function is any function that can be used to map data of arbitrary size to fixed-size values

Information integrity: Describes the spectrum of information and associated patterns of creation, exchange, and consumption in society, where high-integrity information is trustworthy; distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting. (White House, 2022)

Interoperability: "The ability of the user of one member of a group of disparate systems (all having the same functionality) to work with any of the systems of the group with equal ease and via the same interface." (Britannica)

Least Significant Bit: The least significant bit is the lowest bit in binary numbers.

Metadata: "Information describing the characteristics of data including, for example, structural metadata describing data structures (e.g., data format, syntax, and semantics) and descriptive metadata describing data contents (e.g., information security labels)." (NIST SP 800-150 under Metadata, CNSSI 4009-2015)

NCII: Non-consensual intimate imagery

Open information ecosystem: Supports a free exchange of ideas, enables ideas to flow from multiple sources, empowers people to express conflicting perspectives in a constructive manner, and leverages a free market of technologies to distribute information to audiences. (White House, 2022)

Provenance data tracking: records the origin and history for digital content, allowing its authenticity to be determined. It consists of techniques to record metadata as well as overt and covert digital watermarks on digital content. Provenance data tracking can help to establish the authenticity, integrity, and credibility of digital content. (NIST SP 800-161r1 NIST SP 800-218 from NIST SP 800-53 Rev. 5 NIST SP 800-37 Rev. 2)

Software Testing: The evaluation of software that utilizes Verification and validation (also abbreviated as V&V) to check that a product, service, or system meets requirements and specifications and that it fulfills its intended purpose. (Global Harmonization Task Force - Quality Management Systems - Process Validation Guidance (GHTF/SG3/N99-10:2004 (Edition 2) page 3)

Standard: a "document, established by consensus and approved by a recognized body, that provides – for common and repeated use – rules, guidelines or characteristics for activities or for their results, aimed at the achievement of the optimum degree of order in a given context." (ISO)

Steganography: Steganography is a technique which hides a watermark or content information file inside a primary media file. One of the more common types of steganography involve embedding this hidden or secret information in the Least Significant Bit of a media file, which is done by slightly modifying or adding additional information to bytes (or bits on those bytes) of data within pixels in a media file. (Authenticating AI-Generated Content, 2024, NIST SP 800-101 Rev. 1 under Steganography, NIST SP 800-72 under Steganography)

Synthetic content: "information, such as images, videos, audio clips, and text, that has been significantly altered or generated by algorithms, including by AI" (White House AI EO)

Test: (1) activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component; (2) to conduct an activity as in (1); (3) set of one or more test cases and procedures.
https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary citing
https://www.iso.org/obp/ui/en/#iso:std:iso-iec-ieee:24765:ed-2:v1:en